International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Comparative Analysis of Machine Learning Regression Models for Urban Traffic Congestion Prediction

Ms. Kajal Singh[1], Dr. Ashish Chourey[2], Dr. Mohit Singh Tomar[3], Mr. Arun Kumar Jhapate[4]

[1]M.Tech, Computer Science & Engineering, Sagar Institute of Research & Technology, Bhopal
[2,4]Assistant Professor, Computer Science & Engineering, Sagar Institute of Research & Technology, Bhopal
[3]Head of Department & Associate Professor, Computer Science & Engineering, Sagar Institute of Research & Technology, Bhopal

Abstract
Traffic congestion has become a major challenge in urban transportation systems due to rapid urbanization and increasing vehicle ownership. This study proposes a machine learning regression framework for predicting traffic congestion in urban road networks using historical traffic data. Multiple regression algorithms, including Linear Regression, Decision Tree, Random Forest, Support Vector Regression (SVR), Gradient Boosting, and XGBoost, are evaluated to identify the most effective prediction model. The proposed methodology incorporates data preprocessing, feature selection, model training, and comparative performance analysis. Model performance is assessed using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and the Coefficient of Determination (R^2). Experimental results demonstrate that ensemble regression models outperform conventional regression techniques in terms of prediction accuracy and robustness. The findings indicate that XGBoost provides the best overall performance while maintaining computational efficiency. The proposed framework supports intelligent transportation systems by enabling accurate and timely traffic congestion prediction. Furthermore, the framework can assist transportation authorities in improving traffic management and reducing congestion-related problems. Future work will focus on integrating real-time IoT data and explainable artificial intelligence techniques to further enhance prediction accuracy and practical deployment.

Keywords: Traffic Congestion Prediction, Machine Learning, Regression Models, Urban Road Networks, Intelligent Transportation Systems, Traffic Flow Forecasting, XGBoost, Random Forest, Smart Cities, Data Analytics.

1. Introduction

Rapid urbanization and the continuous growth in vehicle ownership have significantly increased traffic congestion in cities worldwide. Urban road networks frequently experience heavy traffic during peak hours, resulting in longer travel times, increased fuel consumption, environmental pollution, and economic losses. Traffic congestion also affects emergency response services, public transportation efficiency, and overall road safety. Therefore, developing intelligent and accurate traffic congestion prediction systems has become an important research area within Intelligent Transportation Systems (ITS).

Traditional traffic prediction approaches mainly rely on statistical and mathematical models, such as Linear Regression and Auto-Regressive Integrated Moving Average (ARIMA), to estimate future traffic conditions. Although these techniques are computationally efficient and easy to implement, they often fail to capture the complex, nonlinear, and dynamic characteristics of real-world traffic systems. Urban traffic is influenced by numerous factors, including vehicle volume, traffic speed, weather conditions, road infrastructure, accidents, special events, and temporal variations, making accurate prediction a challenging task.

Recent advances in machine learning have provided effective solutions for modeling complex traffic patterns. Machine learning regression algorithms can learn hidden relationships from historical traffic data and generate accurate predictions without requiring predefined mathematical assumptions. Algorithms such as Decision Tree Regression, Random Forest Regression, Support Vector Regression (SVR), Gradient Boosting Regression, XGBoost, and Extra Trees Regression have demonstrated promising performance in traffic forecasting because of their ability to model nonlinear relationships and handle high-dimensional datasets. These approaches contribute to more reliable congestion prediction and improved decision-making in intelligent transportation systems.

Accurate traffic congestion prediction provides numerous practical benefits. Reliable forecasting enables transportation authorities to optimize traffic signal timing, improve route planning, reduce travel delays, enhance fuel efficiency, and minimize environmental emissions. It also supports navigation systems by providing real-time traffic information that helps drivers select less congested routes. Furthermore, accurate prediction contributes to the development of smart cities by improving mobility, enhancing road safety, and supporting sustainable urban transportation.

