

PREDICTIVE ANALYSIS OF TRAFFIC VOLUME USING ML

¹ K.Sneha, ² Nagunuri Jitesh Charan, ³ Thummalapalli Yashwanth, ⁴ Gayakwad Praveen, ⁵ Dannada Siddhartha

¹Assistant Professor, ²³⁴⁵Students

Department of ECE

Siddhartha Institute of Technology & Sciences, Narapally

sneha.karrolla@siddhartha.org.in, 24tq1a0430@siddhartha.co.in, 24tq1a0438@siddhartha.co.in,

24tq1a0416@siddhartha.co.in, 24tq1a0462@siddhartha.co.in

ABSTRACT

Road traffic in growing cities has increased far faster than the capacity of the road network, and congestion now causes long travel delays, higher fuel consumption, and increased air pollution. Traffic authorities can manage congestion more effectively if they know in advance how many vehicles are likely to use a road at a given time. This paper presents a predictive analysis system for traffic volume that uses machine learning to forecast hourly vehicle counts from historical sensor data, calendar information, and weather conditions.

The data used by the system comes from roadside vehicle detection equipment such as inductive loop detectors, infrared counters, and camera-based counters, which record the number of vehicles passing a point in each time interval. These counts are combined with calendar attributes including hour of the day, day of the week, month, and public holidays, and with weather attributes including temperature, rainfall, snowfall, cloud cover, and a general weather description. The raw sensor records are checked for gaps, duplicate timestamps, and faulty readings caused by detector malfunction, and are then aggregated into a clean hourly time series.

Feature engineering plays a central role in the system. Cyclical encoding is applied to hour and month so that the model understands that eleven at night is close to midnight, and lag features and rolling averages capture the recent behaviour of the traffic stream. Several regression models are trained and compared, including linear regression, a decision tree regressor, k-nearest neighbours, random forest, gradient boosting, and extreme gradient boosting. The models are evaluated using mean absolute error, root mean square error, and the coefficient of determination on a chronologically separated test period.

The best performing model is used to produce forecasts for the next several hours and for the following day. The results are displayed on a web dashboard that shows the predicted volume curve alongside actual counts, highlights expected peak periods, and classifies each hour into light, moderate, or heavy traffic. Feature importance analysis shows the relative influence of time of day, weekday, holidays, and weather, which helps traffic planners understand the causes of variation rather than only the forecast values.

The system is intended to support signal timing adjustment, deployment of traffic police, scheduling of road maintenance, and advance information to commuters. Because it relies on commonly available sensor counts and public weather data, it can

be applied to other road segments with modest effort. Future work includes real-time streaming of detector data, forecasting across multiple connected junctions, and the use of sequence models for longer forecast horizons.

I. INTRODUCTION

Traffic congestion has become a daily problem in almost every large city. The number of registered vehicles continues to grow each year, while the construction of new roads is limited by space, cost, and time. During peak hours, main corridors operate close to or beyond their capacity, and small disturbances such as a stalled vehicle or a sudden shower can cause queues that take a long time to clear. The resulting delays affect commuters, public transport, freight movement, and emergency services.

Most traffic management today is reactive. Signal timings are fixed according to plans prepared from occasional surveys, and adjustments are made only after congestion has already developed. Traffic police are deployed based on experience rather than on a forecast of expected volume. Yet traffic flow is not random. It follows strong daily and weekly patterns, it changes in predictable ways on holidays and during festivals, and it responds to weather. If these patterns are learned from data, future volume can be estimated with useful accuracy.

From the perspective of electronics and communication engineering, traffic volume data is essentially a time series produced by sensors. Inductive loops embedded in the road surface detect the change in inductance caused by a vehicle's metal body, infrared and microwave sensors detect vehicles passing through a beam, and camera systems count vehicles using image processing. These devices generate continuous streams of counts that must be cleaned, synchronised, and aggregated before analysis, and the quality of any forecast depends heavily on how well this signal is handled.

This paper describes a machine learning system that performs predictive analysis of traffic volume on an urban road segment. It collects hourly counts from detector equipment, merges them with calendar and weather information, engineers features that describe time patterns and recent history, and compares several regression models. The most accurate model is used to forecast volume for the coming hours, and the results are presented through a dashboard intended for traffic control rooms. The aim is to move traffic management from reaction towards anticipation.

