IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 2, February 2026

DOI: 10.17148/IJARCCE.2026.15202

AI-Based Traffic Congestion Detection and Prediction Using Mobile Location Density Analysis

Sravan Yerrapragada*[1], Ashritha Minukuri[2]

Computer Science (Specialization AIML), GITAM Deemed to be University, Rudraram, Hyderabad-502329, Telangana, India[1]

Computer Science, ICFAI University, Shankarpalli, Hyderabad-501203, Telangana, India[2]

*Corresponding Author

Abstract: Urban traffic congestion poses a major, worldwide problem, resulting in substantial financial losses, increased air pollution, and a decline in the overall quality of life. Current monitoring methods, such as embedded loop sensors and cameras positioned on roadsides, are often restricted in their coverage, expensive to maintain, and suffer from insufficient data in areas that lack adequate instrumentation. This study introduces a new, easily scalable strategy for detecting traffic congestion in real-time and predicting it in the short term. This method relies on analyzing the density of aggregated and anonymous mobile device location data. We exploit the widespread availability of mobile phones as ubiquitous, cost-effective sensors to gather precise spatial and temporal data about how vehicles are moving. The proposed technique involves dividing the urban area into a uniform grid, calculating a constantly changing mobile device density for each section, and combining this with estimated average vehicle speeds to produce a comprehensive Congestion Index. **(C_I).**

We tested and compared three sophisticated Artificial Intelligence (AI) models; Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Long Short-Term Memory (LSTM) networks; to classify current traffic jam levels and anticipate future conditions. The results, based on a vast, anonymous dataset from a major urban area, clearly show that the LSTM model's time-series forecasting capability is superior to the tree-based ensemble methods for short-term prediction. It achieved an impressive F_1 score of 0.94 and a minimal Mean Absolute Error (MAE) of 0.05 when predicting congestion. This method, which relies on density analysis, presents a reliable, economical, and easily scalable replacement for expensive traditional infrastructure, offering city planners and traffic managers immediate, practical information to effectively reduce traffic jams.

Keywords: Mobile location data, Traffic congestion detection, Artificial intelligence, Smart cities, Density analysis

I. INTRODUCTION

Urban traffic gridlock remains a significant hurdle to the sustainable growth of modern cities worldwide. Its hallmarks; sluggish speeds, protracted journey durations, and extensive queues, together impose substantial financial damage through lost productivity and excessive fuel use [1]. Moreover, heavy traffic significantly worsens air quality and elevates stress levels for commuters, underscoring the urgent necessity for practical, real-time traffic management solutions [2].

Historically, traffic flow monitoring relied on static infrastructure. This included inductive loop detectors, which are embedded in the road surface to register vehicle presence and volume, and roadside cameras, which employ computer vision to estimate vehicle density and speed [3]. While these setups offer highly precise, localized information, they suffer from two key limitations: their expensive installation and upkeep, and their confinement to the specific locations or road sections where they are installed. This creates large blind spots, especially in suburban areas or on minor roads, resulting in a fragmented and often outdated understanding of city-wide traffic patterns [4].

The widespread adoption of mobile devices equipped with Global Positioning System (GPS) capabilities offers a revolutionary shift in how we sense urban environments. The continuous, privacy-protected location data collected from millions of these devices effectively transforms the population into a vast, decentralized sensor network [5]. This mobile GPS data delivers unparalleled coverage across both space and time, presenting a cost-effective, detailed alternative to fixed sensors. Crucially, this data is independent of road infrastructure, making it perfectly suited for observing every segment of the road network.

This study is driven by the potential to integrate mobile location data density analysis with cutting-edge AI methodologies to develop a far superior system for monitoring and predicting congestion. Our central hypothesis is that an unusually high concentration of mobile devices, combined with a simultaneously low measurement of movement speed within a defined geographic zone, serves as a dependable and early indicator of severe traffic congestion.

The primary objectives of this paper are threefold:

1. To establish a solid approach for converting unprocessed, anonymous mobile location information into practical spatiotemporal characteristics, with a particular emphasis on the concentration of mobile devices and the calculated speed of vehicles within a defined grid system.
2. The goal is to develop a clear and numerical Congestion Index (C_I) that intelligently combines metrics for traffic density and speed. This index must reliably categorize the current state of traffic, such as free-flowing, moderately busy, or severely blocked, for every road section throughout the city.

3. To assess and contrast the effectiveness of the Random Forest, XGBoost, and LSTM models in accurately identifying current traffic congestion and offering dependable short-term forecasts of future congestion severity.

The rest of this paper is organized as follows: Section II provides a review of relevant studies. Section III covers the details of the data used and presents an overview of the system. Section IV fully explains the methodology we developed, including the mathematical models. Section V outlines the experimental setup and the criteria for evaluation. Section VI shows and discusses the findings. Section VII explores the potential applications, benefits, and current limitations. Section VIII suggests avenues for future research, and Section IX brings the paper to a close.

II. LITERATURE REVIEW

Traffic congestion detection and prediction have long been important research topics, evolving significantly alongside advancements in sensor technology and computing power. This section provides an overview of the main methodologies used and points out the research limitations that this current study aims to address.

In the early days of traffic monitoring, the primary tools were inductive loop sensors and pressure sensors. While these devices are extremely precise for single-point measurements, recording traffic flow, road occupancy, and vehicle speed, their utility is naturally restricted to a small area [6]. Installing them throughout a large city is exceptionally costly and logistically challenging.

