© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717886

Machine Learning-Based Traffic Congestion Prediction and Policy Simulation for Sustainable Urban Transportation in Bengaluru

VEDANTH PARIDA[1], PROF. RAKSHITHA B. S[2]
[1] PG Research Scholar, Jain (Deemed-to-be University) Bangalore, India
[2] Assistant Professor, Department of Computer Science and Engineering Jain (Deemed-to-be University), Bangalore, India

Abstract- ***Traffic congestion has become a major urban challenge affecting travel time, fuel consumption, road safety, commuter productivity, and environmental sustainability. Rapid growth in private vehicle usage in Bengaluru has increased pressure on urban road networks, especially in highly active commercial, residential, and employment corridors. This study proposes a machine learning-based framework for traffic congestion prediction and simulation-based evaluation of urban transport policies using Bengaluru traffic datasets. The work includes data cleaning, exploratory data analysis, outlier inspection, feature engineering, model training, policy simulation, residual validation, and explainable artificial intelligence analysis. Three machine learning models, namely Linear Regression, Random Forest Regressor, and XGBoost Regressor, were implemented and evaluated for congestion prediction. Engineered features such as Bus Lane Impact, Transit Efficiency, Carpool Efficiency, and Fuel Wastage Index were developed to support policy-aware analysis. Experimental results showed that XGBoost achieved the best performance with an R2 score of approximately 0.989, outperforming Linear Regression and Random Forest. SHAP-based explainability analysis showed that Traffic Volume was the most influential factor in congestion prediction, while bus lane impact, transit efficiency, and carpooling-related features contributed to policy interpretation. The proposed framework demonstrates how machine learning can support intelligent traffic planning and simulation-based policy evaluation for sustainable urban mobility.***

Index Terms—Traffic Congestion Prediction, Machine Learning, XGBoost, Random Forest, Policy Simulation, Bus Lane Impact, Carpooling, Explainable AI, Bengaluru Traffic

I. INTRODUCTION

Urban traffic congestion has become one of the most critical challenges faced by rapidly growing metropolitan cities across the world. The continuous increase in urban population, economic activity, industrial development, and private vehicle ownership has placed tremendous pressure on existing transportation infrastructure. As cities continue to expand geo-graphically and economically, urban mobility systems struggle to maintain efficient traffic flow. Traffic congestion directly affects commuters by increasing travel delays, fuel consumption, transportation costs, commuter stress, and exposure to air pollution. It also affects overall city productivity because time lost in traffic reduces working hours, increases logistics costs, and lowers the efficiency of urban economic systems.

Bengaluru, widely recognized as one of India's major technology and employment hubs, experiences severe traffic congestion due to rapid urbanization and growing dependency on private vehicles. The city attracts a large number of professionals, students, businesses, and industries, resulting in continuously increasing transportation demand. However, road infrastructure development has not expanded at the same pace as vehicle growth. As a result, several corridors and junctions in Bengaluru frequently experience long vehicle queues, reduced average speeds, high travel time variability, and increased road capacity stress.

The consequences of congestion are not limited to mobility delays. One of the most important outcomes of urban congestion is excessive fuel wastage. Vehicles operating un-der stop-and-go traffic conditions consume fuel inefficiently because engines continue running even when vehicles are moving slowly or idling. This increases petrol and diesel consumption without corresponding useful

movement. In the context of rising global fuel demand, petroleum supply un-certainty, geopolitical conflicts, and fuel price fluctuations, reducing fuel wastage through intelligent traffic management has become highly important. Since petroleum-based fuels are non-renewable resources, improving traffic efficiency can contribute to long-term energy conservation and sustainable urban development.

Traffic congestion also contributes significantly to environ-mental degradation. High vehicle density increases emissions of carbon dioxide, carbon monoxide, nitrogen oxides, and particulate matter. These emissions contribute to poor urban air quality and public health issues. In dense urban areas, traffic-related emissions also contribute to heat generation and noise pollution. Therefore, congestion prediction and traffic policy analysis are not only transportation problems but also environmental and sustainability issues.

Traditional traffic management systems mainly depend on fixed signal timings, manual monitoring, road widening, fly-overs, and reactive policy decisions. While these methods provide basic traffic regulation, they are often insufficient for modern dynamic traffic systems. Urban traffic patterns change due to multiple interacting factors such as traffic volume, vehicle speed, road capacity utilization, accident occurrence, weather conditions, public transport usage, parking demand, pedestrian movement, roadwork activity, and signal compliance. Traditional statistical approaches often struggle to model these nonlinear and interdependent relationships effectively.