II. LITERATURE SURVEY

Traffic flow forecasting has a long history in transportation engineering. Early approaches relied on historical averages, in which the forecast for a given hour was simply the mean volume observed at that hour on similar days in the past. These methods are simple and easy to maintain, but they cannot respond to unusual conditions such as holidays, special events, or bad weather. Statistical time series models were introduced to address this limitation, and the autoregressive integrated moving average model in particular became a standard technique for short-term traffic prediction. Studies using this approach demonstrated that recent observations carry significant information about the next few intervals. However, researchers also found

that these models assume linear relationships and stationary behaviour, and that they perform poorly when traffic changes abruptly or when external factors such as weather have a strong influence. Seasonal variants and the use of exogenous variables improved matters somewhat, but they required considerable manual tuning for each location.

The quality of detector data has been studied as a separate and important problem. Inductive loop detectors, which remain the most widely installed sensing technology on urban roads, are subject to failures such as broken lead-in cables, sensitivity drift, and cross-talk between adjacent loops. Studies of large detector networks have reported that a noticeable fraction of the data at any time can be missing or erroneous. Researchers have proposed rule-based checks that flag counts exceeding physical capacity, long runs of identical values, and zero counts during periods when traffic is known to be present. Imputation methods ranging from linear interpolation to pattern-based filling from neighbouring days and neighbouring detectors have been compared, and the general conclusion is that careful cleaning of the sensor signal improves forecasting accuracy as much as the choice of model does. Similar observations have been made for camera and microwave counters, which are affected by lighting, occlusion, and heavy rain.

Machine learning regression methods have been applied to traffic volume prediction with increasing success. Studies have compared k-nearest neighbours, support vector regression, decision trees, and artificial neural networks against statistical baselines, generally reporting lower errors for the machine learning methods, especially when external variables are included. Ensemble tree methods such as random forest and gradient boosting have become popular because they handle mixed numerical and categorical inputs, capture nonlinear interactions between time and weather, and require relatively little preprocessing. A widely used public dataset of hourly interstate traffic volume with weather and holiday attributes has served as a benchmark in many of these studies, and gradient boosted models have frequently produced among the lowest root mean square errors reported on it. Researchers have also highlighted the importance of chronological train and test splits, since random splits allow information from the future to leak into training and give an overly optimistic result.

Feature engineering for temporal data has received considerable attention. Hour of day and month are cyclical quantities, and researchers have shown that encoding them with sine and cosine components allows models to recognise that the end of one cycle is adjacent to the beginning of the next. Lag features that include the volume one hour, two hours, and one week earlier, together with rolling means over recent windows, provide the model with information about current conditions and have been found to reduce error considerably. Weather attributes such as rainfall and snowfall tend to reduce volume and speed, while holidays shift the daily profile. Studies using feature importance and Shapley value analysis consistently rank hour of day and recent lags as the most influential variables, followed by day of week and weather, which agrees with the practical experience of traffic engineers and supports the use of these features.

More recent research has explored deep learning and network-level forecasting. Recurrent neural networks, long short-term memory networks, and gated recurrent units have been applied to capture longer temporal dependencies, and graph-based neural networks have been used to model the spatial relationships between detectors along a road network. These approaches have achieved strong results on large datasets, but they require substantial training data and computing resources and are harder to interpret. Several authors have noted that for a single road segment with a few years of hourly data, well-tuned ensemble tree models often perform comparably while being faster to train and easier to explain. Studies on deployed systems also emphasise the need for dashboards that present forecasts in a form traffic operators can act on. The system proposed in this paper follows this practical direction by combining careful sensor data cleaning, temporal feature engineering, and ensemble regression with an operator dashboard.

III. SYSTEM ANALYSIS

The traffic volume prediction system receives vehicle counts from roadside detectors installed on a selected road segment, together with calendar and weather information. Counts from inductive loops, infrared counters, or camera-based counters are received at short intervals and aggregated into hourly totals with consistent timestamps. The sensor stream is checked for missing hours, duplicated records, counts above the physical capacity of the lanes, and long runs of constant values that indicate detector failure. Faulty values are removed and gaps are filled using interpolation for short gaps and same-hour values from similar days for longer ones. Weather observations such as temperature, rainfall, snowfall, and cloud cover are aligned to the same hourly grid, and holidays are marked from a public calendar. Temporal features, cyclical encodings, lags, and rolling averages are computed, and several regression models are trained on earlier data and tested on later data. The selected model forecasts volume for upcoming hours, and the dashboard displays predicted and actual curves, peak periods, and congestion levels.

