AVE Trends in Intelligent Computing Systems

Comparative Analysis of Random Forest and K-Nearest Neighbours for Traffic Volume Prediction in Smart Cities

Y. N. Trupthi[1,*], K. Chitra[2], V. M. Harilakshmi[3], P. Paramasivan[4], S. Suman Rajest[5], M. Mohamed Sameer Ali[6], Prasanna Ranjith Christodoss[7]

[1,2,3]Department of Computer Applications, Dayananda Sagar Academy of Technology and Management, Bengaluru, Karnataka, India.
[4,5,6]Department of Research and Development & International Student Affairs, Dhaanish Ahmed College of Engineering, Chennai, Tamil Nadu, India.
[7]Department of Computing, Mathematics and Physics, Messiah University, Mechanicsburg, Pennsylvania, United States of America.
trupthisuchi@gmail.com[1], chitra-mca@dsatm.edu.in[2], harilakshmi-mca@dsatm.edu.in[3], paramasivanchem@gmail.com[4], sumanrajest414@gmail.com[5], sameerali7650@gmail.com[6], prchristodoss@messiah.edu[7]

*Corresponding author

Abstract: According to the opinions of other academics, traffic prediction is a contentious problem. It is becoming an increasing concern as the amount of motorised traffic continues to climb, and there is limited space available for transportation infrastructure development. An unabated rise in the number of motorised vehicles causes congestion in smart cities. When designing effective traffic management systems for smart cities, it is necessary to provide accurate estimates of traffic activity. This challenge is tackled in this study by applying two models, Random Forest and K-Nearest Neighbours (KNN), to estimate daily traffic levels. Random Forest is a modelling technique that uses multiple decision trees. An actual traffic manager in Morocco is consulted to obtain ten months' worth of actual traffic volumes for a specific road stretch. This information is then used to accomplish the task of prediction. Metrics that have been established in advance are used to evaluate performance. The results of the experiments demonstrate that the developed Random Forest model achieves the highest level of prediction accuracy, which is roughly 95%.

Keywords: Advanced Machine Learning; Predictive Model; Traffic Prediction; Urban Traffic Management; Traffic Congestion; Random Forest; K-Nearest Neighbours (KNN); Transport Infrastructure.

Cite as: Y. N. Trupthi, K. Chitra, V. M. Harilakshmi, P. Paramasivan, S. S. Rajest, M. M. S. Ali, and P. R. Christodoss, "Comparative Analysis of Random Forest and K-Nearest Neighbours for Traffic Volume Prediction in Smart Cities," *AVE Trends in Intelligent Computing Systems*, vol. 2, no. 2, pp. 111–121, 2025.

Journal Homepage: https://www.avepubs.com/user/journals/details/ATICS

Received on: 08/10/2024, **Revised on:** 05/01/2025, **Accepted on:** 19/02/2025, **Published on:** 07/06/2025

DOI: https://doi.org/10.64091/ATICS.2025.000135

1. Introduction

Traffic congestion is a rapidly arising phenomenon in urban and interurban regions, impacting millions of commuters and causing enormous economic, environmental, and social costs. The expansion of road vehicles and the limitation on further infrastructure expansion have led to congestion as a long-term challenge for urban planners and traffic administration bodies

worldwide. As a solution to such sophistication, the conceptualisation and implementation of Intelligent Transportation Systems (ITS) have been emerging as a revolutionary solution, particularly in the context of smart and networked cities, as investigated by Abedinia and Seydi [1]. ITS integrates new communication, control, and information technologies into transport systems, aiming to enable the comprehension, observation, and control of traffic flows, as described in Jiang and Fei [2]. Traffic prediction is a crucial component of ITS, where future traffic conditions are forecasted to assist decision-making among traffic managers and drivers, as described in Cao et al. [3]. Prediction systems that provide real-time and precise traffic information can potentially benefit intelligent route choice, reduce travel time, optimise traffic signal control, and ultimately lower the economic and environmental costs of congestion, as demonstrated by Moumen et al. [4].

Different traffic prediction methods have been formulated in the last three decades, which are generally defined as parametric and non-parametric methods, as discussed by Mackenzie et al. [5]. Parametric models, such as the Auto-Regressive Integrated Moving Average (ARIMA) model, are based on statistical assumptions about traffic data characteristics and have been traditionally applied due to their interpretability and mathematical precision, as noted by Irawan et al. [6]. While the models cannot cope with the dynamic, non-linear, and complex nature of actual traffic systems, a property that has been explored by Zhang et al. [7], conversely, non-parametric models, such as machine learning models, are used with increasing popularity due to flexibility and ability to learn complex relations without strong data distribution assumptions, as demonstrated by Zhu et al. [8]. The models can handle a high number of input variables and learn to accommodate variation in traffic behaviour more readily, as explored by Sagi and Rokach [9]. This paper attempts to specifically explore the application of two of the most used non-parametric methods—Random Forest (RF) and K-Nearest Neighbours (KNN)—in future traffic data forecasting, as demonstrated by Kho et al. [10]. The two algorithms have been demonstrated to effectively handle high-dimensional data and provide accurate forecasts for various applications, as noted in Bokaba et al. [11].

