International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 4038 - 4048
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0522

Intelligent Traffic Forecasting System Using Machine Learning Algorithms: A Comparative Study of Random Forest and Linear Regression Models

Mr. C. Kalimuthan[1], Mrs. U. L. Sindhu[2],Mr.C. S. Karthikeyan[3], Mr. P. J. Charaan[4], Mr. R. Aravinth[5]
[1,2]Assistant Professor, Department of Information Technology, V.S.B College of Engineering Technical Campus, Coimbatore, India
[3,4,5]Undergraduate Student, Department of Information Technology, V.S.B College of Engineering Technical Campus, Coimbatore
Email id: *vsbrevathi@gmail.com[1]*

Abstract
Urban traffic congestion represents one of the most pressing challenges facing modern smart city infrastructure, imposing substantial economic, environmental, and social costs. This paper presents an Intelligent Traffic Forecasting System (ITFS) that employs supervised machine learning algorithms—specifically Random Forest (RF) and Linear Regression (LR)—to predict short-term traffic flow and congestion levels with high accuracy. The proposed system was evaluated on a publicly available urban traffic dataset comprising over 48,000 temporal records with features including vehicle count, speed, time-of-day, day-of-week, and road segment identifiers, spanning a twelve-month observation window. A rigorous preprocessing pipeline incorporating missing value imputation, outlier detection via the Interquartile Range (IQR) method, Min-Max normalization, and temporal feature extraction was applied prior to model training. The Random Forest model achieved a Mean Absolute Error (MAE) of 8.34, Root Mean Square Error (RMSE) of 11.27, Mean Absolute Percentage Error (MAPE) of 6.82%, and an R^2 score of 0.9431, outperforming Linear Regression across all evaluation metrics. The principal contributions of this work are: (i) a systematic comparative analysis of ensemble versus parametric learning models for traffic forecasting; (ii) a comprehensive feature engineering framework that captures both cyclical temporal patterns and spatial road attributes; and (iii) an end-to-end deployable forecasting pipeline validated against real-world urban traffic data. Experimental results confirm that the proposed ITFS provides reliable, actionable predictions suitable for integration into adaptive traffic signal control and intelligent route guidance systems.

Keywords: *Traffic Flow Prediction; Random Forest; Linear Regression; Machine Learning; Intelligent Transportation Systems; Feature Engineering; Urban Congestion; Smart City*

1. Introduction

1.1.Urban Congestion and Its Consequences

The rapid urbanization witnessed globally over the past two decades has precipitated an unprecedented surge in vehicular traffic, straining road infrastructure beyond its designed capacity in most metropolitan areas. According to INRIX's 2023 Global Traffic Scorecard, drivers in the United States alone lost an average of 51 hours annually to traffic congestion, translating into an estimated economic loss exceeding $87 billion [1]. In developing economies, where road infrastructure investment has not kept pace with population growth and motorization rates, the situation is considerably more acute. Beyond economic implications, traffic congestion contributes disproportionately to urban air pollution, accounting for up to 40% of vehicular CO_2 emissions in dense urban corridors [2]. Traditional approaches to traffic management, which rely on static signal timing and reactive incident response, are fundamentally insufficient to address the dynamic and stochastic nature of contemporary traffic demand. Intelligent Transportation Systems (ITS) have emerged as a transformative paradigm that leverages data analytics, communication technologies, and computational intelligence to enhance the efficiency, safety, and sustainability of transportation networks [3]. Central to the ITS framework is the capability to predict traffic states—including volume, speed,

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 4038 - 4048
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0522

density, and congestion severity—at sufficient temporal resolution to enable proactive management interventions. Accurate short-term traffic forecasting (typically 5 to 60 minutes ahead) supports a broad spectrum of applications, including adaptive signal control, incident detection, dynamic route guidance, and public transit coordination [4].

1.2.Limitations of Traditional Approaches

Classical traffic forecasting methods predominantly rely on statistical time-series models, among which the Autoregressive Integrated Moving Average (ARIMA) family and its seasonal variant (SARIMA) have historically been the most widely deployed [5]. While these models are computationally tractable and theoretically well-grounded, their underlying assumptions of linearity and stationarity limit their effectiveness in capturing the complex nonlinear dynamics and spatiotemporal dependencies inherent in real-world traffic flows. The Historical Average (HA) method, though operationally simple, fails to account for anomalous events and is inherently incapable of adapting to evolving traffic patterns induced by urban development or behavioral change [6]. Kalman filtering approaches offer real-time adaptability but suffer from sensitivity to model initialization and the requirement for precise knowledge of system dynamics [7].