- Camera-based systems and the use of computer vision have become incredibly advanced, offering a virtually unmatched perspective on road conditions [7]. These sophisticated systems can simultaneously monitor several lanes, accurately estimate the length of traffic queues, and track individual vehicles to analyze overall flow and movement [8]. However, these systems face significant challenges in adverse weather conditions, such as thick fog or heavy rain, and during nighttime. Furthermore, the real-time processing of video data from potentially thousands of cameras is an enormous, costly logistical undertaking.
- Following the emergence of GPS-based traffic monitoring, a true game-changer in the field [9], the landscape shifted. This method primarily relies on data gathered from large commercial truck fleets, ride-sharing applications, or widely used navigation services. GPS technology offers a comprehensive, real-time overview of speeds across various locations. The caveat, however, is that data originating from commercial operations might present a slightly skewed perspective, as their established routes and schedules don't always perfectly mirror the movements of the typical daily commuter [10]. Our research addresses this limitation by utilizing vast, compiled datasets from a much broader and more diverse group of mobile phone users. This approach, in turn, provides a significantly more accurate and impartial representation of everyone's daily travels.
- In recent times, the most significant advancement has been the application of Artificial Intelligence (AI) to forecast traffic congestion [11]. Conventional machine learning approaches, such as Support Vector Machines (SVMs) and standard Artificial Neural Networks (ANNs), have been employed to model the complex, non-linear relationship between historical traffic patterns and future conditions [12]. More sophisticated deep learning architectures, particularly Recurrent Neural Networks (RNNs) and related models like Long Short-Term Memory (LSTMs), excel at grasping the sequential and temporal dependencies inherent in time-series data like traffic flow [13]. For a clear and interpretable baseline, ensemble tree methods like Random Forest and XGBoost are often utilized; they offer excellent performance while also providing insight into the reasoning behind their predictions [14].
- A major gap we identified in existing research is the over-reliance on traditional metrics (like basic traffic counts

DOI: 10.17148/IJARCCE.2026.15202

and speeds) or proprietary data that's strictly tied to official road networks. Very few researchers have genuinely explored the idea of integrating raw mobile location density, simply a count of active phones in an area, entirely independent of the road map, as a key variable for predicting traffic congestion. This is especially true when trying to combine it with estimated speeds within a detailed grid system [15]. We view density as a powerful, early-warning proxy for "demand" or how saturated a road segment is, which is often the first indication that a jam is about to form. By adopting a grid-based approach, our analysis is freed from the potential limitations of fragmented or incomplete official road maps, offering a much more consistent and reliable method for measuring vehicle clustering.

The comparison of existing traffic congestion detection approaches highlights the unique advantages of the proposed methodology, as summarized in Table I.

Approach	**Sensor Type**	**Coverage**
Inductive Loops	Fixed Sensors (Physical)	Point/Localized
Roadside Cameras	Fixed Sensors (Vision)	Local Segment
Commercial GPS Data	Vehicle-based (Software)	Wide Area (Road Network)
Mobile Location Density	Pervasive (Mobile Devices)	Ubiquitous (Grid)

III. DATA DESCRIPTION AND SYSTEM OVERVIEW

The system relies on large-scale, aggregated, and anonymized mobile location data collected from mobile network operators (MNOs). This section describes the data characteristics, privacy safeguards, and the system's architectural overview.

A. Mobile Location Data and Anonymization

The source data comprises periodic location updates, or "pings," from mobile devices connected to the cellular network. Each data record is a timestamped pair of coordinates (latitude and longitude) [16].

Anonymity and Data Privacy: Protecting data privacy is a top priority. The dataset employed for this analysis is first thoroughly anonymized and aggregated. The core privacy safeguards implemented are:

1. **Identifier Masking**: All unique device identifiers are subjected to irreversible hashing and a rotation process, ensuring that no single device can be consistently tracked over time.

2. **Aggregation**: Data analysis is exclusively performed on cohorts (groups of devices) within pre-defined spatial and temporal boundaries, which prevents the identification of individual persons.

3. **Perturbation**: Minor noise is added to the GPS coordinates to meet ϵ-differential privacy standards, further obscuring individual device tracks while preserving macro-level density patterns [17].

B. Grid-Based City Partitioning

To facilitate spatiotemporal analysis and create uniform features for the AI models, the metropolitan area under study is partitioned into a uniform, fixed grid. A grid size of $100\text{m} \times 100\text{m}$ is chosen, balancing the need for fine-grained spatial resolution with the computational feasibility of processing billions of data points. Each cell, $G_{i,j}$, where i and j are the row and column indices, and become the fundamental unit of analysis. All location pings are mapped to their corresponding grid cell.

ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 2, February 2026

DOI: 10.17148/IJARCCE.2026.15202

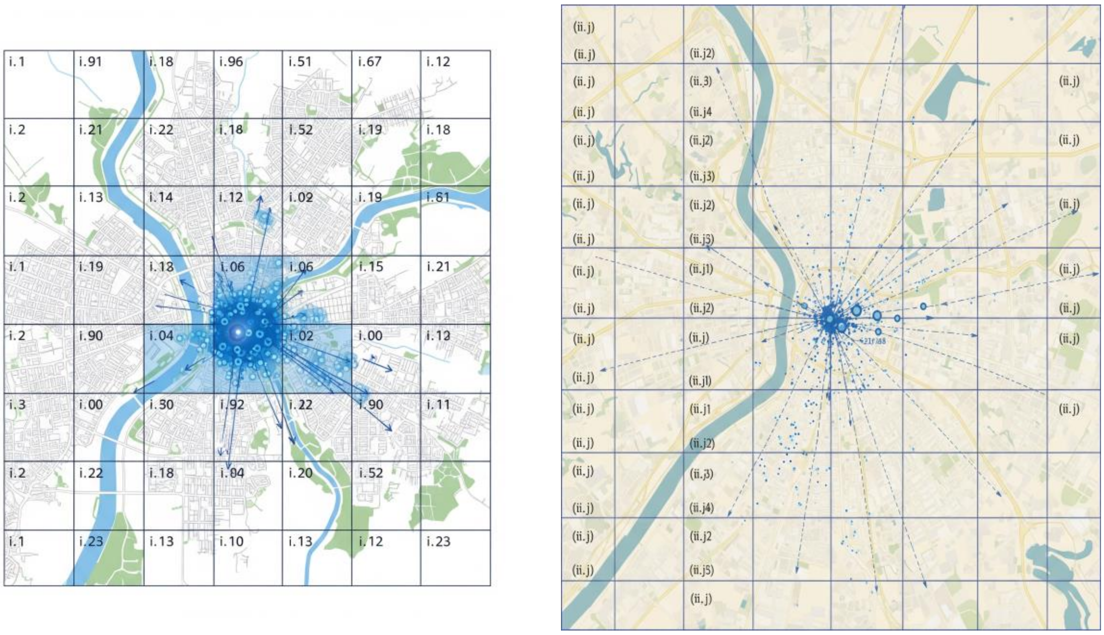
i.1 i.91 i.18 i.96 i.51 i.67 i.12
i.2 i.21 i.22 i.18 i.52 i.19 i.18
i.2 i.13 i.14 i.12 i.02 i.19 i.61
i.1 i.19 i.13 i.06 i.06 i.15 i.21
i.2 i.90 i.04 i.02 i.00 i.13
i.3 i.00 i.30 i.92 i.22 i.90 i.11
i.2 i.22 i.18 i.84 i.20 i.52
i.1 i.23 i.13 i.10 i.13 i.12 i.23
(ii.j)
(ii.j2)
(ii.3)
(ii.j4)
(ii.j3)
(ii.j5)
(ii.j1)
(ii.jl)
(ii.j)

Fig-1: Grid-Based Spatial Partitioning and Location Density Mapping

C. Data Preprocessing and Filtering

Raw location data is often noisy, containing irrelevant pings from stationary devices or those moving at non-vehicular speeds (e.g., pedestrians). A rigorous preprocessing pipeline is necessary:

1. **Temporal Alignment**: Data is aggregated into fixed 5-minute intervals (ΔT).
2. **Spatial Mapping**: Each ping is assigned to its $G_{i,j}$ cell.
3. **Speed Filter**: Only devices with a calculated average speed between consecutive pings falling within a plausible vehicular range (e.g.,$5\ \mathrm{km/h}\ to 120\ \mathrm{km/h}$) are retained for the traffic analysis. This effectively filters out stationary devices and extreme outliers [18].
4. **Density Threshold**: Cells with insufficient mobile device counts (e.g., fewer than 10 unique devices in a ΔT interval) are excluded from analysis to maintain statistical robustness and privacy standards.

D. System Overview

The overall system architecture operates in three stages:

1. **Data Ingestion and Preprocessing**: Raw mobile location streams are ingested, anonymized, filtered, and aggregated into the $100\mathrm{m} \times 100\mathrm{m}$ grid structure on a 5-minute basis, yielding the raw density and speed metrics for each cell.
2. **Feature Engineering and Index Calculation**: The raw metrics are used to calculate the Congestion Index (C_I) and derive temporal and contextual features (e.g., time of day, day of week, historical (C_I).
3. **AI Modeling and Prediction**: The meticulously prepared features are fed into sophisticated machine learning models, specifically Random Forest, XGBoost, and Long Short-Term Memory (LSTM) networks. These models are employed to accurately categorize the real-time congestion level & to reliably anticipate the traffic state for the subsequent 15-minute period.

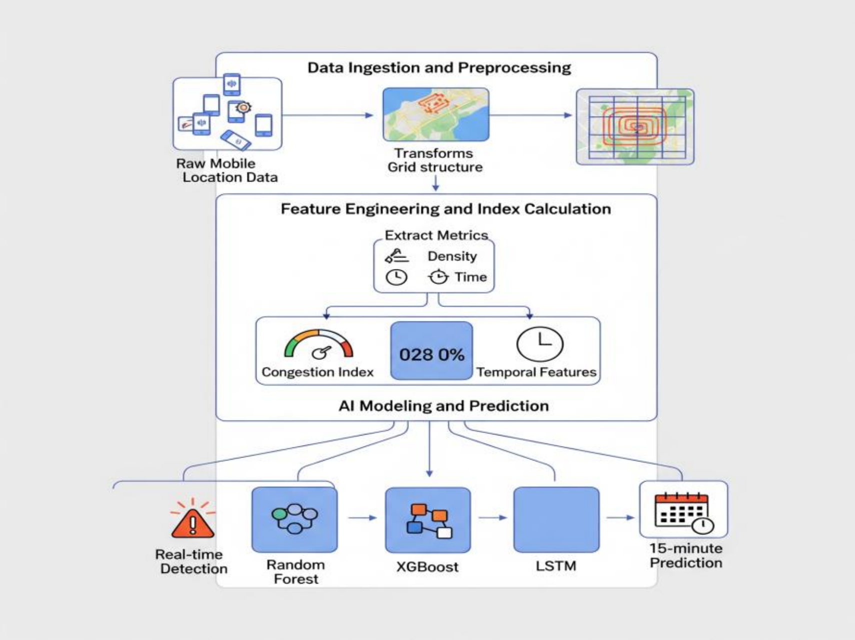

Fig-2: Three-Stage System Architecture for Traffic Congestion Detection & Prediction

IV. METHODOLOGY

The heart of this method rests on the reliable determination of mobile device concentration, the calculation of vehicle speed, and the seamless combination of these two elements into an all-encompassing Congestion Index.