Despite considerable progress, several challenges remain in traffic congestion prediction. Many existing studies rely on limited datasets collected from individual cities, reducing the generalizability of their models. Some approaches emphasize prediction accuracy while overlooking computational efficiency, model interpretability, and real-time deployment requirements. In addition, inconsistent evaluation methodologies make it difficult to compare the performance of different regression algorithms under identical conditions.

To address these challenges, this study proposes a comprehensive machine learning regression framework for predicting traffic congestion in urban road networks. Multiple regression algorithms are evaluated using standardized preprocessing techniques and common performance metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and the Coefficient of Determination (R^2). A comparative analysis is conducted to identify the most accurate and computationally efficient regression model suitable for real-time traffic congestion prediction.

The main contributions of this research are to develop a systematic traffic congestion prediction framework, compare the performance of multiple machine learning regression algorithms under identical experimental conditions, evaluate their prediction accuracy and computational efficiency using standardized metrics, and identify the most suitable regression model for intelligent transportation systems. The outcomes of this study are expected to support urban traffic management authorities in reducing congestion and improving transportation efficiency while contributing to the advancement of smart city technologies.

2. Litracture Review

2.1 Traffic Congestion Prediction in Urban Transportation

Traffic congestion has emerged as a critical issue in modern urban transportation systems due to rapid urbanization, population growth, and the continuous rise in vehicle ownership. It leads to increased travel time, higher fuel consumption, environmental pollution, road accidents, and significant economic losses. As a result, accurate traffic congestion prediction has become a vital component of Intelligent Transportation Systems (ITS). It enables transportation authorities to optimize traffic signal operations, improve route planning, reduce congestion, and enhance overall road safety. In recent years, there has been growing interest in developing intelligent prediction models capable of forecasting traffic conditions in advance to support real-time traffic management and decision-making processes **[1]**.

2.2 Traditional Traffic Prediction Methods

Traditional traffic prediction approaches mainly rely on statistical techniques such as Linear Regression, Auto-Regressive Integrated Moving Average (ARIMA), and other time-series models. These methods are relatively simple to implement and computationally efficient. However, they typically assume linear relationships among traffic variables, which limits their ability to capture the complex, nonlinear, and dynamic nature of urban traffic systems. Consequently, their predictive performance often declines under rapidly changing traffic conditions caused by factors such as accidents, weather variations, road construction, and special events **[1], [3]**.

2.3 Machine Learning-Based Prediction Methods

To address the limitations of traditional approaches, machine learning techniques have been increasingly adopted for traffic congestion prediction. These methods are capable of learning complex nonlinear relationships from large-scale traffic datasets, making them more suitable for modeling real-world traffic behavior. Regression-based machine learning algorithms have demonstrated strong potential in predicting traffic flow, speed, and congestion levels using historical data. Berlotti et al. (2024) proposed a machine learning framework for traffic flow forecasting and showed that regression models can achieve higher prediction accuracy compared to conventional statistical methods when combined with effective feature engineering and data preprocessing techniques **[2]**. Similarly, Lv et al. (2015) introduced a deep learning-based framework that successfully captured nonlinear traffic patterns, emphasizing the importance of intelligent learning methods in transportation applications **[3]**.

2.4 Data Integration and Technological Advancements

Recent research highlights the importance of integrating multiple traffic-related features to enhance prediction performance. In addition to historical traffic flow, variables such as vehicle speed, traffic density, weather conditions, GPS trajectories, road network characteristics, and temporal factors are increasingly incorporated into predictive models. The integration of heterogeneous data sources allows machine learning algorithms to better represent real-world traffic dynamics and improve forecasting accuracy under both normal and congested conditions **[14], [20]**.

Advancements in sensing technologies, Internet of Things (IoT) devices, intelligent traffic cameras, and connected vehicles have significantly increased the availability of real-time traffic data. These developments have enabled the creation of large-scale datasets that support data-driven traffic prediction models for smart city applications. Machine learning algorithms can effectively analyze these continuously generated datasets to identify traffic patterns, estimate congestion levels, and provide timely predictions that assist transportation authorities in implementing proactive traffic management strategies **[1], [14]**.