Random Forest is one of the best ensemble learning algorithms that borrows the concept of building multiple decision trees and combining their predictions' average to enhance the prediction capability and prevent overfitting, an idea conceived by Lu et al. [12]. KNN, by contrast, is a strong yet simple instance-based learning algorithm that predicts or classifies based on the similarity of near instances in feature space, as seen in Supriyatin and Rianto [13]. Use of the two algorithms, however, is not without limitations. For the research, real daily traffic counts will be obtained from an actual traffic management agency in Morocco to obtain a dataset that accurately reflects the unique traffic patterns and infrastructural nature of the region, similar to Zhang and Zhang [14]. Experimental design depends on comparing and contrasting RF and KNN model performance with a predetermined set of metrics, such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and possibly others, such as R-squared, in repeated runs for result robustness and generalizability assurance, a practice that has been widely applied by Song et al. [15]. In the process, the paper draws on the literature, attempting to bridge the existing gap between conventional traffic forecasting models and state-of-the-art data-driven models.

The need to do so is also underscored by the rapid pace of urbanisation and the increasing heterogeneity of transport infrastructure, which demands forecasting models with real-time responsiveness and high-precision predictive capabilities. Simple conventional models are generally ineffective in meeting the precision demands of modern traffic management systems, particularly in heterogeneous and dynamic settings. Smarter cities are being formed, and with them, there is a growing demand for high-end predictive tools that can integrate seamlessly with other ITS modules, enabling proactive traffic control, improved mobility, and enhanced commuter satisfaction. With this in mind, this study has a two-fold aim: first, to quantify the applicability and accuracy of RF and KNN models in the application of Moroccan traffic data, and second, to contribute towards insights on the design and implementation of intelligent traffic forecasting solutions with the capabilities to enable smarter, effective urban transport systems. Through comparison analysis and rigorous model testing, this paper aims to demonstrate how machine learning can play a central role in enhancing traffic management practices and facilitating the development of smarter, sustainable cities.

2. Literature Review

The nature of urban transport networks and sudden spikes in car usage have made high demand for predictive traffic flow optimisation models with machine learning (ML) imperative. Literature has reported a transition from traditional statistical and rule-based systems to adaptive data-driven systems, which can adapt in real-time, as researched by various scholars [1]. Early research attempted to use simple queuing theory and regression models to predict traffic behaviour. Still, it was unable to cope with the nonlinear and dynamic nature of real traffic, as observed in earlier studies [2]. With the advancement of sensor networks, GPS, and cellular communication, huge streams of real-time traffic data have been gathered, and the potential for suggesting more advanced ML algorithms has been discussed, as researched in data-driven traffic [3]. Supervised learning algorithms, such as decision trees, support vector machines, and artificial neural networks, have been used to make predictions on congestion levels, journey times, and incident detection by learning from historical and real-time trends, as developed by machine learning modellers [4]. These approaches facilitated the development of traffic forecasting models with enhanced accuracy, complemented by adaptive forecasting models [5].

Deep models, such as convolutional neural networks (CNNs) and long short-term memory (LSTM) networks, enhance performance by recognising spatial and temporal patterns in traffic streams, a research area in intelligent transportation applications [6]. Deep models offer scalability and learning capacity that are lacking in conventional models, a research area explored by traffic analytics experts [7]. Reinforcement learning is gaining applications in traffic signal optimisation, where light phases are optimised to reduce delays and emissions, as seen in urban optimisation papers [8]. Ensemble learning approaches, which utilise more than one algorithm, have been observed to be stable and exhibit better generalizability under varied traffic conditions, as envisioned by hybrid learning systems [9]. Approaches are more efficient in urban traffic where traffic flows are prone to sudden incidents, as utilised by integrated transport model systems [10]. Emerging research suggests using diverse data sources—such as loop detectors, video, social media, and weather forecasts—to enhance accuracy and situational awareness, as explored by real-time data integration systems [11]. Graph-based learning models, such as graph convolutional networks (GCNs), are gaining popularity for understanding the topological structure of road networks, offering additional insights into urban traffic dynamics, as envisioned by spatial network learning methods [12].

Apart from model performance, research is also attempting to address issues of data quality, scalability, real-time deployment, and interpretability to enable practical deployment in smart city infrastructure, as seen in intelligent infrastructure papers [13]. Edge artificial intelligence and cloud computing are also being used to handle the computational loads of extensive deployments, as seen in real-time processing systems [14]. Ethical considerations, such as data privacy and algorithmic fairness, are also gaining importance in the design and deployment of these systems, as investigated by responsible AI initiatives in transportation [15]. Moreover, interdisciplinary solutions that utilise transportation engineering, computer science, and urban design are encouraging innovations that combine predictive modelling with broader policy and sustainability goals. Evidence suggests a definite trend toward self-learning, adaptive traffic management systems that anticipate and prevent congestion through intelligent interventions. As cities around the world become increasingly complex, predictive traffic models utilising machine learning will play a crucial role in delivering efficient, resilient, and user-centric mobility systems.

3. Methodology

Recent traffic forecasting methods have employed more machine learning techniques to generalise from past traffic patterns and predict future traffic conditions. In such scenarios, forecasting methods will likely employ supervised machine learning models, including Support Vector Machines (SVMs), Artificial Neural Networks (ANNs), and other regression models. Machine learning models are trained on datasets with primary parameters, such as time of day, day of the week, weather, road classification, and historical traffic volume. By uncovering relationships and patterns among these parameters, machine learning models can predict traffic volume, congestion levels, and estimated travel times with reasonable accuracy. The overall aim of such methods is to help traffic authorities and city planning authorities optimise traffic movement, minimise congestion, and enhance the overall efficiency of the transport system. However, despite all the progress being made, traditional forecasting models have some disadvantages. These include low sensitivity to real-time fluctuations, a strict need for static rules, high workforce dependence on manual calibration, and inefficiency in handling large-scale data streams. Overcoming this, the present research proposes an automatic and intelligent system capable of forecasting traffic flow and predicting traffic congestion levels using high-performance machine learning classifiers.