Machine Learning (ML) has emerged as an effective approach for intelligent transportation systems because it can learn patterns from large datasets and generate predictive insights. ML models can analyze historical traffic behavior, identify relationships among traffic variables, and predict con-gestion levels with improved accuracy. Compared to traditional rule-based systems, machine learning models are more flexible and can adapt to complex traffic conditions. Ensemble learning methods such as Random Forest and XGBoost are especially useful for structured tabular datasets because they handle nonlinear feature interactions effectively.

In addition to prediction accuracy, modern intelligent transportation systems require interpretability. Traffic authorities and policymakers need to understand why a model predicts high congestion and which factors contribute most strongly to that prediction. Black-box prediction without explanation may not be useful for decision-making. Therefore, Explain-able Artificial Intelligence (XAI) methods such as SHAP are important because they provide feature-level explanations for model outputs.

This research focuses on developing a machine learning-based framework for traffic congestion prediction and simulation-based evaluation of urban transport policies using Bengaluru traffic datasets. Instead of claiming actual real-world implementation of traffic policies, this study performs policy simulation by modifying engineered policy-related features and observing changes in model-predicted congestion levels. This makes the approach suitable for evaluating hypothetical strategies such as bus lane prioritization, public transport improvement, and carpooling enhancement.

The proposed framework integrates data preprocessing, exploratory data analysis, feature engineering, predictive modeling, model comparison, SHAP explainability, residual validation, and policy simulation. Additional engineered indicators such as Bus Lane Impact, Transit Efficiency, Carpool Efficiency, and Fuel Wastage Index are created to incorporate policy-aware and sustainability-oriented factors into the analysis. The final objective is to demonstrate how machine learning can support data-driven traffic planning and intelligent urban mobility management.

A. Problem Statement

Existing traffic prediction systems mainly focus on forecasting traffic flow, vehicle speed, or congestion using historical data. However, many approaches do not provide a practical mechanism to evaluate how policy-level interventions may affect congestion. In real urban planning, policymakers require more than prediction values; they need insight into whether interventions such as dedicated bus lanes, public transport improvements, and carpooling promotion may help reduce congestion. In addition, many machine learning models lack interpretability, making

it difficult to understand why a model produces a particular congestion prediction. This creates a gap between traffic prediction research and policy-support systems.

B. Objectives

The main objectives of this study are:

- To preprocess and analyze Bengaluru traffic data for congestion-related patterns.
- To develop machine learning models for traffic congestion prediction.
- To compare Linear Regression, Random Forest, and XG-Boost models.
- To engineer policy-aware features related to bus lanes, public transport, and carpooling.
- To perform simulation-based policy evaluation using trained machine learning models.
- To apply SHAP explainability for interpreting model behavior.
- To validate model performance using residual analysis.

C. Contribution

The major contributions of this paper are:

- A complete machine learning pipeline for Bengaluru traffic congestion prediction.
- Feature engineering for policy-aware traffic modeling.
- Comparative model evaluation using MAE, MSE, and R2.
- Policy simulation for bus lane, public transport, and carpooling scenarios.
- Explainable AI analysis using SHAP to interpret feature influence.
- Residual analysis to verify model prediction stability.

II. RELATED WORK

Traffic congestion prediction has been widely studied in intelligent transportation systems. Over time, research has evolved from traditional statistical methods to machine learn-ing, deep learning, graph-based models, and explainable AI approaches. Each generation of methods has contributed to improving prediction performance, but several limitations re-main when these methods are applied to policy-level traffic planning.

A. Traditional Approaches

Early traffic forecasting approaches relied on statistical and time-series models such as Historical Average, ARIMA, and VAR. These models are easy to implement and interpret, making them useful for basic traffic forecasting. However, they assume relatively stable relationships in data and often fail to capture nonlinear traffic behavior. Urban traffic is highly dynamic, and congestion can change rapidly due to road incidents, signal delays, weather, and peak-hour travel demand. Traditional methods usually perform poorly when traffic conditions become irregular or highly congested.

B. Machine Learning Approaches

Machine learning approaches improved traffic prediction by learning nonlinear relationships from historical data. Models such as Support Vector Machines, Decision Trees, Random Forest, Gradient Boosting, and regression-based models have been used for traffic flow, speed, and congestion prediction. These models are effective for structured tabular datasets to traffic volume, speed, congestion, road capacity utilization, incidents, public transport usage, parking behavior, pedestrian movement, weather conditions, and roadwork activity.

B. Machine Learning Approaches

Machine learning approaches improved traffic prediction by learning nonlinear relationships from historical data. Models such as Support Vector Machines, Decision Trees, Random Forest, Gradient Boosting, and regression-based models have been used for traffic flow, speed, and congestion prediction. These models are effective for structured tabular datasets and can handle multiple input features. Random Forest is particularly useful because it combines multiple decision trees and reduces overfitting compared to a single tree. However, some machine learning models may still lack interpretability unless additional explanation methods are used.