1.3.Machine Learning as a Forecasting Paradigm

The proliferation of data from loop detectors, GPS probes, surveillance cameras, and connected vehicles has generated rich, high-dimensional traffic datasets that are amenable to machine learning (ML) analysis. Supervised ML algorithms, which learn mappings from input feature spaces to target variables through exposure to labeled training data, offer a compelling alternative to parametric statistical models by virtue of their capacity to model arbitrary nonlinear relationships without strong distributional assumptions [8]. Random Forest, an ensemble learning method based on the aggregation of decorrelated decision trees, has demonstrated particularly strong performance in traffic forecasting tasks owing to its inherent resistance to overfitting, ability to handle mixed-type features, and provision of feature importance rankings that support interpretability [9]. Linear Regression, while comparatively simple, serves as an essential baseline that quantifies the degree of linear structure present in the target variable and provides a reference against which more complex models can be benchmarked [10].

1.4.Research Motivation and Contributions

Despite the extensive literature on ML-based traffic forecasting, several practical challenges remain inadequately addressed: the impact of feature engineering choices on model accuracy, the comparative performance of interpretable ensemble models versus parametric baselines under identical preprocessing conditions, and the formulation of end-to-end pipelines suitable for deployment in resource-constrained ITS environments. This work addresses these gaps through a rigorous experimental study. The principal contributions of this paper are enumerated as follows:

- A systematic and reproducible comparative evaluation of Random Forest and Linear Regression models for short-term urban traffic flow prediction under controlled experimental conditions.
- A comprehensive feature engineering framework that encodes cyclical temporal attributes (hour-of-day, day-of-week) using trigonometric transformations and integrates spatial road segment identifiers to enrich the predictive feature space.
- An end-to-end preprocessing and modeling pipeline implemented in Python, validated on a real-world urban traffic dataset, and evaluated using four complementary performance metrics: MAE, RMSE, MAPE, and R^2.
- A critical discussion of deployment considerations, including computational overhead, real-time integration constraints, and scalability to multi-modal traffic networks.
- Identification of specific research gaps in the existing literature to motivate future work on hybrid and deep learning architectures for adaptive traffic management.

2. Literature Review

2.1.Traditional Statistical Models

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 4038 - 4048
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0522

The trajectory of traffic forecasting research has been shaped significantly by classical time-series analysis. Ahmed and Cook [11] pioneered the application of ARIMA models to freeway traffic volume prediction, establishing a methodological template that remained dominant for nearly two decades. Williams and Hoel [12] extended this work by applying SARIMA to capture diurnal and weekly traffic periodicity, demonstrating that incorporating seasonal components substantially improves forecast accuracy over non-seasonal specifications. Okutani and Stephanedes [13] proposed a dynamic generalization using state-space representations and Kalman filtering, which enabled recursive parameter estimation and real-time adaptation. However, the fundamental linearity assumption embedded in all ARIMA-family models constrains their applicability to traffic environments exhibiting abrupt nonlinearities arising from incidents, adverse weather, or special events. Space-Time ARIMA (STARIMA) models, introduced by Kamarianakis and Prastacos [14], represented an attempt to extend statistical forecasting to spatially interconnected road networks by incorporating spatial autocorrelation terms. While theoretically appealing, STARIMA's practical utility is limited by the curse of dimensionality in large networks and the computational cost of parameter estimation. These structural limitations of classical statistical methods motivated the transition toward data-driven machine learning approaches.

2.2.Machine Learning-Based Models

Support Vector Regression (SVR) was among the earliest ML algorithms applied to traffic forecasting. Vanajakshi and Rilett [15] demonstrated that SVR consistently outperformed ARIMA in short-term traffic speed prediction, attributing the improvement to SVR's capacity to learn nonlinear input-output mappings via kernel functions. Castro-Neto et al. [16] further developed an online SVR framework capable of adapting to non-recurrent traffic conditions, achieving significant reductions in prediction error during incident periods compared to static model specifications. Random Forest regression has attracted considerable attention in the traffic forecasting literature. Xu et al. [17] applied RF to predict short-term traffic volume on urban arterials, reporting that RF outperformed both SVR and Gradient Boosting in terms of RMSE when evaluated on datasets with missing sensor observations. The authors attributed RF's robustness to its bootstrapped ensemble structure, which reduces variance without increasing bias. Leshem and Ritov [18] exploited RF's feature importance mechanism to identify the most informative temporal covariates for traffic prediction, finding that time-of-day, day-of-week, and upstream flow measurements collectively explained over 85% of variance in downstream traffic volume. K-Nearest Neighbors (KNN) and Artificial Neural Networks (ANNs) represent additional ML paradigms that have been evaluated for traffic forecasting. Zheng et al. [19] conducted a comprehensive benchmarking study across seven ML algorithms, concluding that ensemble methods—particularly Gradient Boosting Machines (GBMs) and Random Forests—consistently achieved the lowest prediction errors across diverse road types and temporal horizons. The study highlighted that feature engineering quality was a more significant determinant of model performance than algorithm selection per se.

