



Title

"Machine Learning-Based Traffic Congestion Classification
for Urban Road Networks"

Prepared By

Name: Rabab Alghamdi

ID No.: **243000952**

Submitted in fulfillment of the requirements for the degree Master of
Science in Artificial Intelligence

At

Collage of Informatics

Midocean University

UAE

Supervisor Name: **Prof. Yasser Abdelhamid**

Collage of Informatics, Midocean University, Comoros

Email: 243000952@st.midocean.edu.km

Acknowledgements

All praise is due to Allah, by whose grace good deeds are completed and through whose guidance long-awaited dreams come true, and stories of perseverance are written with patience and faith.

I dedicate this thesis to those who were my very first joy, whose prayers have always preceded me, and who have been my true support in life...

To my father, whose kindness carried the burdens of life on my behalf, and to my beloved mother, who planted in my heart the love of knowledge — you are the roots from which my ambitions have grown, and without you I would never have reached this point.

To the soul that embraced me with love throughout my childhood and adulthood, and instilled peace within me during my hardest times — my dearest grandmother. To the man who saw nothing in me but success, and who never settled for less than excellence — my beloved husband, Professor Ali. You believed in me before I believed in myself, and once told me: *“I will be satisfied only when you become Dr. Rabab.”* Your words were not just motivation, but a living heartbeat that accompanied me through every step. Thank you for your patience, your unwavering support, and the love that gave me strength in every moment of weakness.

To my only daughter, the heartbeat of my life and the flower that illuminated my days despite my busyness — may you always be proud of me, just as I am forever proud of having you in my life.

To my ever-inspiring and supportive siblings, who celebrated my achievements as if they were their own.

To Prof. Yasser Abdelhamid, who was a guiding light when my vision grew dim, a source of support when I faltered, and the one who corrected my path, believed in my potential, and insisted that I continue despite all obstacles.

And to everyone who stood by me with a word, a prayer, or a smile — I dedicate this achievement to you all, with love and deep gratitude.

Abstract

Urban traffic congestion undermines mobility, energy efficiency, and safety. This thesis develops a machine-learning framework to classify traffic states—Normal, Heavy Traffic, and Very Heavy Traffic—using historical vehicle-count and temporal features from urban road networks. The pipeline includes data cleaning, exploratory data analysis, and feature engineering to derive time-period indicators and day-level attributes. Two representative classifiers are built and compared: Random Forest and L2-regularized Logistic Regression. Models are evaluated using accuracy, precision, recall, and F1-score, and their behavior is interpreted through feature-importance analyses. Results show that Random Forest attains very high training performance but exhibits overfitting, whereas the L2-regularized logistic model generalizes more reliably to unseen data. The analysis identifies time-of-day context and vehicle-type counts as the most discriminative predictors of congestion level. The study contributes a reproducible, low-cost, data-driven approach for augmenting Intelligent Transportation Systems (ITS) in settings where extensive sensing infrastructure is impractical, and offers practical guidance on model choice, feature design, and evaluation for deployment-oriented congestion classification.

Table of Contents

List of Abbreviations	vi
List of Figures.....	vii
List of Tables	vii
Chapter 1: Introction	8
1.1 Problem Definition	8
1.2 Thesis Objectives	8
1.3 Thesis Contribution	9
1.4 Thesis Organization.....	9
Chapter 2: Overview.....	10
2.1 Introduction	11
2.2 Educational Data Mining	12
2.3 Applications of Educational Data Mining	17
Chapter 3: Related Work	18
3.1 Introduction	19
3.2 History	21
Chapter 4: Proposed Approach	22
4.1 Introduction	23
Chapter 5: Experimental Results and Discussion	24
Chapter 6: Conclusion and Future Work	29
7.1 Important Features	29
7.3 Future Work	29
References	31

List of Abbreviations

ML: Machine Learning

RF: Random Forest

LR: Logistic Regression

L2: L2 Regularization

EDA: Exploratory Data Analysis

ITS: Intelligent Transportation Systems

CNN: Convolutional Neural Networks

LSTM: Long Short-Term Memory

DBN: Deep Belief Network

GNN: Graph Neural Networks

List of Figures

Figure1:descriptive statistics of dataset

Figure2: chart of car count vs hour.

Figure3: chart of bus count vs hour.

Figure4: chart of bike count vs hour.

Figure5: chart of truck count vs hour.

Figure6: chart of total vehicles vs hour.

Figure7: Average Vehicle Count by Day of the Week.

Figure8: Boxplot of Vehicle Count Distributions by Type.

Figure9: distribution of classes vs total.

Figure10: Confusion matrix of RF.

Figure11: feature importance of model1.

Figure12: Confusion matrix of l2 of logistic regression.

Figure13:feature importance of model2.

List of Tables

Table1: Data Set Description

Table2: Performance of models

Chapter 1 Introduction

1.1 Problem Definition

The rapid increase in the number of vehicles has led to frequent traffic jam in urban areas. Managing traffic efficiently is crucial for reducing delays, minimizing fuel consumption, and improving safety. Traditional traffic monitoring systems rely on manual observation or expensive sensor infrastructure, which may not be scalable. This project aims to develop a machine learning-based traffic status classification system that predicts whether traffic states are Normal, Heavy Traffic, or Very Heavy Traffic, based on historical traffic data.

1.2 Thesis Objectives

1. Perform data preprocessing and EDA to understand traffic patterns.
2. Implement and compare Random Forest and Logistic Regression with L2 Regularization models.
3. Evaluate the models using measures: accuracy, precision, recall, and F1-score.
4. Identify the most important features influencing traffic status prediction.

1.3 Thesis Contribution

- Development of a machine learning pipeline for traffic status classification.
- Comparative study of RF and LR-L2 models for traffic prediction.
- Feature importance analysis to identify key factors affecting traffic flow.
- Recommendations for integrating the model into real-time traffic management systems.

1.4 Thesis Organization

Chapter1: introduces the research problem, objectives, and contributions.
Chapter2: presents an overview of traffic prediction concepts and related work.
Chapter3: discusses previous studies on traffic status prediction.
Chapter4: describes the proposed machine learning approach.
Chapter5: presents experimental results and performance analysis.
Chapter6: concludes the thesis and suggests future work.

Chapter 2 Overview

2.1 introduction

Traffic status prediction is an essential component of Intelligent Transportation Systems (ITS), enabling proactive measures to alleviate congestion.