C. Deep Learning Approaches

Deep learning methods such as LSTM, GRU, CNN, and Graph Neural Networks have been widely used for spatio-temporal traffic prediction. LSTM and GRU models are effective for sequence learning because they capture temporal dependencies in traffic data. CNN-based models can identify spatial patterns when

traffic networks are represented in grid-like structures. Graph Neural Networks are especially suitable for road networks because roads and junctions naturally form graph structures. However, deep learning models generally require large-scale, high-frequency, and clean spatio-temporal datasets. When the available dataset is structured and tabular rather than high-frequency sensor data, ensemble tree-based models may be more suitable.

D. Hybrid and Recent Approaches

Recent traffic prediction studies have explored hybrid methods, combining graph models, attention mechanisms, transformers, wavelet transforms, and multi-task learning. These models improve prediction accuracy by capturing complex spatial and temporal dependencies. However, many of these approaches focus mainly on accuracy metrics such as MAE, RMSE, and R^2. They often do not evaluate how transportation policies may influence congestion. This creates a research gap in policy-aware traffic modeling.

E. Explainable AI in Traffic Prediction

Explainable AI has become important in traffic prediction because urban planners need transparent decision-support tools. SHAP and feature importance methods help identify which variables are most influential in prediction. In traffic management, this is valuable because policymakers can understand whether congestion is mainly influenced by traffic volume, road capacity, public transport usage, parking pressure, or other factors. This study uses SHAP to interpret the XGBoost model and support policy-level interpretation.

III. DATASET DESCRIPTION

This study uses Bengaluru traffic-related datasets collected from public sources. The primary dataset contains structured traffic observations from different areas of Bengaluru, supporting road accident datasets is used for safety-oriented descriptive analysis. The datasets contain information related to traffic volume, speed, congestion, road capacity utilization, incidents, public transport usage, parking behavior, pedestrian movement, weather conditions, and roadwork activity.

A. Dataset Attributes

TABLE I IMPORTANT DATASET FEATURES

Feature	Description
Area Name	Location or region in Bengaluru where traffic observation is recorded.
Traffic Volume	Number of vehicles or traffic load observed in the area.
Average Speed / Congestion Level	Average movement speed of vehicles. Target variable representing traffic congestion intensity.
Travel Time Index	Indicator of delay compared to normal travel conditions.
Road Capacity Utilization	Percentage usage of available road capacity.
Public Transport Usage	Usage level of public transport systems.
Traffic Signal Compliance	Indicator of signal-following behavior.
Parking Usage / Pedestrian and Cyclist Count	Parking demand or usage level. Non-motorized mobility count.
Weather Conditions / Roadwork Activity	Weather-related contextual factor. Indicator of road construction or maintenance.

B. Data Relevance

The dataset is suitable for congestion prediction because it contains both traffic-flow variables and contextual variables. Traffic Volume, Average Speed, Travel Time Index, and Road Capacity Utilization directly describe traffic behavior. Public Transport Usage, Parking Usage, and Pedestrian and Cyclist Count provide mobility context. Incident Reports and Roadwork Activity help capture disruptions. These features allow the model to learn congestion patterns from multiple dimensions rather than relying only on vehicle count.

C. Supporting Accident Dataset

The accident dataset was used to understand the safety context of urban traffic. High traffic density and congestion-prone areas may also experience higher accident risk due to frequent braking, lane changing,

and driver frustration. Although accident prediction was not the main objective of this study, the accident dataset supported the broader argument that intelligent traffic management is important for both efficiency and safety.

IV. DATA PREPROCESSING AND EXPLORATORY DATA ANALYSIS

Data preprocessing is a critical stage in machine learning-based traffic modeling. Raw traffic datasets may contain missing values, inconsistent formats, duplicate entries, outliers, and hidden formatting issues. These problems can reduce model performance and lead to inaccurate conclusions if not handled carefully.

A. Data Cleaning

The dataset was first inspected for missing values, duplicate records, inconsistent column names, and unsuitable data types. Date fields were converted into proper datetime format where applicable. Column names were standardized to avoid errors caused by hidden spaces or special characters. Duplicate rows were checked and removed where necessary. Missing values were handled based on feature type and availability.

B. Outlier Analysis

Outlier detection was performed using boxplots, interquartile range analysis, and z-score inspection. In traffic datasets, extreme values do not always represent errors because peak-hour congestion and traffic surges are real conditions. Therefore, aggressive outlier removal was avoided. Instead, the analysis focused on identifying whether extreme observations were realistic or corrupted. Most outliers represented genuine high-traffic situations and were retained to preserve realistic congestion patterns.