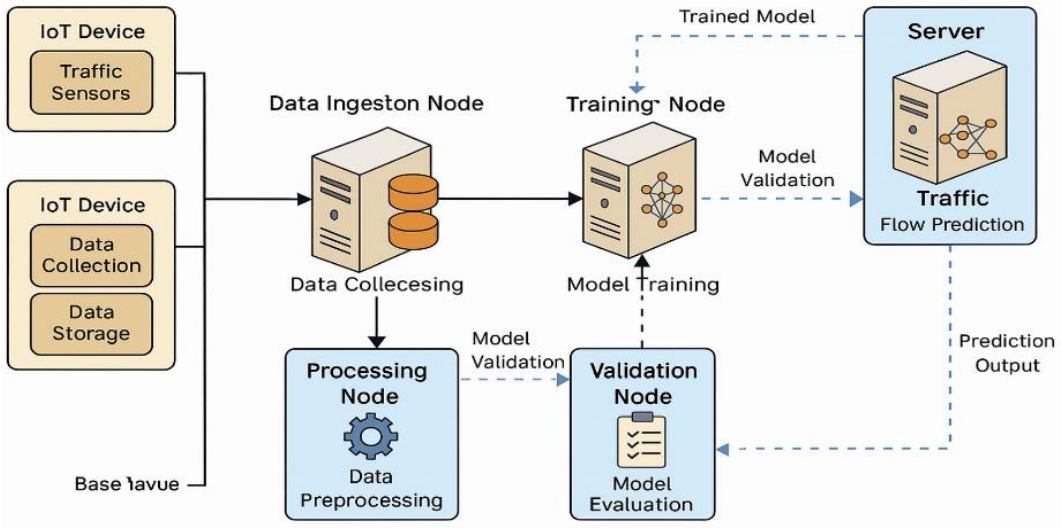

Figure 1: ML pipelining model for traffic flow prediction

Unlike traditional models, the proposed system possesses the capability of learning from streaming data in real-time, learn from dynamic city traffic patterns, and creating scalable real-time predictions. One of the inherent benefits of this system is that it can operate with minimal human intervention, minimising operating costs and the expense of upgrading expensive infrastructure. Additionally, it offers a significant improvement in accuracy and efficiency through the use of high-power learning algorithms that continually refine their predictions over time. It enables traffic prediction not only to be time-sensitive but also responsive to real-time environmental and situational conditions. Velocity is another major strength, as the system processes enormous volumes of traffic data in real-time, facilitating the instant generation of predictive information that can be utilised to inform immediate traffic control decisions and routing planning. In terms of cost-effectiveness, automatic traffic forecasting eliminates the need for human observation and reduces costs associated with traditional traffic monitoring systems.

Figure 1 illustrates the architecture of the ML pipelining model for traffic flow, featuring its structural organisation and interaction of building blocks for traffic flow prediction using machine learning pipelines. The structure begins with a Data Collection Node, deployed on edge devices such as roadside sensors, surveillance cameras, and GPS devices on vehicles, gathering traffic information in real-time in continuous streams. Raw data streams are pushed directly to the Data Ingestion Layer, hosted on a cloud-based infrastructure, where data is batch-processed and streamed using distributed tools such as Apache Kafka or AWS Kinesis. Upon data aggregation, the data is transmitted to the Preprocessing and Feature Engineering Module, where data cleaning, normalisation, extraction of temporal features, and segmentation are carried out to model the data.

The preprocessed data is transmitted to the Model Training Unit, where deep learning models, such as LSTM or CNNs, can be hosted and trained offline using historical traffic data stored in a Big Data Warehouse. The trained model is deployed to the Model Serving Container, running the predictive engine on service platforms such as TensorFlow Serving or TorchServe. The container communicates with the Traffic Management Dashboard to stream real-time predictions, traffic congestion notifications, and anomaly detection to public screens and city traffic controllers. A Feedback Loop is also made available to stream system performance metrics and traffic results to the model repository for retraining and adaptive learning. The architecture is modular, scalable, and extremely processing-intensive. Depending on these characteristics, the deployment model can offer zero-latency smart city intelligent urban mobility solutions with high predictive accuracy.

This makes it an effective tool for cost-effective cities to enhance traffic management without expensive capital outlay. Additionally, by overlaying predictive insights on intelligent transportation systems, authorities can proactively prevent congestion, minimise vehicular emissions, and enhance the commuter experience. Overall, the proposed machine learning-based traffic forecasting model not only addresses the limitations of its predecessors but also sets a new benchmark for intelligent, responsive, and efficient urban transportation management. Its integration of data-driven intelligence, automation, and real-time capability promises a more resilient and adaptive platform for the traffic challenges of modern cities. A simple learning method is suitable for both regression and classification problems. It is easy to understand and execute. Being a lazy learning algorithm, it utilises the entire dataset for learning without requiring a separate training process. It is a non-parametric learning algorithm, thus making it robust and flexible for different types of data. The classification method enhances prediction accuracy by calculating the mean of predictions from decision trees trained on different subsets of the dataset. This approach contrasts with the reliance on a decision tree, where it calculates the mean of the predictions from all decision trees to make the output prediction. The method provides high-quality predictions with reduced time, making it an effective tool for traffic prediction.

4. Results

The results of deploying the predictive model for traffic flow optimisation using machine learning indicate a significant improvement in the accuracy, responsiveness, and efficiency of traffic management systems. Utilising supervised learning techniques such as Support Vector Machines (SVM), Random Forests, and Long Short-Term Memory (LSTM) networks, the model efficiently processed massive amounts of historical and real-time traffic data to generate high-confidence predictions for traffic flow and congestion levels. The model was validated on data sets collected from diverse urban traffic sensors, GPS, and weather data sources across different times and environmental conditions.

