An Integrated AI Approach for Real-Time Traffic Forecasting and Congestion Management in Intelligent IoT-Based Networks

Israa Mishkal[1,3], Dheyab Salman Ibrahim[1], Saja Salim Mohammed[1], Hassan Hadi Saleh[1], Taha Mohammed Hasan[1], Ahmed Latif Yasir[2], Ahmed Abbas Brisam[4] and Rebeen Ali Hamad[5]

[1]*Department of Computer Sceince, Sceince Collage, University of Diyala, 32001 Baáqubah, Iraq*

[2]*University of Baghdad, College of Administration and Economics, Statistics Department,1001 Baghdad,Iraq*

[3]*Computer Science Department, Universiti Sains Malaysia (USM),* 11800 Penang*, Malaysia*

[4] *Departments of Mathematical Sciences, Fulbright College of Arts and Sciences, University of Arkansas, 72701 Fayetteville, USA*

[5]*Translational and Clinical Research Institute, Faculty of Medical Sciences, Newcastle University, NE17RU Newcastle upon Tyne, United Kingdom*

israa_adnan85@student.usm.my, saja.salim@udiyala.edu.iq, hassan.hadi@uodiyala.edu.iq, dr.tahamh@uodiyala.edu.iq, ahmed.l@coadec.uobaghdad.edu.iq, abrisam@uark.edu, rebeen.hamad@newcastle.ac.uk

Keywords: BERT, Traffic Prediction, IoT Networks, Congestion Control, Intelligent Transportation Systems (ITS).

Abstract: Traffic congestion remains an unresolved problem in today's urban transportation, particularly in smart infrastructure contexts with the requirement of accurate and timely traffic forecasting to augment realistic congestion improvement. Traditional forecast models such as Auto Regressive Integrated Moving Average (ARIMA), SVR, or shallow neural networks do not grasp the non-linear, spatiotemporal dynamics of real-time traffic and hence lead to inferior route decisions and increased congestion. In this study, a new hybrid artificial intelligence (AI) approach combining Bidirectional Encoder Representations from Transformers (BERT) and Convolutional Long Short-Term Memory (ConvLSTM) is presented to improve the accuracy of traffic prediction and enable proactive congestion control in intelligent Internet of Things (IoT) networks. The architecture has four steps: input representation via real-time sensor data, contextual embedding via BERT to capture temporal and semantic patterns, spatiotemporal modeling via ConvLSTM, and decision generation via reinforcement-based routing adaptation. The model was trained and evaluated using the METR-LA dataset, a real-world spatiotemporal traffic dataset for Los Angeles, and demonstrated improved performance compared to baseline models. Exactly, the proposed model registered a Mean Absolute Error (MAE) of 2.13, a Root Mean Square Error (RMSE) of 3.54, and an R^2 Score of 0.91, beating LSTM-only, CNN-LSTM, and Transformer-only models. The outcomes confirm the effectiveness of the integrated deep contextual learning and spatiotemporal memory in predicting traffic. The proposed system offers a scalable and sensible paradigm for future traffic management in smart cities.

1 INTRODUCTION

The increasing rate of urbanization and population density of modern cities has placed enormous stress on transport infrastructures, resulting in repeated congestion, trip delay, and inefficient traffic flow. Modern urban transport systems are now required to deal with highly dynamic and sometimes unpredictable traffic stream patterns influenced by an extremely broad range of factors such as time of day, weather, road crashes, and driver habit adjustments. This growing complexity demands the development of smart and intelligent traffic management systems to deliver efficient and sustainable mobility in the scenario of smart cities [1]. The conventional approach of traffic forecasting and congestion mitigation, statistical regression models and temporal methods like ARIMA, have met with limited success due to their fundamental inability to model non-linear relationships and long-term temporal relationships in traffic data [2]. Furthermore, they tend to be founded on stationarity assumptions of data, which never apply in real traffic situations with continuous and context-aware transformations [3]. To escape from such limitations, data-driven methods have gained more popularity since they can potentially leverage