Existing System

In the existing arrangement, traffic on most urban roads is managed with fixed-time signal plans that are prepared from manual traffic surveys conducted once every few years. These surveys count vehicles for a limited number of days and cannot capture seasonal changes, festivals, or the effect of weather. Where detectors are installed, their data is often stored without systematic cleaning, and faulty readings from damaged loops or obstructed cameras remain in the records. Control rooms typically display live counts or camera feeds, but they do not provide any forecast of what traffic will look like in the next few hours. Deployment of traffic personnel and planning of roadworks are therefore based on experience and general assumptions about peak hours. Simple historical averages are sometimes used for planning, but they fail on holidays, during rain, and during special events. Commuters receive information about congestion only after it has formed, through navigation applications, and authorities have little analytical insight into the factors that drive variation in volume on a particular road.

Disadvantages of Existing System

- Signal plans are based on infrequent manual surveys rather than continuous data.
- Detector data is stored without systematic cleaning of faulty readings.
- Control rooms show live counts but provide no forecast of upcoming volume.
- Traffic personnel are deployed from experience rather than from expected demand.
- Historical averages fail on holidays, festivals, and days with bad weather.
- The effect of rain, temperature, and other weather factors is not quantified.
- Congestion is recognised only after it has already formed.
- Roadwork scheduling does not use predicted low-volume periods.
- Commuters receive no advance information about expected traffic conditions.
- The factors that drive volume on a particular road are not analysed.

Proposed System

The proposed system is a machine learning platform for predicting hourly traffic volume on an urban road segment. Vehicle counts are collected from loop detectors, infrared counters, or camera counters and aggregated into an hourly series. A sensor data cleaning stage removes impossible and constant readings and fills gaps using interpolation and similar-day patterns. Weather observations and a holiday calendar are merged with the counts. Feature engineering creates hour, weekday, month, and weekend indicators, sine and cosine encodings for cyclical time, lag features at one hour, two hours, twenty-four hours, and one week, and rolling means over recent windows. Linear regression, a decision tree, k-nearest neighbours, random forest, gradient boosting, and extreme gradient boosting are trained on earlier data and evaluated on a later period using mean absolute error, root mean square error, and the coefficient of determination. The best model forecasts volume for the next hours and the following day, and a dashboard shows forecast curves, peak alerts, congestion levels, and feature importance.

Advantages of Proposed System

- Forecasts hourly traffic volume in advance instead of only reporting current counts.
- Cleans detector data so that faulty readings do not distort the forecasts.
- Includes weather and holiday effects that historical averages ignore.
- Cyclical and lag features capture daily, weekly, and recent traffic behaviour.
- Compares several regression models and selects the one with the lowest error.
- Chronological testing gives a realistic estimate of forecast accuracy.
- Highlights expected peak periods to support signal timing and staff deployment.
- Identifies low-volume periods suitable for scheduling road maintenance.
- Explains the relative influence of time, weather, and holidays on volume.
- Uses commonly available sensor and weather data, so it can be extended to other roads.

IV. METHODOLOGY

The methodology begins with the collection of hourly vehicle counts from roadside detectors over a period of several years, together with matching weather observations and a list of public holidays. The timestamps are converted to a common time zone, duplicate rows are removed, and missing hours are inserted explicitly so that gaps can be seen. Readings above lane capacity or repeated unchanged for many hours are treated as detector faults and replaced through interpolation or similar-day filling. Categorical weather descriptions are encoded, and outliers in temperature and rainfall are corrected. Calendar features, cyclical encodings, lag values, and rolling statistics are computed, taking care that each feature uses only information available before the hour being predicted. Exploratory analysis with hourly profiles, weekday plots, and correlation heatmaps guides feature selection. The data is split chronologically into training, validation, and test periods, and models are tuned with time series cross-validation. The best model is evaluated on the test period using MAE, RMSE, and R-squared, residuals are examined by hour and weather type, and the model is deployed to the dashboard.