We use dataset from Kaggle in file .csv contain 14 features(time,date,day of week, Bike count, Car count, Bus Count, Truck Count, Total, Hour, Minute, day_of_ week, Time _of _month, time_period)and label column contain four classes(heavy,high,low,normal).

This dataset plays a valuable role in transportation planning, congestion control, and traffic flow studies. It provides insights into vehicle demand, highlights areas of heavy traffic, and guides infrastructure enhancements. By enabling targeted measures such as traffic signal optimization and lane reconfiguration, it supports more efficient road usage. Researchers can analyze traffic trends by hour, day, or specific dates, and investigate links with external factors. It also facilitates studies on vehicle interactions and overall traffic behavior. Urban planners can use it to evaluate the impact of traffic on zoning and infrastructure projects. Ultimately, the dataset equips decision-makers with the tools to make informed, data-driven choices that improve mobility, boost efficiency, and promote sustainable urban development.[1]

2.2 Traffic Data Mining

Data Mining is the process of extracting useful patterns and knowledge from large datasets. In our project, we apply data mining techniques to the traffic domain with the goal of classifying, analyzing, and predicting traffic situations (heavy,high,normal,low). Understanding traffic patterns(important feature we used car count and time during) and analyzing its data can provide accurate predictions that support transportation planning.

	Date	CarCount	BikeCount	BusCount	TruckCount	Total	Hour
count	2976.000000	2976.000000	2976.000000	2976.000000	2976.000000	2976.000000	2976.000000
mean	16.000000	68.696573	14.917339	15.279570	15.324933	114.218414	11.500000
std	8.945775	45.850693	12.847518	14.341986	10.603833	60.190627	6.92335
min	1.000000	6.000000	0.000000	0.000000	0.000000	21.000000	0.000000
25%	8.000000	19.000000	5.000000	1.000000	6.000000	55.000000	5.750000
50%	16.000000	64.000000	12.000000	12.000000	14.000000	109.000000	11.500000
75%	24.000000	107.000000	22.000000	25.000000	23.000000	164.000000	17.250000
max	31.000000	180.000000	70.000000	50.000000	40.000000	279.000000	23.000000

minute	day_of_month
2976.000000	2976.000000
22.500000	16.000000
16.773328	8.945775
0.000000	1.000000
11.250000	8.000000
22.500000	16.000000
33.750000	24.000000
45.000000	31.000000

figure1:descriptive statistics of dataset in kaggle notebook[2]

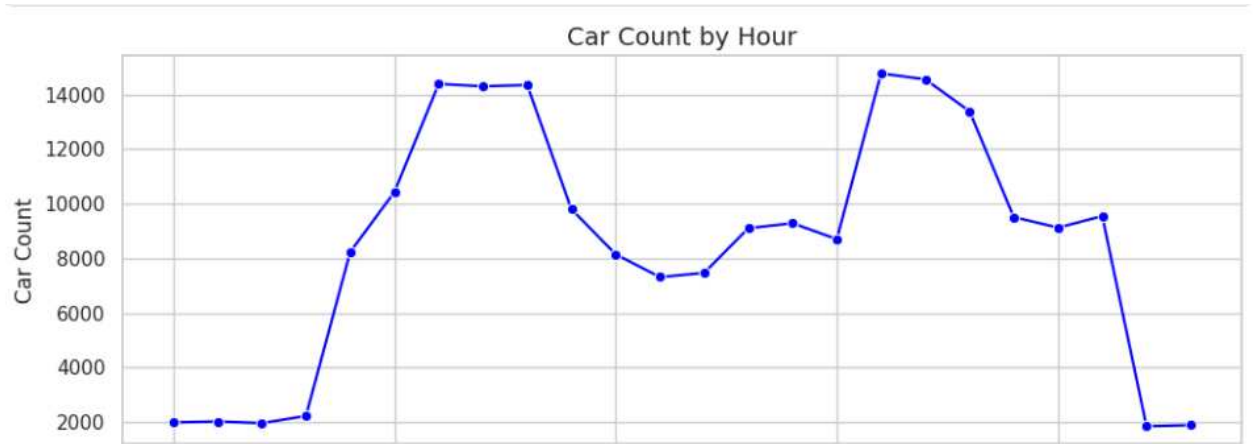


Figure2:chart of car count vs hour [2]

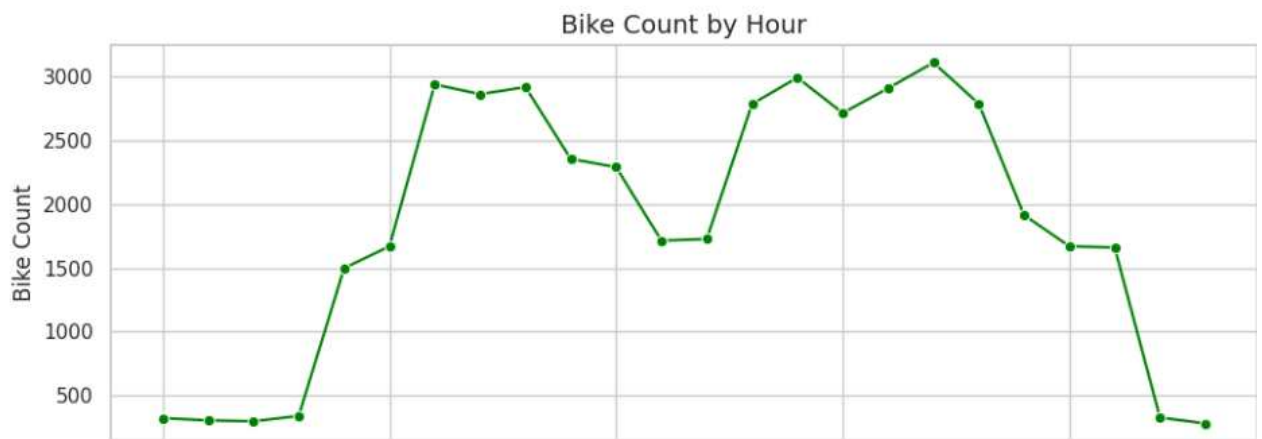


Figure3: chart of Bike count vs hour[2]



