Received: 29 August 2024 | Revised: 1 July 2025 | Accepted: 17 October 2025 | Published online: 1 November 2025

RESEARCH ARTICLE

Artificial Intelligence and Applications
2025, Vol. 00(00) 1-10
DOI: 10.47852/bonviewAIA52024199

BON VIEW PUBLISHING

Optimizing Traffic Signal Control Using Machine Learning and Environmental Data

Edidiong E. Akpan[1], Oluwatobi Akinlade[2], Oluwaseyi O. Alabi[3,*], Oluwasesan A. David[4], Oluwaseyi F. Afe[5] and Sunday Adeola Ajagbe[6,7]

[1] *School of Computing and Informatics, University of Louisiana at Lafayette, USA*

[2] *Department of Computer Science, Birmingham City University, UK*

[3] *Department of Mechanical Engineering, Lead City University, Nigeria*

[4] *Department of Computer Science, Nottingham Trent University, UK*

[5] *Department of Computer Science, Lead City University, Nigeria*

[6] *Department of Computer Engineering, Abiola Ajimobi Technical University, Nigeria*

[7] *Department of Computer Science, University of Zululand, South Africa*

Abstract: Traffic signal control is a critical component of urban transportation management, and optimizing its performance can significantly reduce congestion, decrease travel times, and improve air quality. This study proposes a novel approach to optimizing traffic signal control using machine learning and environmental data. This work focuses on the interplay between smart city infrastructure and environmental data to provide a novel method for traffic pattern prediction. Mitigating traffic congestion is a pressing concern in urbanized societies and emerging smart cities. This study explores leveraging publicly available air pollution data as an environmental indicator to enhance urban mobility and predict traffic patterns. Taking into account factors including vehicle emissions, weather patterns, and topographical features, the study will look at possible connections between air pollution and traffic congestion. The goal of this project is to develop a prediction model that uses real-time air quality data for traffic forecasting by utilizing big data analytics and machine learning approaches. According to our research, the K-nearest neighbors (KNN) model performs better than any other regression model examined. According to experimental findings, the KNN model considerably lowers the error rate in traffic congestion prediction by over 30%.

Keywords: machine learning, forecasting, environmental data, urbanization, traffic, air pollution

1. Introduction

Traffic signal control is a crucial aspect of urban transportation management, aiming to minimize congestion, reduce travel times, and decrease environmental impact. Traditional traffic signal control systems rely on fixed timers or simple sensor-based approaches, which often lead to inefficient traffic flow and increased emissions. With the rapid growth of urban populations and the resulting rise in traffic congestion, there is a pressing need for innovative solutions to optimize traffic signal control. Traffic congestion is a serious issue with significant negative impacts on both the economy and the environment. One of the primary contributors to urban traffic congestion is improperly operated traffic signals [1]. The diversity and unpredictable nature of traffic have surpassed the effectiveness of conventional traffic light systems, which depend on predetermined, fixed-time plans for junction control, despite major advances in online adaptive traffic signal control. When cars are present at a junction, online adaptive signal control can react by modifying the signal timings in real-time in response to shifting traffic patterns, in contrast to pre-timed fixed signal control, which repeats a predetermined regime. However, current adaptive control techniques struggle to effectively manage congestion. These methods often rely on systems that fail to accurately simulate traffic flow or are based on application-specific heuristics, which are inadequate due to the highly unpredictable nature of real-world traffic, such as sudden accidents that obstruct traffic flow [2, 3]. To improve the effectiveness of the traffic light management systems in place today, automated agents with the ability to learn, self-configure, and self-optimize must be put into place. As global urbanization accelerates, the number of vehicles on the road increases, exacerbating traffic congestion. This growth is placing unprecedented strain on existing traffic infrastructures, contributing to congestion, air pollution, and increased travel times in cities. Traffic signal control, a fundamental component of urban traffic management systems, plays a vital role in regulating the flow of vehicles. Traditionally, traffic signals have been controlled by static or rule-based systems, where the timing of signal phases is predetermined based on historical traffic patterns [4, 5]. These systems, although effective to some degree, are inherently limited in their ability to adapt to real-time fluctuations in traffic conditions and environmental factors, leading to inefficiencies and contributing to environmental and economic costs. The advent of machine learning (ML) technologies offers a promising avenue to enhance the efficiency of traffic signal control systems. ML algorithms can learn from large volumes of real-time traffic data, predict traffic conditions, and dynamically adjust signal timings to optimize traffic flow [2, 6]. This capability represents a significant departure from conventional systems, as it enables a more responsive

***Corresponding author:** Oluwaseyi O. Alabi, Department of Mechanical Engineering, Lead City University, Nigeria. Email: alabi.oluwaseyi@lcu.edu.ng

and adaptive approach to managing traffic, ultimately helping to reduce congestion, improve travel times, and lower emissions. In addition to traffic data, incorporating environmental data into traffic signal control strategies offers a powerful tool for addressing sustainability challenges in urban environments. Environmental data, such as air quality metrics, weather conditions, and noise levels, provide crucial context that can inform more holistic traffic management strategies [7, 8]. For instance, adjusting traffic signals in response to poor air quality or adverse weather conditions can mitigate the negative impacts of vehicle emissions and enhance the overall livability of urban areas.

With the advent of ML, there is an opportunity to revolutionize how traffic signals are managed. By leveraging ML algorithms, traffic signals can dynamically adjust based on real-time traffic data, leading to more efficient traffic flow [9]. Additionally, incorporating environmental data such as weather conditions and air quality can further enhance these systems [3]. This data-driven approach can help reduce emissions, improve air quality, and create more sustainable urban environments. The use of ML as a direct technique to attain adaptive optimum control in nonlinear systems has grown in popularity. ML agents carry out tasks by using perception to keep an eye on their environment, acting to change it, and then analyzing the results to gain knowledge and get better [10, 11]. Sequential decision-making control issues may now be effectively tackled with the help of deep reinforcement learning (DRL) [12, 13]. In high-dimensional, dynamic, and complicated settings like Atari games, DRL has shown to be incredibly successful [4]. A DRL agent must constantly interact with its surroundings to do a task, picking up on the qualities that are essential for every assignment. Understanding the connection between the agent's activities and their ultimate effects on the environment is a vital component of this interaction.