C. Exploratory Data Analysis

Exploratory Data Analysis was performed to understand the behavior of traffic variables. Congestion distribution was analyzed to identify whether most observations belonged to low, moderate, or high congestion ranges. Area-wise analysis was conducted to identify regions with higher average congestion. Monthly trends were examined to understand temporal variation. Fuel wastage patterns were analyzed using the engineered Fuel Wastage Index. Accident distribution was used to support the safety relevance of traffic management.

The analysis showed that Traffic Volume and Congestion Level were strongly related. Area-wise analysis indicated that regions such as Koramangala and M.G. Road showed higher congestion levels compared to areas such as Electronic City. This supports the need for location-sensitive traffic planning rather than a single uniform policy for all regions.

D. Correlation Analysis

Correlation analysis was used to examine relationships among numerical variables. Strong correlation between Traffic Volume and Congestion Level confirmed that vehicle density is a major congestion driver. Negative relationships between Average Speed and congestion-related variables indicated that speed decreases as congestion increases. These findings are consistent with real-world traffic behavior and support the validity of the dataset for congestion prediction.

V. FEATURE ENGINEERING

Feature engineering was performed to support congestion prediction and policy simulation. In addition to original dataset variables, derived indicators were created to represent sustainability and policy-related factors.

A. Fuel Wastage Index

Fuel wastage was estimated using congestion level, travel time index, and average speed. Higher congestion and travel delay increase fuel wastage, while higher speed generally indicates smoother traffic movement.

$$FWI = \frac{Congestion\ Level \times Travel\ Time\ Index}{Average\ Speed} \quad (1)$$

B. Bus Lane Impact

Bus Lane Impact was created as a composite indicator using Public Transport Usage, Congestion Level, and Traffic Volume. The logic behind this feature is that bus lane policies become more relevant in areas where public transport demand, congestion, and traffic volume are high.

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717886

$$BLI = \frac{Public\ Transport\ Usage \times Congestion\ Level \times Traffic\ Volume}{10000}$$

C. Carpool Efficiency

Carpool Efficiency was designed to represent traffic pressure under low-speed and high-volume conditions. Higher vehicle volume with lower speed suggests that shared mobility strategies may be useful.

$$CE = \frac{Traffic\ Volume}{Average\ Speed + 1} \quad (3)$$

D. Transit Efficiency

Transit Efficiency was created to represent the effectiveness of public transport movement under congestion conditions. Higher public transport usage and speed improve transit efficiency, while high congestion reduces it.

$$TE = \frac{Public\ Transport\ Usage \times Average\ Speed}{Congestion\ Level + 1} \quad (4)$$

E. Importance of Engineered Features

The engineered features allow the model to go beyond simple traffic prediction. They help represent transportation policy conditions inside the machine learning framework. This makes it possible to perform simulation-based policy analysis by modifying these features and observing changes in predicted congestion.

VI. PROPOSED METHODOLOGY

The proposed framework consists of six major stages: data collection, preprocessing, feature engineering, model training, explainability analysis, and policy simulation. The methodology is designed to predict congestion accurately and also provide interpretable support for policy-level traffic management.

Traffic Dataset
↓
Data Cleaning and EDA
↓
Feature Engineering
↓
ML Models: LR, RF, XGBoost
↓
Model Evaluation: MAE, MSE, R²
↓
SHAP Explainability
↓
(2) Policy Simulation

Fig. 1. Proposed machine learning-based traffic congestion prediction and policy simulation architecture.

A. Workflow Explanation

The workflow begins with traffic data collection and preprocessing. After cleaning and preparing the dataset, exploratory data analysis is performed to understand congestion trends and feature relationships. Next, feature engineering is applied to create policy-aware indicators. Machine learning models are then trained using selected features. The best-performing model is interpreted using SHAP analysis. Finally, policy simulation is performed by modifying engineered features such as Bus Lane Impact, Transit Efficiency, and Carpool Efficiency.

B. Machine Learning Models

Three models were implemented in this study.

1) Linear Regression:

Linear Regression was used as a baseline model. It assumes a linear relationship between input features and the target variable. Although traffic congestion is often nonlinear, Linear Regression provides a simple benchmark for comparison.

2) Random Forest Regressor:

Random Forest is an ensemble learning method that builds multiple decision trees and averages their predictions. It reduces overfitting and handles nonlinear relationships better than a single decision tree. Random Forest is suitable for tabular traffic datasets because it can model complex interactions among traffic variables.

3) XGBoost Regressor:

XGBoost is a gradient boosting algorithm that builds trees sequentially. Each new tree corrects the errors of previous trees. It is efficient, scalable, and highly effective for structured datasets. XGBoost was expected to perform well because traffic congestion depends on nonlinear relationships among multiple variables.