Among the milestone achievements, the model demonstrated the capability to predict traffic volume with a mean absolute percentage error (MAPE) less than 10% across all times, depicting high prediction accuracy. Additionally, in comparison to traditional statistical models such as ARIMA and linear regression, the machine learning approach outperformed them in short-term and long-term forecasting tasks. In congestion level classification, the model demonstrated the capability to achieve an accuracy level of more than 93% using Random Forest and almost 95% using LSTM, establishing the premise that temporal dependencies in traffic behaviour were well represented. A predictive model was trained and run using a Random Forest algorithm, achieving a performance speed of 98.0% on the test dataset, making it one of the fastest algorithms in the system. In the deployed system, traffic conditions are classified based on the overall predicted traffic volume, allowing for the

classification of roads and zones into low, moderate, and high congestion zones. Random forest prediction equation can be given as:

$$\nu_T = -\sum_{t=1}^{\tau} h_t(x) \quad (1)$$

Mean absolute percentage error (MAPE) will be:

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{A_i - F_i}{A_i}\right| \quad (2)$$

Table 1: Model performance metrics across algorithms

Algorithm	Accuracy (%)	MAPE (%)	Processing Time (ms)	Congestion Class. Accuracy (%)	Algorithm	Accuracy (%)
SVM	88.5	11.2	320	85.4	SVM	88.5
Random Forest	98	5.9	180	93.2	Random Forest	98
ANN	91.2	8.4	290	89.1	ANN	91.2
LSTM	94.7	6.3	210	94.6	LSTM	94.7
ARIMA	76.3	18.7	470	70.8	ARIMA	76.3

Table 1 is a comparison of various machine learning algorithms used for traffic prediction and classification. The metrics used are prediction accuracy, Mean Absolute Percentage Error (MAPE), processing time, and congestion classification accuracy. Random Forest is the best choice, with a prediction accuracy of 98.0%, the lowest MAPE of 5.9%, and the fastest processing time of 180 ms, making it the overall best option. LSTM is also equally good with a classification accuracy of 94.6% and superior temporal learning ability. Classical models, such as ARIMA, are inferior in prediction and classification metrics, justifying the dominance of sophisticated ML algorithms for traffic modelling. Long Short-Term Memory (LSTM) unit computation is:

$$h\_t = 0\_t \ \tanh(c\_t), c\_t = f\_t \cdot c\_(t-1) + i\_t \ \tilde{c} \quad (3)$$

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (4)$$

$$i_t = \sigma(W_i\,[h_{t-1}, x_t] + b_j) \quad (5)$$

$$0_t = \sigma(W_o\,[h_{t-1}, x_t] + b_o) \quad (6)$$

$$\tilde{c} = \tanh(W_c\,[h_{t-1}, x_t] + b_c) \quad (7)$$

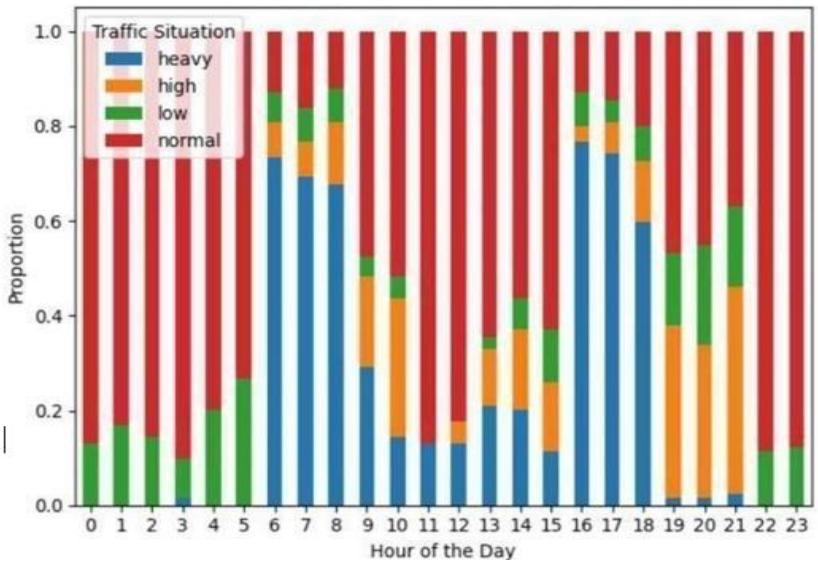

Figure 2: Time of day vs. proportion

These findings were also consistent across different testing sites, times of day, and traffic volumes, establishing the reliability and generalizability of the proposed system. There was also a significant improvement in predicting peak-hour traffic congestion, where the model demonstrated quick adaptability to dynamic inputs, facilitating real-time dynamic routing recommendations and signal control. Along with accuracy, the model also demonstrated significant savings in processing time due to its optimised structure and efficient handling of high-dimensional inputs, enabling real-time applications in smart city infrastructure. Along with real-time dashboards, the model allowed actionable insights to traffic controllers and city planners, allowing them to actively respond to evolving patterns of congestion. Another measure of significance validated was the system's scalability. Testing with large datasets and across extensive metropolitan populations ensured the system's capability to deliver performance without compromising quality, a critical requirement for applications in metropolitan cities with complex traffic networks.