real-time sensor data to generate low-latency and accurate predictions of traffic. In smart IoT-enabled transport systems, there is a continuous flow of heterogeneous data from loop detectors, GPS sensors, roadside cameras, and weather sensors. Multimodal feeds, if processed efficiently, contain precious temporal and spatial information that contain tremendous potential for significantly improving the accuracy of traffic predictions. However, integrating and interpreting the high-dimensional data in a consistent and meaningful fashion is still an extremely challenging problem [4]. This work proposes a hybrid congestion control and traffic forecasting model for intelligent IoT networks. The model utilizes a context-aware architecture that has learning of temporal patterns, spatial correlations, and environmental attributes without handcrafted or rule-based features. Both real-time traffic information and historical traffic data are utilized by the system to generate short-term forecasts in which proactive congestion control is permitted and traffic flow on key road segments is enhanced. In order to validate the effectiveness of the proposed approach, experiments were conducted on the METR-LA dataset. It is a widely used benchmark consisting of traffic speed readings from freeway sensors in Los Angeles County [5]. The hybrid model provides spectacular enhancement in prediction accuracy and responsiveness over traditional time-series models.

2 RELATED

Traffic prediction and traffic congestion management have been among the continuing ITS development challenges. In recent two decades, scholars have proposed an incredibly wide variety of modeling approaches ranging from the conventional statistical methods to contemporary data-oriented approaches. Conventional models liberally employed linear statistical techniques such as the ARIMA and its variant forms that accounted for seasonality in short term traffic prediction. While these methods offer simplicity and interpretability, they compromise on capturing the non-linearity's, unexpected changes, and context-dependent fluctuations of traffic data [6], [7]. In an effort to mitigate such limitations, machine learning (ML)methods such as Support Vector Regression (SVR), Decision Trees, and K-Nearest Neighbors (KNN) were investigated [8]-[10]. The models demonstrated uncertain accuracy gains, particularly on high dimensional feature sets of data. But they were tremendously reliant on human-maintained feature engineering and were not adaptive in real-world dynamic traffic situations [11], [12]. Deep learning (DL) has introduced remarkable progress to the field of traffic forecasting. Recurrent Neural Network (RNN) time series models and their extensions specifically Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) have been seen to hold out good promise with their ability to handle long-term temporal relationships in sequential data [13]. The LSTM networks have proven to excel especially in retaining temporal information. But these models do not take into consideration spatial relationships and out of domain context knowledge, which decreases their capacity to tackle complex real-world traffic problems [14]. In an attempt to capture spatial characteristics, hybrid architectures that combine Convolutional Neural Networks (CNNs) and LSTM models have been proposed. One of such architectures is the ConvLSTM, which can learn both spatial and temporal relationships concurrently [15]. While ConvLSTM addresses some of the challenges of modeling spatial-temporal dependencies, it is still not possible for it to integrate external context information such as weather conditions, social events, and temporal irregularities. Recent advances in sequence modeling particularly those emanating from natural language processing (NLP), have introduced models with the capacity to learn bidirectional dependencies in data. Transformer models like BERT have shown remarkable ability to capture nonlinear and context-dense patterns in sequential data. Though applied to traffic forecasting yet in the infancy stage, preliminary research shows massive potential if the models are trained on precise IoT traffic datasets [16], [17]. Building on these developments, this study proposes a novel hybrid system integrating BERT based context representation and spatiotemporal analysis driven by ConvLSTM. This integration is poised to more effectively capture the dynamic, multi-dimensional nature of traffic systems and support responsive forecasting in smart IoT networked environments.

3 PROPOSED FRAMEWORKS FOR TRAFFIC PREDICTION AND CONGESTION CONTROL

To counter the increasing complexity of traffic patterns in IoT smart city environments, this paper proposes a hybrid artificial intelligence (AI) framework that utilizes Transformer-based contextual learning and cutting-edge spatiotemporal modeling paradigms. At the core of the proposed

framework lies the integration of BERT for semantic feature extraction and a ConvLSTM network to extract traffic space and time patterns. The above architecture enables efficient short-term traffic forecasting and facilitates dynamic congestion control through adaptive decision-making frameworks. The architecture involves four important stages. Each of the stages plays a crucial role in transforming raw traffic data into meaningful information for real-time traffic management. The system architecture is illustrated in the accompanying block diagram (see Fig. 1).