C. Evaluation Metrics

Model performance was evaluated using Mean Absolute Error, Mean Squared Error, and R^2Score.

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717886

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y_i} | \qquad (5)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y_i})^2 \qquad (6)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (7)$$

MAE measures average prediction error, MSE penalizes large errors, and R^2indicates how well the model explains variance in congestion level.

VII. EXPERIMENTAL RESULTS

The experimental evaluation compared Linear Regression, Random Forest, and XGBoost models using MAE, MSE, and R^2Score metrics. The target variable was Congestion Level. The dataset was divided into training and testing sets using an 80:20 split.

TABLE II MODEL PERFORMANCE COMPARISON

Model	MAE	MSE	R^2
Linear Regression	4.924553	36.606936	0.932331
Random Forest	2.674420	16.565614	0.969210
XGBoost	1.539835	5.776934	0.989321

```
              Model       MAE        MSE  R2 Score
0  Linear Regression  4.924553  36.606936  0.932331
1      Random Forest  2.674420  16.656614  0.969216
2            XGBoost  1.539385   5.776934  0.989321
```

Fig. 2. Model performance comparison output.

The Linear Regression model achieved an R^2score of 0.932, indicating that the relationship between selected traffic features and congestion level is strong. However, its higher MAE and MSE show that it is less effective in capturing nonlinear traffic behavior.

The Random Forest model improved prediction performance with an R^2score of 0.969. This improvement shows that ensemble tree-based models are better suited for congestion prediction because they can handle nonlinear feature interactions.

The XGBoost model achieved the best predictive performance with an R^2score of approximately 0.989. It also produced the lowest MAE and MSE values, indicating high prediction accuracy and strong generalization capability.

A. Actual vs Predicted Analysis

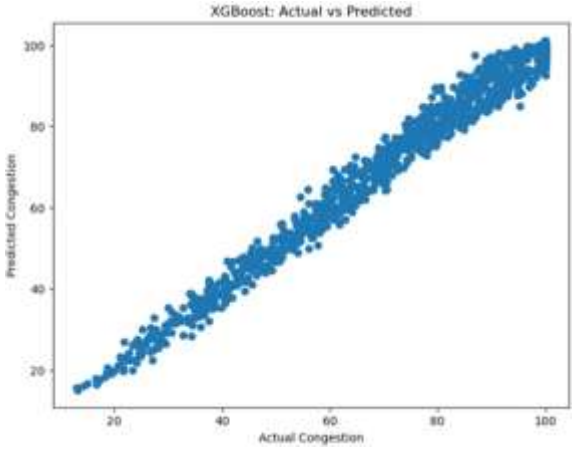

Fig. 3. Actual vs Predicted Congestion using XGBoost

The Actual vs Predicted scatter plot demonstrates that predicted congestion values closely align with actual congestion values. The points are concentrated near the diagonal trend, indicating that the model makes accurate predictions across different congestion levels.

B. Residual Analysis

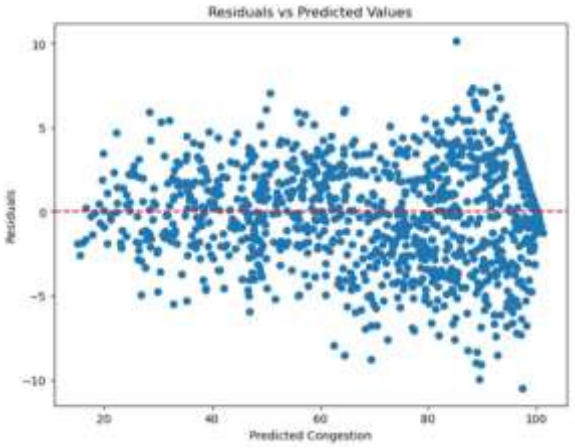

Fig. 4. Residuals vs Predicted Values for XGBoost model.

Residual analysis was performed to evaluate prediction errors. The residuals are distributed around zero without major systematic patterns, indicating stable model performance. Some variation is observed

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717886

at high congestion levels, which is expected because extreme congestion conditions are more difficult to predict.

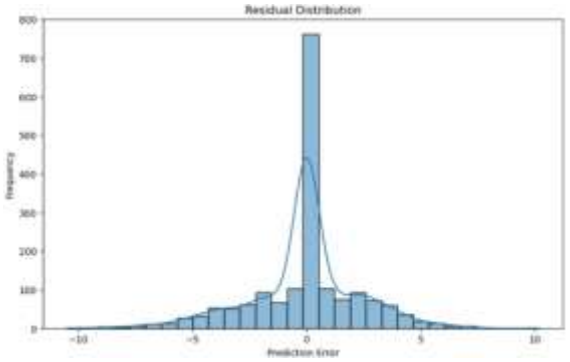

Fig. 5. Residual Distribution of XGBoost prediction errors.