Medical experts assert that pollution and poor air quality, which are primarily caused by traffic congestion, are the primary causes of the higher-than-average death rates in metropolitan areas of big cities [5]. The 115 biggest cities in the European Union, home to around 40 million people together, find it difficult to maintain the high criteria for air quality. Many cities are putting sensor networks along their highways to monitor air pollution levels from traffic and traffic movement to solve this problem. An extended period of traffic congestion causes cars to use more gasoline, which raises the emissions of hydrocarbons (HC), nitrogen dioxide (NO_2), carbon dioxide (CO_2), and other pollutants [14, 15]. Numerous health issues, like as respiratory infections and diseases, heart disease, lung cancer, and other ailments, are linked to these emissions. What is more worrisome is that if drivers are aware of alternate routes or times to avoid traffic, they may reduce traffic bottlenecks. This has the potential to improve health outcomes and lessen air pollution.

Several studies have explored the use of traffic data analysis to predict and simulate air quality [6]. This study mostly used long short-term memory (LSTM) methods, which are well-known for outperforming many other Deep Learning (DL) models in terms of performance. Quantifying important pollutants including O_2, CO, NO_2, and CO_2 was made possible in large part by LSTM models [5]. Weather, car emissions, pollution levels, and traffic data were among the five different combinations of measures and components examined during the experiment [16]. It is important to keep in mind that the impacts of high traffic volume were not considered in the study. Agrahari et al. [17] proposed a stochastic adaptive traffic signal control system utilizing reinforcement learning to effectively prevent traffic congestion. This system enhances the standard intersection model by incorporating real-world complexities like turning fractions and lane configurations.

The study highlights the importance of traffic awareness for travelers' comfort and reduced stress, emphasizing that traffic management systems are an essential component of smart cities [8]. A critical aspect of comprehensive traffic management services is the smart mobility component. Traffic congestion not only causes inconvenience in many large cities but also contributes to various health issues and consumes significant amounts of time [5]. The key to minimizing the harmful health consequences of traffic-related air pollution is to put well-managed programs into place to reroute traffic onto less crowded routes in addition to lowering air pollution levels. Given the complexity and dynamic nature of road networks, accurately and efficiently predicting traffic flow is difficult. Urban growth, ease of travel, and mobility are all critical components of traffic management in smart cities, and they are intimately related to intelligent solutions for reducing congestion. In contrast to other research that mostly relied on transport data to predict air pollution, this study emphasizes the critical role that air quality data plays in predicting traffic intensity, proving that air pollution data might be a useful tool for precise road traffic forecasting.

DL is one of the most well-known subfields within ML, which comprises multiple subfields. An essential component of artificial intelligence (AI) is DL, which makes use of algorithms meant to get better over time. At its foundation, DL uses artificial neural networks (ANNs), in contrast to standard ML, which is predicated on more straightforward ideas [9]. By mimicking cognitive processes seen in the human brain, these ANNs replicate human cognition and learning. An age of intricate and multipurpose neural networks has been ushered in by improvements in processing power and the introduction of Big Data technologies. Computers can now recognize patterns, learn from them, and solve difficult problems more quickly than humans ever could thanks to this ground-breaking advancement [10].

Significant improvements have been achieved in various domains, including image classification, language translation, and speech recognition, thanks to DL. Without requiring human assistance, DL has demonstrated leadership in several disciplines, including speech recognition, picture categorization, and pattern detection [18, 19]. Many layers, each utilizing the potential of DL, form the basis of an ANN [20, 21]. Deep neural networks, a subset of neural networks, have layers capable of interpreting complex patterns related to image analysis and textual data [22, 23]. As the field of ML continues to expand quickly, more businesses are utilizing this ground-breaking technology to create creative models [24, 25].

Optimizing traffic signal control to minimize congestion, reduce environmental impact, and improve traffic flow, while considering real-time environmental factors such as air quality, weather, and traffic volume. The main goal of the study is to assess how well the suggested strategy works to reduce traffic congestion and produce the intended results. It also aims to assess the effectiveness of the models used in this study and their ability to reduce reliance on different types of traffic sensors installed on roadways, while the objective of this research is to develop an intelligent traffic Signal control system using ML algorithms and environmental data to enhance responsiveness to dynamic traffic conditions. To run and maintain these sensors, a substantial number of resources is needed. A traffic forecasting model that solely uses data on air pollution might be developed if it becomes feasible to rely mostly on commonly used sensors made for intricate urban traffic situations, thereby making the intricate network of traffic sensors unnecessary. This research contributes to the development of intelligent transportation systems by presenting a novel approach to optimizing traffic signal control using ML and environmental data. The proposed method integrates real-time environmental data with ML algorithms to adapt traffic signal control to current conditions, reducing congestion and minimizing environmental impact. The gap in this research lies in the integration of environmental data, such as air quality and weather conditions, with ML algorithms to optimize traffic signal control. This approach differs from existing methods that solely rely on traffic volume and timing data, providing a more comprehensive and sustainable solution for traffic management. Additionally, the use of ML enables real-time adaptation to changing conditions, improving the efficiency and effectiveness of traffic signal control

This study is organized as follows: Section 2 presents the methodology, the schematic framework, sample dataset data preprocessing and diagnosis model, parameters, and metrics explained. Section 3 presents experimentation results and are discussed. Section 4 presents the conclusion and future works.

