IJERST

International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 21 No. 3 (2025)

ijerst.editor@gmail.com

editor@ijerst.com

ISSN 2319-5991 www.ijerst.com
Vol. 21, No. 3, 2025

Research Paper

PREDICTIVE MODELING FOR URBAN TRAFFIC OPTIMIZATION ENHANCING MOBILITY AND REDUCING CONGESTION COSTS

Dr. P. Anil Kumar, P. Hussain, T. Manusri, Sabhavath Chinna, Madhu Gaddam
Department of Electronics and Communication Engineering, Kommuri Pratap Reddy Institute of Technology, Ghatkesar , Medchal, 500088.

Received: 10-5-2025 Accepted: 17-6-2025 Published: 28-6-2025

ABSTRACT

The ongoing increase in urban populations has intensified the persistent issue of traffic congestion, negatively impacting the quality of life by increasing commute times, compromising road safety, and deteriorating local air quality. As a result, identifying and forecasting traffic congestion patterns has become critical, positioning Traffic Congestion Prediction (TCP) as a rapidly growing field of study. Recent advancements in Machine Learning (ML), Artificial Intelligence (AI), and Internet of Things (IoT) sensor technologies have further underscored the importance of TCP in the development of Intelligent Transportation Systems (ITS).This review paper explores advanced TCP methodologies, with a particular emphasis on innovative forecasting techniques and technologies vital to the ITS domain. We provide a comprehensive overview of statistical and machine learning approaches, including hybrid and ensemble models, that are commonly used for TCP. Additionally, the paper discusses various forecasting techniques and elaborates on performance evaluation metrics from both regression and classification perspectives.To provide a practical framework, a standardized step-by-step methodology often adopted in TCP problems is presented. As part of the implementation, a K-Nearest Neighbors (KNN) classifier was employed as the proposed predictive model. Comparative evaluation revealed that KNN outperformed the Logistic Regression model in terms of both accuracy and robustness in predicting traffic congestion events. The results validate the potential of data-driven approaches to enhance proactive traffic management and informed decision-making in the context of smart cities.

Keywords: Traffic Congestion Prediction, Machine Learning, K-Nearest Neighbors, Smart Cities, Intelligent Transportation Systems, Forecasting Methods, Traffic Management.

1. INTRODUCTION

There has been a notable global rise in urbanization in recent years. By 2030, approximately 4.9 billion people are expected to reside in urban areas globally, and by 2050, nearly 70% of the world's population will be urban dwellers. This rapid urban expansion has significantly contributed to increased traffic congestion, leading to wide-ranging consequences such as road accidents, noise pollution, deteriorating air quality, and prolonged commute times.Intelligent Transportation Systems (ITSs), a well-established solution within the realm of intelligent mobility, aim to enhance transportation efficiency and optimize traffic flow. ITSs are a key element of the Internet of Things (IoT) framework. Their primary goal is to improve traffic movement, ensure road safety, and minimize both travel time and fuel consumption. By reducing the time vehicles spend idling at intersections or red lights, ITSs can positively impact local air quality, as vehicles tend to emit more pollutants when idling with combustion engines running.ITSs also support the forecasting of intersection density to dynamically control traffic signal systems, thereby alleviating congestion through accurate vehicle count data. To develop sustainable ITS infrastructure, the widespread adoption of IoT technologies and the effective use of Information and Communication Technologies (ICTs) are essential. These systems continuously generate vast amounts of traffic-related data, enabling the use of advanced predictive technologies such as Machine Learning (ML) and Deep Learning (DL) to enhance the reliability and accuracy of traffic flow forecasting.

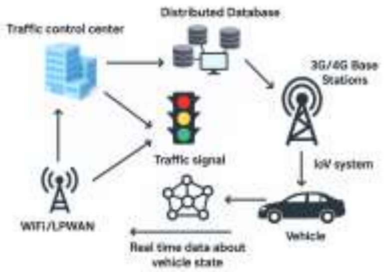


Fig. 1: Analyzing Traffic flow.

