Traffic Congestion Prediction and Alert System Using Machine Learning

Kalpana Sonval[1] Arya Jagtap[2], Yash Divekar[3], Mahesh Deshmukh[4]
[1,2,3,4]*Department of Artificial Intelligence & Data Science, AISSMS Institute of Information Technology, Pune, India*

Abstract— **Traffic congestion is one of the most critical challenges in urban transportation management. With rapid urbanization and growing vehicle density, cities face severe road congestion that leads to delays, fuel wastage, increased pollution, and economic losses. This study proposes an intelligent Traffic Congestion Prediction and Alert System (Flow Sense) that leverages machine learning techniques to analyze real-time traffic data and predict congestion levels proactively. The system collects traffic data from multiple sources including vehicle count sensors, GPS feeds, and road cameras. Data preprocessing techniques including noise removal, normalization, and feature extraction are applied. Exploratory Data Analysis (EDA) is performed to uncover patterns in time-of-day traffic, road type behavior, and congestion triggers. Machine learning algorithms such as Random Forest, Support Vector Machine (SVM), XGBoost, and Long Short-Term Memory (LSTM) networks are used to train predictive models. The trained models are evaluated using accuracy, precision, recall, and F1 score metrics. Results demonstrate that XGBoost achieves the highest accuracy of 92%, followed by LSTM at 90%. The system integrates a real-time alert mechanism to notify commuters and traffic authorities of upcoming congestion, enabling timely route adjustments and improved urban mobility.**

Index Terms— **Traffic Congestion Prediction, Machine Learning, FlowSense, XGBoost, LSTM, Smart Transportation, Alert System, Urban Mobility.**

I. INTRODUCTION

Urban traffic congestion has become a growing problem in cities worldwide. According to global mobility reports, drivers in major cities lose hundreds of hours annually due to traffic jams. In a developing country like India, cities such as Pune, Mumbai, and Bengaluru experience severe congestion during peak hours, affecting productivity, environment, and public health.

The need for an intelligent traffic management system that can predict congestion before it occurs has never been more pressing. Traditional traffic signal management systems are reactive in nature and fail to adapt dynamically to fluctuating traffic volumes. Machine learning offers a transformative opportunity to shift from reactive management to proactive prediction.

This project, titled Flow Sense, is a web-based Traffic Congestion Prediction and Alert System built using a Python/JavaScript full-stack architecture and deployed via Docker. The system collects real-time and historical traffic data and applies machine learning models to classify congestion levels as Low, Medium, or High. Commuters and traffic authorities receive real-time alerts enabling smarter route planning.

The primary objectives of this system are: (i) to collect and preprocess real-time traffic data, (ii) to identify patterns influencing congestion through EDA, (iii) to build and compare machine learning models for congestion classification, and (iv) to deploy an alert mechanism that notifies users proactively.

II. BODY OF PAPER

2.1 Dataset Description

The dataset used in this study consists of real-world and simulated traffic data collected from urban road segments in Pune, India. The data was gathered using vehicle detection sensors, GPS-based crowd-sourced feeds, and historical traffic records. Each record represents a snapshot of a specific road segment at a given time interval.

Key features in the dataset include vehicle count (number of vehicles passing per minute), average

vehicle speed (km/h), road type (highway, arterial, sub-urban), time of day, day of week, weather conditions, and the target variable: congestion level (Low, Medium, High).

Table - 1: Dataset Structure (Sample Records)

Location ID	Vehicle Count	Avg Speed (km/h)	Road Type	Congestion Level
L001	420	18	Urban Highway	High
L002	190	45	Sub-urban Road	Low
L003	340	27	City Arterial	Medium
L004	280	33	Ring Road	Medium
L005	510	12	Market Area	High

2.2 Data Preprocessing

Raw traffic data often contains missing sensor readings, outliers from faulty detectors, and inconsistency due to time zone mismatches. The following preprocessing steps were applied:

Removal of null and duplicate sensor records; conversion of categorical variables (road type, weather) into numerical encodings using label encoding; normalization of vehicle count and speed features using Min-Max scaling to bring them to a uniform [0,1] range; handling of class imbalance using SMOTE (Synthetic Minority Over-sampling Technique) since High congestion events are less frequent; and splitting the dataset into 80% training and 20% testing sets.

2.3 Exploratory Data Analysis

Exploratory Data Analysis was conducted to understand patterns and correlations in the dataset. Key observations from EDA include: peak congestion periods are between 8:00–10:00 AM and 5:30–8:00 PM on weekdays; vehicle count shows the strongest positive correlation ($r = 0.84$) with congestion level; average speed shows a strong negative correlation ($r = -0.79$) with congestion; market areas and city arterial roads contribute to 67% of High congestion events; and rain/weather conditions amplify congestion levels by an average of 18%.