2. Methodology

2.1. Predictive analytics in urban environments

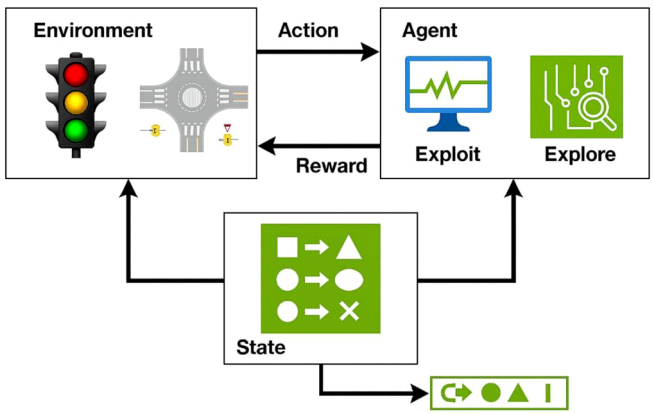
Among the several subfields that make up AI is ML. ML approaches comprise a range of models and algorithms that learn from and adapt to the data they analyze, imitating human decision-making. Based on the data they are trained on these ML models gradually pick up new talents and improve their learning capacity. In the context of traffic management, the control unit at an isolated junction can be viewed as an agent that engages in closed-loop Markov Decision Process interaction with the traffic environment, as shown in Figure 1 [26, 27]. Traffic circumstances (e.g., waiting time, queue length, and total delay) are mapped to the control policy to identify the best course of action, which may include phase shift, cycle length adjustment, and green time extension [28, 29].

On the training set and the testing set, the suggested model showed an accuracy of 72.25% and 85.03%, respectively. The data had a mean absolute error (MAE) of 0.28 and a root mean square error (RMSE) of 0.46. The Naïve Bayes (NB) classifier model's results demonstrate its efficacy in predicting the impact of weather on traffic patterns. The objective of this strategy was to create an Advanced Traveler Information System and an Advanced Traffic Management System for the city of Dhaka. This would allow cars to choose less congested routes, therefore reducing traffic congestion.

The approach of supervised learning regression and classification issues are addressed by K-nearest neighbors (KNN), which assumes that comparable objects are near to one another. Khan et al. [11] tackle the issue of inflexible model designs that fail to account for interactions dependent on time and space. For short-term traffic flow prediction, they suggest the Adaptive-STKNN model, which is based on Adaptive Space and Time and utilizes KNN methodology. The spatiotemporal weights, adaptive spatial neighbors, and time intervals included in this model allow it to fully account for spatial variations in urban traffic. Cross-correlation and autocorrelation functions were employed to determine the optimal spatial and temporal dependencies for each road segment, enabling an accurate assessment of traffic impacts. Then, to improve the effectiveness of candidate neighbor search strategies, distance functions are coupled with adaptive spatiotemporal weights.

Figure 1

An example of how several techniques were used to forecast traffic flow



After several potential neighbors and a weighting variable in the prediction module, this method produces an adaptive spatiotemporal model that incorporates real-time changes in traffic circumstances. In Xu et al. [12], with an emphasis on time series data about traffic conditions on the roads, the Kernel KNN approach is developed. To ascertain the dynamic aspects of traffic, the process entails gathering data on road traffic flow and using reference sequences. To analyze time series data on road traffic conditions, a kernel module is developed that compares and matches data sequences from reference and current data, with a focus on the use of automobiles for transportation.

2.2. Environmental data sources

Incorporating environmental data into traffic signal control optimization is a forward-thinking strategy that aims to balance traffic efficiency with environmental sustainability. Various environmental data sources provide valuable insights into the external conditions that can affect traffic flow and the environmental impact of vehicles. These data sources range from air quality measurements and weather data to noise pollution levels and even real-time emissions data. Understanding these sources and their potential applications is essential to developing smarter, more adaptive traffic control systems. Air quality data is one of the most critical environmental data sources for optimizing traffic signal control in urban areas. Poor air quality is often associated with high levels of vehicle emissions, particularly in densely populated areas. By monitoring air quality in real-time, traffic management systems can adjust signal timings to reduce congestion and emissions in areas where air quality is deteriorating. Air quality data is typically collected using sensors that measure the concentration of various pollutants in the air, such as nitrogen dioxide (NO_2), carbon monoxide (CO), particulate matter (PM2.5 and PM10), and ground-level ozone (O_3). These pollutants are primarily generated by vehicle emissions and can have severe health impacts on urban populations, particularly those with respiratory conditions. Several organizations provide air quality data through networks of sensors and monitoring stations. For instance, the Environmental Protection Agency in the United States operates the Air Quality System, which collects data from thousands of monitoring stations across the country. Other countries have similar systems in place, such as the European Environment Agency's Air Quality e-Reporting system. By integrating air quality data with traffic signal control systems, ML algorithms can be used to prioritize traffic flow in ways that minimize emissions. For example, during periods of poor air quality, the system could give priority to public transportation or low-emission vehicles, while reducing the frequency of signals for high-emission vehicles. Alternatively, traffic could be rerouted away from areas with particularly poor air quality to minimize exposure to harmful pollutants.

Weather conditions have a significant impact on traffic patterns and vehicle performance. Rain, snow, fog, and extreme temperatures can all affect driver behavior, vehicle speed, and road safety. By incorporating weather data into traffic signal control systems, cities can improve traffic flow and safety under various weather conditions. Weather data is collected from a variety of sources, including weather stations, satellites, and radar systems. National meteorological agencies, such as the National Weather Service in the United States and the European Centre for Medium-Range Weather Forecasts, provide real-time weather data and forecasts that can be integrated into traffic management systems. ML algorithms can analyze weather data in conjunction with traffic data to predict how weather conditions will impact traffic flow. For instance, during periods of heavy rain, traffic signal timings could be adjusted to allow for longer stopping distances and slower speeds. Similarly, during periods of extreme heat, traffic signals could be optimized to reduce the amount of time vehicles spend idling at intersections, thereby reducing the risk of overheating and

improving fuel efficiency. In addition to real-time adjustments, weather data can also be used to inform long-term traffic signal planning. For example, historical weather data can be analyzed to identify patterns in traffic behavior during different seasons, allowing for the development of more effective traffic management strategies for specific weather conditions.