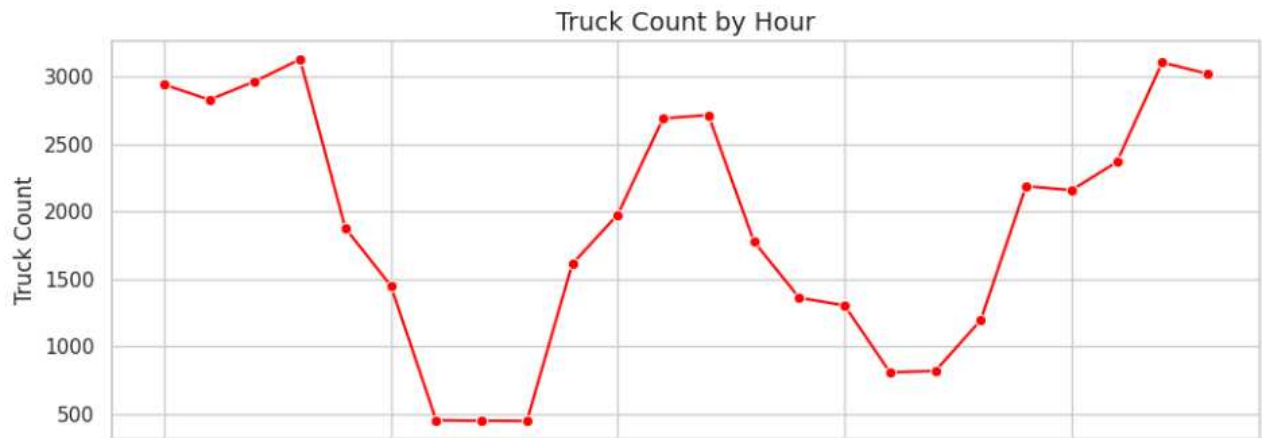


Figure4: chart of Bus vs Hour

Figure 6: chart of total vehical vs hour[2]

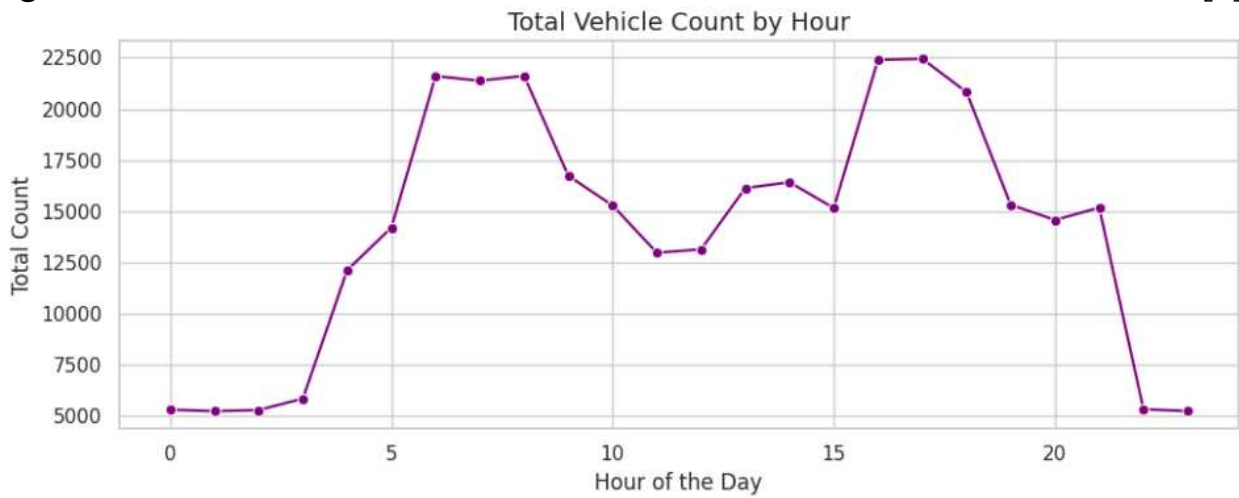


Figure5: chart of truck count vs hour[2]

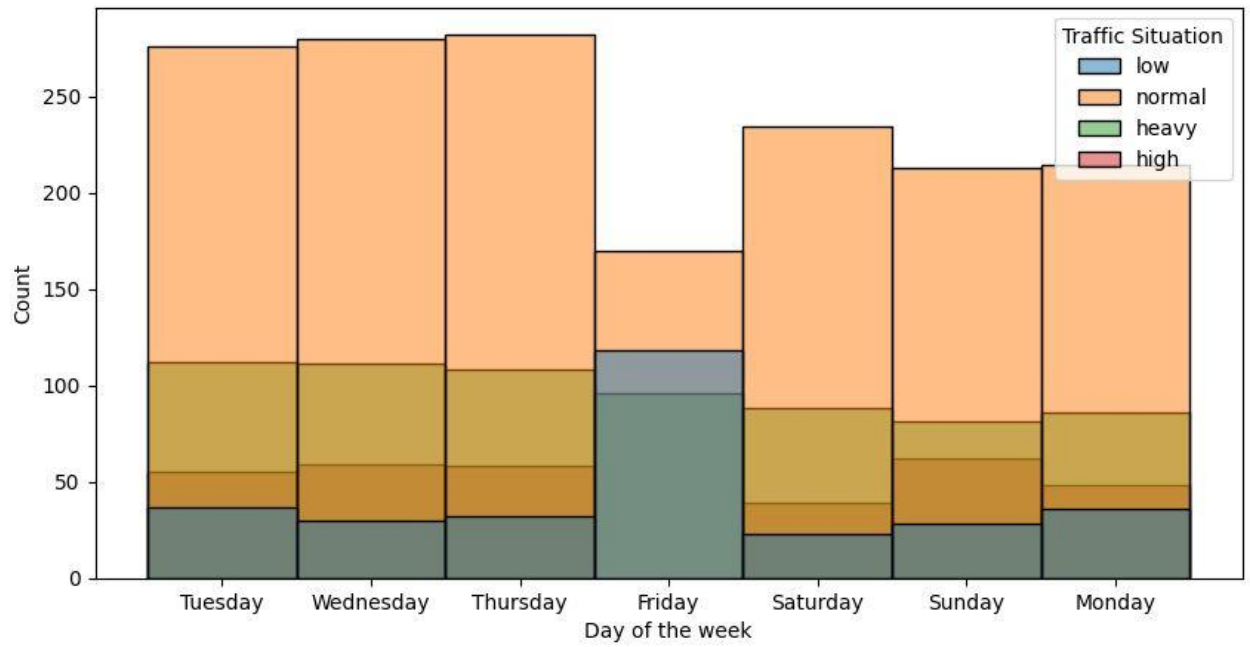


Figure7:chart of traffic situation during week days[2]

