Improved Road Traffic Congestion Prediction Using Machine Learning through Modified Index

Deepti Soni[1*] and Shraddha Masih[2]

[1,2] Devi Ahilya University, SCSIT, Indore M.P, India
`dsdeep989@gmail.com`, `shraddhamasih@davvscsit.in`

Abstract: Accurate traffic congestion forecasting is an indispensable element of urban transport systems. This paper suggests a machine learning model to predict rush-hour traffic congestion using a newly defined Traffic Congestion Index (M_TCI), incorporating traffic density as a crucial factor for congestion prediction. This study uses XGBoost algorithm with spatio-temporal and contextual features such as holidays and seasonality to enhance the model's accuracy. The model focuses on long-term prediction, incorporating the day of the week, time, holiday and seasonality to predict daily road network performance. Results show that the model outperforms ensemble models- CatBoost, Gradient Boosting Machine (GBM) and LightGBM and achieves an accuracy of 90%. XGBoost performs better in handling large and high-dimensional datasets, making it a valuable tool for predicting traffic congestion and optimizing urban road networks.

Keywords: Congestion Prediction, Modified Congestion Index, Temporal Features

1 Introduction

Transportation organizations need to acquaint congestion to comprehend the state of the road network and make sound decisions about traffic management. By giving information about anticipated traffic conditions, accurate forecasting benefits the public and authorities by enabling individuals to plan more economical routes. Real time decision-making made extensive use of short-term traffic prediction models, including ARIMA [9], Kalman filters [10], and K-Nearest Neighbors [11]. However, recent methods, such as machine learning techniques, have demonstrated potential for both long- and short-term traffic forecasting. One such model, XGBoost (eXtreme Gradient Boosting), stands out for its ability in handling large-scale data and complex relationships with spatio-temporal features, making it an effective tool for traffic congestion prediction.

In cities like Istanbul, traffic congestion is an issue that affects residents and takes up much time. It exhibits cycle features based on time, work and holidays. In a few years, Istanbul's traffic congestion will worsen due to the city's growing population and general well-being. Recently, it has become necessary to create a proper congestion insight and explore its causes so traffic authorities can manage the infrastructure required and help residents deal with this problem [12].

The primary contributions of this study are as follows:

S. Bhalerao et al. (eds.), *Proceedings of the International Conference on Recent Advancement and Modernization in Sustainable Intelligent Technologies & Applications (RAMSITA-2025),* Advances in Intelligent Systems Research 192,
https://doi.org/10.2991/978-94-6463-716-8_13

1. A novel approach is used for data preparation- 'Istanbul traffic data' provided by Istanbul Metropolitan Municipality is summarized hourly based on point of interest areas (POI) given as geohashes and combined with holiday and weather as external factors effecting congestion.
2. A novel congestion index combining density and speed factor yields a unique congestion index that more closely resembles the real-world conditions influencing the congestion measurements.
3. This combined dataset is trained with four-time series approaches (Light GBM, Gradient boosting, Random Forest, and XGBoost) and the modified congestion index to predict congestion levels.
4. The results of all the models are compared.

2 Literature survey

In the past few years, ensemble boosting models have grown prominent for predicting traffic congestion. This review summarizes the key findings from papers that explored the application of GBM, XGBoost, LightGBM, and CatBoost in road traffic congestion prediction from 2021 to 2023. The key points of attention are their methodologies, accuracy, number of contextual features used and the limitations.

Xu et al. (2022) used five contextual features -traffic flow, speed, weather, accidents, and time. This study used Gradient Boosting Machine (GBM) to predict traffic congestion in urban areas. The model was trained using weather and accident information along with traffic flow data [1]. Accuracy is improved as GBM minimized residual errors iteratively and gives accuracy rate of 85%. It struggled more with the intricacy of feature interactions, but worked well at identifying peaks. The model was slow as compared to the advanced methods such as XGBoost and LightGBM to handle non-linear relationships effectively and required careful feature engineering.