Noise pollution is another environmental factor that can be incorporated into traffic signal control optimization. High levels of noise pollution are often associated with heavy traffic and can have negative health effects, including stress, sleep disturbances, and cardiovascular issues. By monitoring noise levels in real-time, traffic management systems can adjust signal timings to reduce noise pollution in sensitive areas, such as residential neighborhoods or near schools and hospitals. Noise pollution data is typically collected using sensors that measure sound levels in decibels (dB). These sensors can be placed at strategic locations throughout a city to monitor noise levels in real-time. In some cases, noise pollution data may also be available from mobile devices or crowd-sourced platforms, where users can report noise levels in their area. By integrating noise pollution data with traffic signal control systems, cities can develop strategies to reduce noise in high-traffic areas. For example, during periods of high noise pollution, traffic signals could be adjusted to reduce the speed of vehicles in sensitive areas, or traffic could be rerouted away from these areas altogether. Additionally, traffic signals could be synchronized to reduce the number of stop-and-go movements, which are a significant source of vehicle noise.

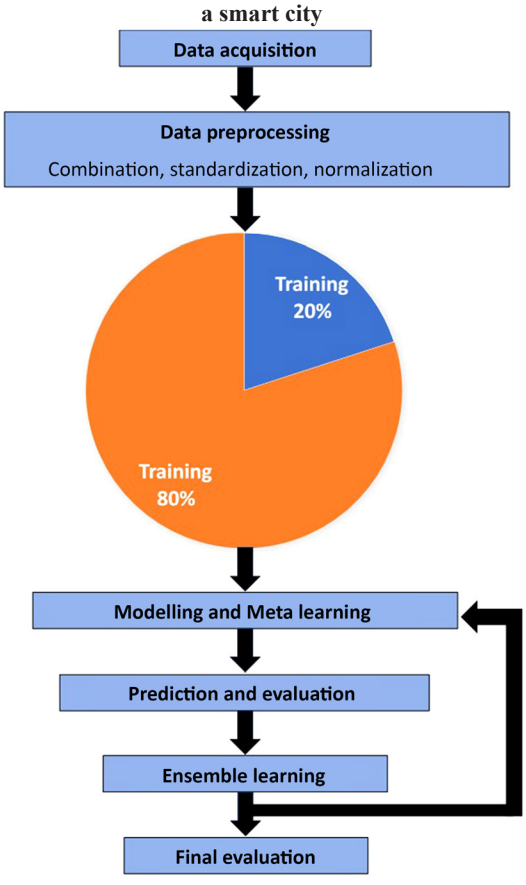
Real-time emissions data provides insights into the environmental impact of vehicles on the road. By monitoring emissions from individual vehicles or groups of vehicles, traffic management systems can adjust signal timings to reduce overall emissions and improve air quality. Emissions data is typically collected using sensors that measure the concentration of pollutants in vehicle exhaust. These sensors can be placed at strategic locations throughout a city, such as at intersections or along major roadways, to monitor emissions in real-time. In some cases, emissions data may also be available from vehicles equipped with onboard diagnostic systems, which can report emissions directly to traffic management systems. By integrating real-time emissions data with traffic signal control systems, cities can develop strategies to reduce emissions in areas with high levels of pollution. For example, during periods of high emissions, traffic signals could be adjusted to prioritize the flow of low-emission vehicles, such as electric cars or public transportation. Additionally, traffic signals could be synchronized to reduce the number of idling vehicles at intersections, which are a significant source of emissions.

In recent years, the proliferation of mobile devices and apps has opened up new opportunities for collecting environmental data through crowd-sourced platforms. Apps that allow users to report traffic conditions, air quality, and noise pollution in real-time provide a valuable supplement to traditional sensor networks. By leveraging crowd-sourced data, traffic management systems can gain a more comprehensive and granular understanding of environmental conditions across a city. Mobile data can be collected through a variety of apps, including those designed specifically for environmental monitoring, as well as more general navigation or health apps. For instance, users of navigation apps like Waze or Google Maps can report traffic incidents, congestion, and other conditions in real-time, providing valuable data that can be used to optimize traffic signal control. Additionally, some apps are designed to collect environmental data passively, such as air quality monitoring apps that use the sensors in smartphones to measure pollutant levels in the air. By integrating crowd-sourced and mobile data with traditional environmental data sources, traffic management systems can develop more responsive and adaptive strategies for optimizing traffic flow and reducing environmental impacts. The performance of the model is greatly impacted by the vital duties of data processing and gathering in any technique. This study made use of a sizable, openly accessible dataset that City Pulse in Aarhus, Denmark, gathered in real time. Two datasets were primarily used: pollution data and traffic intensity data [13]. The city has numerous sensors deployed that gather data on passing cars every five minutes. The air dataset includes details on pollutants emitted by these vehicles, such as particulate matter, ozone (O_3), carbon monoxide (CO), and sulfur dioxide (SO_2). The statistics on traffic, vehicle density, and pollution included 96,000 occurrences of data spanning more than a year. Each instance had characteristics including O_3, CO_2, SO_2, and NO_2 levels, as well as a date indicating the arrival of a vehicle and a vehicle count. Because traffic flow and pollution were used to predict each other, separate statistics were provided for each. By comparing the timestamps of the two databases, pollution levels and vehicle traffic data from the Aarhus website were merged. This study provided insightful information on the dynamic pulse of Aarhus by using real-time data from publicly accessible, open-source sensor datasets.

In addition to pollution statistics (NO_2, CO, SO_2, O_3, latitude, longitude, and information on the distance between two locations), the traffic datasets utilized in the investigations contained information on traffic intensity. The associations between the dataset's properties were examined using a correlation matrix graph (Figure 2). This graph indicates the degree of correlation between the various qualities.

This research does not directly use the vehicle data from the traffic dataset. Rather, they were timestamp-based, therefore they were incorporated into the pollution dataset. The fact that the sensors utilized to collect the data were positioned along the same paths made the combining of these datasets conceivable. There is a direct relationship between the quantity of pollutants released and the number of cars on the road; larger traffic volumes result in higher emissions of CO_2, SO_2, and NO_2.