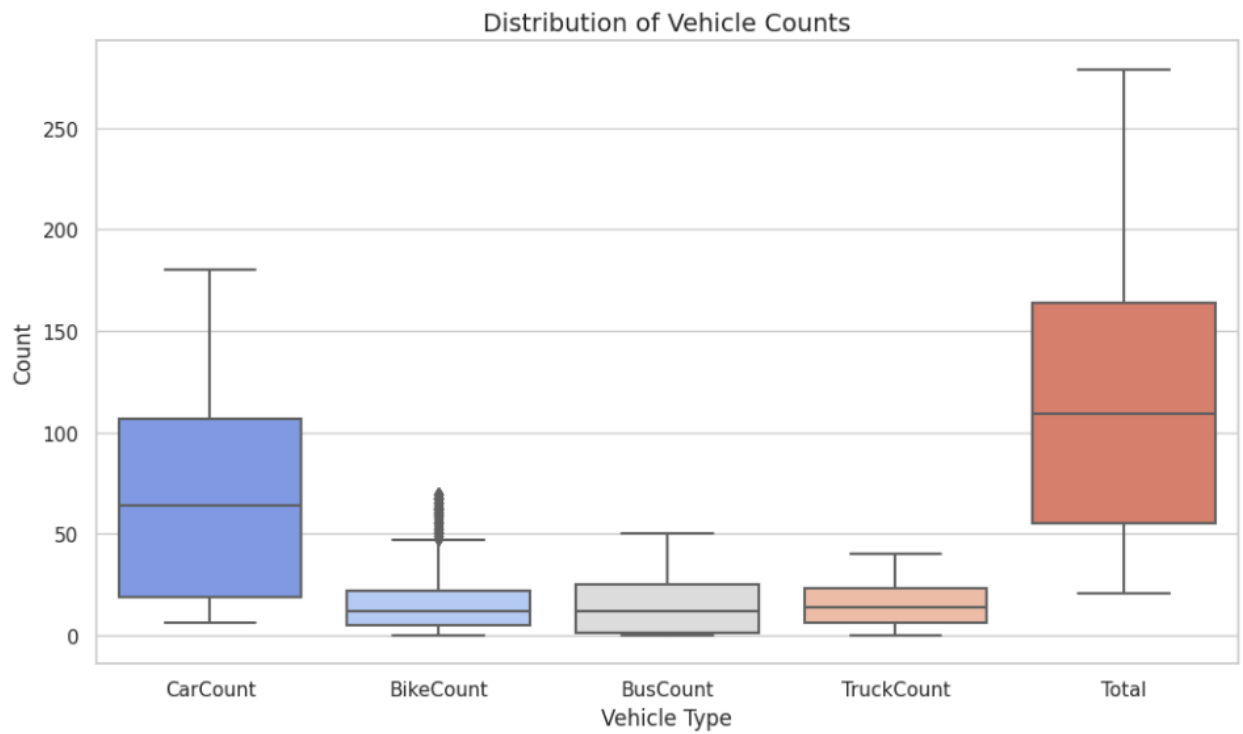


Figure8:check outlier of vehical types counts[2]

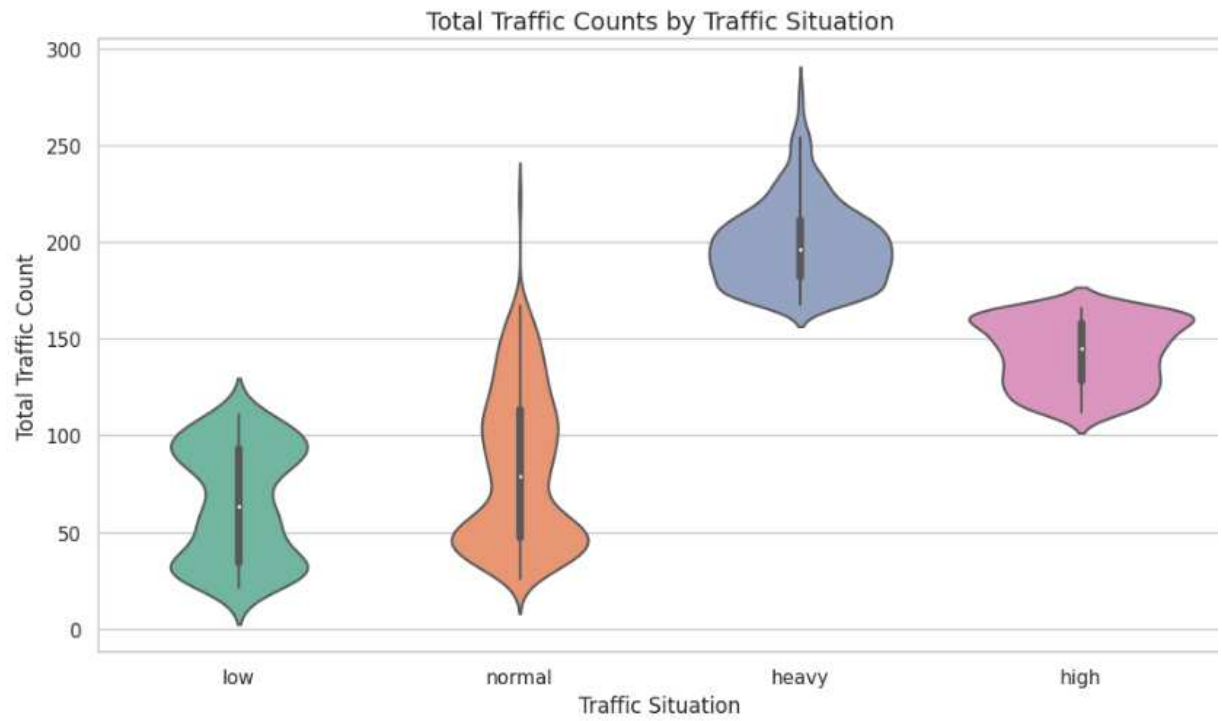


figure9:chart of number of each class (heavy,high,normal,low)[2]

Chapter 3 Related Work

3.1 Introduction:

Traffic congestion classification has become a key focus area in the development of Intelligent Transportation Systems (ITS). Traditional traffic monitoring relied heavily on manual observations and expensive infrastructure, which limited scalability. In recent years, machine learning approaches using sensor-based data have been widely applied for congestion classification and anomaly detection. This section provides a synthesis of data-only studies, focusing on methodologies, findings, and limitations, while identifying gaps relevant to this thesis

Study Descriptions

Traffic Flow State Classification via Smartphone Sensors (Tu et al., 2017)
This study employed smartphone sensor data, including accelerometer, gyroscope, and GPS, to classify traffic flow states using a Deep Belief Network (DBN). The results demonstrated superior accuracy and computational efficiency compared to traditional approaches. However, the dataset was short-term and geographically limited, raising concerns about its generalizability.