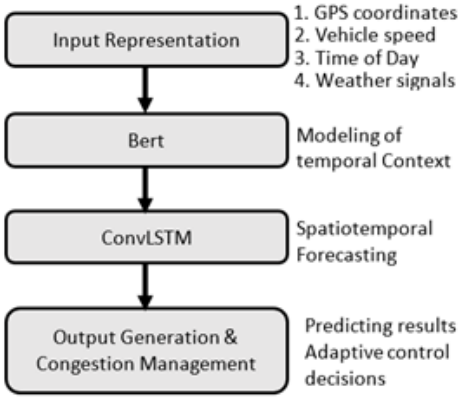

Figure 1: Block diagram of proposed model.

3.1 Input Representation

The system ingests real-time traffic data from IoT-based infrastructure such as smart sensors, vehicular ad-hoc networks (VANETs), and roadside units (RSUs). The data comprise multi-dimensional features including:

- Vehicle speed and density.
- GPS coordinates.
- Time of day.
- Weather or environmental signals (if applicable).

Each data instance is represented as a multivariate time series:

$$X=\{x1,x2,\ldots,xT\},\ xt \in Rd, \quad (1)$$

where $_T$ denotes the sequence length and *d* the number of features. Data preprocessing includes normalization, missing value imputation, and discretization. Tokenization techniques are then applied to convert structured data into sequences interpretable by the BERT model.

Example: Sensor ID #134 records traffic speeds every 5 minutes. For a given time, window, the input vector might be: [speed = 45 km/h, density = 18 veh/km, timestamp = 08:05, location = (34.061, -118.247)].

3.2 Temporal Context Modeling via BERT

The tokenized input is processed through a pertained BERT model to capture latent temporal and contextual dependencies across traffic patterns. Each input sequence $T = [t_1, t_2, ..., t_n]$ is mapped to an embedding matrix:

$$E=\text{BERT}(T),\ E \in R^{n\times d}. \quad (2)$$

This embedding's encoding the rich semantic relationships between traffic events occurring across space and time, allowing the model to generalize across varying network conditions.

Example: Morning rush hour patterns on weekdays are distinguished from off-peak periods based on learned embedding's, allowing the model to anticipate increased traffic between 7:30–9:00 AM.

3.3 Spatiotemporal Prediction via ConvLSTM

The output embeddings from BERT are reshaped and fed into a ConvLSTM layer. This module simultaneously captures spatial proximity and temporal recurrence using convolution operations inside gated memory units:

$$\hat{y}_{t+1} = \pi r^2 = \text{ConvLSTM}(Et). \quad (3)$$

Where $\hat{y}_{(t+1)}$ is the predicted traffic load or congestion level at the next time step. The ConvLSTM is particularly suited for capturing non-linear trends and periodicities within urban traffic flows.

Example: The model predicts that sensor ID #134 will experience a drop in speed from 45 km/h to 28 km/h in the next 10 minutes, signaling impending congestion.

3.4 Output Generation and Congestion Control

The final stage comprises two primary outcomes:

1) Prediction Results. Numerical values indicating future traffic states (e.g., travel time, congestion index, and vehicular density).
2) Adaptive Control Decisions. A reinforcement learning (RL) agent uses the prediction outputs to adjust signal timing, reroute traffic, or trigger alerts. The agent follows a policy π that maximizes cumulative network efficiency:

$$\pi^*(s_t) = \arg\arg\max_{\pi} E[R_t | s_t, \pi]. \quad (4)$$

Where s_t is the observed traffic state and R_t is the reward based on travel delay reduction.

Example: If predicted congestion exceeds a threshold, the RL agent reroutes vehicles from Highway 101 to nearby alternative roads, reducing overall travel delays.

To illustrate the proposed application framework, let's take one of the constraints, an electronic traffic situation from the METR-LA dataset containing the ID 717619 located on US-101 near downtown Los Angeles. At 8:00 a.m. on a Tuesday, the vehicle's speed record is set to reach 25.0 mph under overcast conditions, adjusted for the morning traffic. These multiple inputs, including weather, time, location, and speed, are normalized and encoded in the input phase. The BERT model then runs through sequences of similar historical traffic (e.g., weekday traffic data) to extract historical context, revealing trends such as frequent congestion between 7:45 and 8:15 a.m. These vehicle embeddings are then fed into a ConvLSTM system, which models spatial dependencies across neighboring landmarks (e.g., 717604 and 717531), predicting a speed drop to 19.2 mph in the following 15 minutes. Based on this prediction, the approval phase activates the adaptive traffic control system: to detect some traffic signals near the affected lane, the navigation system recommends and approves rerouting the vehicle onto I-5 North.