Figure 2
Leverage on the ML and ensemble models for traffic forecasting in a smart city



Rather than focusing on minute details, the study primarily uses pollution data to develop a model suited for broader metropolitan monitoring, which helps reduce infrastructure costs associated with traffic flow evaluation. By relying on pollution data for generalized predictions, the approach eliminates the need for specialized sensors, offering a financial advantage through reduced sensor requirements. Table 1 provides a sample of the dataset, and Figure 2 illustrates the ensemble models used for smart city traffic forecasting.

2.3. Ensemble methods for regression modeling

In this stage, a three-step procedure was used to create an ensemble model, as shown in Figure 3. The first step involved dividing the data into several bootstrap samples, or samples of size "*B*," by randomly choosing 'B' observations from an initial dataset of size "*N*." Further actions were then done in the following ways:

$$A^n b = a^1 1, a^1 2, \ldots a^1 B, \ldots a^2 1, a^2 2, \ldots a^2 B, \ldots . a^3 1, a^3 2, \ldots a^3 B. \quad (1)$$

After using bootstrapping, *N* independent weak learners were fitted on each dataset in the following phase:

$$W^L = w^1, w^2, w^3, \ldots, W^L, \quad (2)$$

After fitting, the outcomes of every *N*-independent weak model were merged using the following formula to produce an ensemble model with low variance:

$$A^N = 1/n \sum_{i=1}^{N} W^L \quad (3)$$

A^N was an aggregated result after the ensemble.

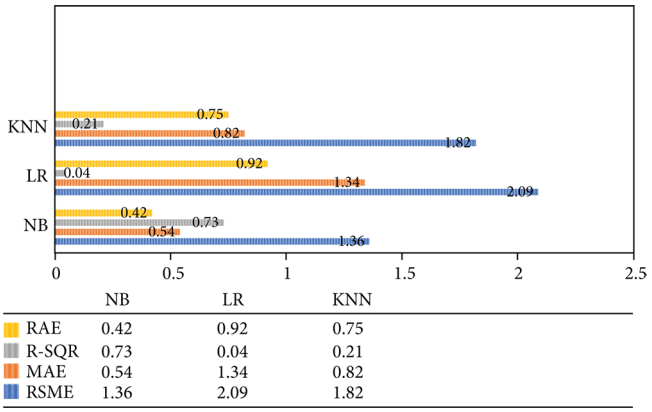
To improve prediction accuracy, this work integrated many models using a variety of bagging ensemble strategies. The final model improved its predictions by incorporating the knowledge from multiple inferior models. The following is a summary of this model's workflow: The bootstrap technique was initially used to divide the dataset into numerous *B*-sized samples. Next, many weak models were trained on various samples from the original dataset at the same time. Conclusions were reached by combining the predictions produced by these weak models using techniques such as averaging.

The following three combinations of bagging ensembles were used in the study:

1) KNN ensemble
2) Random forest ensemble
3) Multi-layer perceptron ensemble

Because it had the lowest error rate out of all three models, the KNN ensemble performed best.

Figure 3
Performance comparison between proposed model and other baseline model



2.4. Evaluation metrics

To evaluate the models, we employed established measures. The most often utilized assessment measures that were employed were *R*-squared (*R*-SQR), MAE, relative absolute error (RAE), and RMSE.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |G_i - GP_i| \quad (4)$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} (G_i - GP_i)^2 \quad (5)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (G_i - GP_i)^2} \quad (6)$$

$$ME = Max|G_i - GP_i| \quad (7)$$

$$R_{Squared} = 1 - \frac{SS_{regression}}{SS_{total}} \quad (8)$$

3. Results and Discussion

3.1. Comparative analysis of regression methods

This study compared the performance of three regression methods—Linear Regression (LR), The KNN, NB, and KNN combined with Logistic Regression (KNN-LR) in predicting continuous outcomes. Our results show that LR outperformed KNN and NB in terms of coefficient of determination (R-squared) and mean squared error. LR achieved an RAE value of 0.92, compared to 0.75 for KNN and 0.42 for NB. Additionally, LR had the highest RMSE of 2.09, indicating better fit and predictive accuracy. The superiority of LR can be attributed to its ability to handle collinearity and feature selection,

Table 1
A sample of the dataset

Ozone	Particular matter	Carbon monoxide	Sulfur dioxide	Nitrogen dioxide	Time stamp	Vehicle count
55	38	31	51	82	8 January 2014 6:45	0
55	42	30	54	79	8 January 2014 6:50	0
50	38	29	51	82	8 January 2014 6:55	0
47	36	28	56	80	8 January 2014 7:00	0
42	41	32	54	75	8 January 2014 7:05	0
41	37	27	54	79	8 January 2014 7:10	0
37	42	24	57	81	8 January 2014 7:15	0

which is critical in high-dimensional data. NB, on the other hand, showed poor performance due to over-regularization. These findings suggest that LR is a suitable choice for regression analysis, especially when dealing with complex data. However, further research is needed to validate these results and explore the applications of LR in different domains. To identify the best model for traffic forecasting—that is, a model that can accurately estimate traffic flow or vehicle count this section compares the several regression models that were employed in the experiment. The R-squared (R-SQR) values of the models are used to assess how well they work. These values show how much of the response variable's variation is explained by the regression model. A perfect match is shown by an ideal R-SQR value of 1, which shows very little fluctuation between observed and anticipated values. The KNN model performed better in this investigation, as evidenced by its greatest R-SQR score.

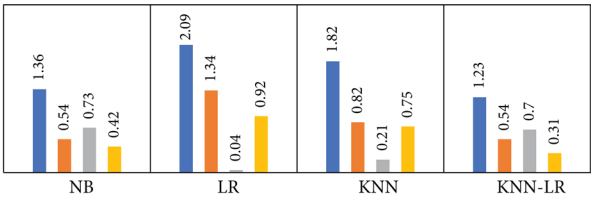
The use of two widely accepted error measures, MAE and RMSE, to evaluate the model's performance about multiple baseline models is demonstrated in Figure 3. Lower RMSE and MAE values indicate improved model accuracy. These measures quantify the difference between expected and actual values. The model performed better than the other baseline models, as seen in Figure 4, with the lowest error rates in comparison.