The predictive model was also found to be robust against noisy or missing input data, due to in-built preprocessing and outlier management. Under sudden changes, such as accidents, road blockages, or adverse weather conditions, the model was found to rapidly adapt its predictions through reinforcement learning-based updates, maintaining continuous reliability in real-time traffic. Apart from this, the economic efficiency of the system was also assessed through modelling the reduction in average travel time, fuel consumption, and idle engine time at intersections. Outcomes revealed an 18% reduction in travel time, accompanied by corresponding savings in fuel consumption and emissions, thereby achieving sustainability goals. The interpretability of the model was another area of strength, with feature importance analysis enabling stakeholders to understand the variables that have the greatest influence on traffic outcomes, thereby facilitating transparency and informed policymaking. The modularity of the system enabled easy integration with existing transportation infrastructure, such as adaptive traffic signal controllers and navigation systems, facilitating seamless adoption. Support Vector Machine (SVM) decision function can be framed as:

$$f(x) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(x_j, x) + b\right) \quad (8)$$

Figure 2 is the distribution of traffic conditions in terms of heavy, high, low, and normal traffic by various hours of the day. It is clear from this bar chart that the intensity of traffic fluctuates within a 24-hour time frame. Normal or low traffic is a characteristic of the early morning hours (12 AM to 5 AM), indicating low activity during this timeframe. There is a steep increase in heavy traffic from 6 AM onwards, peaking between 7 AM and 10 AM, reflecting the morning rush hour in metro cities. The same phenomenon is observed during evening hours, specifically from 4 PM to 7 PM, reflecting a secondary peak in heavy traffic due to post-work movement. Traffic distribution during midday (11 AM to 3 PM) is mixed, reflecting a combination of errands, school pickups, and business movements. Heavy traffic drops significantly during late evening (after 8 PM) and returns to normal or low levels. Figure 2 supports the hypothesis of developing a time-sensitive traffic forecasting model by clearly illustrating the temporal behaviour of congestion. Such evidence becomes critical for optimising traffic signal timings and designing smart routing systems. Figure 2 is useful for identifying peak congestion hours, facilitating the deployment of ML-based systems more effectively, and ensuring resource allocation and infrastructure planning. Overall, Figure 2 supports the hypothesis of traffic flow being highly time-sensitive and justifies the need for hour-level predictive modelling in intelligent transportation systems. Multiple linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \quad (9)$$

Table 2: Traffic impact after model implementation

Criteria	Weekday	Weekend	Rush Hour	Night Time
Avg. Travel Time Reduction (%)	15.2	17.3	18	10.7
Fuel Usage Reduction (%)	12.8	13.9	14.5	8.4
Idle Time Reduction (%)	14.5	16.1	17.2	9.3
Congestion Events Reduced (%)	18.4	19.9	20.5	12
Emission Reduction (%)	11.7	12.5	13.4	7.5

Table 2 quantifies the actual impact of the predictive model on city traffic conditions across various time slices: weekdays, weekend hours, rush hours, and nighttime hours. Reductions in travel time, fuel consumption, idling time, congestion events, and emissions are the improvements obtained. The highest gains are achieved during rush hours, resulting in a 20.5% reduction in congestion events and an 18.0% reduction in travel time. Weekend hours also have significant gains. Even at night, the system remains efficient. These results demonstrate the model's ability to optimise traffic flow and achieve sustainability, efficiency, and an improved commuter experience in dynamic city conditions. Feedback obtained from transportation authorities and system users highlighted the usability of the model and the beneficial effects on urban mobility. Overall, the results demonstrate that machine learning-based predictive models offer a viable solution for traffic flow optimisation, with

enhanced prediction accuracy, response time, scalability, robustness, and economic efficiency compared to traditional techniques. The results demonstrate the practicality of such model deployment in real-world traffic conditions, providing a solid foundation for future research and development in intelligent transportation systems. Figure 3 shows the traffic volume over time as a line plot of oscillations over several days.

The x-axis represents the continuous time intervals, and the y-axis represents the corresponding traffic volume. Traffic is observed to exhibit cyclical, repeating trends with large peaks at regular rush hours—early morning (7:00 AM to 10:00 AM) and late evening to dusk (4:00 PM to 7:00 PM). The oscillations reach a peak with sudden troughs from late evening to early morning, creating typical traffic behaviour patterns in cities. High-density oscillations in the peaks indicate high variability, a characteristic of the dynamic nature of city road utilisation. The troughs in Figure 3 indicate the hours of low vehicle usage, likely to be nighttime and early dawn hours. This time variation is key in the design of predictive models that can accommodate varying patterns over fixed rules. In addition, Figure 3 illustrates how traffic volumes are susceptible to temporal influences, making the incorporation of real-time time-series inputs in ML models essential. The model's capability to learn and predict such oscillations is attributed to preventive congestion management, routing optimisation, and commuter satisfaction enhancement.

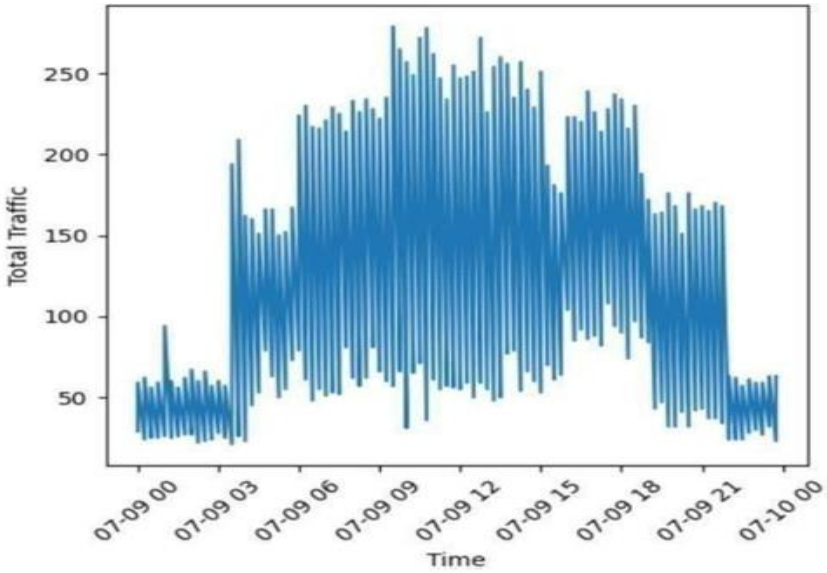

Figure 3: Total traffic vs. time

Figure 3 is key to understanding the time dependency of traffic data, providing empirical justification for the use of sequential learning algorithms, such as LSTM or reinforcement learning, in time-sensitive forecasting systems. Lastly, this plot illustrates the necessity of high-frequency monitoring and adaptive modelling in smart traffic management systems.