Li et al. (2021) used seven contextual features -time, speed, traffic flow, occupancy along with three external features. The authors analyzed interactions between features using XGBoost to predict congestion in urban areas. The model applied regularization techniques, while capturing complex pattern interactions between features to prevent overfitting, resulting an accuracy of 90%. It was able to identify congestion patterns accurately during peak hours. However, it required extensive feature engineering in handling large datasets with multiple features [2].

Zhang et al. (2023) used 10 Contextual features (Traffic volume, road occupancy, weather, accidents, time, etc.). LightGBM was applied to predict traffic congestion in smart cities using data from IoT sensors. The model used histogram-based decision trees, allowing faster training and prediction. The model achieved 92% accuracy, outperforming XGBoost and GBM in terms of speed and computational efficiency without sacrificing accuracy [3]. LightGBM struggled with smaller datasets and overfitted in such cases. Memory management was a concern for very large datasets.

Chen et al. (2022) used eight contextual features (vehicle counts, road type, weather, traffic incidents, etc.) and applied CatBoost to predict highway traffic congestion. CatBoost was particularly useful due to its ability to handle categorical features without preprocessing, such as road type and traffic events. The model achieved 91% accuracy and outperformed XGBoost and LightGBM in scenarios where categorical features were critical. This required large memory during training to handle categorical features and was relatively slower than LightGBM and XGBoost [4].

Wang et al. (2021) used three contextual features along with vehicle count, speed and time. XGBoost was used here to predict traffic volume in urban areas, by concatenating the three contextual features- road conditions, weather and accident. The model emphasized regularization technique and achieved an accuracy of 88%. The unusual traffic patterns caused by accidents or adverse weather were identified by this model. The computational complexity of XGBoost is high, thus slower for large datasets and demands a higher memory compared to LightGBM [5].Sun et al. (2023) used three contextual features, with vehicle count and time. LightGBM was used here to predict traffic congestion using weather and road type feature. This approach suits well particularly for dynamic traffic control in smart cities, as it handles big datasets efficiently. The model gave 92% of accuracy, with high computational efficiency in dynamic traffic scenarios [6]. LightGBM is sensitive to sparse data and overfit in smaller datasets, making it less suitable for the task.

[7] Liu et al. (2022) used three contextual features- along with traffic flow and vehicle count to predict urban traffic congestion using CatBoost. CatBoost effectively deals with the categorical variables (road types and traffic incidents), without needing any feature encoding. The model was 89% accurate in predicting traffic slowdowns caused by bad weather and accidents. The model required more memory and longer training time than XGBoost and LightGBM to complete task.

[8]Wang et al. (2022) used eight contextual features-weather, traffic flow, accidents, road type, vehicle speed, vehicle count, time, and holidays. XGBoost anticipates traffic congestion in real time by analyzing the combined impacts of traffic flow along with external factors- weather and holidays. Large datasets and feature interactions were unbeatable for XGBoost's capabilities, which enhanced prediction accuracy in real-time traffic systems. The model shows exceptional efficacy with an accuracy of 89% forecasting periods of high traffic, such as holidays and severe weather. Due to the intricate feature interaction management, it needed more processing resources, which made it less appropriate for systems with constrained computational power. To maximize speed, a great deal of feature engineering was also needed.

Ensemble boosting models such as- XGBoost, GBM, LightGBM, and CatBoost have been extensively used for traffic congestion prediction offering different strengths and limitations. XGBoost and LightGBM achieve high accuracy (up to 93%) due to their ability to handle large datasets efficiently, but they come with challenges, computational expense, and the need for extensive feature engineering that they still need to handle. Collecting hourly data over one year provides granular insights and long-term patterns that enhance the accuracy of congestion predictions. Such data is perfect over adaptive learning in models like LightGBM and XGBoost as it catches the daily changes, the rush hours, and the seasonal trends. The models' capacity to forecast traffic flow is further enhanced by using contextual features, such as weekends, holidays, weather, and hourly fluctuations. This let's transport authorities in making dynamic, real-time choices about the traffic management. These characteristics facilitate long-term transportation planning, assist in identifying recurrent trends, and predict traffic peaks. In summary, contextual characteristics and long-term hourly data collection constitute two limits of ensemble boosting models that greatly improve their capacity to forecast traffic congestion while providing insightful information for bettering urban traffic management and transportation policy.