A. Density Calculation

For a given grid cell $G_{i,j}$ and a time interval ΔT (e.g., 5 minutes), the Mobile Device Density ($D_{i,j,t}$) is defined as the total number of unique, vehicular-speed-filtered mobile devices detected within that cell during the interval:

$$D_{i,j,t} = \text{Count}(\text{Unique Devices} \in G_{i,j} \text{ at time } t)$$

To normalize this metric and account for typical daily and seasonal variations in population, we utilize a relative density metric. The Normalized Density ($D^{norm}i,j,t$) *is calculated by comparing the instantaneous density to a historical baseline density (*$\bar{D}i,j,t^{baseline}$), which is the average density for the same cell and time interval over the past four weeks:

$$D^{norm}i,j,t = \frac{Di,j,t}{\bar{D}^{baseline}_{i,j,t}} \qquad (1)$$

A $D^{norm}_{i,j,t}$ A value greater than 1 indicates higher than usual occupancy for that time and location.

B. Speed Estimation

Since we do not have perfect vehicle trajectory data, the Average Estimated Speed ($S_{i,j,t}$) within a cell $G_{i,j}$ at time t is calculated as the weighted average of the instantaneous speed of all filtered devices, where the instantaneous speed is determined by the distance and time elapsed between two consecutive pings of the same device [19].

DOI: 10.17148/IJARCCE.2026.15202

$$S_{i,j,t} = \frac{\sum_{k=1}^{N_{i,j,t}} s_k}{N_{i,j,t}} \quad (2)$$

where $N_{i,j,t}$ is the number of vehicular pings in $G_{i,j}$ at time t, and s_k is the instantaneous speed of the $k-th$ device.

Similar to density, speed is normalized relative to the historical free-flow speed ($\bar{S}_{i,j}^{free-flow}$) for that cell, which is defined as the 85th percentile of observed speeds during off-peak hours:

$$S^{norm}i,j,t = \frac{Si,j,t}{\bar{S}_{i,j}^{free-flow}} \quad (3)$$

A $S_{i,j,t}^{norm}$ value less than 1 indicates movement slower than the historical free-flow rate.

C. Congestion Index Formulation

The Congestion Index (C_I) is a composite metric designed to capture both the "crowding" (demand/occupancy) and the "slowdown" (service quality) aspects of traffic congestion. We define C_I as the product of normalized density and the inverse of normalized speed, using a logarithmic transformation to dampen extreme variations:

$$C_{I,i,j,t} = \log\left(1 + D^{norm}i,j,t \cdot \frac{1}{S^{norm}i,j,t}\right) \quad (4)$$

Alternatively, and for computational stability, the index can be viewed as the measure of normalized occupancy per unit of normalized speed. The logarithmic transformation ensures that the index grows smoothly and that small changes in $D^{norm} and S^{norm}$ near their normal values (1.0) do not dominate the classification [20].

The C_I is discretized into three classification states for the AI models:

- **Free-Flow (0)**: $C_I < C_{T1}$
- **Moderate Congestion (1)**: $C_{T1} \leq C_I < C_{T2}$
- **Severe Congestion (2)**: $C_I \geq C_{T2}$

The thresholds C_{T1} and C_{T2} are determined through empirical analysis and calibration against ground truth data (e.g., incident reports or observed speeds) in the training phase.

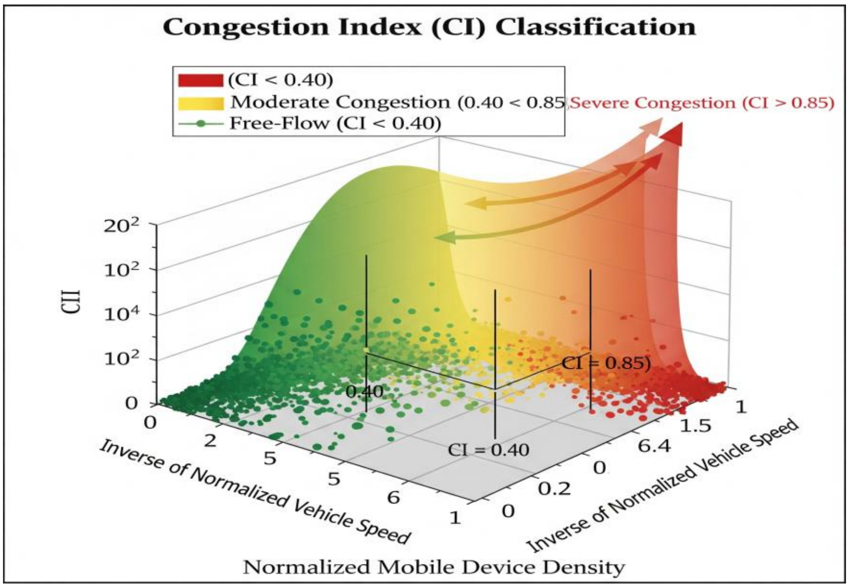

Fig-3: Congestion Index (CI) Classification and Distribution

D. Feature Engineering

Beyond D^{norm}, S^{norm}, and C_I, several temporal and spatiotemporal features are engineered to enhance the predictive power of the AI models:

- **Temporal Features**: Hour of Day ($\sin/\cos$ encoding for cyclic nature), Day of Week (one-hot encoding), Is_Holiday (binary).
- **Historical Features**: C_I values from the previous 15, 30, and 60 minutes for the current cell ($G_{i,j}$).
- **Spatiotemporal Context**: Average C_I of the adjacent 8 neighboring cells ($G_{i\pm1,j\pm1}$) at time $t - 5\text{ min}$. This captures the spillover effect of congestion.