The study's results were also compared with those of the foundational paper [14]. Findings indicated that bagging an ensemble instead of boosting one in the study by Jereb et al. [14], decreased by more than 30% the mistake rate. The two research are comparable because they both used the same dataset and focused on related topics. Most ML initiatives start with comparable data pre-processing procedures. Jereb et al. [14] employed the KNN and Elastic Net models instead of ANNs and Decision Trees. Results from the base publication are included in Figure 4 for reference, making it easier to compare the effectiveness of the suggested strategy with previous studies.

3.2. Sources of systematic error

This research is currently limited to using data from Lagos City. It is important to understand that the switch to electric vehicles will not happen quickly, even though an increase in their presence is predicted to dramatically lower pollution levels in large cities. The model might need to adjust to new traffic and pollutant patterns during this phase of transition. The world is moving towards electric cars due to air quality issues. For example, starting in 2020, several nations have set aggressive goals to sell more than seven million electric vehicles yearly [14]. Because the model used in this study is dependent on the pollution that conventional cars generate, gradually substituting electric vehicles for conventional ones might potentially undermine the model's long-term viability. Optimizing traffic signal control using ML and environmental data holds great promise for improving urban traffic efficiency and sustainability. However, achieving reliable and accurate optimization is not without challenges. Systematic errors recurring biases or inaccuracies in the data or models can undermine the effectiveness of these systems. This discussion explores key sources of systematic error that can impact the optimization process and suggests potential mitigation strategies. ML models are only as good as the data used to train them. One of the most significant sources of systematic error is bias in the training data. In the context of traffic signal optimization, if the historical traffic and environmental data used to train the model are not representative of all conditions, the model may make inaccurate predictions. For example, if the training data predominantly reflects traffic patterns during fair weather, the model may perform poorly under adverse weather conditions, such as heavy rain or snow. To mitigate this, it is crucial to ensure that the training data is diverse and representative of a wide range of scenarios, including different times of day, days of the week, seasons, weather conditions, and special events. Data augmentation techniques, such as artificially generating scenarios or under-sampling overrepresented cases, can also help balance the dataset. Another common source of systematic error comes from inaccuracies in the environmental data used for decision-making. Sensors that monitor air quality, weather, and emissions may malfunction, degrade over time, or provide readings with inherent inaccuracies due to calibration issues or environmental interference. For instance, air quality sensors may provide biased data due to local sources of pollution (e.g., nearby industrial emissions), while weather stations may not accurately capture microclimates within a city. To reduce systematic errors from environmental data inaccuracies, sensor maintenance and calibration must be prioritized. Implementing sensor redundancy deploying multiple sensors in critical areas can help cross-validate data and ensure reliability. Additionally, using predictive models to interpolate or smooth data from nearby sensors can mitigate gaps caused by sensor failures or outlier readings. Temporal misalignment between different data sources can introduce systematic errors into the optimization process. Traffic signal control systems often rely on real-time data to adjust timings dynamically. However, if there is a delay in receiving environmental data (e.g., due to communication lags or processing times), the system may base its decisions on outdated or incomplete information. This can result in suboptimal traffic signal adjustments, particularly in rapidly changing conditions such as during sudden weather shifts or traffic incidents. Addressing temporal misalignment requires synchronizing the data streams from various sources as closely as possible. Implementing real-time data processing pipelines and reducing latency in data transmission can help ensure that the traffic control system responds to the most current conditions. Additionally, ML models can be trained to account for potential delays by predicting short-term future states rather than reacting solely to the present. Overfitting occurs when a ML model becomes too complex and begins to "memorize" the training data rather than generalizing from it. This can lead to systematic errors, particularly when the model encounters real-world scenarios that deviate from the patterns in the training data. In traffic signal optimization, overfitting could result in a model that performs well under specific conditions but fails to adapt to unusual traffic patterns, such as those caused by accidents, construction, or public events. To prevent overfitting, regularization techniques such as L1/L2 regularization or dropout can be employed to limit model complexity. Cross-validation methods can also be used to evaluate the model's performance on unseen data, ensuring that it generalizes well to different scenarios. Furthermore, maintaining a dynamic model that is periodically retrained on new data can help the system adapt to evolving traffic patterns and environmental conditions. In real-world traffic management systems, human operators may occasionally override automated decisions made by ML algorithms, particularly

Figure 4
Performance comparison of baseline models with the ensemble ML model



	NB	LR	KNN	KNN-LR
RSME	1.36	2.09	1.82	1.23
MAE	0.54	1.34	0.82	0.54
R-SQR	0.73	0.04	0.21	0.7
RAE	0.42	0.92	0.75	0.31

during emergencies or special events. Although human intervention can be valuable in certain situations, inconsistencies in when and how these interventions occur can introduce systematic errors into the optimization process. For example, if operators frequently intervene in specific types of scenarios, the ML model may not have sufficient opportunities to learn from these situations, leading to degraded performance in similar future events. To address this, a clear protocol for human intervention should be established, with operators logging the reasons for their actions. This data can then be fed back into the ML model to improve its decision-making capabilities. Additionally, providing operators with decision support tools that highlight the rationale behind the algorithm's suggestions can reduce unnecessary interventions and allow the model to operate autonomously more effectively. Systematic errors in optimizing traffic signal control using ML and environmental data arise from several sources, including biases in training data, inaccuracies in environmental data, temporal misalignment, overfitting, and inconsistent human intervention. Addressing these issues requires careful attention to data quality, model robustness, and system integration. By mitigating these sources of error, traffic signal optimization systems can more reliably improve traffic flow, reduce emissions, and enhance the sustainability of urban transportation networks.

3.3. Comparison of models on different dataset sizes

An 85,000-item dataset was used to train and test the bagging KNN model, which turned out to be the top-performing model. To determine whether the model's performance held up with fewer instances, a random subset of 20,000 instances was chosen from the dataset. Models were then trained on this smaller dataset. After that, the model was put to the test, and Figures 5 and 6 show the individual findings for 85,000 and 20,000 cases, respectively. The results show that even with a lower dataset, the model continues to function as expected. The model's ability to learn new patterns as it processes more data can be used to explain any little variance in the findings and improve its overall performance.