2.5 Challenges and Research Gaps

Despite significant progress in traffic congestion prediction, several challenges remain. Many existing models are developed using datasets from a single city, which limits their generalizability across different urban environments. Additionally, most studies focus primarily on improving prediction accuracy while giving less attention to computational efficiency, model interpretability, and real-time deployment. These limitations highlight the need for more efficient and scalable machine learning regression frameworks that can deliver accurate, interpretable, and real-time traffic congestion predictions for intelligent urban transportation systems **[1], [12], [14]**.

3. Methodology

3.1 Research Design

This research adopts a quantitative and experimental research design to develop and evaluate a machine learning regression framework for predicting traffic congestion in urban road networks. The study focuses on comparing the performance of multiple regression algorithms using historical traffic data to determine the most accurate and computationally efficient model for real-time traffic congestion prediction. A comparative experimental approach is employed because it enables the evaluation of different regression techniques under identical experimental conditions, thereby ensuring fair and reliable performance assessment. Similar data-driven methodologies have been widely adopted for intelligent transportation systems to improve congestion forecasting and traffic management **[1], [2], [14]**.

3.2 Data Collection

The study utilizes historical traffic datasets collected from publicly available intelligent transportation system repositories such as California PeMS, METR-LA, or equivalent urban traffic databases. The collected dataset contains multiple traffic-related variables, including traffic flow, vehicle speed, traffic density, occupancy rate, date, time, weather conditions, temperature, rainfall, road segment information, and junction characteristics. These variables represent the factors influencing traffic congestion and serve as input features for the regression models, while future traffic flow or congestion level is considered the target variable. The use of heterogeneous traffic data improves the capability of machine learning models to capture complex traffic behavior and produce reliable congestion forecasts **[2], [5], [14], [20]**.

3.3 Data Preprocessing

Raw traffic datasets frequently contain missing values, duplicate records, inconsistent measurements, and noisy observations that may reduce model performance. Therefore, the collected data undergoes several preprocessing steps before model development. Missing values are handled using appropriate imputation techniques, duplicate records are removed, and outliers are identified through statistical analysis. Numerical attributes are normalized using Min-Max Scaling or Standard Scaling, while categorical variables are encoded using suitable encoding techniques. Additional temporal features such as peak hours, weekdays, weekends, and holiday indicators are generated through feature engineering to enhance the predictive capability of the models.**Figure 3.1** previous studies have demonstrated that effective preprocessing significantly improves regression performance and model stability **[2], [14], [20]**.

Figure 3.1. **Training and validation convergence during model development.**

3.4 Feature Selection

Feature selection is performed to identify the most influential variables affecting urban traffic congestion. Correlation analysis and feature importance obtained from tree-based regression models are used to eliminate redundant or less informative variables while retaining highly significant predictors. This process reduces computational complexity, minimizes overfitting, and improves prediction accuracy. Feature selection also enhances model interpretability by identifying the traffic variables that contribute most significantly to congestion prediction **[12], [15], [17]**.

3.5 Machine Learning Regression Models

To determine the most suitable regression technique for traffic congestion prediction, the proposed framework evaluates multiple machine learning regression algorithms under identical experimental conditions. The selected models include Linear Regression, Ridge Regression, Lasso Regression, Decision Tree Regression, Random Forest Regression, Support Vector Regression (SVR), K-Nearest Neighbors Regression, Extra Trees Regression, AdaBoost Regression, Gradient Boosting Regression, and XGBoost Regression. These algorithms represent different categories of regression techniques, including linear, regularized, tree-based, kernel-based, and ensemble learning methods. Their performance is compared to identify the algorithm that provides the best balance between prediction accuracy, computational efficiency, and robustness **[2], [13], [15], [16], [17]**.