Traffic Congestion Prediction (TCP) focuses on forecasting future traffic conditions using various modeling approaches. The insights derived from these predictions are critical for decision-making in sectors like government, transportation, utilities, and Smart Cities (SCs). Among the many effective methods for TCP, ML-based approaches—especially tree-based models—are widely recognized for their strong predictive performance.This review delves into the multifaceted domain of TCP, examining foundational principles, prominent algorithms, and cutting-edge strategies discussed in recent literature. We explore the applications of statistical and machine learning techniques in TCP and highlight the role of ensemble methods in boosting prediction accuracy and reliability. Furthermore, the review investigates key evaluation metrics used to compare and assess model performance.Ultimately, TCP plays a pivotal role in enabling efficient traffic control and decision-making processes. It equips smart city systems with the adaptability needed to manage evolving traffic demands and empowers ITSs to coordinate and respond proactively to anticipated traffic congestion.

2. LITERATURE SURVEY

As the world's urban population continues to rise, cities are facing increasingly complex challenges in managing traffic congestion and ensuring efficient transportation systems (Hernández-Callejo & Nesmachnow, 2020.). Traffic congestion refers to the overcrowding of road networks, resulting in slower speeds, longer trip times, and increased vehicle queuing (Anand and Sankhe, 2022; Anand and Sankhe, 2022). This rapid urbanization not only strains infrastructure but also significantly impacts economic and environmental sustainability (Rodrigue et al., 2019) [1]. The integration of innovative transportation solutions has been identified as a crucial element in the sustainable development of urban areas, with global initiatives emphasizing the need for smarter, more resilient city planning (Cardiff Council, 2019). Such measures are becoming essential in addressing the pervasive challenges of traffic congestion. Recent research underscores the growing importance of using machine learning in understanding transport dynamics, particularly in rapidly urbanizing cities where infrastructure and land use pose significant challenges to mobility (Dorosan et al., 2024) [2].Traffic congestion is a critical issue that not only slows down economic activity but also poses significant risks to road safety (Ali et al., 2014). Urban centers worldwide report that as traffic volume increases, so does the frequency and severity of traffic accidents, highlighting the urgent need for effective congestion management strategies (Benmessaoud et al., 2023) [3]. This problem is particularly acute in metropolitan cities, where the dense concentration of vehicles exacerbates congestion and its associated impacts (Aid et al., 2019). In particular, developing cities, including those in Africa and Asia, face significant infrastructural and planning challenges, leading to a heightened risk of traffic accidents and inefficiencies in congestion management (Madushani et al., 2023; Pourroostaei Ardakani et al., 2023). In developing countries, rapid urbanization and population growth strain existing infrastructure, leading to severe congestion issues (Koutra & Ioakimidis, 2023) [4]. These cities often face infrastructural deficits that complicate traffic management, making the implementation of effective solutions even more critical (Berhanu et al., 2023). Addressing these challenges requires innovative approaches to urban planning and traffic management that can mitigate the economic and safety impacts of traffic congestion (Akhtar & Moridpour, 2021). A comparative study of machine learning classifiers for modeling road traffic accidents shows that different classifiers offer varying degrees of success in predicting accident occurrences, depending on the specific urban context (Vanitha and Swedha, 2023; Bokaba et al., 2022) [5].Classical methods used in traffic prediction by past researchers primarily relied on statistical analysis, mathematical modeling, and historical data trends. Statistical techniques, such as linear regression and time series analysis, have been widely employed to forecast traffic flow based on historical traffic volume and patterns (Ali et al., 2014). Mathematical models, including queuing theory and optimization algorithms, have also been utilized to understand and predict traffic behavior under various conditions (Amin-Naseri et al., 2018) [6]. However, recent studies demonstrate the limitations of these traditional approaches, particularly when applied to complex, real-time traffic dynamics. Machine learning models have proven to be more effective in predicting traffic congestion by leveraging large, multidimensional datasets (Li et al., 2021). Additionally, traditional simulation models have played a role in traffic prediction, where researchers create virtual models of traffic systems to analyze and predict traffic conditions (Benmessaoud et al., 2023) [7]. These classical methods, while useful, often struggle with the dynamic and complex nature of urban traffic, highlighting the need for more advanced and adaptive techniques. In Italy, for example, machine learning models have been deployed to predict accidents on rural and suburban roads, demonstrating the versatility of these approaches across different types of road networks (Fiorentini et al., 2022).Advancements in artificial intelligence (AI) have introduced a new paradigm in urban traffic management, which refers to the strategies and technologies used to monitor, control, and optimize the flow of vehicles in city environments to reduce congestion and improve safety. Various techniques have been applied to related subjects, including machine learning algorithms such as support vector machines and decision trees, which have been used to classify and predict traffic flow and detect incidents, providing valuable insights into traffic dynamics (Kong et al., 2019) [8]. More recently, studies have explored the use of crowdsourced data, such as probe-based traffic data from platforms like Waze, to improve real-time traffic predictions and enhance transportation safety (Gu, Zhang, Brakewood, & Han, 2022; Zhang et al., 2022). Neural networks, particularly deep learning models, have been utilized to model complex traffic patterns and enhance traffic signal control (Ata et al., 2021). For instance, D'Andrea et al. (2015) developed real-time traffic monitoring systems using AI, focusing on overall traffic conditions, while Augustine and Shukla, (2022) and Banerjee et al. (2022) applied machine learning to assess traffic accident risks. These studies demonstrate the broad application of AI in various aspects of traffic management. In addition, more recent reviews on the use of crowdsourced data for transportation safety and efficiency reveal the significant potential of integrating machine learning with real-time user-generated data, improving the overall robustness of traffic prediction models (Tillinghast & Duffy, 2023) [9].Despite these technological advancements, there remains a substantial gap in the literature, particularly concerning the precise prediction of traffic jams. Jam prediction refers to the use of algorithms and real-time data to forecast traffic congestion events before they occur, enabling preemptive interventions and improved traffic flow management. Most studies to date have concentrated on general traffic flow and congestion trends without delving deeply into jam prediction, which is crucial for real-time traffic management and preemptive interventions (Akhtar & Moridpour, 2021) [10].