The residual distribution approximately follows a normal distribution centered around zero. This indicates minimal prediction bias and supports the reliability of the XGBoost model.

VIII. EXPLAINABLE AI ANALYSIS

Explainable Artificial Intelligence techniques were used to interpret the predictions generated by the XGBoost model. SHAP analysis was performed to identify the contribution of each feature toward congestion prediction.

A. SHAP Feature Importance

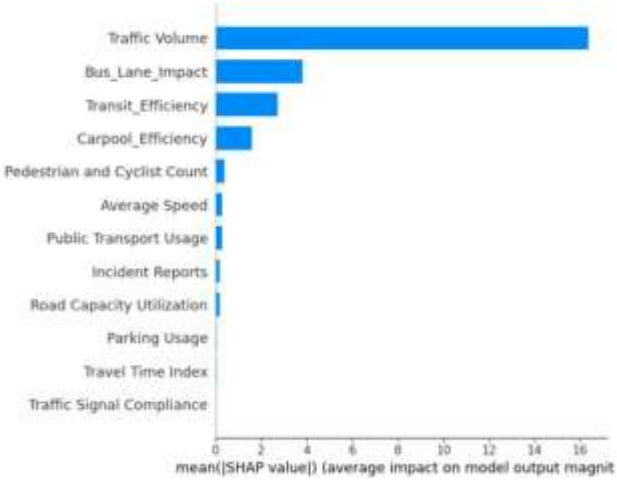

Fig. 6. SHAP Feature Importance Analysis.

The SHAP feature importance plot shows that Traffic Volume is the most influential feature affecting congestion prediction. This result is expected because higher vehicle volume directly increases road occupancy and reduces movement efficiency. Bus Lane Impact, Transit Efficiency, and Carpool Efficiency also contribute to the prediction process, although their impact is smaller than Traffic Volume.

B. XGBoost Feature Importance

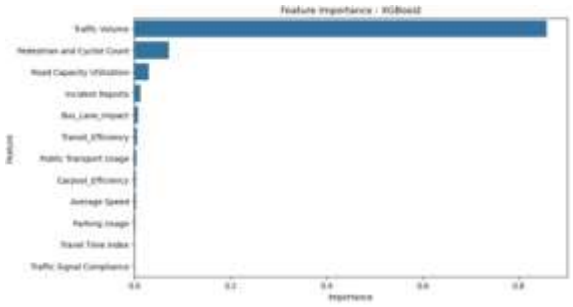

Fig. 7. Feature Importance obtained from XGBoost model.

The XGBoost feature importance output confirms that Traffic Volume dominates congestion prediction. Other features such as Pedestrian and Cyclist Count, Road Capacity Utilization, Incident Reports, and Bus Lane Impact provide additional contextual contribution.

C. SHAP Summary Analysis

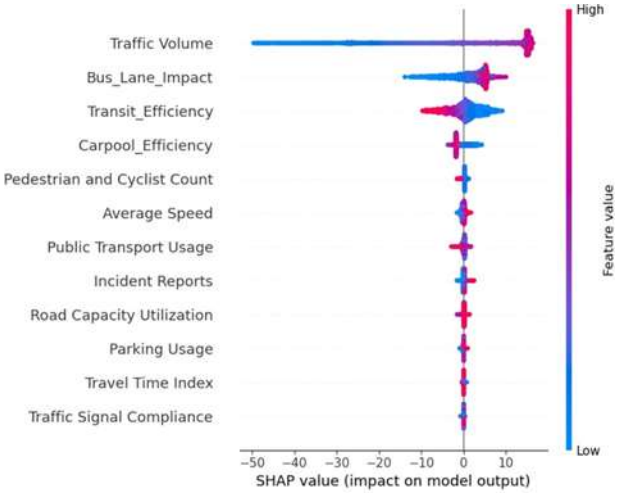

Fig. 8. SHAP Summary Plot for XGBoost model interpretation.

The SHAP summary plot demonstrates how individual feature values influence the predicted congestion level. Higher Traffic Volume values increase congestion prediction significantly. Transit Efficiency and Bus Lane Impact provide policy-related interpretability by showing how public transport-oriented features influence congestion prediction.

IX. POLICY SIMULATION

Policy simulation was performed by modifying engineered features and observing changes in predicted congestion. The simulated scenarios included bus lane prioritization, public transport improvement, and carpooling efficiency. The purpose of policy simulation is not to claim real-world implementation effects, but to estimate possible congestion changes under hypothetical traffic management strategies.