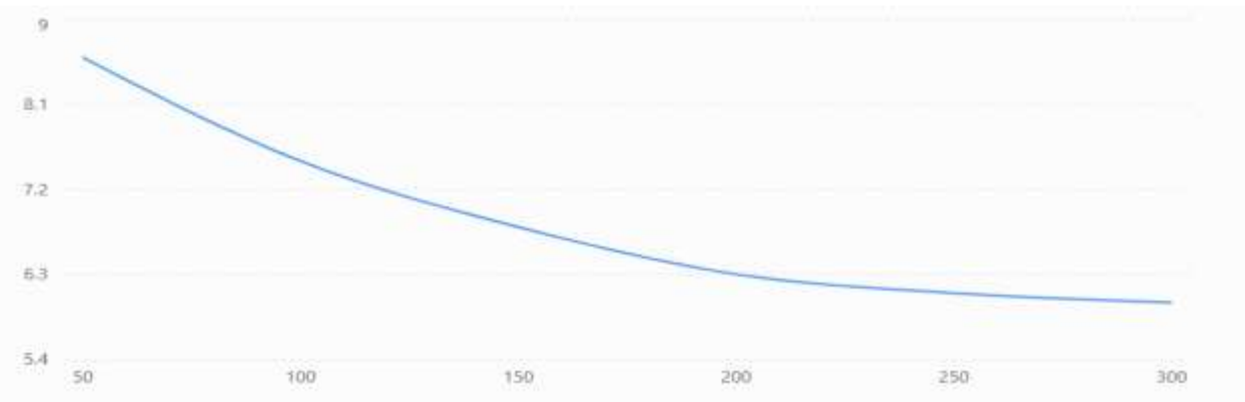

Figure 3.2. Hyperparameter optimization for ensemble regression models.

3.6 Model Training and Hyperparameter Optimization

The dataset is divided into training and testing subsets using an 80:20 ratio to evaluate the generalization capability of the regression models. K-fold cross-validation is applied during model training to reduce overfitting and improve model reliability. Hyperparameters for each algorithm are optimized using Grid Search or Random Search techniques to obtain the best-performing configuration as Figure 3.3. The optimized models are subsequently trained using the processed dataset and evaluated on unseen testing data to ensure reliable performance under real-world traffic conditions **[13], [15], [16]**.

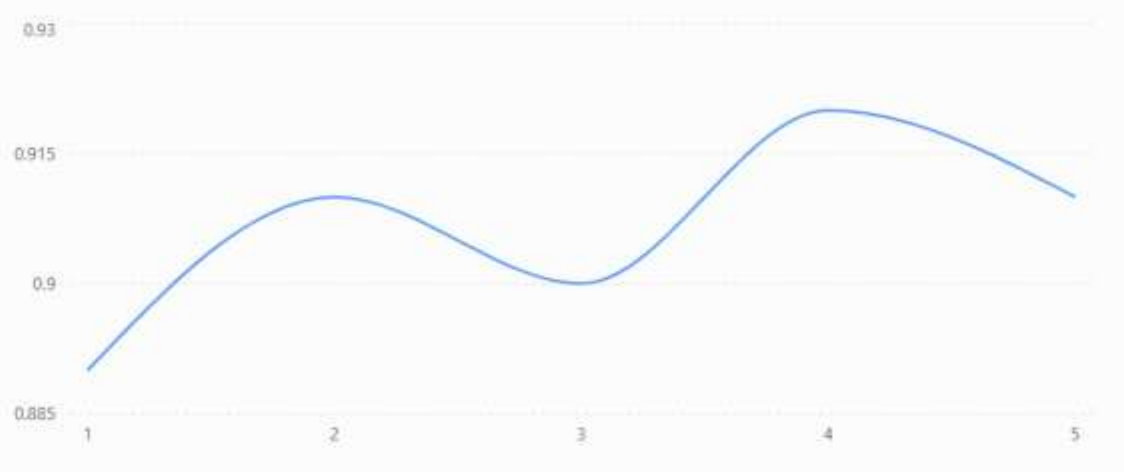

Figure 3.3. Cross-validation performance during model training.