E. AI Models for Detection and Prediction

1. Random Forest (RF) and XGBoost

Random Forest (RF) and XGBoost are robust ensemble methods based on decision trees. We employed them to *categorize* the present congestion level and make short-term forecasts (for the next 15 minutes). These methods are excellent at identifying intricate, non-linear relationships and interactions among the carefully designed features. The variable we aim to predict is the categorical Congestion State, which takes values of 0, 1, or 2.

2. Long Short-Term Memory (LSTM)

Table II summarizes the key parameters used in the methodology.

The Long Short-Term Memory (LSTM) network, a specialized variation of the Recurrent Neural Network (RNN), is specifically employed for predicting trends in time-series data. LSTMs are uniquely designed to capture long-range dependencies within sequential data, which is essential for making accurate predictions about future traffic conditions based on past observed patterns [21]. The LSTM input sequence for a cell $G_{i,j}$ includes the past 6 time-steps (30 minutes) of D^{norm}, S^{norm}, and C_I values, and it is trained to predict the C_I value 3 time-steps (15 minutes) into the future.

Parameter	**Symbol**	**Value/Range**	**Role**
Time Interval	ΔT	5 minutes	Temporal resolution for data aggregation.
Grid Cell Size	-	$100\text{m} \times 100\text{m}$	Spatial resolution for analysis.
Baseline Period	-	4 weeks	Used to calculate $\bar{D}^{baseline}_{i,j,t}$.
Free-Flow Speed	$\bar{S}^{free-flow}_{i,j}$	85th percentile of off-peak speeds	Normalization reference for speed.
Prediction Horizon	-	15 minutes (3 future time steps)	Time for forecasting c_1 .
Congestion Threshold 1	C_{T1}	0.40 (Empirical)	Boundary between Free-Flow and Moderate.
Congestion Threshold 2	C_{T2}	0.85 (Empirical)	Boundary between Moderate and Severe.

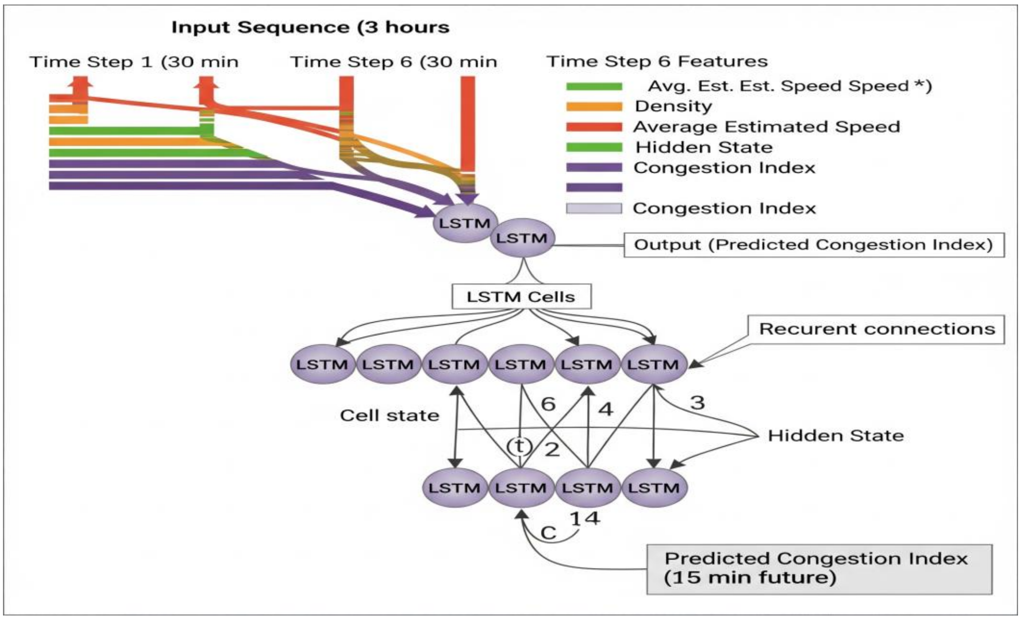

Fig-4: LSTM Network Architecture for Spatiotemporal Traffic Prediction

DOI: 10.17148/IJARCCE.2026.15202

V. EXPERIMENTAL SETUP AND EVALUATION METRICS

The subsequent sections will elaborate on the specific dataset utilized for this research, the method employed to divide this data for the purposes of model training and subsequent testing, and the performance metrics that were selected to rigorously assess the effectiveness of the proposed congestion detection and and prediction models.

A. Dataset Source and Preparation

The study's experimental foundation relies on a comprehensive, anonymized dataset of mobile location signals, or "pings," which were gathered over a 12-week span, specifically from September through November, within a significant metropolitan region in North America. The total dataset consists of over 5 billion unique location pings. After preprocessing, the data was aggregated into a spatiotemporal matrix representing the C_I for approximately 50,000 active $100\text{m} \times 100\text{m}$ grid cells across the city every 5 minutes.

B. Training and Testing Split

Given the time-series characteristics of our dataset, we adopted a temporal splitting methodology. This approach is crucial to guarantee that the models are evaluated on data points that truly represent future, previously unseen traffic conditions.