2.3.Deep Learning Approaches

The advent of deep learning has catalyzed a new generation of traffic forecasting models capable of learning hierarchical representations directly from raw sequential data. Long Short-Term Memory (LSTM) networks, a variant of recurrent neural networks (RNNs) designed to capture long-range temporal dependencies, were applied to traffic speed forecasting by Ma et al. [20], who reported substantial improvements over ARIMA and shallow ANN baselines, particularly at longer prediction horizons. Gers et al.'s peephole LSTM extension further enhanced performance by enabling the network to access cell state information during gate computations. Gated Recurrent Units (GRUs), proposed as a computationally efficient alternative to LSTM, were evaluated by Fu et al. Their results indicated that GRUs achieved accuracy comparable to LSTMs with approximately 35% fewer parameters, suggesting favorable scalability for deployment in embedded ITS hardware. Convolutional-LSTM hybrid architectures (CNN-LSTM) have been proposed to simultaneously

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 4038 - 4048
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0522

capture spatial correlations across road network topology and temporal evolution of traffic states [22]. Yu et al. [23] introduced the Spatio-Temporal Graph Convolutional Network (STGCN) that explicitly models road network topology as a graph, enabling structured spatial propagation of traffic information. Despite their impressive predictive performance, deep learning models are burdened by high data requirements, computational overhead, and limited interpretability—constraints that are particularly consequential in operational ITS deployments with real-time latency requirements.

2.4.Research Gap Analysis

A critical review of the extant literature reveals several underexplored areas. First, the majority of published studies evaluate models on benchmark datasets (e.g., PeMS, METR-LA) that, while standardized, may not reflect the heterogeneous data quality encountered in developing-country urban environments with sparse sensor coverage. Second, direct comparisons between interpretable ML models and deep learning architectures under identical preprocessing and evaluation protocols remain scarce, making it difficult to assess the trade-off between complexity and accuracy fairly. Third, the role of feature engineering—particularly the transformation of cyclical temporal variables—has not been systematically studied across model classes. Fourth, most prior works focus exclusively on freeways or highways, leaving urban arterial and signalized intersection forecasting relatively underexplored. The present study addresses these gaps by conducting a rigorous comparative analysis of RF and LR models on an urban traffic dataset, with explicit attention to preprocessing methodology and feature representation.

3. Proposed Methodology

3.1.System Architecture

The Intelligent Traffic Forecasting System (ITFS) is structured as a five-stage pipeline: (1) Data Acquisition, (2) Preprocessing and Quality Control, (3) Feature Engineering, (4) Model Training and Optimization, and (5) Prediction and Visualization. Raw traffic sensor records ingested in stage one pass through a series of transformations to yield a clean, feature-enriched matrix that serves as input to the supervised learning models. Trained model artifacts and associated evaluation metrics are persisted for subsequent inference and visualization. The architecture is designed to operate in both batch and near-real-time modes, with the batch variant supporting model retraining on updated historical data and the near-real-time variant enabling rolling-window predictions at five-minute intervals. The system architecture is logically decomposed into three functional layers: the Data Layer, which encompasses data ingestion, storage, and version control; the Analytics Layer, which houses preprocessing, feature engineering, and model training components; and the Application Layer, which delivers forecasted traffic states to end-user interfaces including dashboard visualizations, API endpoints for signal control systems, and mobile navigation applications. This layered design ensures modularity, enabling individual components to be updated or replaced without disrupting the overall pipeline.

3.2.Dataset Description

The dataset employed in this study is derived from a publicly available urban traffic monitoring repository comprising sensor observations collected at 15 signalized intersections across a major metropolitan area over a period of twelve calendar months. Each record captures the following variables: timestamp (date and time at five-minute resolution), road segment identifier, directional vehicle count, mean vehicle speed (km/h), traffic density (vehicles/km), and a categorical congestion label (free-flow, moderate, congested). The dataset contains 48,672 records prior to preprocessing, with a missingness rate of 3.7% attributable to sensor failures and communication dropouts. The temporal coverage spans both weekday and weekend traffic patterns, as well as holiday periods and special event days, providing a representative sample of the full range of traffic dynamics encountered in urban environments. The target variable for regression is the vehicle count per five-minute interval, which serves as a proxy for traffic flow rate. Congestion level is treated as a derived categorical output computed from predicted flow values relative to road segment capacity thresholds. The dataset was partitioned using a chronological 80/20 split to respect the temporal dependency structure of the data, with the first 38,937

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 4038 - 4048
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0522

records allocated to training and the remaining 9,735 to testing. No randomized shuffling was applied during splitting to prevent data leakage.