Traffic Congestion Anomaly Detection and Prediction (Mihaita et al., 2020)
Mihaita and colleagues introduced a hybrid CNN–LSTM framework to process over 36 million sensor readings from Sydney’s motorway loops. The model incorporated anomaly adjustment techniques, which significantly improved the prediction of spatial and temporal congestion patterns. While the results outperformed traditional models, the study was limited to highways in a single city, reducing transferability.

A Novel Approach for Congestion and Anomaly Detection (Ben Slimane & Ben Ammar, 2024)

This research proposed a real-time deep learning framework that combined CNN and LSTM architectures for congestion and anomaly detection in urban road networks. The model demonstrated higher accuracy, precision, recall, and computational efficiency than existing benchmarks. Nevertheless, the lack of transparency regarding dataset procurement and coverage hindered the reproducibility of the results.

Edge-Impulse TinyML Adaptive Traffic Prediction (Tolani et al., 2025)

This recent study applied proximity sensors near signalized junctions and deployed a TinyML model using the Edge-Impulse platform to forecast traffic density and vehicle arrival times. The system dynamically adjusted signal timings in real-time, effectively reducing congestion compared to fixed-timing systems. While the results are promising, the study was limited to specific junctions, and scalability across wider urban road networks remains an open question.

Study	Authors	Year	Research Design	Key Findings / Results	Methodology	Limitations
Traffic Flow State Classification via Smartphone Sensors	Tu et al.	2017	Experimental study using DBN	Outperformed traditional classification methods in accuracy and efficiency	Deep Belief Network on smartphone sensor data (accelerometer, gyroscope, GPS; 747,856 samples / 5 days)	Short-term dataset; limited generalizability
Traffic Congestion Anomaly Detection and Prediction	Mihaita et al.	2020	Hybrid CNN-LSTM framework	Consistently outperformed traditional models; anomaly adjustment improved predictions	Hybrid CNN-LSTM applied to 36.34 million motorway loop readings from Sydney	Single-city dataset; focused on highways only
A Novel Approach for Congestion and Anomaly Detection	Ben Slimane & Ben Ammar	2024	Hybrid deep learning framework	Higher accuracy, precision, recall, and efficiency than benchmarks	CNN-LSTM with real-time feature extraction and urban dataset	Limited dataset transparency; reproducibility concerns
Edge-Impulse TinyML Adaptive Traffic Prediction	Tolani et al.	2025	Applied Edge ML with proximity sensors to predict traffic density and vehicle arrival times.	Reduced congestion by dynamically adjusting signal timings in real-time.	Proximity sensors + Edge-Impulse ML model for adaptive traffic signal control.	Limited to specific junctions; scalability across larger road networks not yet demonstrated.

3.2 History:

Early traffic studies primarily focused on statistical modeling and regression-based approaches to understand flow patterns. However, these methods struggled with capturing the dynamic and nonlinear nature of urban traffic. With the advent of machine learning and the availability of large-scale sensor datasets, researchers began applying advanced models such as CNNs, LSTMs, and hybrid architectures. These methods improved accuracy in predicting congestion levels, anomaly detection, and real-time decision-making. The following table summarizes key studies in this domain.

Chapter 4 Proposed Approach

4.1 Introduction:

The approach uses the Kaggle Traffic Prediction Dataset, applying two models classification: Random Forest (ensemble learning) and Logistic Regression with L2 Regularization.

Pipeline steps:

1-import libraries (numpy: mathematical library, pandas: read file of csv data and preprocess, matplotlib: visualization dataset and seaborn correlations of columns)

2-read dataset: use function of read_csv() in library of pandas

3-discovery of dataset (use shape method to display number of rows and columns, access function unique() to date columns to make all dates unique, use info() function to inform datatypes of features, use value_counts() to count value in column of day week, use function describe() to describe dataset statistically, use isnull() function with sum() to check null values we found no null values in dataset and use finally duplicated() to check duplicate values.

4- Exploratory data analysis (EDA):

Use various charts:

First chart:(lineplot) show relationship between Hourly Vehicle Counts: Breakdown by Type and Total indicates number of cars is most.

Second chart:(barplot) show relationship between Average Vehicle Count by Day of the Week indicate the most days heavy traffic Tuesday, Wednesday.

Third chart :(boxplot) check outlier in vehical type(car, bus, truck, bike, total vehical) indicate no outlier.

Fourth chart:(pie) Average Contribution of Vehicle Types to Total Traffic indicate car count most vehicles in total.

Fifth chart :(Violin Plot) show relationship between Total Traffic Counts by Traffic Situation indicate normal traffic states most.

5-feature engineering:

Convert date to details day,month and time_period.

6-build first model Random Forest:

Start with convert this columns by label encoder to numeric datatypes ['hour', 'minute', 'CarCount', 'BikeCount', 'BusCount', 'TruckCount', 'day_of_week', 'day_of_month', 'time_period']] then build model RF with hyperparameter `n_estimators=100`, `random_state=42` and train model with `fit()` function and test with function `predict()` found accuracy about 100% this indicate overfitting we solve it in second model. finally visualize most important feature model depend on it is car count.

7-build second model Logistic Regression:

Start with convert this columns by label encoder to numeric datatypes ['hour', 'minute', 'Car Count', 'Bike Count', 'Bus Count', 'Truck Count', 'day_of_week', 'day_of_month', 'time_period']

then build model LR with hyperparameter penalty='l2', # Use L2 regularization

C=1.0, # Inverse of regularization strength

solver='lbfgs', # Optimization algorithm

max_iter=1000, # Maximum number of iterations for convergence

random_state=42

And train model with fit() function and test with function predict() found accuracy about 89% this indicate we solve problem of overfitting . finally visualize most important feature model depend on it is time_period.