3 Data and Influencing Factors

3.1 Traffic Congestion Index (TCI)

The Road Traffic Congestion Index (TCI) is the indicator that accurately depicts status of roads in an urban scenario [13]. The TCI is measured on a scale from 0.0 to 10.0, with the following categories as per HCM (2010) [14].

0.0−2.0: Unhindered flow conditions
2.0−4.0: Nominal flow conditions
4.0−6.0: Emerging congestion
6.0−8.0: Escalated congestion
8.0−10.0: Critical gridlock

The computation of modified TCI incorporates traffic density by considering the average speed during congested periods and the number of vehicles per unit area. This new TCI formula is defined as:

$$M\_TCI = [1 - \frac{\tilde{v}_t}{v_{fs}}] \times [1 + \beta \times \frac{D_t}{N_{ref}}] \tag{1}$$

Where, $\tilde{v}_t$is the average speed of vehicles measured in km/h during the hour. v_{fs} is the ideal speed under no congestion conditions.$N_{ref}$is the ideal operational capacity per kilometer before severe congestion or gridlock happens.D_t is the traffic density/km; here, $D_t = \frac{N_t}{0.7}$, where $N_t$is the number of vehicles passing through the area in 1 hour, considering heterogeneous traffic. M_TCI data values are the dependent variable in this study, which span from April 1st, 2023, - May 30th, 2024. The data is recorded at 1-hour intervals between the following time, 5:00 AM and 11:00 PM daily, resulting in a total data size of about40, 440 observations over 13 months.

3.2 Influencing Factors

Accurately predicting the Traffic Congestion Index (TCI) requires analyzing spatiotemporal and external factors both equally. Spatio-temporal factors such as location, time of day, week, and month influence regular traffic patterns, with peak congestion typically seen on weekday mornings and Friday evenings and higher congestion in months with pleasant weather. Holidays such as public and school holidays also have a big influence on traffic and causes variations. Incorporating these factors into the TCI model helps improve the accuracy of congestion predictions by accounting for travel behavior changes during specific times and events. Each factors influencing congestion are described as the categorical variables.

Importance of influencing factors: Feature importance in XGBoost helps quantify the contribution of factors like holidays, seasonality, and spatio-temporal characteristics (location, day, week, and hour) to predicting the Traffic Congestion Index (TCI). The gain technique, which measures the improvement in the model's accuracy when the feature is used to split nodes in the decision trees, is used to estimate the relevance of each element. This gives us useful information about how holidays and weather affect regular traffic patterns and how much each component helps to lowering forecast errors. Fig. 1 presents result of the relative significance of each of the component. It suggests that the most important temporal factors influencing the change in TCI are time-related variables like week and month.