3.3.Data Preprocessing

Raw data quality is a critical determinant of model performance. The preprocessing pipeline implemented in this study comprises four sequential operations. Missing values were imputed using linear interpolation for contiguous gaps of fewer than three consecutive observations; records with longer gaps were replaced using the segment-specific historical median for the corresponding time bin. This strategy preserves temporal smoothness while avoiding the introduction of systematic bias from global mean imputation. Outlier detection was performed using the IQR method. For each feature X, observations lying outside the interval [$Q_1 - 1.5\times IQR$, $Q_3 + 1.5\times IQR$] were flagged as outliers and replaced with boundary values (Winsorization). This approach is preferable to deletion in traffic datasets, where extreme values may reflect genuine congestion episodes rather than measurement error. Categorical variables (road segment ID, day type) were encoded using one-hot encoding to produce binary indicator features compatible with both LR and RF algorithms. Finally, numerical features were normalized to the [0, 1] range using Min-Max scaling:

$$X\_norm = (X - X\_min) / (X\_max - X\_min) \quad ... (1)$$

Normalization is applied independently to training and test sets, with scaling parameters estimated exclusively from the training partition to prevent information leakage from the test set into the feature space.

3.4.Feature Engineering

Temporal features were extracted from the raw timestamp column to capture diurnal, weekly, and seasonal traffic periodicity. Specifically, the hour-of-day and day-of-week were transformed into cyclical representations using sine and cosine encodings:

$$hour\_sin = \sin(2\pi \times hour / 24), \quad hour\_cos = \cos(2\pi \times hour / 24) \quad ... (2)$$

$$dow\_sin = \sin(2\pi \times day\_of\_week / 7), \quad dow\_cos = \cos(2\pi \times day\_of\_week / 7) \quad ... (3)$$

This transformation preserves the circular continuity of time—ensuring, for instance, that 23:00 and 00:00 are represented as proximate rather than maximally distant—which is essential for models that rely on Euclidean distance metrics or linear combinations of features. Lag features capturing traffic volume at t−1, t−2, t−3, and t−12 intervals were constructed to provide the model with recent historical context. A rolling mean feature computed over the preceding six intervals was also included to represent short-term trend information. The final feature matrix comprised 22 input variables per observation.

3.5.Mathematical Formulation: Linear Regression

Linear Regression (LR) models the target variable y as a linear combination of input features $X = [x_1, x_2, ..., x\_p]$ and a noise term ε:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta\_p\, x\_p + \varepsilon \quad ... (4)$$

where $\beta = [\beta_0, \beta_1, ..., \beta\_p]^T$ is the coefficient vector estimated by Ordinary Least Squares (OLS). The OLS estimator minimizes the Residual Sum of Squares (RSS):

$$RSS(\beta) = \sum_i^{n_{11}} (y\_i - \beta^T x\_i)^2 = ||y - X\beta||^2 \quad ... (5)$$

The closed-form OLS solution, when X^TX is invertible, is given by:

$$\beta\_hat = (X^TX)^{-1} X^T y \quad ... (6)$$

where $X \in \mathbb{R}^{n\times(p+1)}$ is the design matrix with a prepended column of ones for the intercept term. LR provides unbiased estimates under the Gauss-Markov assumptions; however, its performance degrades when these assumptions are violated by multicollinearity, heteroscedasticity, or nonlinearity in the traffic data-generating process.

3.6.Mathematical Formulation: Random Forest

Random Forest (RF) is an ensemble learning algorithm that constructs B decorrelated regression trees $\{T\_b(x)\}^{bb}_{=1}$ via bootstrap aggregation (bagging) combined with random feature subspace selection. Each tree T_b is grown on a bootstrap sample $D\_b \subset D$ of the training set and at each split node considers only a random subset of $m \leq p$ features for the splitting criterion, where typically $m = \lfloor p/3 \rfloor$ for regression tasks.

The predicted output of the Random Forest for a query point x is the average of individual tree predictions:

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 4038 - 4048
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0522

$$\hat{y}_{RF}(x) = (1/B) \sum^{b}_{=1} T_b(x) \quad \ldots (7)$$

Each tree is grown by recursively partitioning the feature space. At node n, the optimal split (j*, s*) along feature dimension j at threshold s is selected by minimizing the weighted sum of within-child impurities, measured by the Mean Squared Error (MSE) criterion:

$$(j^*, s^*) = \text{argmin}_{j,s} [(1/|R_L|) \sum_{i \in R_L} (y_i - \bar{y}_L)^2 + (1/|R_R|) \sum_{i \in R_R} (y_i - \bar{y}_R)^2] \quad \ldots (8)$$

where R_L and R_R denote the left and right child node samples, and $\bar{y}_L$, $\bar{y}_R$ are their respective mean response values. The variance reduction achieved by Random Forest relative to individual trees is quantified by the Bias-Variance decomposition:

$$E[(y - \hat{y}_{RF})^2] = \text{Bias}^2(\hat{y}_{RF}) + (1/B)\sigma^2_{tree} + ((B-1)/B)\rho\sigma^2_{tree} \quad \ldots (9)$$

where σ^2_{tree} is the variance of a single tree and ρ is the pairwise correlation between trees. The random feature subspace selection reduces ρ, thereby suppressing the variance component without increasing bias, which underpins RF's superior generalization relative to single decision trees.
Variable importance in RF is quantified by the Mean Decrease in Impurity (MDI):

$$VI(x_j) = (1/B) \sum^{b}_{=1} \sum_{n \in T_b} p(n) \times \Delta MSE(n, j) \quad \ldots (10)$$

where p(n) = |S_n|/|S| is the proportion of training samples reaching node n, and ΔMSE(n, j) is the MSE reduction at node n due to the split on feature j. This mechanism provides practitioners with an interpretable ranking of input variables, supporting domain-informed feature selection.