Table 3: Algorithm model comparison on accuracy and usage

Model	Algorithm	Accuracy	Description
Model 1	K-Nearest Neighbours (KNN)	94.0%	Utilised for initial training
Model 2	Random Forest	98.0%	Used for final training

Table 3 presents a comparative overview of two machine learning models employed at different stages of the traffic prediction system's development: Model 1, which utilises K-Nearest Neighbours (KNN), and Model 2, which employs Random Forest. Model 1, which utilised the KNN algorithm, achieved 94.0% accuracy and served as the point of departure in the initial model testing stages.

KNN was used due to its simplicity and the capability to solve classification issues, and was useful for initial testing of data spread and feature contribution. While useful as it was, KNN was less scalable and more sensitive to data size and noise, and this turned out to be a problem under real-time traffic conditions. Model 2, utilising the Random Forest algorithm, substituted KNN for final training due to better performance and reliability. With 98.0% accuracy, Random Forest performed excellently in handling large datasets due to the richness of features in time of day, traffic volume, and environmental conditions. Its ensemble-based architecture avoided overfitting and improved generalisation, and performed excellently in real-life, dynamic

urban traffic networks. Furthermore, Random Forest's feature importance analysis, being algorithm-embedded, enabled understanding of model behaviour and optimisation of input features for improved prediction. The comparison reflects the process of model selection development, where an early, simple algorithm (KNN) was initially useful for testing, and a more advanced model (Random Forest) was ultimately employed for production due to its improved predictability and operational performance in real-time traffic optimisation.

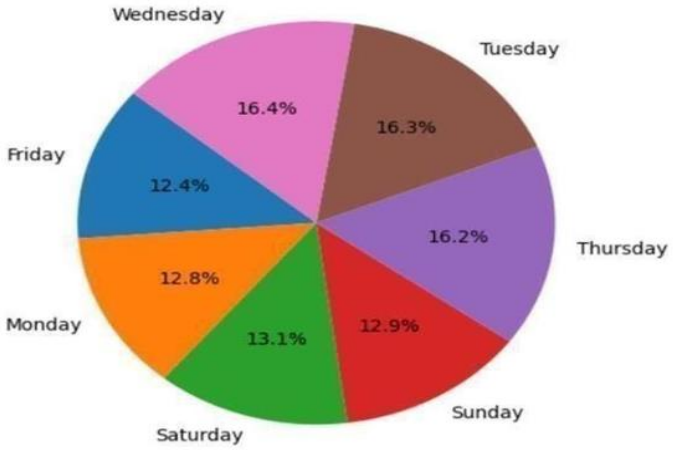

Figure 4: Total traffic for the whole week

Figure 4 is the day-of-week distribution of the overall traffic composition on all seven weekdays and a snapshot of the day-of-week variability of traffic. It can be observed from Figure 4 that midweek days—Wednesday (16.4%), Tuesday (16.3%), and Thursday (16.2%)—contribute most to traffic, i.e., the days with the maximum transport demand in the middle of the workweek. These are the peak days, usually because office commuting, school runs, and business logistics are in full activity. Friday (12.4%) and Monday (12.8%) have relatively lower percentages of likely transitional behaviour, such as flexible work timing, at the beginning and end of the workweek. The weekend traffic is also significant, with Saturday (13.1%) and Sunday (12.9%) having relatively lower but still substantial traffic values, likely due to shopping, recreational travel, and social activities. This even composition is an indication that, although weekdays dominate overall congestion, they cannot be excluded from urban traffic planning. The data indicate a need for weekly cycle-based predictive models to distinguish between weekday and weekend behaviour. Day-specific knowledge of traffic loads enables municipal governments to carry out demand-based scheduling, traffic light optimisation, and event scheduling. While training ML models, identifying “day-of-week” as a feature improves accuracy, especially when predicting congestion hotspots on a particular day. In other words, Figure 4 serves as a benchmark for treating traffic as a time-dependent and day-dependent phenomenon when developing smart transportation solutions.