A. Bus Lane Policy Simulation

Bus lane simulation was represented through the Bus Lane Impact feature. The idea is that dedicated bus lanes may improve public transport movement and reduce the number of private vehicles if commuters shift toward buses. In this study, the simulation modifies the policy-related feature and observes the model-predicted congestion response. This provides an approximate understanding of how bus-lane-oriented planning may influence congestion under modeled conditions.

B. Public Transport Improvement Simulation

Public transport improvement was represented through Transit Efficiency. Higher transit efficiency indicates better public transport usage and smoother movement. Simulation results showed that improvements in public transport-related indicators can contribute to lower predicted congestion. This supports the idea that public transport strengthening is important for sustainable traffic management.

C. Carpooling Simulation

Carpooling simulation was represented through Carpool Efficiency. Carpooling reduces the number of individual private vehicles by increasing passenger occupancy per vehicle. The simulation showed that carpooling has a smaller but relevant influence. This suggests that carpooling alone may not solve congestion but can support broader traffic management strategies when combined with public transport improvements and lane prioritization.

D. Interpretation of Policy Simulation

The simulation results indicate that traffic volume remains the dominant congestion factor. Therefore, isolated policy interventions may produce limited impact unless they reduce total vehicle volume or improve road capacity utilization. This finding is important because it suggests that sustainable congestion management requires integrated strategies rather than single-policy solutions.

X. DISCUSSION

The findings show that machine learning can effectively predict congestion using structured Bengaluru traffic data. XGBoost performed best because it can model nonlinear relationships and interactions among traffic variables. The high R^2 score and low prediction error values indicate that the selected features are useful for congestion prediction.

SHAP analysis confirmed that Traffic Volume dominates congestion prediction, which aligns with real-world traffic behavior. When the number of vehicles increases, road capacity becomes saturated, vehicle speed decreases, and congestion rises. This validates the model from a transportation perspective.

The policy simulation component adds practical relevance to the research. It allows policymakers and transport planners to test hypothetical interventions before real-world deployment. However, the simulation results should be interpreted carefully because the dataset does not contain actual before-and-after bus lane implementation data. Therefore, the policy results represent model-based estimates rather than causal proof.

The study also highlights the importance of explainability. Without SHAP analysis, the model would only provide pre-diction values. With SHAP, it becomes possible to understand which variables influence congestion the most. This improves trust and makes the model more useful for decision support.

XI. APPLICATIONS

The proposed framework can be applied in several real-world scenarios. Urban traffic authorities can use the model to predict congestion-prone areas and support planning decisions. Smart city systems can integrate such models into intelligent transportation platforms. Public transport agencies can use policy simulation to understand whether improving bus movement or transit efficiency may reduce congestion.

The framework can also support carpooling and shared mobility planning. Ride-sharing platforms can use congestion predictions to optimize routes and encourage shared trips in high-congestion areas. Environmental agencies can use fuel wastage indicators to identify areas where congestion causes excessive emissions.

In academic research, the framework contributes to policy-aware traffic prediction by combining machine learning, feature engineering, explainable AI, and simulation-based policy evaluation. This makes the work useful for future research in intelligent transportation systems and sustainable urban mobility.

XII. LIMITATIONS

The study has several limitations. First, the dataset is not real-time sensor data. It does not contain high-frequency traffic observations such as minute-wise or second-wise vehicle flow. Second, hourly traffic granularity is limited, so detailed peak-hour prediction could not be performed. Third, real-world bus lane deployment data is unavailable. Therefore, bus lane analysis is performed through simulation using engineered features rather than actual before-and-after field measurements.

Fourth, policy simulation is predictive and not causal. The model estimates possible outcomes based on learned patterns, but it cannot prove that a policy will produce the same result in real life. Fifth, external factors such as fuel prices, festivals, public events, school schedules, metro disruptions, and GPS trajectory data were not included. These factors can significantly affect urban traffic patterns.

XIII. FUTURE SCOPE

Future work can extend this research using real-time GPS data, traffic camera feeds, and IoT sensor data. High-frequency temporal data would allow the use of deep learning models such as LSTM, GRU, and Graph Neural Networks. Road network graph data can enable spatio-temporal modeling of congestion propagation between connected road segments.

Future studies can also integrate reinforcement learning for adaptive traffic signal control. Another direction is real-world policy validation using actual bus lane implementation or public transport improvement data. The framework can be extended to other cities such as Mumbai, Delhi, Chennai, Hyderabad, and Pune to test generalizability.

Additional sustainability metrics such as carbon emissions, fuel cost, vehicle occupancy, and public transport accessibility can also be included. Integration with smart city dashboards can help authorities monitor congestion and evaluate policies in real time.