3.4. Threat to validity

As of right now, this analysis has only used data from Lagos City. Even though more cars are being driven, which should significantly lower pollution levels in and around large cities, it is important to understand that the switch to electric vehicles will not happen quickly. In this phase of transition, the model might adjust to new patterns. Air pollution concerns are driving a global shift in the usage of electric vehicles. For example, starting in 2020, certain nations have set high goals to sell over seven million electric vehicles yearly. Because the plan outlined in this study depends on the pollution that conventional vehicles emit, the progressive substitution of electric vehicles for them may provide a problem for their long-term viability.

Figure 5
Comparative of different bagging KNN model of 85,000 instances datasets

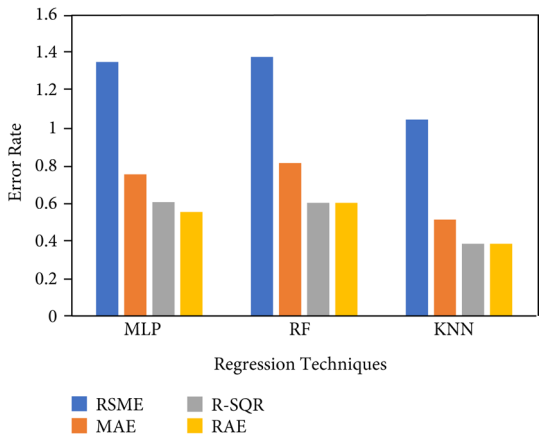
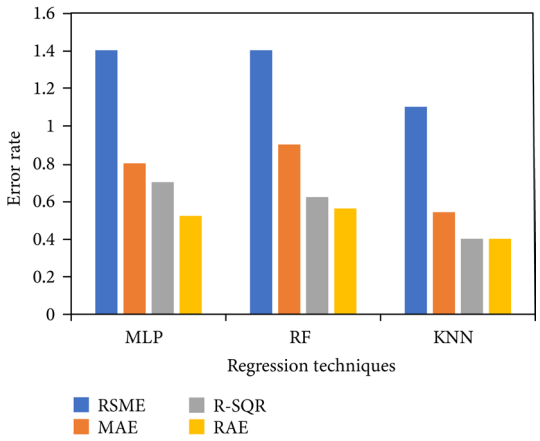


Figure 6
Comparative of different bagging KNN model of 20,000 instances datasets



4. Conclusions

Accurate traffic flow prediction is a key objective for smart cities and can significantly enhance drivers' ability to manage their trips. This study utilized pollution and traffic data from Aarhus, Germany, to estimate traffic flow. The dataset was subjected to a variety of conventional ML techniques to determine the best accurate model; KNN produced the lowest values of MAE and RMSE. After assessing the output of conventional models, bagging and stacking ensemble techniques were used to increase accuracy even more. Using bootstrapping with replacement, the dataset was separated into samples, which were subsequently utilized to train several homogenous models. Combining the results from these models resulted in a robust bagging ensemble model, with the KNN bagging ensemble being the most accurate combination. The capacity of KNN can handle nonlinear data is responsible for its outstanding performance. But if there are too few nearest neighbors (K) or too many, KNN may underfit or overfit. The suggested bagging ensemble strategy reduced error rates by thirty percent when this experimental study was contrasted with earlier research that employed boosting for traffic flow prediction in smart cities.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Author Contribution Statement

Edidiong E. Akpan: Conceptualization, Supervision, Project administration. **Oluwatobi Akinlade:** Data curation, Writing – original draft. **Oluwaseyi O. Alabi:** Methodology, Formal analysis, Resources, Data curation, Visualization. **Oluwasesan A. David:** Software, Validation. **Oluwaseyi F. Afe:** Software, Investigation. **Sunday Adeola Ajagbe:** Conceptualization, Resources, Writing – review & editing.

References

[1] Shahid, N., Shah, M. A., Khan, A., Maple, C., & Jeon, G. (2021). Towards greener smart cities and road traffic forecasting using air pollution data. *Sustainable Cities and Society*, *72*, 103062. https://doi.org/10.1016/j.scs.2021.103062

[2] Garg, D., Chli, M., & Vogiatzis, G. (2019). A deep reinforcement learning agent for traffic intersection control optimization. In *2019 IEEE Intelligent Transportation Systems Conference*, 4222–4229. https://doi.org/10.1109/ITSC.2019.8917361

[3] Kővári, B., Knáb, I. G., Esztergár-Kiss, D., Aradi, S., & Bécsi, T. (2024). Distributed highway control: A cooperative reinforcement learning-based approach. *IEEE Access*, *12*, 104463–104472. https://doi.org/10.1109/ACCESS.2024.3434965

[4] Khozam, S., & Farhi, N. (2023). Deep reinforcement Q-learning for intelligent traffic control in mass transit. *Sustainability*, *15*(14), 11051. https://doi.org/10.3390/su151411051

[5] Qin, H., Zhang, W., & Tian, R. (2024). Collaborative control method of transit signal priority based on cooperative game and reinforcement learning. In *2024 IEEE 4th International Conference on Electronic Technology, Communication and Information*, 537–542. https://doi.org/10.1109/ICETCI61221.2024.10594507

[6] Zohuri , B. & Rahmani, F. M. (2019). Artificial intelligence driven resiliency with machine learning and deep learning components. *Journal of Communication and Computer*, *15*(1), 1–13.