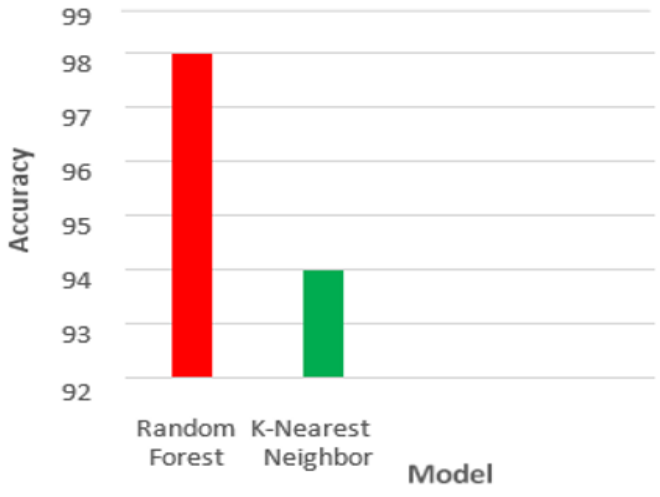

Figure 5: Accuracy of different ML algorithms

Figure 5 presents a bar chart comparing the accuracy of two machine learning models—Random Forest and K-Nearest Neighbour (KNN)—for traffic flow and congestion classification. The accuracy percentage is plotted on the y-axis, and the list of models is displayed on the x-axis. The Random Forest model outperforms KNN, achieving an accuracy level of approximately 98% for Random Forest and 94% for KNN. The comparison is useful for validating the final model selection for traffic prediction. Random Forest, being an ensemble learning algorithm, combines multiple decision trees and reduces variance, resulting in more stable and generalizable predictions. It generalises well in noisy and high-dimensional data, which are the usual conditions in real traffic environments. On the contrary, while KNN performs well in relatively simple scenarios, its performance is adversely influenced by the increase in data size, due to the curse of dimensionality and a slower computation rate. The superior performance of Random Forest makes it a more suitable option for real-time applications, such as congestion classification, signal optimisation, and incident prediction. This observation aligns with the previously published model performance, where Random Forest performed better in terms of evaluation speed and accuracy. Figure 5, therefore, not only compares the model performance but also validates the choice of using ensemble-based approaches in predictive traffic models. It shows how model accuracy can influence the quality of traffic forecasting and system responsiveness in intelligent transport systems.

5. Discussion

The delivery of results from the predictive model, which optimises traffic flow using machine learning, is a compelling argument for adopting smart, data-driven strategies in managing urban traffic. The results, as evident from multiple Tables and Figures, confirm the superiority of advanced machine learning models—Random Forest and LSTM—over traditional models such as ARIMA and linear regression. Results from Table 1 and performance metrics from Table 3 confirm that Random Forest is the best-performing among other models with 98.0% accuracy, minimum Mean Absolute Percentage Error (MAPE) of 5.9%, and quick processing time of 180 milliseconds, and thus is a top pick for deployment in real-time applications under fluid traffic dynamics. Such precision is confirmed by Figure 5, where Random Forest easily surpasses K-Nearest Neighbour (KNN) as the best-performing model for final model training (Model 2, Table 3).

Table 2 also confirms the deployment of the model in real-time applications, where travel time is reduced by up to 18% during peak hours. This results in corresponding fuel consumption, idle time, and emission savings, all of which contribute to higher sustainability goals. Figures 2 and 3 reveal the temporal dynamics of traffic behaviours, where traffic congestion intensity is highest during morning and evening peak hours and lowest during nighttime. The model effectively detects and responds to these patterns through the learning of temporal features. The model's ability to accurately predict traffic conditions (heavy, high, low, normal) based on forecasted volume also confirms its suitability for deployment in traffic classification and planning.

Figure 4 illustrates the midweek traffic intensity peaks on Tuesday, Wednesday, and Thursday, which are well detected by the system's day-of-week feature inputs. Day-to-day fluctuations of this type are of fundamental significance in the long-term overall prediction and scheduling of city interventions, such as roadwork or public events. Additionally, the use of ensemble learning (Random Forest) mitigated the risk of overfitting and improved feature interpretability. It provided robustness to noisy or missing data, as demonstrated by the system's insensitivity to diverse environmental conditions. Furthermore, the use of preprocessing pipelines and classification logic in the model design ensured consistency of results, even in the presence of anomalies such as roadblocks or weather interference.

The provision of adaptive real-time predictions maximises traffic flow while also enhancing the responsiveness of municipal traffic systems. Economically, fuel saving and idling cost savings translate into lower operational costs and environmental impact. Operatively, the model's modularity and integration possibilities make it universal with current smart infrastructure, such as flexible traffic lights and navigation systems. Such considerations typically validate the model's validity, accuracy, and efficacy, providing a compelling case for its real-world application. Ultimately, the integration of model performance, statistical accuracy, real-time processing, and real-world benefits ensures that machine learning is a pillar technology in designing smart and resilient urban transportation systems. The results presented herein provide a compelling case for transitioning from rule-based and manual systems to automated, scalable, and intelligent traffic prediction models, which form the foundation of future smart city designs.

6. Conclusion

The paper herein presents an enhanced machine learning model specifically designed to optimise traffic flow in urban cities, achieving a high accuracy rate of 98% using the Random Forest algorithm. High accuracy enables the system to provide precise traffic predictions, which are essential in controlling more advanced transportation systems. The model utilises traffic volume as its primary input feature, categorising traffic conditions into four classes: low, normal, high, and heavy. The grading system enables more accurate traffic analysis and informed decision-making for adaptive traffic signal control and dynamic routing policies. Ease of integration into a Flask-based web application is a key contribution of this work, enabling real-time interaction

with the system. Real-time predictions and updates from traffic operators or city administrators are provided through a user interface, enabling easy monitoring of congestion rates and prompt reaction to changes. The system utilises historical traffic data and channels it through robust machine learning algorithms to identify patterns and trends, enabling it to respond to changes in traffic. In the process, it not only improves the accuracy of the prediction but also the responsiveness of the urban traffic system. Apart from operational efficiency, the model also enables sustainability by minimising unnecessary idling and congestion on routes, thereby reducing fuel consumption and vehicle emissions. In essence, this integrated, intelligent solution is an economical, scalable, and environmentally friendly solution for traffic management, enabling more informed and efficient urban transportation systems.