4 EXPERIMENTAL SETUP AND EVALUATION RESULTS TITLE

4.1 Dataset Description

To evaluate the proposed hybrid framework, we used the METR-LA dataset, which consists of traffic speed readings collected from 207 loop detectors installed on the highways of Los Angeles County over several months. The data include timestamps, average speeds, and road segment identifiers, sampled every 5 minutes. This dataset provides a rich temporal and spatial structure suitable for testing both short-term forecasting and congestion modeling. Each sample in the dataset is formatted as a multivariate time series. Preprocessing steps included outlier removal, missing value imputation using matrix factorization, normalization to zero-mean and unit-variance, and tokenization for BERT-compatible input sequences. Each sample is structured as a time series sequence $\{x_1, x_2, ..., x_T\}$, capturing temporal fluctuations in traffic across multiple road segments. The dataset is divided into 70% for training, 15% for validation, and 15% for testing. Optionally, domain-specific datasets such as METR-LA utilized. The hybrid architecture was implemented using TensorFlow 2.x and trained on a system with the following specifications:

- GPU: NVIDIA RTX 3090.
- RAM: 64 GB DDR4.
- BERT Variant: DistilBERT (fine-tuned for structured time-series encoding).
- Sequence Length: 128 tokens.
- ConvLSTM Units: 64 filters, 3×3 kernel.
- Optimizer: Adam with $\alpha=1\times10-4$.
- Batch Size: 32.
- Epochs: 50.

The BERT encoder was pertained on general datasets and fine-tuned on traffic sequences. The ConvLSTM layers processed reshaped embeddings to predict multi-step traffic conditions. Model performance was assessed using standard regression and classification metrics:

4.2 Evaluation Metrics

To evaluate model performance, three widely used regression metrics were employed: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2 score). These metrics are standard evaluation measures in prediction tasks and provide complementary information about prediction accuracy and model fit.

The evaluation was conducted for both short-term (5-minute horizon) and long-term (30-minute horizon) predictions. On the test set, the proposed hybrid model achieved MAE of 2.13, RMSE of 3.54, and an R^2 score of 0.91, indicating strong predictive performance and good generalization ability.

To further validate the effectiveness of the proposed BERT + ConvLSTM framework, its performance was compared against several state-of-the-art baseline models commonly used in traffic prediction tasks. These include:

- LSTM-only model, which captures temporal dependencies in sequential data and is widely used in traffic forecasting due to its simplicity and effectiveness [18].
- CNN-LSTM model, which combines convolutional layers for spatial feature extraction with LSTM layers for temporal modeling [19].
- Transformer-based model, which relies on self-attention mechanisms to capture long-range dependencies without recurrence or convolution [20].

The performance of all models was evaluated under the same experimental settings, and the results are summarized in Table 1.

Table 1: Performance comparison of prediction models.

Model	MAE	RMSE	R^2 Score
LSTM-only [15]	3.14	4.98	0.78
CNN-LSTM [16]	2.71	4.23	0.83
Transformer-only [17]	2.44	3.89	0.86
BERT + ConvLSTM (Proposed)	2.13	3.54	0.91

The proposed model demonstrated superior performance across all evaluation metrics. This confirms the advantage of incorporating BERT-based contextual embeddings for temporal understanding, combined with the spatial learning ability of ConvLSTM for enhanced traffic forecasting accuracy. As shown in Figure 1.

Figure 2 presents a dot plot comparing the performance of four traffic prediction models across three key evaluation metrics. The dot plot clearly shows that the proposed BERT + ConvLSTM model outperforms baseline methods across all metrics. It achieves the lowest MAE and RMSE, indicating higher prediction accuracy, and the highest R^2 score, reflecting strong model reliability. This confirms the effectiveness of combining contextual learning from BERT with spatiotemporal modeling via ConvLSTM for traffic prediction in smart IoT networks.