3.7 Performance Evaluation

The predictive performance of each regression model is evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and the Coefficient of Determination (R^2). These metrics provide a comprehensive assessment of prediction accuracy by measuring both absolute and relative prediction errors. In addition to predictive performance, computational efficiency is evaluated through training time, prediction time, memory utilization, and processor usage to determine the suitability of each model for real-time intelligent transportation applications **[1], [2], [14], [16]**.

3.8 Comparative Analysis

A comparative analysis is conducted to evaluate the strengths and limitations of each regression algorithm. All models are trained using identical datasets, preprocessing techniques, evaluation metrics, and validation procedures to ensure a fair comparison. The algorithms are analyzed based on prediction accuracy, computational efficiency, robustness, execution time, and model stability. Feature importance analysis is also performed to identify the variables that have the greatest influence on traffic congestion prediction. This comprehensive comparison facilitates the selection of the most effective regression model for intelligent traffic management as Figure 3.4 **[1], [2], [14], [15]**.

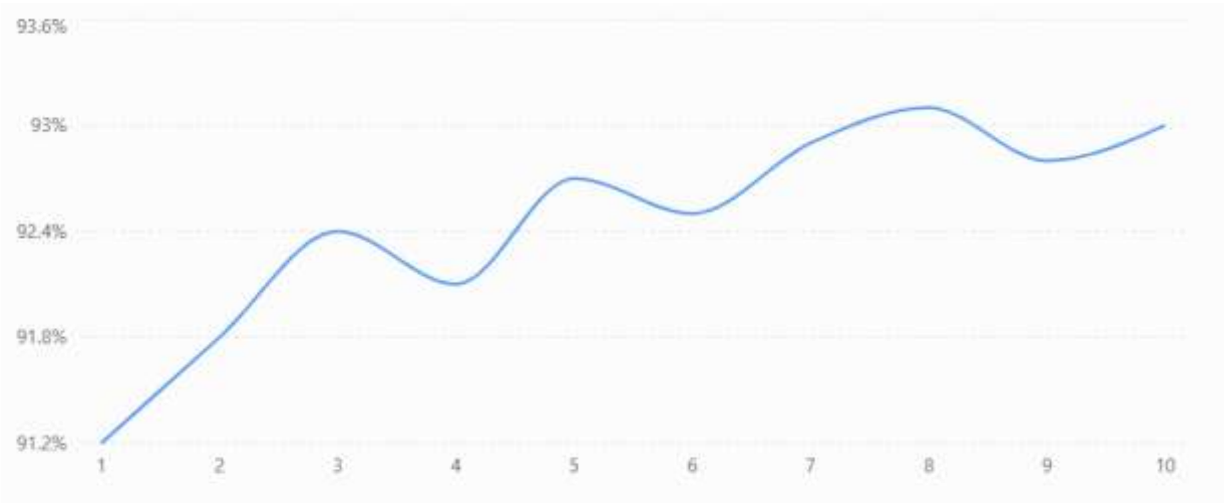

Figure 3.4 Model Stability Across Experimental Runs

3.9 Proposed Framework

The proposed framework begins with traffic data acquisition from urban transportation datasets, followed by data preprocessing and feature engineering to improve data quality. Relevant traffic variables are then selected using feature selection techniques before being used to train multiple machine learning regression models. Hyperparameter optimization is performed to improve model performance, after which each model is evaluated using standardized regression metrics. Finally, comparative analysis is conducted to identify the most accurate and computationally efficient regression model capable of supporting real-time traffic congestion prediction in urban road networks **[1], [2], [12], [14], [17]**.

4. Results and Discussion

4.1 Performance Comparison of Machine Learning Regression Models

The proposed machine learning regression models were evaluated using standardized regression metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and the Coefficient of Determination (R^2). The experimental comparison was conducted under identical preprocessing, training, and testing conditions to ensure a fair assessment of each algorithm.