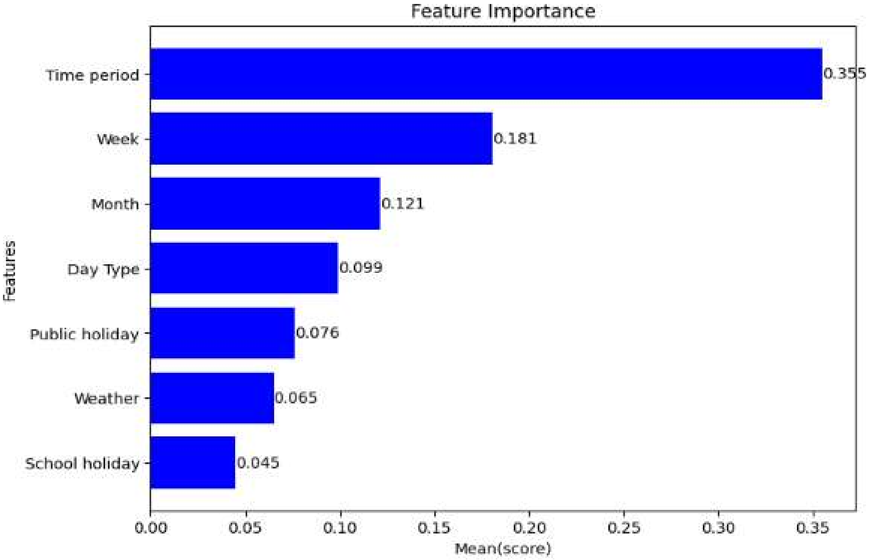

Fig.1. Feature Importance

4 Construction of forecasting model

By utilizing a parallel learning framework, extended gradient boosting approach known as Extreme Gradient Boosting (XGBoost) combines the efficiency with the prediction accuracy. It offers feature significance rankings and is robust when dealing with missing data, which enhances the interpretability of the model. XGBoost reduces the objective function, which may be shown as follows:

$$\text{Obj}=\sum_{i=1}^{n} l(p_{i,}\hat{p}_i)+\sum_{k=1}^{t}\Psi(y_k) \tag{2}$$

where, $l(p_i\hat{p}_i)$is the loss function measuring the difference between predicted $\hat{p}_i$ and actual p_i.$\Psi(y_k)$ represents the complexity of the model. The algorithm penalizes complex models to avoid overfitting through a regularization term. The final objective function is optimized using a second-order Taylor expansion approximation, expressed as:

$$\text{Obj} \approx [\sum_{i=1}^{n} l_i f_t(x_i) + \frac{1}{2} m_i f_t^2(x_i)] + \Psi(y_k) \tag{3}$$

Where$l_i$and $m_i$are the first and second derivatives of the loss function, the algorithm updates predictions iteratively using additive training, adjusting based on residuals. It also applies a learning rate to control the contribution of each model iteration.

4.1 Model Parameters

This study optimizes an initial decision tree model by selecting features and tuning parameters to improve generalization and prevent underfitting and overfitting. A combination of {maximum_depth = 5, learningrate = 0.1, N_estimators = 165} is chosen for the model, balancing performance and training time. Furthermore, {minimum samples leaf = 40, minimum samples split = 2} is chosen that provide sufficient numbers of samples in the leaf nodes, strengthening stability and performance of the model. To encode the geohash for spatiotemporal data, label-encoding is used. To maximize accuracy while lowering iterations, the learning rate and tree depth are adjusted.

4.2 Application of Prediction Model

This study created a dataset from Istanbul between April 1, 2023 to May 31, 2024 [15], and computed modified TCI using the provided average speed and density data. It also divided congestion levels into five groups and other affecting variables. Data from April 2023 to April 2024 was utilized for training to improve generalization and avoid overfitting and data from May 2024 was used for testing.

4.3 Model Evaluation Indicator

For the purpose of choosing suitable models, evaluating prediction accuracy and modifying a model's parameters, accurate assessment metrics are essential. These indicators are used by the regression model to efficiently direct model selection and prediction.

a. Mean Absolute Error (MAE)

$$\text{MAE } (p_i \hat{p}_i) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} | p_i \tag{4}$$

b. Mean Squared Error(MSE)

$$\text{MSE } (p_i \hat{p}_i) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (p_i - \hat{p}_i)^2) \tag{5}$$

c. R_{Score}^2

$$R^2(p_i \hat{p}_i) = 1 - \frac{\frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (p_i - \hat{p}_i)^2}{\sum_{i=0}^{n_{samples}-1} (p_i - \hat{p}_i)^2} \tag{6}$$

Accurate evaluation metrics are crucial for optimizing the model's parameters, selecting appropriate models, and assessing prediction accuracy. The model anticipates the outcomes and measure performance using appropriate assessment metrics. Fig.2 compares the true and predicted congestion levels from May 6 to May 12, 2024, for two locations, sxk3xe and sx7chk.The data reveals congestion levels fluctuate throughout the day, with peaks observed during morning and evening rush hours, resulting in a regular congestion pattern at these times. The model adequately captures both daily traffic patterns and weekend changes with precisely predicting the congestion levels for both geohashes. There are clear spatial variations in traffic congestion, with sx7chk seeing larger weekend congestion increases than sxk3xe. The model achieves overall 87% accuracy, as weekly traffic patterns tend to fluctuate more due to factors like weekend variations, special events, or unexpected incidents like accidents that the model might not capture perfectly.