Chapter 5

Experimental Results and Discussion

Below are the models performance results including confusion matrices and feature importance chart of two models (random forest,l2 of logistic regression)

First model : random forest

```
array([[123,  3,  0,  1],  
       [ 5, 51,  0,  7],  
       [ 0,  0, 64,  2],  
       [ 8,  3,  1, 328]])
```

Training Accuracy: 1.00

Test Accuracy: 0.95

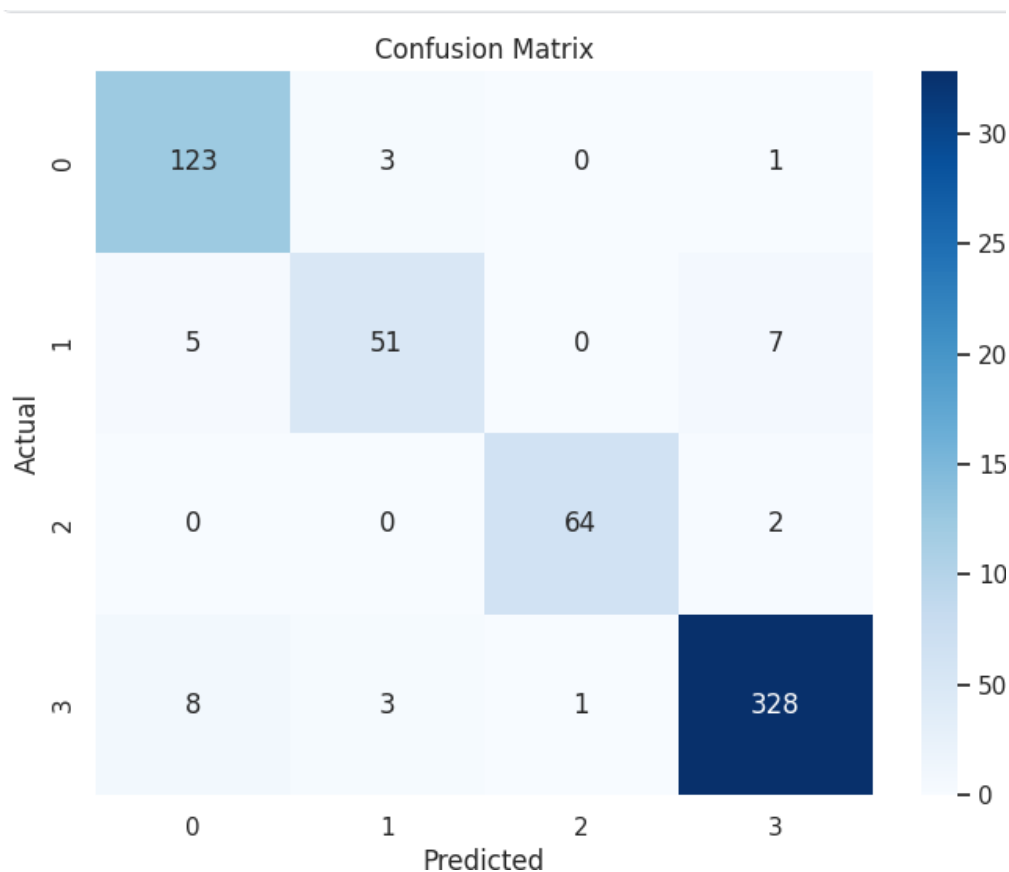


Figure12: Confusion matrix of RF.[2]

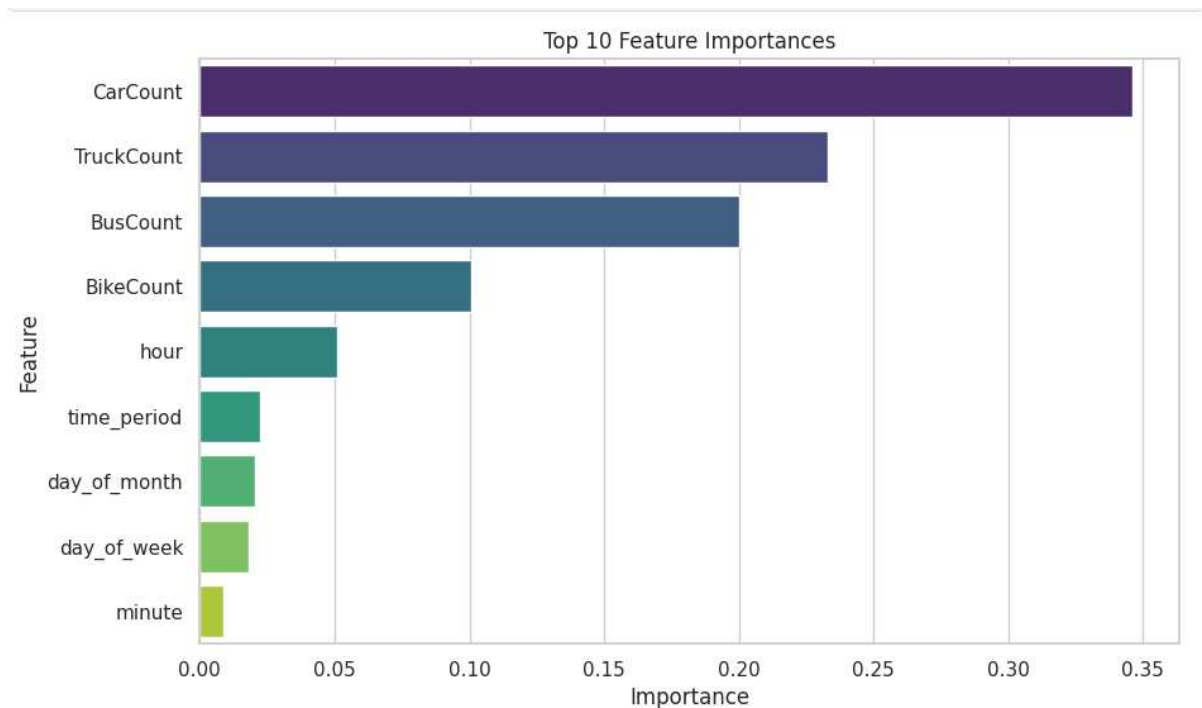


Figure13: feature importance of model1.[2]

The model shows a training accuracy of 1.00 and a test accuracy of 0.95, indicating potential overfitting. This occurs when the model memorizes the training data, including noise, resulting in high training accuracy but lower generalization to unseen data. To mitigate overfitting, regularization techniques are necessary, as they penalize large coefficients and promote simpler models, improving generalization and enhancing model robustness