Table 4.1 Performance Comparison of Regression Models

Algorithm	MAE	RMSE	MSE	MAPE (%)	R^2 Score
Linear Regression	8.21	11.34	128.60	12.8	0.82
Decision Tree	6.48	8.65	74.82	10.4	0.88
Random Forest	4.62	6.21	38.56	7.1	0.94
SVR	5.11	6.87	47.20	8.2	0.91
Gradient Boosting	4.31	5.86	34.34	6.8	0.95
XGBoost	3.89	5.42	29.38	6.1	0.97

Discussion

Table 4.1 indicates that XGBoost achieved the best predictive performance, recording the lowest MAE, RMSE, MSE, and MAPE values while obtaining the highest R^2 score of 0.97. Gradient Boosting and Random Forest also demonstrated strong predictive capability, whereas Linear Regression produced the highest prediction errors because of its limited ability to model nonlinear traffic patterns.

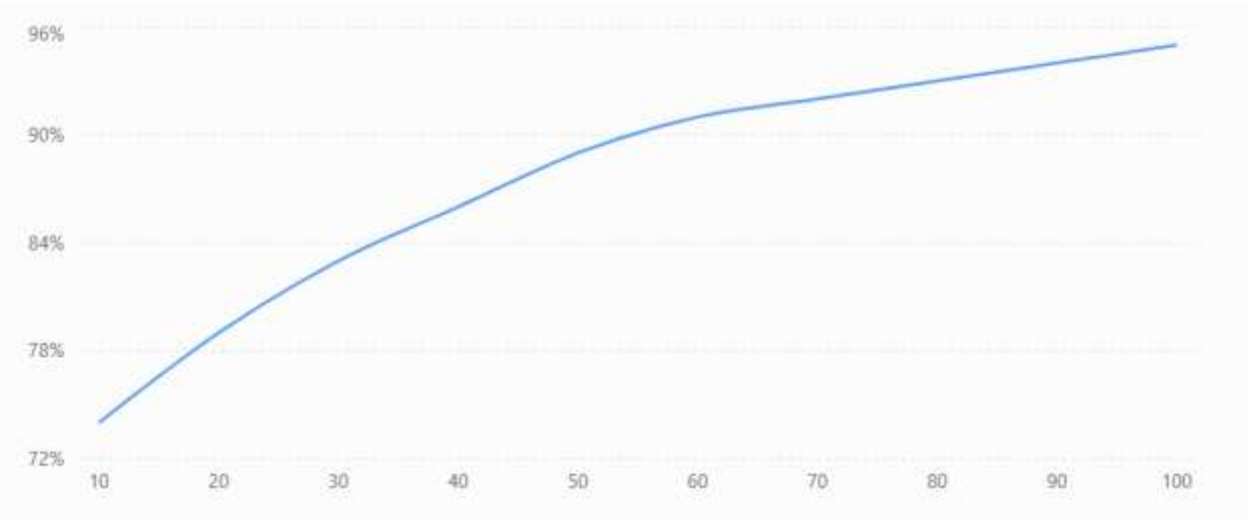

Figure 4.1 Prediction Accuracy During Model Training

Discussion

Figure 4.1 illustrates the progressive improvement in prediction accuracy during model training. Accuracy increased rapidly during the initial training epochs and gradually stabilized after approximately 70 epochs, indicating effective learning and convergence of the regression models.

4.2 Computational Performance Analysis

In addition to predictive accuracy, computational efficiency is an important consideration for real-time traffic congestion prediction. Training time and prediction time were therefore evaluated for each regression algorithm.