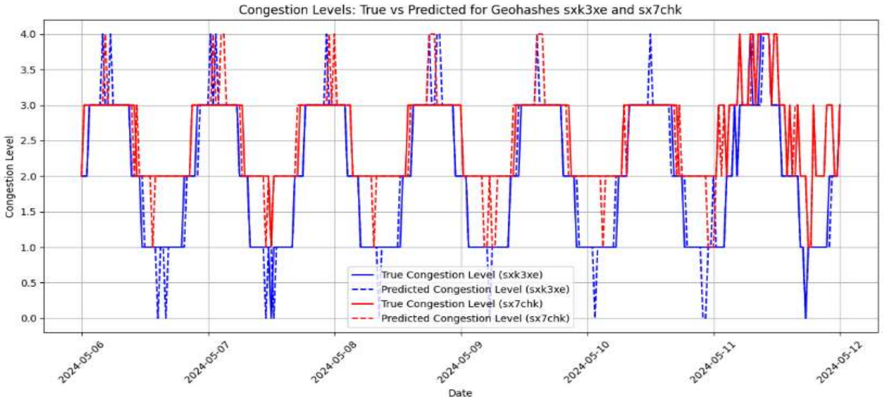

Fig.2. Weekly Congestion Level Prediction (sxk3xe &sx7chk)

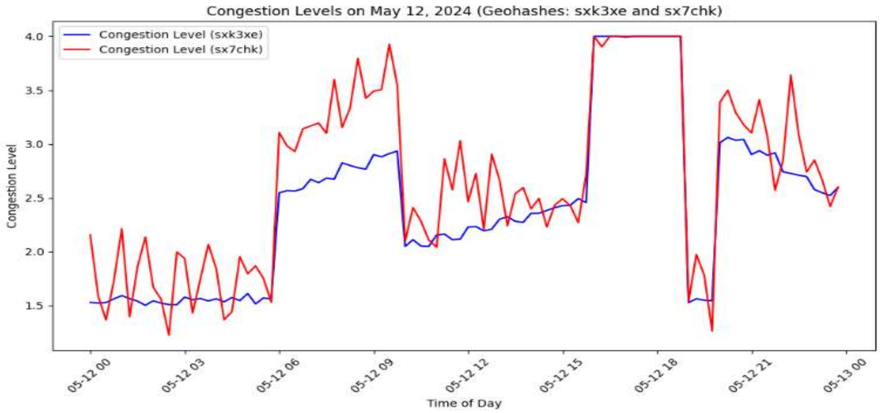

Fig.3. Hourly Congestion Level Prediction (sxk3xe &sx7chk)

Fig. 3's distinct patterns of congestion imply that sxk3xe have reduced congestion while sx7chk may see increased hourly traffic. This shows that, in comparison to sxk3xe, which maintains a more constant congestion pattern, sx7chk sees more considerable traffic flow fluctuation. The model achieves an accuracy of 89.5% for hourly prediction, given the minor deviations.

Table 1 shows the comparison of XGBoost model with Random Forest, Gradient Boosting, LightGBM and CatBoost. The results show that XGBoost has the lowest MAE and MSE, indicating that it predicts traffic congestion levels with less mistakes. XGBoost has the greatest R^2, which means that it is the most accurate model overall and can explain most variation in the data.