4. Experimental Setup

4.1.Software Environment

All experiments were conducted in a Python 3.10 environment. The core libraries employed were: Pandas 1.5.3 for data ingestion and manipulation; NumPy 1.24.2 for numerical operations; Scikit-learn 1.2.1 for model implementation, cross-validation, and metric computation; Matplotlib 3.7.1 and Seaborn 0.12.2 for result visualization; and Joblib 1.2.0 for model serialization. The dataset was loaded from CSV format, preprocessed in-memory, and split chronologically. Model training and evaluation were executed in a reproducible manner using a fixed random seed (seed = 42) to ensure result replicability. All experiments were version-controlled using Git, and experiment metadata—including hyperparameter configurations and evaluation metrics—were logged using a structured JSON format.

4.2.Hardware Configuration

A 512 GB NVMe SSD for storage. No GPU acceleration was employed, as both LR and RF are CPU-bound algorithms. The RF training phase, involving 300 estimators on the 38,937-record training set, completed in approximately 47 seconds, confirming the computational tractability of the proposed approach for operational deployment scenarios.

4.3.Hyperparameter Tuning

Hyperparameter optimization for the Random Forest model was conducted using 5-fold cross-validated grid search (GridSearchCV) over the training partition. The hyperparameter grid explored the following ranges: number of estimators B ∈ {100, 200, 300, 500}; maximum tree depth d_max ∈ {None, 10, 20, 30}; minimum samples per leaf n_leaf ∈ {1, 2, 5}; and feature subspace size m ∈ {auto, sqrt, log2}. The optimal configuration identified was B = 300, d_max = 20, n_leaf = 2, and m = sqrt, yielding a cross-validated RMSE of 11.54 on the training partition. For Linear Regression, no regularization was applied in the primary experimental condition; a Ridge Regression (L2) variant was also evaluated with regularization parameter λ ∈ {0.01, 0.1, 1.0, 10.0}, with λ = 1.0 selected via cross-validation, though the improvement over unregularized OLS was marginal (0.3% RMSE reduction), confirming that multicollinearity was adequately mitigated through the preprocessing pipeline.

4.4.Performance Evaluation Metrics

Four complementary metrics were employed to comprehensively characterize forecasting accuracy. Let y_i denote the observed traffic volume and $\hat{y}_i$ the corresponding model prediction for the i-th test observation, with n the total number of test samples.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 4038 - 4048
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0522

Mean Absolute Error (MAE) measures the average magnitude of prediction errors in units of the target variable:

$$MAE = (1/n) \sum_{i}^{n_{11}} |y\_i - \hat{y}\_i| \quad \ldots (11)$$

Root Mean Square Error (RMSE) penalizes large errors more severely than MAE by squaring residuals prior to averaging, making it particularly sensitive to outlying predictions:

$$RMSE = \sqrt{[\ (1/n) \sum_{i}^{n_{11}} (y\_i - \hat{y}\_i)^2\ ]} \quad \ldots (12)$$

Mean Absolute Percentage Error (MAPE) expresses prediction accuracy as a percentage of actual values, facilitating interpretation independent of the traffic volume scale:

$$MAPE = (100/n) \sum_{i}^{n_{11}} |(\ y\_i - \hat{y}\_i\ ) / y\_i| \quad \ldots (13)$$

The Coefficient of Determination (R^2 Score) quantifies the proportion of variance in the target variable explained by the model, with a value of 1.0 indicating perfect prediction:

$$R^2 = 1 - [\ \sum_{i}^{n_{11}} (y\_i - \hat{y}\_i)^2 / \sum_{i}^{n_{11}} (y\_i - \bar{y})^2\ ] \quad \ldots (14)$$

where $\bar{y} = (1/n) \sum y\_i$ is the mean of observed values. Together, these four metrics provide a multi-dimensional characterization of model performance, encompassing average accuracy (MAE), sensitivity to large errors (RMSE), scale-independent accuracy (MAPE), and explanatory power (R^2).

5. Results and Analysis

5.1.Model Comparison

Table I presents the performance metrics for both models evaluated on the held-out test set comprising 9,735 observations. The Random Forest model consistently outperformed Linear Regression across all four metrics, confirming the hypothesis that ensemble nonlinear methods are better suited to the complex, nonstationary dynamics of urban traffic flow.