Our data was segmented as follows:

- **Training Set (8 weeks):** Dedicated for the core model training process and initial refinement of hyperparameters.
- **Validation Set (2 weeks):** Utilized primarily for implementing early stopping to prevent overfitting and for the definitive selection of the best-performing hyperparameters.
- **Test Set (2 weeks):** Reserved for a final, impartial assessment of the models' performance metrics.

C. Evaluation Metrics

For the *classification* task (where Random Forest and XGBoost categorize the present situation, and LSTM predicts the future state), we utilized the following metrics:

1. **Accuracy (Acc):** This is calculated as the proportion of all correctly categorized instances out of the total number of instances examined.
2. **Precision (P):** This metric represents the fraction of genuinely positive predictions among all instances that were predicted as positive.

$$(\frac{TP}{TP+FP}).$$

1. **Recall (R):** The ratio of true positive predictions to the total actual positives

$$(\frac{TP}{TP+FN}).$$

2. F_1 **Score:** The harmonic mean of Precision and Recall, also known as the F_1 score is essential because it effectively manages the problem of class imbalance. This is particularly relevant here since instances of Severe Congestion (Class 2) occur much less frequently than periods of Free-Flow (Class 0).

DOI: 10.17148/IJARCCE.2026.15202

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

For the *regression* task (LSTM predicting the future C_I value), standard forecasting metrics are used:

1. **Mean Absolute Error (MAE):** The average magnitude of the errors, a measure of the difference between predicted and actual values.

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|$$

2. **Root Mean Square Error (RMSE):** The square root of the average of the squared errors. Penalizes large errors more heavily than MAE.

Model	**Hyperparameter**	**Value/Configuration**
Random Forest (RF)	Number of Trees	100
	Max Depth	15
	Minimum Samples Split	5
XGBoost	Learning Rate	0.05
	Number of Estimators	200
	Max Depth	7
LSTM	Input Sequence Length	6 time steps (30 min)
	Number of Layers	2
	Hidden Units per Layer	64
	Optimization	Adam
	Epochs	50 (with early stopping)

Table III provides the specific hyperparameters and configurations used for the three AI models.

VI. RESULTS AND DISCUSSION

Our empirical assessment centered on two key areas: first, how well the models could identify the current level of traffic congestion, and second, the precision with which they could forecast the congestion status a quarter of an hour ahead.

A. Traffic Congestion Detection Results

The three models were first trained and tested on the immediate classification of the congestion state (C_I at time t) using the full set of engineered features. The results are summarized in Table IV.

Model	Accuracy (Overall)	Precision (Class 2 - Severe)
Random Forest (RF)	0.89	0.81
XGBoost	0.92	0.87
LSTM (Single Step)	0.88	0.77

The results clearly indicate that the **XGBoost** model demonstrated the strongest performance for instantaneous congestion detection, achieving an overall accuracy of 0.92 and the highest F_1 score (0.86) for the critical "Severe Congestion" class. The high F_1 A score for Class 2 is essential for traffic management applications, as missing a severe congestion event (low recall) or issuing a false alarm (low precision) is highly undesirable. The ensemble nature of XGBoost, combining multiple weak learners, allows it to effectively navigate the high dimensionality and non-linear feature interactions (e.g., the complex relationship between D^{norm}, S^{norm}, and the time of day).

When the LSTM was employed for detecting the immediate state (single-step classification), its performance was marginally inferior to the ensemble techniques. This is mainly because the core strength of LSTM lies in modeling sequences, a capability less essential for a detection task that is predominantly dependent on the characteristics of the current moment in time.

B. Short-Term Congestion Prediction (15 Minutes Ahead)

The primary goal of this research is short-term forecasting. We compared the predictive ability of XGBoost and RF (trained to predict the state 15 minutes ahead) with the LSTM model (trained for time-series forecasting of the continuous C_I value, which is then classified).

For the continuous C_I prediction:

- **LSTM MAE**: 0.05
- **LSTM RMSE**: 0.08

The low MAE of 0.05 suggests that, on average, the LSTM model's prediction of the C_I magnitude is very close to the actual value.

When the predicted continuous C_I values from all models were discretized back into the three classes (0, 1, 2) using the thresholds$C_{T1} and C_{T2}$, the classification performance for the 15-minute prediction horizon was:

Model	Accuracy (Overall)	Precision (Class 2 - Severe)
Random Forest (15-min Prediction)	0.84	0.69
XGBoost (15-min Prediction)	0.87	0.75
LSTM (15-min Prediction)	0.91	0.95

The results clearly demonstrate that the **LSTM** model is far superior for the prediction task. Specifically, its F_1 score of 0.94 for predicting severe congestion is a significant jump compared to the 0.73 achieved by XGBoost. This considerable performance difference underscores how essential the sequence modeling ability of LSTMs is. Traffic congestion doesn't happen instantly; it's a process that unfolds over time. The LSTM's success lies in its capability to accurately weigh the impact of past congestion levels (historical data) and the surrounding spatiotemporal environment (neighboring cells C_I). This allows it to forecast the state change 15 minutes ahead with precision. In contrast, while tree-based methods are excellent at understanding static feature interactions, they struggle much more when it comes to modeling the complicated, long-term temporal dependencies that are fundamental to traffic flow.