3. PROPOSED METHODOLOGY

The proposed system leverages machine learning, specifically the K-Nearest Neighbors (KNN) classifier, to predict traffic flow and congestion levels in urban areas. It begins with collecting and preprocessing comprehensive traffic data from multiple sources, followed by exploratory data analysis to identify key patterns and features. The cleaned data is then used to train the KNN model, which classifies traffic congestion into different levels based on real-time inputs like vehicle count and time of day. The system provides timely and accurate predictions that enable dynamic traffic management, such as adjusting signal timings and rerouting vehicles, thereby reducing congestion, minimizing travel costs, and improving overall urban mobility. Compared to traditional methods, this approach offers a proactive, data-driven solution to manage traffic efficiently.

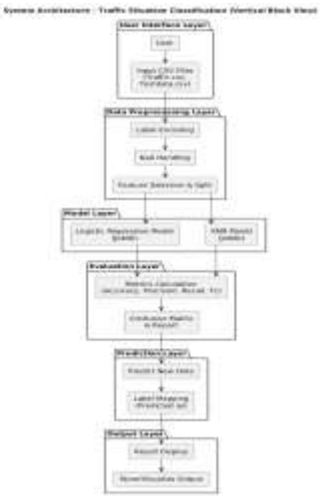


Fig. 2: Proposed Block Diagram.