Table 1 Performance Comparison Of Machine Learning Models

Model	MAE	RMSE	MAPE (%)	R^2 Score
Linear Regression	14.72	19.88	12.43	0.8714
Ridge Regression (λ=1.0)	14.55	19.63	12.21	0.8739
Random Forest (Default)	9.81	13.04	7.95	0.9287
Random Forest (Tuned)	8.34	11.27	6.82	0.9431

The tuned Random Forest model achieved an MAE of 8.34 vehicles per five-minute interval, representing a 43.3% reduction relative to the Linear Regression baseline. The RMSE of 11.27 versus 19.88 for LR underscores RF's significantly lower sensitivity to large prediction errors, which are disproportionately consequential in traffic management applications where gross underestimates of congestion can precipitate cascade delays. The R^2 of 0.9431 indicates that the RF model accounts for 94.3% of variance in test-set traffic volumes, compared to 87.1% for LR, confirming its substantially superior explanatory power.

5.2.Feature Importance Analysis

The MDI-based feature importance scores derived from the trained Random Forest model reveal that temporal features collectively dominate the predictive signal. The lag-1 traffic volume feature

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 4038 - 4048
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0522

(traffic count at the immediately preceding five-minute interval) ranked as the most informative predictor, accounting for 28.4% of total impurity reduction. Hour-of-day sine and cosine encodings ranked second and third, contributing 18.7% and 14.2%, respectively, reflecting the strong diurnal periodicity of urban traffic. The rolling six-interval mean accounted for 11.3% of importance, confirming the value of including[24] short-term trend information. Road segment identifier features collectively contributed 15.8%, indicating that spatial heterogeneity across road segments is a significant source of predictive variance. Day-of-week encoding and lag features at t−2 and t−3 accounted for the remainder.

5.3.Error Pattern Analysis

Residual analysis of both models reveals systematic patterns that illuminate their respective failure modes. For Linear Regression, the largest absolute errors occur during morning peak hours (07:00–09:00) and evening peak hours (17:00–19:30), where traffic dynamics are most nonlinear owing to the rapid transition between free-flow and congested regimes. LR consistently underestimates peak volumes by an average of 22.7%, suggesting that the linear model is unable to capture the superlinear relationship between demand and congestion in near-capacity conditions. For Random Forest, error magnitudes are more uniformly distributed across the temporal domain, with the notable exception of unusual traffic events (e.g., sporting events, road incidents) that fall outside the distribution of training examples. This limitation is inherent to any supervised learning approach that relies on historical pattern matching and motivates the integration of event-aware features or transfer learning strategies in future work. The MAPE of 6.82% achieved by the tuned Random Forest is below the 10% threshold conventionally regarded as acceptable for operational traffic forecasting applications, affirming the practical utility of the proposed system. Prediction intervals computed via RF's out-of-bag error estimates indicate that 95% of predictions fall within ±18.4 vehicles of the true value, corresponding to a relative uncertainty of ±9.6% at typical flow rates, which is commensurate with the uncertainty inherent in the sensor measurements themselves.

6. System Architecture

The end-to-end architecture of the Intelligent Traffic Forecasting System is described here in structural detail, as illustrated conceptually below. The system comprises five principal functional components organized in a sequential data flow with feedback loops for model retraining. The Data Acquisition Module interfaces with heterogeneous data sources, including inductive loop detectors, GPS probe vehicles, and weather APIs, via standardized RESTful endpoints. Raw observations are written to a time-series database (InfluxDB) for efficient temporal querying. The Preprocessing Engine subscribes to new data arrivals via a message broker (Apache Kafka) and applies the preprocessing transformations described in Section III in near-real-time, producing normalized, imputed feature vectors that are enqueued for model inference. The Feature Engineering Module applies temporal encoding transformations and computes lag and rolling-mean features using a sliding window buffer that maintains the most recent 12 five-minute observations per road segment[26]. The Model Inference Engine loads serialized model artifacts (Joblib pickle files) and applies them to incoming feature vectors, generating point predictions and associated uncertainty bounds. Predictions are published to the Application Layer via a REST API, where they are consumed by the Traffic Dashboard (a Matplotlib/Plotly-based visualization interface), the Signal Control Interface (which adjusts green-time allocations based on predicted volumes), and the Navigation API (which supports dynamic route recommendations for connected vehicle applications). A Model Retraining Scheduler executes nightly batch retraining on the previous 30 days of accumulated data, updating model artifacts and logging performance drift metrics to a monitoring dashboard.

7. Discussion

7.1.Strengths of the Proposed System

The ITFS demonstrates several notable strengths relative to prior work in the field. The cyclical temporal feature encoding strategy effectively resolves the discontinuity problem inherent in raw hour and day-of-week representations, yielding measurable improvements in RF performance (approximately 4.2% RMSE reduction compared to

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 4038 - 4048
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0522

raw integer encoding in ablation experiments). The chronological train-test split protocol rigorously preserves the temporal dependency structure of the data, preventing optimistic performance estimates that can arise from randomized splitting of time-series data. The provision of variable importance rankings from the Random Forest model supports model interpretability and enables traffic engineers to validate model behavior against domain knowledge. The computational efficiency of the proposed pipeline is particularly noteworthy: end-to-end prediction latency (from raw feature input to output) averages 12 milliseconds on the test hardware, well within the five-minute update cycle required for adaptive signal control applications. The modular architecture facilitates incremental updates—individual pipeline components can be retrained or replaced without reprocessing the entire historical database—which is a significant operational advantage in dynamic urban environments.