XIV. CONCLUSION

This paper presented a machine learning-based framework for traffic congestion prediction and policy simulation using Bengaluru traffic data. The study implemented Linear Regression, Random Forest, and XGBoost models and compared their performance. XGBoost achieved the best performance with an R2 score of approximately 0.989.

Feature engineering enabled policy-aware analysis through Bus Lane Impact, Transit Efficiency, Carpool Efficiency, and Fuel Wastage Index indicators. SHAP explainability showed that Traffic Volume was the dominant congestion factor, while policy-related variables contributed to interpretation and simulation. The proposed framework demonstrates that machine learn-ing can support predictive traffic management and simulation-based policy evaluation. Although real-world policy deployment data was unavailable, the study provides a practical and interpretable approach for analyzing urban congestion and sustainable transport strategies.

REFERENCES

[1] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[2] B. Yu, H. Yin, and Z. Zhu, "Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2020.

[3] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang, "Graph WaveNet for Deep Spatial-Temporal Graph Modeling," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

[4] C. Zheng, X. Fan, C. Wang, and J. Qi, "GMAN: A Graph Multi-Attention Network for Traffic Prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 1, pp. 1234–1241, 2021.

[5] J. Chen, Y. Lin, Z. Zhao, and L. Wang, "A Survey on Traffic Prediction Using Deep Learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 5, pp. 4821–4840, 2023.

[6] Y. Lv, Y. Duan, W. Kang, Z. Li, and F. Wang, "Traffic Flow Prediction with Big Data: A Deep Learning Approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 865–873, 2015.

[7] X. Ma, Z. Tao, Y. Wang, H. Yu, and Y. Wang, "Long Short-Term Memory Neural Network for Traffic Speed Prediction Using Remote Microwave Sensor Data," *Transportation Research Part C: Emerging Technologies*, vol. 54, pp. 187–197, 2015.

[8] J. Ke, H. Zheng, H. Yang, and X. Chen, "Short-Term Forecasting of Passenger Demand Under On-Demand Ride Services: A Spatio-Temporal Deep Learning Approach," *Transportation Research Part C*, vol. 85, pp. 591–608, 2017.

[9] J. Zhang, Y. Zheng, and D. Qi, "Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.

[10] Y. Yuan and X. Li, "Traffic Congestion Prediction Based on Gradient Boosting Decision Trees," *Journal of Advanced Transportation*, vol. 2018, pp. 1–10, 2018.

[11] J. Wang, Q. Gu, Y. Wu, G. Liu, and Z. Xiong, "Traffic Speed Prediction and Congestion Source Exploration: A Deep Learning Method," in *IEEE International Conference on Big Data*, 2019.

[12] D. Xu, Y. Wang, L. Peng, and H. Song, "Real-Time Road Traffic State Prediction Based on Kernel-KNN," *Transportation Research Part C*, vol. 117, pp. 1–15, 2020.

[13] A. Ranjan and P. Sahu, "Urban Traffic Congestion Prediction Using Random Forest and XGBoost Algorithms," *International Journal of Intelligent Transportation Systems Research*, vol. 19, no. 4, pp. 433–445, 2021.

[14] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

[15] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[16] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[17] S. Lundberg and S. Lee, "A Unified Approach to Interpreting Model Predictions," in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[18] C. Molnar, *Interpretable Machine Learning* (2nd ed.). Lulu Publications, 2020.

[19] F. Doshi-Velez and B. Kim, "Towards a Rigorous Science of Interpretable Machine Learning," *arXiv preprint arXiv:1702.08608*, 2017.

[20] M. Treiber and A. Kesting, *Traffic Flow Dynamics: Data, Models and Simulation.* Springer, Berlin, Germany, 2013.

[21] H. Chourabi, T. Nam, S. Walker, J. R. Gil-Garcia, S. Mellouli, K. Nahon, T. Pardo, and H. J. Scholl, "Understanding Smart Cities: An Integrative Framework," in *45th Hawaii International Conference on System Sciences*, pp. 2289–2297, 2012.

[22] P. Newman and J. Kenworthy, *Sustainability and Cities: Overcoming Automobile Dependence.* Island Press, Washington, DC, USA, 1999.

[23] D. Banister, "The Sustainable Mobility Paradigm," *Transport Policy*, vol. 15, no. 2, pp. 73–80, 2008.

[24] Kaggle, "Bangalore City Traffic Dataset." Available: https://www.kaggle.com/datasets/preethamgouda/banglore-city-traffic-dataset/data

[25] OpenCity, "Bengaluru Traffic Signal and Traffic Police Open Data." Available: https://data.opencity.in/dataset/?organization=bengaluru-traffic-police

[26] OpenCity, "Bengaluru City Traffic Signal Data." Available: https://data.opencity.in/dataset/bengaluru-city-traffic-signal-data