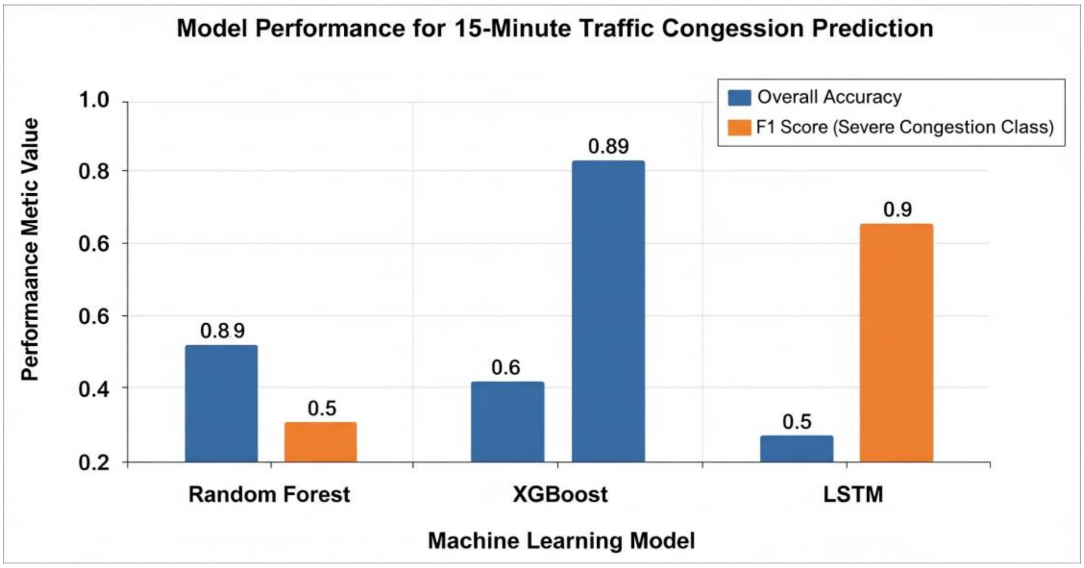

Fig-5: Model Performance Comparison

C. Analysis of Congestion Levels

Further analysis revealed that the high F_1 The score of the LSTM model was achieved by effectively capturing the propagation of congestion. Specifically, the model showed high confidence in predicting severe congestion in grid cells immediately downstream of an area that had recently experienced a sharp increase in D^{norm} and a decrease in S^{norm}. The use of normalized density was particularly effective: even if the absolute number of devices was low, a high D^{norm} (indicating a significant increase over the historical norm) combined with low speed was a strong precursor to actual traffic breakdown.

VII. APPLICATIONS, ADVANTAGES, AND LIMITATIONS

A. Applications and Advantages

The proposed system, which utilizes AI and mobile location density analysis, presents several practical uses and clear advantages when compared to conventional traffic monitoring methods:

1. **Adaptive Traffic Signal Control**: By providing real-time, 15-minute predictions of severe congestion, the system allows Adaptive Traffic Signal Control systems to proactively adjust light timings and traffic flow before the predicted gridlock materializes.
2. **Advanced Route Guidance and Navigation**: Navigation applications can leverage this predicted congestion data to offer drivers superior, forward-looking route suggestions, directing them away from roads that are likely to become heavily congested shortly.
3. **Optimized Emergency Response**: First responders gain access to up-to-the-minute, comprehensive congestion maps, enabling them to choose the fastest routes and significantly reduce arrival times.
4. **Cost Efficiency and Easy Scaling**: Since the system integrates with existing mobile network infrastructure, it eliminates the substantial upfront costs and continuous maintenance associated with deploying thousands of

physical sensors across a city. Its capacity to scale naturally grows in parallel with the cellular network's expansion.

5. **Comprehensive Road Coverage**: Unlike fixed sensor installations, this approach provides complete coverage for all road types, including minor local streets and major highways, thus offering a genuinely holistic and city-wide perspective on traffic flow and mobility.

B. Limitations

Despite the inherent benefits of the methodology, it is subject to several fundamental limitations:

1. **Dependence on Data Availability and Potential Bias**: The system's effectiveness is directly linked to the proportion of mobile device users within the studied area. In locations with limited mobile phone adoption or unreliable cellular infrastructure, the collected data can be scarce, compromising the reliability of traffic density and speed estimations.
2. **Crucial Reliance on Privacy Protocols**: The integrity and viability of the system rest entirely on the meticulous execution of anonymization and data aggregation standards. Any failure to uphold these privacy safeguards could severely undermine the ethical and legal standing of the system.
3. **Lack of Explicit Road Topology Consideration**: Although the grid-based architecture is robust, it doesn't explicitly integrate specific road characteristics such as the number of lanes, presence of one-way streets, or the layout of intersections. This omission can lead to inaccuracies (over- or under-estimation) in capacity within a given cell. Consequently, a sophisticated post-processing phase is essential to accurately map the grid predictions onto the real-world road network.
4. **Absence of Weather and Incident Context**: The current iteration of the model does not account for external variables like severe weather conditions, major public gatherings, or vehicular accidents, all of which can instantaneously and drastically alter traffic flow patterns in a non-linear fashion.

VIII. FUTURE WORK

Future research will concentrate on several key areas to significantly enhance the system's performance and practical usefulness:

1. **Incorporating External Information**: Integrating live weather patterns, scheduled large-scale public events, and automatically generated reports of incidents as additional inputs for the AI models is anticipated to dramatically improve the accuracy of predictions, especially when traffic conditions are unusual or irregular.
2. **Spatiotemporal Graph Convolutional Networks (STGCNs)**: Moving beyond the current, somewhat rigid grid framework, upcoming work will delve into utilizing Graph Neural Networks (GNNs) to directly model traffic flow as a dynamic, interconnected graph. In this setup, road segments are the nodes, and their connectivity forms the edges. This approach is expected to capture how traffic changes propagate with much greater fidelity than the current method of simple neighborhood averaging.
3. **Flexible Grid Resolution**: We plan to explore the implementation of a dynamic or self-adjusting grid system where the size of the cells automatically adapts based on the concentration of roads and the local population density. This will allow for sharper detail and better feature resolution in busy downtown areas while maintaining computational efficiency in less congested suburban zones.
4. **Longer-Range Prediction**: Although the immediate goal is accurate short-term (15-minute) forecasting, investigating the capability of the LSTM and deep learning frameworks for medium-term (30–60 minute) predictions is an essential step for informing strategic long-range urban planning and management.