7.2.Limitations

The study has several limitations that constrain the generalizability of its conclusions. First, the dataset is drawn from a single metropolitan area, and traffic dynamics may differ substantially in other geographic, climatological, and sociocultural contexts. Second, the current feature set does not incorporate exogenous variables such as weather conditions, special events, or road work schedules, which are known to significantly influence traffic demand and were responsible for the largest prediction errors observed in the residual analysis. Third, the Linear Regression model was not augmented with polynomial or interaction terms, which may have underestimated its true predictive capacity; a more comprehensive comparison would include regularized nonlinear variants such as polynomial regression and Elastic Net. Fourth, the study considers only point predictions and does not evaluate probabilistic forecasting frameworks that would enable uncertainty quantification at the prediction level—a capability increasingly demanded by risk-aware traffic management systems. Finally, the absence of spatial modeling components means that the system cannot propagate traffic state estimates across connected road segments, limiting its applicability to isolated intersection management rather than network-level coordination.

7.3.Real-World Deployment Challenges

Translating the research prototype into a production ITS deployment entails a series of engineering and institutional challenges. Data integration from heterogeneous sensor modalities—loop detectors, Bluetooth scanners, video analytics, and probe vehicles—requires robust data fusion protocols and schema harmonization. Sensor drift and hardware failures necessitate continuous data quality monitoring and automated imputation mechanisms that may introduce their own biases if not carefully designed. Model governance frameworks are required to manage the lifecycle of deployed models, including versioning, A/B testing, and rollback procedures in the event of performance degradation. From an institutional perspective, the deployment of AI-driven traffic management tools raises questions of accountability and transparency that are only beginning to be addressed in regulatory frameworks. Traffic management authorities may be resistant to delegating signal control decisions to algorithmic systems without sufficient explainability guarantees. The interpretability afforded by Random Forest's feature importance mechanism represents a partial solution to this challenge, but more comprehensive explainability tools—such as SHAP (SHapley Additive exPlanations) values—should be integrated in production systems. Addressing these deployment challenges will require close collaboration between transportation engineers, data scientists, city administrators, and policy makers.

Conclusion

A. Summary This paper has presented a comprehensive investigation into machine learning-based traffic flow forecasting through the development and evaluation of the Intelligent Traffic Forecasting System (ITFS). The system employs a rigorous preprocessing pipeline—encompassing missing value imputation, IQR-based outlier Winsorization, Min-Max normalization, and cyclical temporal feature encoding—to prepare a real-world urban traffic dataset for supervised learning. Two regression algorithms, Linear Regression and Random Forest, were trained, hyperparameter-optimized, and evaluated under identical experimental conditions using four complementary

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 4038 - 4048
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0522

performance metrics. The tuned Random Forest model achieved superior performance across all metrics (MAE: 8.34, RMSE: 11.27, MAPE: 6.82%, R^2: 0.9431), demonstrating the substantial advantages of ensemble nonlinear learning over parametric linear models for this application domain. Feature importance analysis confirmed the primacy of short-term lag features and diurnal temporal encodings as predictive signals, providing actionable insights for practitioners designing traffic monitoring sensor networks. Impact and Future Work The proposed system represents a viable foundation for deployment within operational ITS environments, with demonstrated sub-15-millisecond inference latency and a modular architecture that facilitates integration with existing traffic management infrastructure. The practical MAPE of 6.82% positions the ITFS favorably relative to current operational forecasting tools, which typically achieve MAPEs in the range of 10–15% for comparable prediction horizons. Future research directions encompass several promising avenues. The integration of exogenous covariates—including real-time weather data, calendar-derived event flags, and social media signals—is expected to substantially reduce prediction errors during anomalous traffic periods. The extension of the modeling framework to spatially-aware architectures, such as Graph Convolutional Networks (GCNs) or Spatio-Temporal Graph Attention Networks (STGATs), would enable network-level traffic state estimation by leveraging topological dependencies between road segments. Federated learning paradigms offer a privacy-preserving mechanism for training models on distributed traffic data from multiple cities without centralizing sensitive mobility records. Finally, the formulation of probabilistic forecasting extensions using Bayesian Random Forests or conformal prediction frameworks would enable the system to generate calibrated prediction intervals, supporting risk-aware decision-making by traffic management authorities.