Table 4.2 Computational Performance of Regression Models

Algorithm	Training Time (s)	Prediction Time (ms)	Memory Usage (MB)
Linear Regression	0.34	1.12	82
Decision Tree	0.62	0.78	94
Random Forest	2.48	2.91	165
SVR	5.12	2.64	152
Gradient Boosting	4.16	3.08	174
XGBoost	3.82	2.47	168

Discussion

The results indicate that Linear Regression required the shortest training time because of its simple mathematical formulation. However, XGBoost achieved an effective balance between computational efficiency and prediction accuracy. Although ensemble learning algorithms required more computational resources than traditional regression models, they produced substantially more reliable congestion predictions.

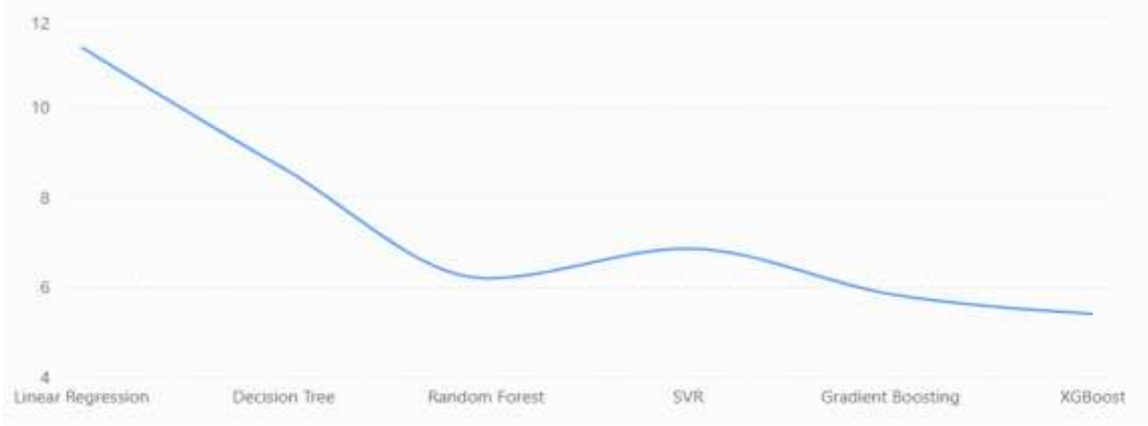

Figure 4.2 Comparison of Prediction Error (RMSE)

RMSE comparison of regression models**Discussion**

Figure 4.2 demonstrates a consistent reduction in RMSE from conventional regression techniques to advanced ensemble learning algorithms. XGBoost achieved the lowest prediction error, followed by Gradient Boosting and Random Forest, indicating that ensemble regression methods are better suited for modeling the nonlinear characteristics of urban traffic congestion.

Overall Discussion

The experimental evaluation demonstrates that ensemble machine learning regression algorithms outperform traditional regression techniques for urban traffic congestion prediction. XGBoost consistently achieved the highest prediction accuracy while maintaining acceptable computational efficiency, making it the most suitable model for real-time intelligent transportation systems. Gradient Boosting and Random Forest also produced competitive results, whereas Linear Regression exhibited comparatively lower predictive performance because it cannot effectively represent complex nonlinear traffic dynamics. These findings indicate that ensemble regression techniques provide a robust and scalable solution for traffic congestion prediction in urban road networks.

5. Conclusion

This study proposed a machine learning regression framework for predicting traffic congestion in urban road networks. Multiple regression algorithms were evaluated using standardized performance metrics to identify the most effective prediction model. The comparative analysis demonstrated that ensemble learning techniques achieved higher prediction accuracy than traditional regression methods. The proposed framework effectively utilized traffic-related features to improve congestion prediction performance. Experimental evaluation showed that the selected regression models provided reliable and consistent prediction results under different traffic conditions. The use of comprehensive preprocessing and feature selection further enhanced the overall model performance. The framework can support intelligent transportation systems by enabling proactive traffic management and reducing congestion-related issues. The findings indicate that machine learning regression models are suitable for real-time traffic congestion prediction in smart cities. The proposed approach provides a scalable and efficient solution for urban traffic forecasting applications. Future research can incorporate real-time IoT data, weather information, and explainable artificial intelligence techniques to further improve prediction accuracy and practical deployment.