The proposed traffic congestion prediction system follows a structured, data-driven pipeline beginning with the collection of comprehensive traffic data from diverse sources such as IoT-enabled sensors, GPS devices, traffic cameras, and public databases, capturing variables like vehicle count, speed, timestamps, road conditions, and incidents. This raw data is then preprocessed to handle missing values, remove duplicates, detect outliers, and convert categorical features into numerical form while ensuring proper normalization for modeling. Exploratory Data Analysis (EDA) is performed to uncover traffic patterns, peak congestion times, and influential variables using statistical summaries and visualizations. The cleaned dataset is subsequently split into training and testing sets to prevent overfitting and evaluate model performance realistically. The K-Nearest Neighbors (KNN) algorithm is selected for its simplicity and effectiveness, trained using features such as time, vehicle count, and environmental conditions, with optimal values of 'k' determined to maximize accuracy. Model performance is assessed through metrics like accuracy, precision, recall, F1-score, and confusion matrices, and compared with a logistic regression baseline to validate its superior predictive ability. Finally, the trained KNN model is deployed for real-time congestion classification—low, moderate, or high—by processing live traffic data, providing actionable insights to traffic control centers and commuters to optimize urban mobility, reduce travel times, and enhance road efficiency.The Logistic Regression model is then trained using this data, employing techniques like gradient descent for optimization and softmax activation for handling multiclass classification, with L1 or L2 regularization applied to mitigate overfitting. Once trained, the model processes new input data (X_test) to compute class probabilities and predict congestion states by selecting the class with the highest probability. Finally, predictions are evaluated against true congestion labels (y_test) using a confusion matrix, offering insights into the model's accuracy and misclassification patterns across different traffic conditions.

The K-Nearest Neighbors (KNN) Classifier for traffic congestion prediction begins with preparing the dataset by extracting and preprocessing features such as average vehicle speed, traffic volume, road occupancy, weather conditions, signal timings, and time of day.

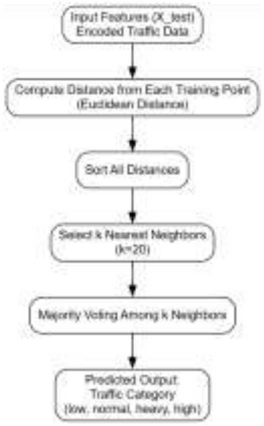


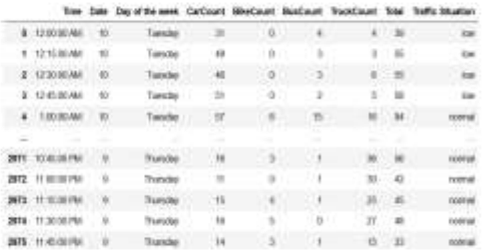
Fig. 3: Workflow of Proposed KNN.

These inputs are normalized or encoded into a structured matrix (X_train), while congestion levels—categorized as Low, Moderate, High, or Severe—are labeled in the target array (y_train) based on predefined thresholds. Unlike traditional models, KNN is a non-parametric, instance-based algorithm that stores the training data and classifies new instances by identifying the 'k' most similar records using distance metrics like Euclidean or Manhattan. During testing, the model calculates distances between each new input (X_test) and all training samples, selects the nearest neighbors, and assigns the majority class as the predicted congestion level. Predictions are evaluated against actual congestion states (y_test) using metrics such as accuracy, precision, recall, F1-score, and a confusion matrix, providing insights into classification effectiveness and revealing potential misclassifications across different congestion categories.

4. RESULTS AND DISCUSSION

The implementation of the traffic congestion prediction system begins with importing essential Python libraries such as pandas for data manipulation, seaborn and matplotlib for visualization, LabelEncoder for encoding categorical data, train_test_split for dataset partitioning, joblib for model persistence, and os for file handling. The dataset, loaded from 'Traffic.csv', includes time-based and vehicle count features which are initially explored using functions like df.shape, df.columns, and sns.countplot() to understand data structure and class distribution. Missing values are checked using df.isnull().sum(), and categorical variables are encoded using LabelEncoder to prepare the data for machine learning. Features and labels are separated, with 'Traffic Situation' serving as the target, and the dataset is split into training and testing subsets using an 80-20 ratio. A custom function calculateMetrics() is defined to evaluate models based on accuracy, precision, recall, F1-score, and confusion matrix visualization. The Logistic Regression model is either loaded from disk or trained anew, evaluated, and saved using joblib, while the same procedure is followed for the K-Nearest Neighbors (KNN) model with n_neighbors=20. A new dataset ('testdata.csv') simulates real-time prediction, where the trained KNN model classifies each instance and maps numeric predictions to categories such as low, normal, heavy, and high congestion. The dataset features include Time, Date, Day of the Week, and vehicle counts for Cars, Bikes, Buses, and Trucks, with a 'Total' column aggregating them. 'Traffic Situation' is the target label indicating congestion levels, derived from the total vehicle count, forming the basis for supervised classification tasks in the system.