SYSTEM ARCHITECTURE

The architecture of the traffic volume prediction system is arranged in layers that move data from roadside sensors to forecasts used by traffic operators. The Sensor Layer consists of inductive loop detectors, infrared counters, and camera-based counters installed on the road segment. The Data Acquisition Layer receives counts from these devices, together with weather observations and holiday information, and stores them with timestamps. The Signal Cleaning Layer detects faulty readings, removes duplicates, and fills gaps in the count series. The Aggregation and Alignment Layer converts the counts to an hourly series and aligns weather and calendar data to the same time grid. The Feature Engineering Layer computes cyclical time encodings, lags, and rolling means. The Model Training Layer trains and tunes regression models using chronological validation. The Evaluation Layer compares models on MAE, RMSE, and R-squared and stores the best one. The Forecasting Layer produces hourly predictions and congestion levels, and the Dashboard Layer displays forecasts, actual counts, peak alerts, and feature importance to traffic control staff.

Predictive analysis of traffic volume using ML

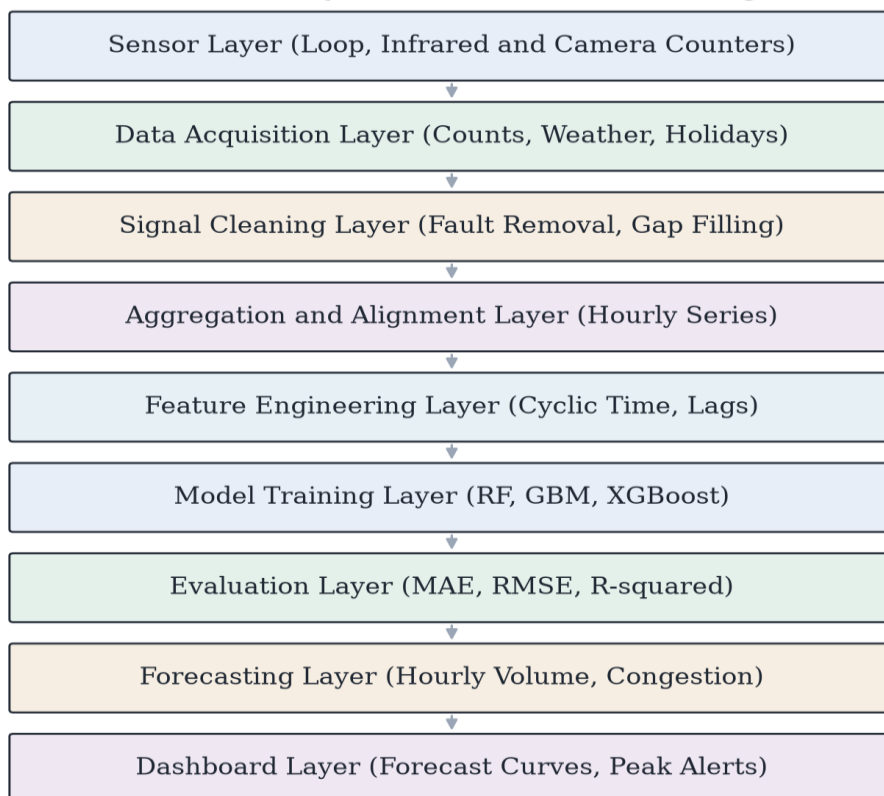
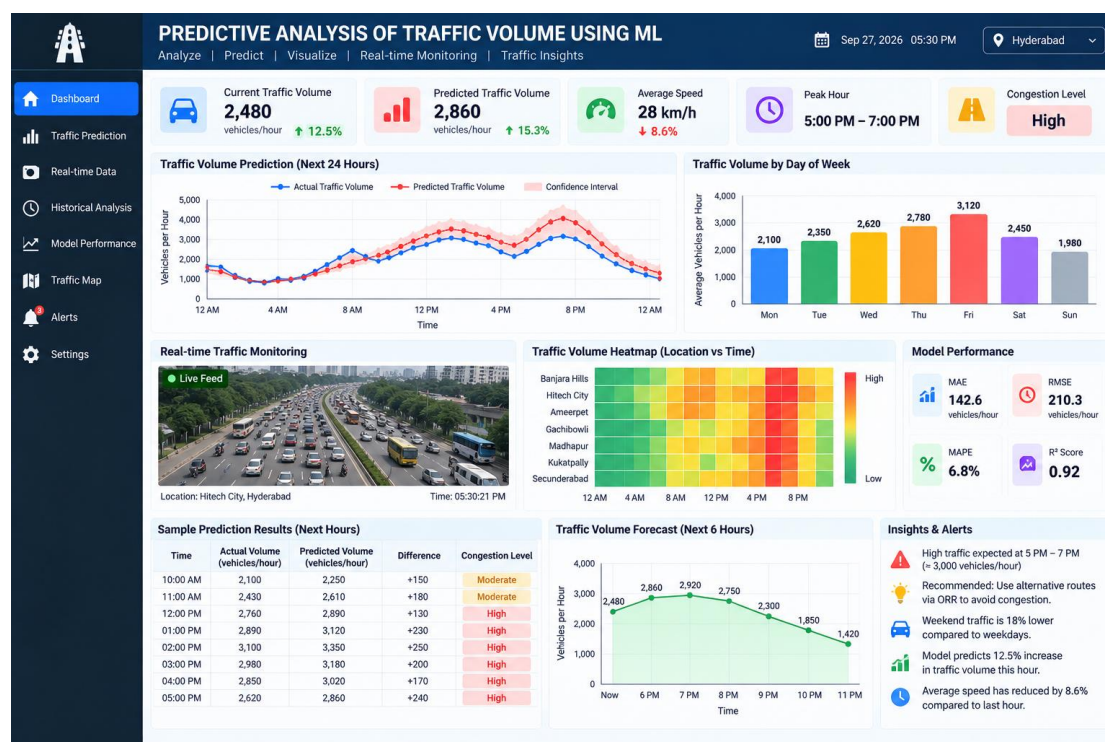