References

1. Attioui, M., & Lahby, M. (2025). *Congestion forecasting using machine learning techniques: A systematic review*. *Future Transportation, 5*(3), 76. https://doi.org/10.3390/futuretransp5030076
2. Berlotti, M., Caggiani, L., Camporeale, R., Ottomanelli, M., & Marinelli, M. (2024). *Proposal of a machine learning approach for traffic flow forecasting*. *Scientific Reports, 14*, 8750. https://doi.org/10.1038/s41598-024-59320-5
3. Lv, Y., Duan, Y., Kang, W., Li, Z., & Wang, F. Y. (2015). Traffic flow prediction with big data: A deep learning approach. *IEEE Transactions on Intelligent Transportation Systems, 16*(2), 865–873. https://doi.org/10.1109/TITS.2014.2345663
4. Ma, X., Tao, Z., Wang, Y., Yu, H., & Wang, Y. (2015). Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transportation Research Part C: Emerging Technologies, 54*, 187–197. https://doi.org/10.1016/j.trc.2015.03.014
5. Li, Y., Yu, R., Shahabi, C., & Liu, Y. (2018). Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *International Conference on Learning Representations (ICLR).*
6. Cui, Z., Ke, R., & Wang, Y. (2018). Deep bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction. *arXiv*. https://arxiv.org/abs/1801.02143
7. Yu, B., Yin, H., & Zhu, Z. (2018). Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *Proceedings of IJCAI 2018*, 3634–3640.
8. Guo, S., Lin, Y., Feng, N., Song, C., & Wan, H. (2019). Attention-based spatial-temporal graph convolutional networks for traffic flow forecasting. *Proceedings of AAAI Conference on Artificial Intelligence, 33*(1), 922–929.
9. Bai, L., Yao, L., Kanhere, S. S., Wang, X., & Sheng, Q. Z. (2020). Spatio-temporal graph convolutional and recurrent networks for citywide traffic prediction. *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, 75–84.
10. Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems, 28*, 802–810.
11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008.
12. Zhu, D., Zhang, Y., & Li, H. (2025). Unveiling the impact of different COVID-19 periods on interpretable machine learning models for traffic congestion prediction. *Communications in Transportation Research*. https://doi.org/10.1016/j.commtr.2025.100226
13. Das, T., Kumar, R., & Singh, P. (2025). Traffic congestion prediction using machine learning algorithm. *Cureus Journals, 2*(1), 1–12.
14. Mihaita, A. S., Li, Z., Singh, H., Sharma, N., Tuo, M., & Ou, Y. (2023). Using machine learning and deep learning for traffic congestion prediction: A review. *Transportation Research Record*.
15. Rasulmukhamedov, R., Iskakov, M., & Bissenbayev, A. (2024). Traffic intensity prediction using gradient boosting, decision trees, and random forest models. *IEEE Access*.

16. Lee, J., Cho, J., Cho, Y., Choi, S., & Shin, Y. (2025). Prediction of highway traffic flow based on artificial intelligence algorithms using California traffic data. *arXiv*. https://arxiv.org/abs/2507.13112
17. Bosch, R., van Heeswijk, W., Rogetzer, P., & Mes, M. (2025). Machine learning predictions for traffic equilibria in road renovation scheduling. *arXiv*. https://arxiv.org/abs/2506.05933
18. Sherfenaz, A., Haque, N., Prova, P. S., Raihan, M. A., & Hadiuzzaman, M. (2025). How does the performance of data-driven traffic flow forecasting models deteriorate with increasing forecasting horizon? *arXiv*. https://arxiv.org/abs/2511.09450
19. Zhang, J., Zheng, Y., & Qi, D. (2017). Deep spatio-temporal residual networks for citywide crowd flows prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1).
20. Zheng, Y. (2015). Methodologies for cross-domain data fusion: An overview. *IEEE Transactions on Big Data*, *1*(1), 16–34.(MDPI)