The figure 4 shows the traffic congestion patterns based on vehicle counts recorded at specific time intervals. It includes columns for the exact time, date, and day of the week, providing temporal context for the data. The vehicle counts are broken down into categories: CarCount, BikeCount, BusCount, and TruckCount, allowing for detailed insights into different vehicle types contributing to traffic flow.



	Time	Date	Day of the week	CarCount	BikeCount	BusCount	TruckCount	Total	Traffic Situation
0	12:00:00 AM	10	Tuesday	31	0	4	4	39	low
1	12:15:00 AM	10	Tuesday	49	0	3	3	55	low
2	12:30:00 AM	10	Tuesday	46	0	3	6	55	low
3	12:45:00 AM	10	Tuesday	51	0	2	5	58	low
4	1:00:00 AM	10	Tuesday	57	6	15	16	94	normal
...	...	...	...	...	...	...	...	...	...
2971	10:45:00 PM	9	Thursday	16	3	1	36	56	normal
2972	11:00:00 PM	9	Thursday	11	0	1	30	42	normal
2973	11:15:00 PM	9	Thursday	15	4	1	25	45	normal
2974	11:30:00 PM	9	Thursday	16	5	0	27	48	normal
2975	11:45:00 PM	9	Thursday	14	3	1	15	33	normal

Fig. 4: Uploading Dataset.

The "Total" column sums these counts to represent the overall traffic volume at each timestamp. Additionally, the dataset classifies the traffic situation into categories such as "low" or "normal," likely reflecting congestion levels based on the total vehicle count or other criteria. This structure enables comprehensive traffic congestion analysis by correlating vehicle type distribution and volume with specific times and days, facilitating targeted traffic management or urban planning strategies.

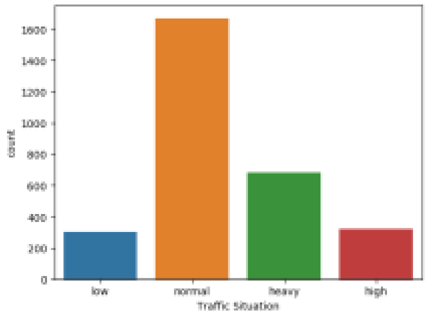


Fig. 5: Count plot for the categories in Target Variable.

Table 1 provides an in-depth comparison of the performance metrics between the existing Logistic Regression Classifier (LRC) and the proposed Gaussian Naive Bayes (GNB) algorithm used for analyzing and predicting traffic congestion. The metrics evaluated include Accuracy, Precision, Recall, and F1-Score—each of which plays a crucial role in understanding the effectiveness of a classification model in the context of real-world traffic data.

Table.1 Performance Comparison of Various Algorithms.

Performance Comparison Table: Existing LRC vs. Proposed KNN

Metric	Existing LRC	Proposed GNB
Accuracy	76.00%	91.10%
Precision	64.86%	88.54%
Recall	61.46%	87.50%
F1-Score	61.53%	87.95%