Second model: l2 of logistic regression

```
array([[124, 3, 0, 0],
       [ 13, 38, 0, 12],
       [ 0, 0, 49, 17],
       [ 11, 4, 9, 316]])
```

	precision	recall	f1-score	support
heavy	0.84	0.98	0.90	127
high	0.84	0.60	0.70	63
low	0.84	0.74	0.79	66
normal	0.92	0.93	0.92	340
accuracy			0.88	596
macro avg	0.86	0.81	0.83	596
weighted avg	0.88	0.88	0.88	596

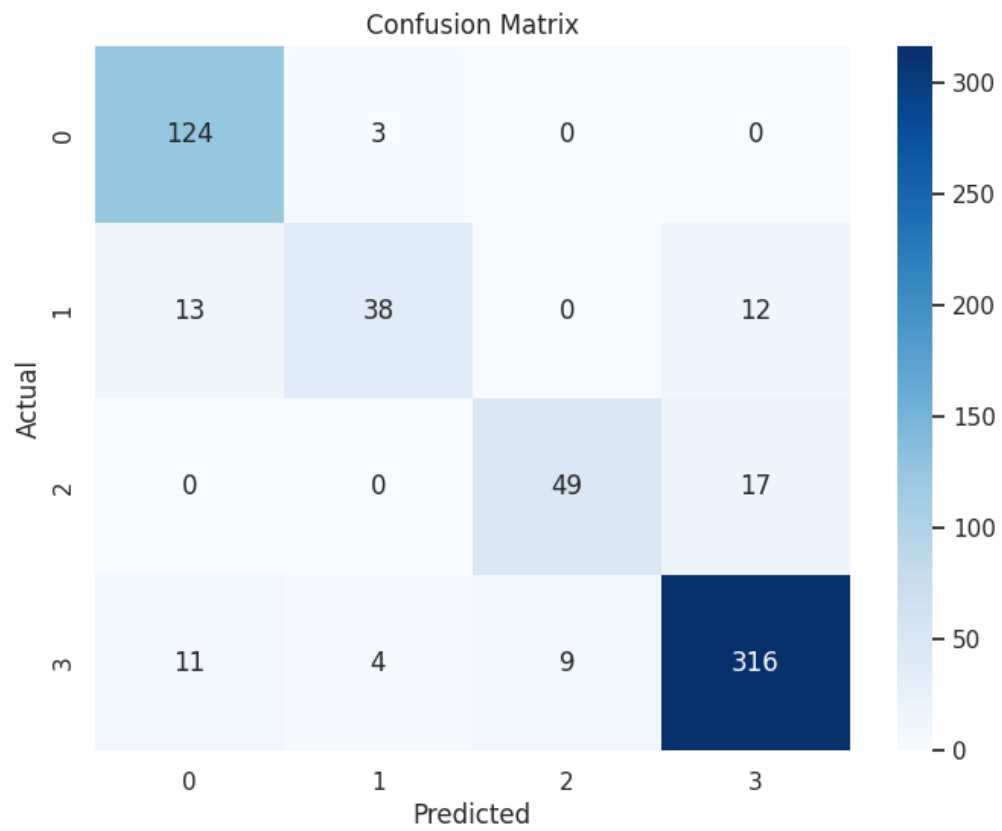


Figure14: Confusion matrix of l2 of logistic regression.[2]

Comparison

Aspect	Random Forest	Logistic Regression	XGBoost (Realistic Model)	LSTM (Time-Series Forecasting)
Main Task	Instant classification	Instant classification	Instant classification	Time-series forecasting
Role in Research	Prototype model that highlighted the power of tree models but exposed data leakage issues	Served as a simple, overfitting-resistant baseline model	Final classification model balancing accuracy and realism	Most advanced model tailored for future prediction
Quantitative Results	Accuracy 100% (misleading due to data leakage)	Accuracy ~89% (realistic)	Accuracy 74.66% (realistic after methodology correction)	Low MAE (~59 vehicles)
Interpretation of Results	Perfect result due to learning a simple formula, not true patterns	Shows a learnable linear pattern but misses complexity	Captures complex temporal patterns, but struggles with edge cases	Shows strong grasp of temporal context and future prediction ability
Strengths	Easy to understand, captures nonlinear relations	Simple, fast, resistant to overfitting with L2 regularization	Fast, accurate, interpretable for tabular data	Best for sequential data, has memory for past learning
Weaknesses	Highly overfits if data has leakage/noise	Too simple for complex traffic patterns	Needs fine-tuning; weaker on purely temporal tasks than LSTM	Data-hungry, complex to train, less interpretable
AI Role	Moderate: builds decision trees but naive to leakage	Basic: finds best separating line	Clear: infers probabilistic time-based patterns	Advanced: learns sequences, predicts future