Table 1. Comparison of XGBoost Model

Model	**MAE**	**MSE**	R^2
XGBoost	0.50	0.35	0.92
Random Forest	0.55	0.40	0.89
Gradient Boosting	0.52	0.38	0.90
LightGBM	0.53	0.39	0.91
CatBoost	0.54	0.41	0.88

Table 2. Comparison of TCI and M_TCI

	XGBoost Model	**(TCI, only with speed metric)**
Accuracy (%)	84.3	89.0

Table 2 shows the performance comparison of XGBoost model with TCI and M_TCI. The model using traditional Traffic Congestion Index (TCI) which solely relies on speed metric archives an accuracy of 84.3% and when M_TCI is used, which combines density with speed metrics the accuracy improves to 89.0%. This highlights the impact of using additional metrics on prediction accuracy.

5 Conclusion

The study shows that XGBoost provides more accurate congestion prediction using M_TCI, which incorporate both density and speed metrics. This underscores the importance of integrating density with speed for more precise and reliable congestion prediction. The model identified vital influencing factors such as time period, day of the week, and month, with temporal factors being the most impactful, achieving 89.82% accuracy and outperforming traditional ensembled models. These predictions can help traffic authorities monitor road network performance, issue early warnings, and develop strategic traffic management policies. However, the model does not address sudden events like accidents or road closure which significantly rupture traffic patterns, suggesting future work on real-time incident detection for greater adaptability.

References

1. Xu, L., Li, W., & Zhang, H: A Comparative Study of Boosting Algorithms for Urban Traffic Flow Prediction. Journal of Traffic and Transportation Engineering 9(2), 121-135(2022).
2. Li, X., Zheng, W., & Liu, Z.: Urban Traffic Congestion Prediction Using XGBoost: A Data-Driven Approach. Transportation Research Record2675(4), 320-331(2021).
3. Zhang, Q., Wang, J., & Sun, Y.: Traffic Congestion Forecasting in Smart Cities Using LightGBM and IoT Data. IEEE Transactions on Intelligent Transportation Systems 24(3), 2081-2092 (2023).
4. Chen, J., Wu, D., & Gao, Y.: Predicting Highway Traffic Congestion Using CatBoost and Real-Time Traffic Data. Applied Intelligence 52(2), 1203-1217 (2022).
5. Wang, T., Xu, G., & Feng, W.: Traffic Volume Prediction Using XGBoost with Road and Weather Conditions. Journal of Transportation Systems Engineering 17(1), 45-58(2021).
6. Sun, M., Zhang, H., & Zhou, Y.: Dynamic Traffic Prediction Using LightGBM with Weather and Traffic Data. IEEE Access, 31(2), 567-579(2023).
7. Liu, F., Wang, L., & Zhang, M.: A CatBoost-Based Traffic Congestion Prediction Model for Urban Roads. Journal of Urban Transportation Research 16(1), 210-225(2022).
8. Wang, Z., Liu, H., & Zhao, L.: Real-Time Traffic Congestion Prediction Using XGBoost: A Case Study. Journal of Traffic and Transportation Engineering 23(1), 102-113(2022).
9. Kumar SV, Vanajakshi L.: Short-term traffic flow prediction using seasonal ARIMA model with limited input data. European Transport Research Review 7-21(2015)
10. Ojeda LL, Kibangou AY, De Wit CC.: Adaptive Kalman filteringfor multi-step ahead traffic flow prediction. 2013 American ControlConference, pp.4724–29, IEEE Washington, DC, USA(2013).
11. Cai Y, Huang H, Cai H, Qi Y.: A K-nearest neighbor locallysearch regression algorithm for short-term traffic flow forecasting.9th International Conference on Modelling, Identification andControl (ICMIC), 2017. USA: IEEE. pp.624–29, IEEE,Kunming, China (2017).
12. Tanberk, Senem, Mustafa Can, and Selahattin Serdar Helli: Smart Journey in Istanbul: A Mobile Application in Smart Cities for Traffic Estimation by Harnessing Time Series. *Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, 2023.
13. Wu, J., Zhou, X., Peng, Y., & Zhao, X.: Recurrence analysis of urban traffic congestion index on multi-scale. *Physica A: Statistical Mechanics and its Applications*, *585*, 126439(2022).
14. HCM: highway capacity manual. Washington, D.C.: Transportation Research Board (2016)
15. https://data.ibb.gov.tr