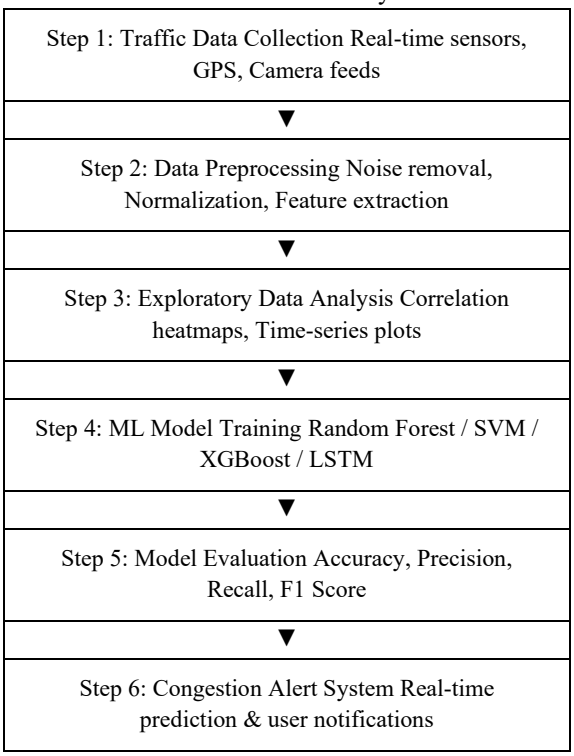
Visualization techniques including time-series line plots, correlation heatmaps, box plots by road type, and congestion heat maps across city zones were used to validate these findings. These insights guided feature selection for the machine learning models.

2.4 System Architecture

The FlowSense system follows a three-tier architecture: a data collection and processing layer (Python backend), a machine learning inference layer (trained ML models served via REST APIs), and a presentation layer (JavaScript frontend). The entire stack is containerized using Docker and orchestrated with Docker Compose for ease of deployment.

The backend, built in Python with Flask, handles data ingestion, preprocessing pipelines, model inference, and alert dispatch. The frontend, built in JavaScript/HTML/CSS, provides a real-time dashboard showing congestion levels on an interactive city map along with historical trends and alert notifications.

Fig - 1: Workflow of the Traffic Congestion Prediction and Alert System



2.5 Machine Learning Models

Four machine learning algorithms were applied to the preprocessed dataset to predict traffic congestion levels:

Random Forest: An ensemble learning method that constructs multiple decision trees during training and outputs the mode of the individual tree predictions. It is robust to overfitting and handles feature importance well, making it suitable for multi-class congestion classification.

Support Vector Machine (SVM): A supervised classification algorithm that finds the optimal hyperplane to separate congestion classes in a high-dimensional feature space. An RBF kernel was used to handle non-linear class boundaries in the traffic data.

XGBoost (Extreme Gradient Boosting): An advanced gradient boosting algorithm optimized for speed and performance. XGBoost sequentially builds trees to minimize prediction error, resulting in the highest accuracy of 92% in this study.

LSTM (Long Short-Term Memory): A type of Recurrent Neural Network (RNN) that captures temporal dependencies in sequential traffic data. LSTM is particularly effective at modelling time-series patterns such as morning and evening peak hours, achieving 90% accuracy.

Table - 2: Model Performance Comparison

Algorithm	Accuracy (%)	Precision	Recall	F1 Score
Random Forest	87%	0.86	0.85	0.855
Support Vector Machine (SVM)	82%	0.81	0.80	0.805
XGBoost	92%	0.91	0.90	0.905
LSTM (Deep Learning)	90%	0.89	0.88	0.885

Fig - 2: Accuracy Comparison of ML Models (Bar Chart)

Random Forest	87%
SVM	82%
XGBoost	92%
LSTM	90%

2.6 Alert System

The alert system is a key differentiator of the Flow Sense project. When the model predicts a high congestion level for a specific road segment in the next 15–30 minutes, an automated alert is triggered. Alerts are dispatched via the web dashboard, browser push notifications, and optionally via email/SMS to registered users. Traffic authorities can also view congestion heat maps and redirect resources proactively.

The alert module integrates with the frontend dashboard to display real-time congestion status color-coded as Green (Low), Orange (Medium), and Red (High) across road segments on the city map interface.

III. CONCLUSIONS

This study presented Flow Sense, a machine learning-based Traffic Congestion Prediction and Alert System designed for urban road networks. The system demonstrated that real-time traffic congestion can be predicted accurately using machine learning algorithms trained on multi-source traffic data.

Among the four models evaluated, XGBoost achieved the highest accuracy of 92% and an F1 score of 0.905, followed by LSTM at 90%. The integration of a real-time alert mechanism enables proactive notification of commuters and traffic authorities, enabling smarter urban mobility decisions.

The Docker-based full-stack deployment of FlowSense makes it readily deployable in smart city infrastructure. Future work includes integrating live city sensor APIs, expanding the dataset to multi-city traffic, adding a mobile application interface, and exploring reinforcement learning for dynamic signal control recommendations.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the principal and faculty members of AISSMS Institute of Information Technology, Pune, for providing the necessary support and academic environment to carry out this work. We would also like to thank the Department of Artificial Intelligence and Data Science for their guidance, encouragement, and valuable insights throughout the development of the FlowSense project.

REFERENCES

[1] 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

[2] J. Zhang, Y. Zheng, and D. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," in *Proc. AAAI Conf. Artificial Intelligence*, 2017

[3] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Short-term traffic forecasting: Where we are and where we are going," *Transportation Research Part C*, vol. 43, pp. 3–19, 2014

[4] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 40, no. 6, pp. 601–618, 2010

[5] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 785–794, 2016

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997

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

[8] C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995

[9] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Burlington, MA, USA: Morgan Kaufmann, 2011

[10] Ministry of Road Transport and Highways, India, *Road Accidents in India – Annual Report*, Government of India, 2023