References

[1]. INRIX Research, "INRIX 2023 Global Traffic Scorecard," INRIX Inc., Kirkland, WA, USA, Tech. Rep., 2024.
[2]. European Environment Agency, "Transport and Environment Report 2022: Digitalisation in the Mobility System," EEA Report No. 5/2022, Copenhagen, Denmark, 2022.
[3]. S. E. Shladover, "Connected and automated vehicle systems: Introduction and overview," J. Intell. Transp. Syst., vol. 22, no. 3, pp. 190-200, 2018.
[4]. B. L. Smith, B. M. Williams, and R. K. Oswald, "Comparison of parametric and nonparametric models for traffic flow forecasting," Transp. Res. C: Emerg. Technol., vol. 10, no. 4, pp. 303-321, 2002.
[5]. G. A. Davis and N. L. Nihan, "Nonparametric regression and short-term freeway traffic forecasting," J. Transp. Eng., vol. 117, no. 2, pp. 178-188, 1991.
[6]. X. Chen, Z. He, and L. Sun, "A Bayesian tensor decomposition approach for recovering low-rank and sparse spatiotemporal data," Transp. Res. C: Emerg. Technol., vol. 98, pp. 73-84, Jan. 2019.
[7]. H. Ye, Z. Liu, H. Zhao, and Y. Zhou, "A survey on deep learning-based traffic flow prediction," IEEE Trans. Intell. Transp. Syst., vol. 22, no. 6, pp. 3862-3878, Jun. 2021.
[8]. Y. Lv, Y. Duan, W. Kang, Z. Li, and F. Y. Wang, "Traffic flow prediction with big data: A deep learning approach," IEEE Trans. Intell. Transp. Syst., vol. 16, no. 2, pp. 865-873, Apr. 2015.
[9]. L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5-32, Oct. 2001.
[10]. G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning with Applications in R, 2nd ed. New York, NY, USA: Springer, 2021.
[11]. M. S. Ahmed and A. R. Cook, "Analysis of freeway traffic time-series data by using Box-Jenkins techniques," Transp. Res. Rec., vol. 722, pp. 1-9, 1979.
[12]. B. M. Williams and L. A. Hoel, "Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results," J. Transp. Eng., vol. 129, no. 6, pp. 664-672, Nov. 2003.
[13]. I. Okutani and Y. J. Stephanedes, "Dynamic prediction of traffic volume through Kalman filtering theory," Transp. Res. B: Methodol.,

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 4038 - 4048
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0522

vol. 18, no. 1, pp. 1-11, Feb. 1984.

[14]. Y. Kamarianakis and P. Prastacos, "Space-time modeling of traffic flow," Comput. Geosci., vol. 31, no. 2, pp. 119-133, Mar. 2005.

[15]. L. Vanajakshi and L. R. Rilett, "A comparison of the performance of artificial neural networks and support vector machines for the prediction of traffic speed," in Proc. IEEE Intell. Vehicles Symp., Parma, Italy, Jun. 2004, pp. 194-199.

[16]. M. Castro-Neto, Y. S. Jeong, M. K. Jeong, and L. D. Han, "Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions," Expert Syst. Appl., vol. 36, no. 3, pp. 6164-6173, Apr. 2009.

[17]. F. Xu, H. He, Z. Shen, and W. Zhang, "Random Forest for short-term traffic volume prediction in urban arterials," J. Adv. Transp., vol. 2017, pp. 1-10, 2017.

[18]. G. Leshem and Y. Ritov, "Traffic flow prediction using adaboost algorithm with random forests as a weak learner," in Proc. World Acad. Sci. Eng. Technol., vol. 19, 2007, pp. 193-198.

[19]. W. Zheng, D. H. Lee, and Q. Shi, "Short-term freeway traffic flow prediction: Bayesian combined neural network approach," J. Transp. Eng., vol. 132, no. 2, pp. 114-121, Feb. 2006.

[20]. 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," Transp. Res. C: Emerg. Technol., vol. 54, pp. 187-197, May 2015.

[21]. R. Fu, Z. Zhang, and L. Li, "Using LSTM and GRU neural network methods for traffic flow prediction," in Proc. 31st Youth Acad. Annu. Conf. Chin. Assoc. Autom., Xi'an, China, Nov. 2016, pp. 324-328.

[22]. S. Xingjian, Z. Chen, H. Wang, D. Y. Yeung, W. K. Wong, and W. C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in Adv. Neural Inf. Process. Syst., vol. 28, 2015, pp. 802-810.

[23]. B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," in Proc. 27th Int. Joint Conf. Artif. Intell. (IJCAI), Stockholm, Sweden, Jul. 2018, pp. 3634-3640.

[24]. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Adv. Neural Inf. Process. Syst., vol. 30, 2017, pp. 5998-6008.

[25]. Z. Wu, S. Pan, G. Long, J. Jiang, X. Chang, and C. Zhang, "Connecting the dots: Multivariate time series forecasting with graph neural networks," in Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., Virtual, Aug. 2020, pp. 753-763.