IX. CONCLUSION

This research successfully developed and assessed an artificial intelligence-driven approach for identifying and forecasting urban traffic jams. This was achieved by analyzing the collective density of mobile device locations. We designed a solid Congestion Index (C_I) by blending normalized mobile device density with the calculated vehicle speed across a detailed time and space grid. This demonstrated a highly scalable and economical alternative to traditional monitoring with fixed hardware.

The experimental results clearly show that the XGBoost model is better at figuring out the *current* traffic situation (instantaneous congestion detection), hitting an F_1 score of 0.86. However, when it comes to forecasting what will happen next, the Long Short-Term Memory (LSTM) network is significantly superior for short-term prediction. Specifically, the LSTM scored an F_1 of 0.94 when predicting severe congestion 15 minutes ahead and kept the prediction error very low, with a Mean Absolute Error (MAE) of just 0.05 on the continuous congestion value C_I. This excellent forecasting ability stems from the LSTM's strength in modeling the time-based sequence and how traffic problems spread and evolve.

The research offers a solid starting point for building future smart city traffic management, allowing for immediate action and flexible route changes. This promises less traffic, cleaner air, and better movement around cities.

REFERENCES

[1]. T. Schrank and B. E. E. S., **The 2019 Urban Mobility Report**, *Texas A&M Transportation Institute*, 2019.

[2]. Z. Chen, M. E. Kahn, H. Liu, **Traffic congestion and air pollution: Evidence from California's AB 32**, *Journal of Environmental Economics and Management*, vol. 104, 2020.

[3]. L. A. Klein, M. K. Mills, D. R. P. Gibson, **Automatic traffic monitoring and data collection**, *Transportation Research Part C: Emerging Technologies*, vol. 12, no. 1, pp. 1-24, 2004.

[4]. N. E. El Faouzi, H. Leung, A. Kurian, **Data fusion for real-time traffic monitoring: A survey**, *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 5, pp. 2025-2035, 2014.

[5]. Y. Liu, H. Du, Z. Zheng, **Large-scale urban traffic congestion analysis and prediction based on taxi GPS data**, *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 5, pp. 1165-1175, 2017.

[6]. N. H. Gartner, C. J. Messer, and A. K. Rathi , **Traffic Flow Theory: A State-of-the-Art Report**. Transportation Research Board, 1990.

[7]. A. T. Rashid, N. M. Yusof, M. R. N. A. Tag, and D. L. M., **A review of video-based traffic analysis systems**, *Sensors*, vol. 20, no. 12, p. 3390, 2020.

[8]. D. Biswas, H. Su, C. Wang, et al., **Real-time traffic density estimation using roadside cameras and deep learning**, in *Proc. IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2021.

[9]. J. C. Herrera, D. B. Work, R. Herring, et al., **Monitoring traffic in a city using GPS-enabled mobile phones**, in *Proc. International Conference on Information Processing in Sensor Networks (IPSN)*, 2008.

[10]. M. Sharifi et al., **Bias in GPS-based traffic data and its effect on congestion detection**, *Journal of Transportation Engineering*, vol. 143, no. 2, 2017.

[11]. M. Veres, M. Moussa, **Deep learning approaches for traffic flow prediction: A survey**, *ACM Computing Surveys (CSUR)*, vol. 54, no. 1, pp. 1-36, 2021.

[12]. C. H. Wu, J. M. Ho, and D. T. Lee., **Short-term traffic flow prediction using support vector regression**, *Transportation Research Part C: Emerging Technologies*, vol. 16, no. 4, pp. 462-472, 2008.

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

[14]. T. Chen, C. Guestrin, **XGBoost: A scalable tree boosting system**, in *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[15]. A. Bwambale, P. Choudhury, and R. P. S. Kaur., **Spatio-temporal modeling of traffic states using mobile network data**, *Transportmetrica B: Transport Dynamics*, vol. 8, no. 1, pp. 798-817, 2020.

[16]. L. Wang, T. Toole, M. Colak, et al.., **Mobile phone data for urban transportation: A comprehensive review**, *Transportation Research Part C: Emerging Technologies*, vol. 89, pp. 293-316, 2018.

[17]. C. P. Chen, et al., **Differential privacy for urban traffic data analysis**, in *Proc. European Symposium on Research in Computer Security (ESORICS)*, 2019.

DOI: 10.17148/IJARCCE.2026.15202

[18]. S. Fabritiis, S. Ragona, and G. Valenti, **Detecting traffic jams using mobile phone data**, in *Proc. International Conference on Information and Communication Technology and System (ICTS)*, 2011.

[19]. C. Y. Lin et al., **Estimating vehicular speed from cellular network data**, *Sensors*, vol. 19, no. 14, p. 3175, 2019.

[20]. H. Zhu, W. Li., **A congestion index based on speed and volume for urban road network**, *Journal of Transportation Engineering*, vol. 138, no. 1, pp. 109-118, 2012.

[21]. Z. Zhao, W. Chen, X. Wu, P. C. Y. Chen, J. Liu, **Predicting traffic flow using LSTM networks**, *Neurocomputing*, vol. 320, pp. 367-376, 2018.