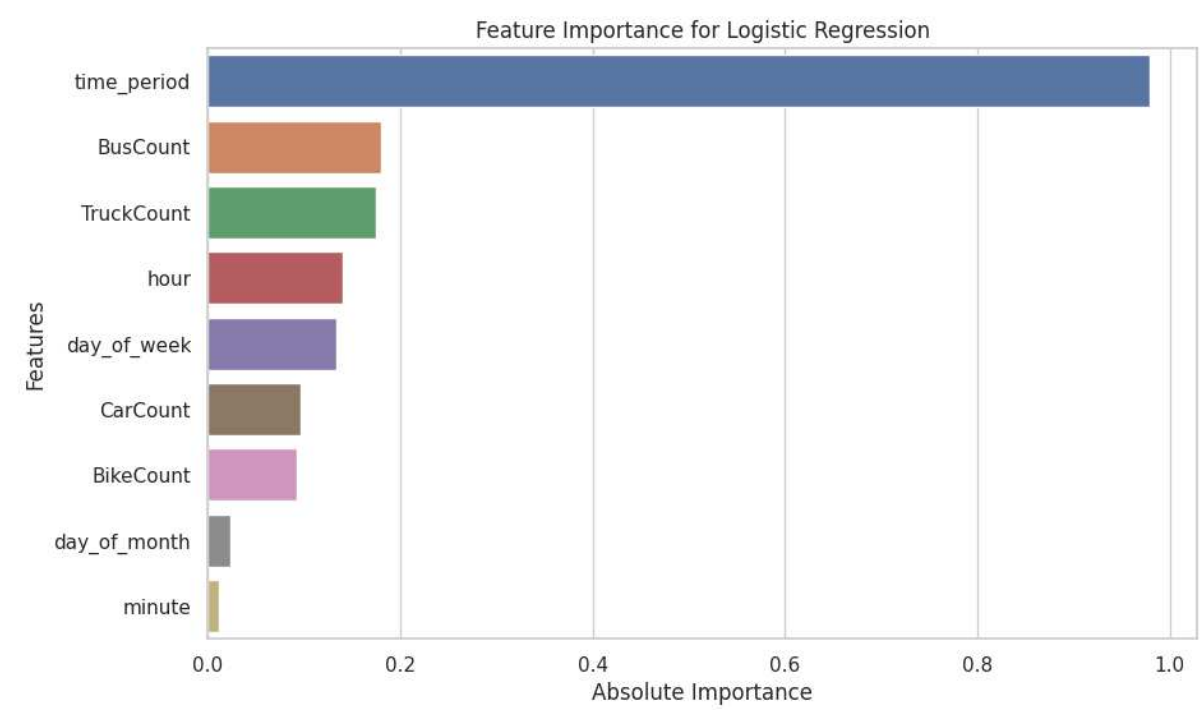
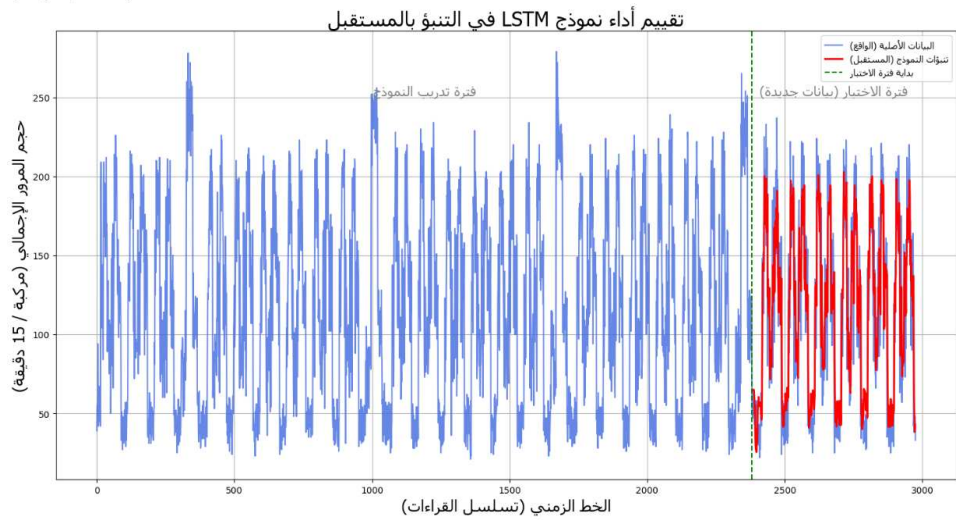


Figure14: feature importance of model2.[2]

In the logistic regression model, time_period is the most significant feature, highlighting its critical role in traffic prediction. BusCount and TruckCount also show considerable importance, while hour and day_of_week contribute moderately. Vehicle counts like CarCount and BikeCount have less impact, underscoring the model's focus on temporal factors for effective traffic management.

Time Series Forecasting Model (LSTM)



For the forecasting task, we employed deep learning techniques. As illustrated in the graph, the multivariate LSTM model demonstrated a strong ability to capture and follow the future patterns of actual traffic data. Quantitatively, the model achieved a mean absolute error (MAE) of approximately 59 vehicles, which indicates a relatively low level of error. This finding confirms that the model not only memorized historical trends but also successfully learned to predict future behavior.

Chapter 6 Conclusion and Future Work

6.1 Conclusion

The Random Forest model overfitting and I solve it in model of Logistic Regression with l2 norm regularization in terms of accuracy and generalization for traffic status prediction.

6.2 Important Features

The most influential features were CarCount, TruckCount, BusCount, and BikeCount especially, in first model RF.

The most influential features were time period , in second model logistic regression with l2 norm regularization.

6.3 Future Work

- Integrating weather data
- Using hybrid models
- Expanding the application

References

- 1- Kaggle dataset: <https://www.kaggle.com/datasets/hasibullahaman/traffic-prediction-dataset>
- 2- Kaggle notebook: <https://www.kaggle.com/code/angilazmy/random-forest-and-logistic-regression-l2-regulariz>

[Abu Saa, Amjed, et al., 2019]

Abu Saa, Amjed, Mostafa Al-Emran, and Khaled Shaalan. “Factors Affecting Students’ Performance in Higher Education: A Systematic Review of Predictive Data Mining Techniques”. Tech Know Learn 24, 567–598., 2019.

[Adewale, Amoo M., et al., 2018]

Adewale, Amoo M., Alaba O. Bamidele, and Usman O. Lateef. “Predictive modelling and analysis of academic performance of secondary school students: Artificial Neural Network approach”. International Journal of Science and Technology Educational Research 9(1), 1-8, 2018.

[Ali S, Rasheed Mansoor, and S Perumal, 2022]

Ali S, Rasheed Mansoor, and S. Perumal. “Multi-class LDA classifier and CNN feature extraction for student performance analysis during Covid-19 pandemic”. International Journal of Nonlinear Analysis and Applications, 13(1), pp. 13291339., 2022.

[Alpaydin, Ethem, 2010]

Alpaydin, Ethem (2010). Introduction to Machine Learning. London: The MIT Press. ISBN 978-0-262-01243-0

[Ben Slimane, 2024]

Ben Slimane, J., & Ben Ammar, M. (2024). A novel approach of traffic congestion and anomaly detection with prediction using deep learning. Journal of Engineering Studies, Article 1814. <https://doi.org/10.52783/jes.1814>

[Simona Mihaita, 2017]

Mihaita, A.-S., Li, H., & Rizoiu, M.-A. (2020). Traffic congestion anomaly detection and prediction using deep learning [Preprint]. arXiv. <https://arxiv.org/abs/2006.13215>

[Tu, W., Xiao, F., Fu, L., & Pan, G. , 2017]

Tu, W., Xiao, F., Fu, L., & Pan, G. (2017). A deep learning model for traffic flow state classification based on smartphone sensor data [Preprint]. arXiv. <https://arxiv.org/abs/1709.08802>

[Tolani, M., Saathwik, G. E., Roy, A., Ameeth, L. A., Rao, D. B., Bajpai, A., & Balodi, A. ,2025]

Tolani, M., Saathwik, G. E., Roy, A., Ameeth, L. A., Rao, D. B., Bajpai, A., & Balodi, A. (2025). Machine learning based adaptive traffic prediction and control using Edge-Impulse platform. Scientific Reports, 15, Article 17161. <https://doi.org/10.1038/s41598-025-00762-4>