[7] Zai, W., & Yang, D. (2023). Improved deep reinforcement learning for intelligent traffic signal control using ECA_LSTM network. *Sustainability*, *15*(18), 13668. https://doi.org/10.3390/su151813668

[8] Zhang, K., Xu, H., Pan, B., & Zheng, Q. (2024). Explicit coordinated signal control using soft actor–critic for cycle length determination. *IET Intelligent Transport Systems*, *18*(8), 1396–1407. https://doi.org/10.1049/itr2.12519

[9] Liu, D., & Li, L. (2023). A traffic light control method based on multi-agent deep reinforcement learning algorithm. *Scientific Reports*, *13*(1), 9396. https://doi.org/10.1038/s41598-023-36606-2

[10] Bhardwaj, A., Iyer, S. R., Ramesh, S., White, J., & Subramanian, L. (2023). Understanding sudden traffic jams: From emergence to impact. *Development Engineering*, *8*, 100105. https://doi.org/10.1016/j.deveng.2022.100105

[11] Sreejith, K., Mathi, S., & Pradeep, P. (2024). Beyond sensors: Intellisignal's map-integrated intelligence in traffic flow optimization. *IEEE Access*, *12*, 39028–39040. https://doi.org/10.1109/ACCESS.2024.3375335

[12] Noaeen, M., Naik, A., Goodman, L., Crebo, J., Abrar, T., Abad, Z. S. H., ..., & Far, B. (2022). Reinforcement learning in urban network traffic signal control: A systematic literature review. *Expert Systems with Applications*, *199*, 116830. https://doi.org/10.1016/j.eswa.2022.116830

[13] Affolder, N. (2019). Transnational environmental law's missing people. *Transnational Environmental Law*, *8*(3), 463–488. https://doi.org/10.1017/S2047102519000190

[14] Jereb, B., Stopka, O., & Skrúcaný, T. (2021). Methodology for estimating the effect of traffic flow management on fuel consumption and CO2 production: A case study of celje, slovenia. *Energies*, *14*(6), 1673. https://doi.org/10.3390/en14061673

[15] Qu, G., Wu, H., Li, R., & Jiao, P. (2021). DMRO: A deep meta reinforcement learning-based task offloading framework for edge-cloud computing. *IEEE Transactions on Network and Service Management*, *18*(3), 3448–3459. https://doi.org/10.1109/TNSM.2021.3087258

[16] Abdurrahman, M. I., Chaki, S., & Saini, G. (2020). Stubble burning: Effects on health & environment, regulations and management practices. *Environmental Advances*, *2*, 100011. https://doi.org/10.1016/j.envadv.2020.100011

[17] Agrahari, A., Dhabu, M. M., Deshpande, P. S., Tiwari, A., Baig, M. A., & Sawarkar, A. D. (2024). Artificial intelligence-based adaptive traffic signal control system: A comprehensive review. *Electronics*, *13*(19), 3875. https://doi.org/10.3390/electronics13193875

[18] Adekunle, T. S., Alabi, O. O., Lawrence, M. O., Adeleke, T. A., Afolabi, O. S., Ebong, G. N., ..., & Bamisaye, T. A. (2024). An intrusion system for internet of things security breaches using machine learning techniques. *Artificial Intelligence and Applications*, *2*(3), 165–171. https://doi.org/10.47852/bonviewAIA42021780

[19] Khan, A., Fouda, M. M., Do, D.-T., Almaleh, A., & Rahman, A. U. (2023). Short-term traffic prediction using deep learning long short-term memory: Taxonomy, applications, challenges, and future trends. *IEEE Access*, *11*, 94371–94391. https://doi.org/10.1109/ACCESS.2023.3309601

[20] Akande, T. O., Alabi, O. O., & Oyinloye, J. B. (2024). A review of generative models for 3D vehicle wheel generation and synthesis. *Journal of Computing Theories and Applications*, *1*(4), 368–385. https://doi.org/10.62411/jcta.10125

[21] Chen, X., Chen, H., Yang, Y., Wu, H., Zhang, W., Zhao, J., & Xiong, Y. (2021). Traffic flow prediction by an ensemble framework with data denoising and deep learning model. *Physica A: Statistical Mechanics and Its Applications*, *565*, 125574. https://doi.org/10.1016/j.physa.2020.125574

[22] Akhtar, M., & Moridpour, S. (2021). A review of traffic congestion prediction using artificial intelligence. *Journal of Advanced Transportation*, *2021*(1), 8878011. https://doi.org/10.1155/2021/8878011

[23] Jiang, W., & Luo, J. (2022). Graph neural network for traffic forecasting: A survey. *Expert Systems with Applications*, *207*, 117921. https://doi.org/10.1016/j.eswa.2022.117921

[24] Nunez, I., & Nehdi, M. L. (2021). Machine learning prediction of carbonation depth in recycled aggregate concrete incorporating SCMs. *Construction and Building Materials*, *287*, 123027. https://doi.org/10.1016/j.conbuildmat.2021.123027

[25] Li, Z., Yu, H., Zhang, G., Dong, S., & Xu, C.-Z. (2021). Network-wide traffic signal control optimization using a multi-agent deep reinforcement learning. *Transportation Research Part C: Emerging Technologies*, *125*, 103059. https://doi.org/10.1016/j.trc.2021.103059

[26] Khan, N. U., Shah, M. A., Maple, C., Ahmed, E., & Asghar, N. (2022). Traffic flow prediction: An intelligent scheme for forecasting traffic flow using air pollution data in smart cities with bagging ensemble. *Sustainability*, *14*(7), 4164. https://doi.org/10.3390/su14074164

[27] Xu, D., Wei, C., Peng, P., Xuan, Q., & Guo, H. (2020). GE-GAN: A novel deep learning framework for road traffic state estimation. *Transportation Research Part C: Emerging Technologies*, *117*, 102635. https://doi.org/10.1016/j.trc.2020.102635

[28] Wilding, M., Benmore, C., Weber, R., Parise, J., Lazareva, L., Skinner, L., ..., & Tamalonis, A. (2015). Exploring the structure of high temperature, iron-bearing liquids. *Materials Today: Proceedings*, *2*, S358–S363. https://doi.org/10.1016/j.matpr.2015.05.050

[29] Nimesh, V., Sharma, D., Reddy, V. M., & Goswami, A. K. (2020). Implication viability assessment of shift to electric vehicles for present power generation scenario of India. *Energy*, *195*, 116976. https://doi.org/10.1016/j.energy.2020.116976

How to Cite: Akpan, E. E., Akinlade, O., Alabi, O. O., David, O. A., Afe, O. F., & Ajagbe, S. A. (2025). Optimizing Traffic Signal Control Using Machine Learning and Environmental Data. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA52024199