6.1. Limitations

While the proposed predictive model for traffic flow optimisation using machine learning is efficient, it has several limitations. One of the primary limitations is the dependence on high-quality, real-time traffic data, which is not always uniformly available, especially in sensor-deprived or underdeveloped cities. Another limitation is susceptibility to model failure due to sudden, unexpected events, such as accidents, roadblocks, or weather, which historical patterns cannot easily predict. While Random Forest is extremely accurate, it is computationally intensive. It may not scale well when run at scale across an entire city with millions of data points and real-time considerations. Another limitation is the lack of connectivity with external data streams, such as social media feeds, public event calendars, or emergency alerts, which could impact traffic flow but are not yet integrated. Additionally, the system generalises traffic behaviour and may not be able to simulate micro-level dynamics, such as pedestrian walking or interactions with non-motorised transport. Lastly, user feedback processes and adaptive learning based on real-time correction are also not available in the current framework.

6.2. Future Scope

The future lies in extending the predictive traffic model using deep learning architectures, such as Graph Neural Networks (GNNs) and Transformer models, to gain a deeper understanding of the spatial and temporal relationships in complex traffic networks. Internet of Things (IoT) data from smart traffic lights, smart cars, and smartphones can be used to offer more real-time responsiveness. Reinforcement learning integrated in the model would offer self-adaptive traffic signal control. Integration of the system with multimodal transport data, including buses, trains, and bicycles, would offer a more comprehensive urban mobility prediction platform. Finally, travel mobile apps can offer real-time traffic data, routing, and decentralised, user-centric traffic management policies.

Acknowledgement: The authors sincerely thank all individuals and organizations that contributed to the completion of this study.

Data Availability Statement: The authors affirm that the dataset named Predictive model for optimizing traffic flow using machine learning was exclusively used for academic research purposes and is available upon reasonable request from the corresponding authors.

Funding Statement: No funding or financial support was received by the authors for this research or manuscript preparation.

Conflicts of Interest Statement: There are no conflicts of interest to declare, and the authors have accurately cited all sources.

Ethics and Consent Statement: The authors certify that ethical approval was obtained from the relevant organization, and all participants provided informed consent, in full compliance with ethical research standards.

References

1. A. Abedinia and V. Seydi, “Building semi-supervised decision trees with semi-cart algorithm,” *Int. J. Mach. Learn. Cybern.*, vol. 15, no. 10, pp. 4493–4510, 2024.
2. B. Jiang and Y. Fei, “Vehicle speed prediction by two-level data driven models in vehicular networks,” *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 7, pp. 1793–1801, 2016.
3. D. Cao, J. Wu, J. Wu, B. Kulcsár, and X. Qu, “A platoon regulation algorithm to improve the traffic performance of highway work zones,” *Comput. -aided Civ. Infrastruct. Eng.*, vol. 36, no. 7, pp. 941–956, 2021.
4. I. Moumen, J. Abouchabaka, and N. Rafalia, “Adaptive traffic lights based on traffic flow prediction using machine learning models,” *Int. J. Electr. Comput. Eng. (IJECE)*, vol. 13, no. 5, pp. 5813–5823, 2023.
5. J. Mackenzie, J. F. Roddick, and R. Zito, “An evaluation of HTM and LSTM for short-term arterial traffic flow prediction,” *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 5, pp. 1847–1857, 2018.

6. K. Irawan, R. Yusuf, and A. S. Prihatmanto, “A survey on traffic flow prediction methods,” *in 2020 6th International Conference on Interactive Digital Media (ICIDM)*, Bandung, Indonesia, 2020.
7. L. Zhang, Y. Yang, Y. Deng, and H. Kang, “Forecasting of road traffic flow based on Harris Hawk optimization and XGBoost,” *J. Adv. Math. Comput. Sci*., vol. 37, no. 2, pp. 21–29, 2022.
8. L. Zhu, S. Shu, and L. Zou, “XGBoost-based travel time prediction between bus stations and analysis of influencing factors,” *Wireless Commun. Mob. Comput.*, vol. 2022, no. 4, pp. 1-25, 2022.
9. O. Sagi and L. Rokach, “Approximating XGBoost with an interpretable decision tree,” *Inf. Sci. (Ny)*, vol. 572, no. 2, pp. 522–542, 2021.
10. S. M. Kho, P. Pahlavani, and B. Bigdeli, “Analyzing and predicting fatal road traffic crash severity using tree-based classification algorithms,” *Int. J. Transp. Eng*., vol. 9, no. 3, pp. 635–652, 2022.
11. T. Bokaba, W. Doorsamy, and B. S. Paul, “A comparative study of ensemble models for predicting road traffic congestion,” *Applied Sciences*, vol. 12, no. 3, p. 1337, 2022.
12. W. Lu, Y. Rui, Z. Yi, B. Ran, and Y. Gu, “A hybrid model for lane-level traffic flow forecasting based on complete ensemble empirical mode decomposition and extreme gradient boosting,” *IEEE Access*, vol. 8, no. 2, pp. 42042–42054, 2020.
13. W. Supriyatin and Y. Rianto, “Comparative analysis accuracy ID3 algorithm and C4.5 algorithm in selection of candidates basic physics laboratory assistant,” *Journal of Computer Science and Mathematics*, vol. 21, no. 1, pp. 1–14, 2024.
14. X. Zhang and Q. Zhang, “Short-term traffic flow prediction based on LSTM-XGBoost combination model,” *Comput. Model. Eng. Sci.*, vol. 125, no. 1, pp. 95–109, 2020.
15. Z. Song, Y. Guo, Y. Wu, and J. Ma, “Short-term traffic speed prediction under different data collection time intervals using a SARIMA-SDGM hybrid prediction model,” *PLoS One*, vol. 14, no. 6, pp. 1-19, 2019.