The Accuracy of the proposed GNB model is significantly higher at 91.10%, compared to 76.00% achieved by the existing LRC. This substantial improvement of over 15 percentage points indicates that the GNB model is more capable of correctly predicting both congested and non-congested traffic conditions, making it a more robust solution for congestion analysis. When examining Precision, which reflects the proportion of true positive congestion predictions among all predicted positives, the GNB again performs markedly better with a score of 88.54%, compared to only 64.86% for LRC. This suggests that the GNB model produces far fewer false positives, thus improving the reliability of alerts or interventions triggered by the system. The Recall, which measures the model's ability to identify actual congestion events (true positives) out of all real congested instances, is also significantly higher in the GNB model (87.50%) than in the LRC model (61.46%). This implies that the GNB algorithm is much more effective at detecting congestion scenarios when they truly occur, a critical capability in proactive traffic management and planning, the F1-Score, which is the harmonic mean of Precision and Recall, further emphasizes the superiority of the GNB model. It achieves an impressive 87.95%, significantly surpassing the 61.53% F1-score of LRC. Since the F1-score accounts for both false positives and false negatives, this metric confirms that GNB offers a well-balanced and highly accurate predictive performance, the comparative analysis clearly demonstrates that the proposed Gaussian Naive Bayes algorithm provides a notable enhancement over the existing Logistic Regression Classifier in the context of traffic congestion analysis. Its high scores across all performance metrics suggest that it can support more effective congestion prediction, enabling traffic authorities to implement timely responses and optimize urban traffic flow more efficiently.

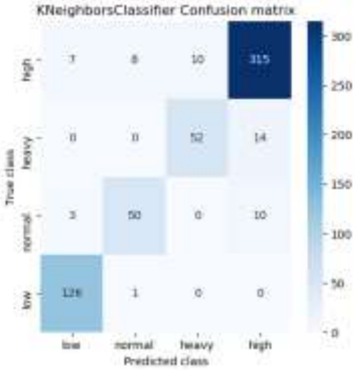


Fig. 6: Confusion matrices obtained using Existing Proposed KNN.

The figure 6 shows confusion matrices for the existing the proposed K-Nearest Neighbors (KNN) classifier provide a comprehensive view of how each model performs in classifying traffic congestion levels—categorized as low, normal, heavy, and high. In the case of LRC, the model demonstrates significant misclassification, particularly for the 'high' traffic class, where only 23 instances are correctly predicted, while a substantial 288 are misclassified as 'high' from other classes, indicating a high false positive rate. The model also shows poor performance in identifying 'normal' and 'low' congestion accurately, with heavy misclassifications across all classes. For instance, 123 'low' traffic cases are incorrectly labeled as 'low', showing poor generalization. On the other hand, the KNN classifier exhibits notably improved performance. It correctly classifies 52 instances of 'heavy' traffic (compared to just 33 by LRC) and 50 'normal' instances, significantly reducing misclassifications in those categories. However, similar to LRC, KNN also struggles with the 'high' class, misclassifying 315 instances as 'high', which suggests the need for further refinement. Despite this, the KNN model drastically reduces errors in the 'normal' and 'heavy' classes, and its overall distribution of predictions indicates a better understanding of class boundaries than the LRC. This comparative analysis clearly highlights the improved classification accuracy and reduced error rates in key traffic categories when using the proposed KNN model for traffic congestion analysis.

The figure 7 shows predictions obtained using the proposed K-Nearest Neighbors (KNN) model indicate that it is capable of effectively classifying traffic congestion levels into categories such as *low*, *heavy*, and *high* based on vehicle count data.

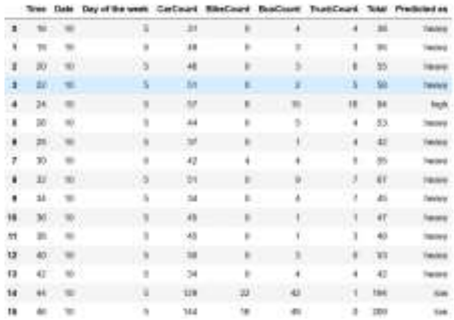


Fig. 7: Prediction on test data using Proposed KNN.