Fig. 1 System Architecture

V. RESULT AND OUTPUT



VI. CONCLUSION

This paper has presented a predictive analysis system for traffic volume that uses machine learning to forecast hourly vehicle counts on an urban road segment. By combining cleaned detector data with calendar and weather information, the system provides traffic authorities with an estimate of future demand rather than only a view of current conditions, which makes it possible to plan ahead instead of reacting to congestion after it has formed.

The work showed that careful handling of the sensor signal is essential. Detector faults, gaps, and duplicated records are common in real traffic data, and removing or correcting them before modelling improves forecast accuracy significantly. Temporal feature engineering, particularly cyclical encoding of time and the use of lag and rolling features, allowed the models to capture the strong daily and weekly patterns of traffic flow as well as its recent behaviour.

Among the regression models compared, ensemble methods such as random forest and gradient boosting achieved the lowest errors on the chronologically separated test period, while remaining fast to train and reasonably easy to interpret. Feature importance analysis confirmed that hour of day, recent volume, weekday, and weather are the dominant influences. The dashboard presents forecasts, peak alerts, and congestion levels in a form that control room staff can use when adjusting signals, deploying personnel, or scheduling roadworks.

Future work can extend the system to real-time streaming of detector data so that forecasts are updated every few minutes. Forecasting across a network of connected junctions, using spatial relationships between detectors, would allow congestion to be anticipated as it spreads. Sequence models such as long short-term memory networks can be explored for longer horizons, and integration with adaptive signal controllers would allow forecasts to influence signal timing directly.

REFERENCES

- [1] Kumar, R. D., Prudhviraaj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.
- [2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.
- [3] Sv satyakrishna, shirisha rangu, bhargavi nalacheruve. (2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm
- [4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.
- [5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, "Real-Time Object Detection in Drone Surveillance Using YOLOv5," in Proc. 2025 3rd Int.

Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.

[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, “Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks,” in *Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment*, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.

[7] R. D. Kumar, V. N. S. Manaswini, “Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology,” in *Blockchain for Smart Cities*, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.

[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, “An advanced movie recommender using collaborative filtering and sentiment analysis,” *International Research Journal of Modernization in Engineering Technology and Science*, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.

[9] Ravi Kumar Banoth, Ramana Murthy B V, “Automatic crop recommendation system using LightGBM and decision tree machine learning models,” *Journal of Machine and Computing*, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.

[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, “Smart agriculture through IoT and machine learning for analyzing carbon footprints,” in *Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE)*, Apr. 2025.

[11] Ravi Kumar Banoth, B. V. Ramana Murthy, “Soil image classification using transfer learning approach: MobileNetV2 with CNN,” *SN Computer Science*, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.