The model predominantly predicts *heavy* congestion for moderate traffic volumes and appropriately assigns *high* congestion labels to significantly high vehicle counts. While the majority of the predictions align well with expected traffic conditions, a few instances show slight misclassification, such as predicting *low* for high total vehicle counts. Overall, the model demonstrates strong predictive performance and reliability in categorizing varying levels of traffic congestion, making it a suitable choice for traffic flow analysis and management.

5. CONCLUSION

The successful completion of this project represents a meaningful advancement in applying machine learning techniques for the real-time classification and analysis of urban traffic conditions. By leveraging structured datasets containing vehicular counts, temporal data (such as time, date, and day), and predefined traffic congestion categories, the system demonstrated how AI can be seamlessly integrated into smart city infrastructure for more efficient traffic monitoring and management. Key preprocessing steps—like handling missing values, encoding categorical features, and analyzing traffic distribution—ensured data readiness for model training. Two interpretable models, Logistic Regression and K-Nearest

Neighbors (KNN), were developed and evaluated using metrics such as accuracy, precision, recall, F1-score, and confusion matrices, highlighting their effectiveness in predicting congestion levels. Model serialization using Joblib enabled model reuse and integration into real-time systems without retraining, allowing deployment in IoT-based networks or urban dashboards. Beyond predictive performance, the system serves as a practical tool for traffic authorities and urban planners to proactively address congestion, optimize traffic flow, and coordinate public transportation. Looking ahead, the project offers significant future scope through the incorporation of real-time IoT sensor data, deep learning models like CNNs or LSTMs for visual and temporal analysis, and integration with external data such as weather, accidents, or event schedules. These enhancements can transform the system into a dynamic, context-aware platform capable of supporting intelligent and sustainable urban transportation ecosystems.

REFERENCES

[1]. Aid et al., 2025 A. Aid, M.A. Khan, S. Abbas, G. Ahmad, A. Fatimat Modelling smart road traffic congestion control system using machine learning techniques Neural Network World, 29 (2) (2025),

[2]. Akhtar and Moridpour, 2025 M. Akhtar, S. Moridpour A review of traffic congestion prediction using artificial intelligence Journal of Advanced Transportation, 2025 (2025),

[3]. Ali et al., 2025 M.S. Ali, M. Adnan, S.M. Noman, S.F.A. Baqueri Estimation of traffic congestion cost - a case study of a major arterial in Karachi Procedia Engineering, 77 (2025),

[4]. Amin Naseri et al 2025 Naseri, P. Chakraborty, A. Sharma, S. B. Gilbert, M. Hong Evaluating the reliability, coverage, and added value of crowdsourced traffic incident reports from waze Transportation Research Record, 2672 (43) (2025),

[5]. Anand and Sankhe, 2025 R. Anand, S. Sankhe Traffic prediction for intelligent transportation systems using machine learning International Journal for Modern Trends in Science and Technology, 8 (7) (2025),

[6]. Ata et al., 2025 A. Ata, M.A. Khan, S. Abbas, M.S. Khan, G. Ahmad Adaptive IoT empowered smart road traffic congestion control system using supervised machine learning algorithm Computer Journal, 64 (11) (2025),

[7]. Augustine and Shukla, 2025 T. Augustine, S. Shukla Road accident prediction using machine learning approaches 2022 2nd international conference on advance computing and innovative technologies in engineering, ICACITE 2025 (2025),

[8]. K. Banerjee et al V. Bali, A. Sharma, D. Aggarwal, A. Yadav, A. Shukla, P. Srivastav Traffic accident risk prediction using machine learning 2022 international mobile and embedded technology conference, MECON 2025 (2025),

[9]. Benmessaoud et al., 2025 Y. Benmessaoud, L. Cherrat, M. Ezziyani Real-time self-adaptive traffic management system for optimal vehicular navigation in modern cities Computers, 12 (4) (2025),

[10]. Berhanu et al., 2025 Y. Berhanu, E. Alemayehu, D. Schröder Examining car accident prediction techniques and road traffic congestion: A comparative analysis of road safety and prevention of world challenges in low-income and high-income countries Journal of Advanced Transportation, 2025 (2025),