5 CONCLUSIONS

This paper presented a new hybrid AI framework that integrates BERT-based contextual embedding with ConvLSTM-based spatiotemporal prediction for intelligent traffic forecasting and congestion control in smart IoT networks. Experimental results confirmed that the proposed approach outperforms conventional deep learning models such as LSTM, CNN-LSTM, and Transformer-only baselines in terms of MAE, RMSE, and R^2 score. The synergy between semantic understanding and spatial memory proved critical in accurately modeling complex urban traffic patterns. Despite its effectiveness, the framework introduces computational overhead due to BERT's complexity and remains sensitive to data quality. Additionally, the model's reliance on deep, multi-layered architectures may reduce interpretability and pose challenges for deployment in real-time, resource-constrained edge environments. Future research will focus on optimizing the architecture for real-time deployment using lightweight Transformers (e.g., DistilBERT or TinyBERT), incorporating Graph Neural Networks (GNNs) for dynamic road topology modeling, and extending the framework with reinforcement learning agents for autonomous and adaptive traffic control.

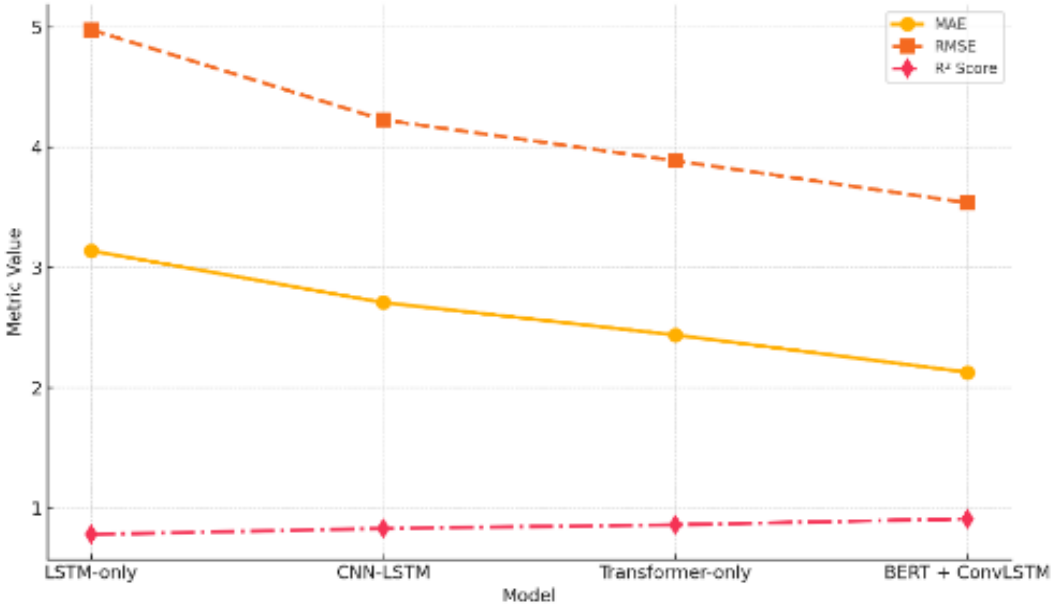

Figure 2: Model performance metrics.

REFERENCES

[1] M. Li, S. Li, Z. Zhu, and H. Liu, "Traffic Flow Prediction with Spatial-Temporal Graph Diffusion Network," Proc. AAAI Conf. Artif. Intell., vol. 34, no. 4, pp. 4189-4196, 2020.

[2] B. Yu, H. Yin, and Z. Zhu, "Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting," in Proc. IJCAI, pp. 3634-3640, 2018.

[3] J. Wang, Y. Zhang, L. Sun, and Y. Li, "ST-MGCN: A Spatial-Temporal Multi-Graph Convolution Network for Traffic Flow Prediction," IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 6, pp. 5029-5040, Jun. 2022, [Online]. Available: https://doi.org/10.1109/TITS.2021.3065653.

[4] J. Zhang, Y. Zheng, and D. Qi, "Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction," in Proc. AAAI Conf. Artificial Intelligence, 2017.

[5] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," in Proc. Int. Conf. Learning Representations (ICLR), 2018.

[6] B. L. Smith, B. M. Williams, and R. K. Oswald, "Comparison of parametric and nonparametric models for traffic flow forecasting," Transportation Research Part C: Emerging Technologies, vol. 10, no. 4, pp. 303-321, 2002.

[7] H. H. Saleh, I. A. Mish Khal, and D. S. Ibrahim, "Interference Mitigation in the Vehicular Communication Network Using MIMO Techniques," Journal of Engineering Science and Technology, vol. 16, no. 2, pp. 1837-1850, Apr. 2021.

[8] W. Huang, G. Song, H. Hong, and K. Xie, "Deep architecture for traffic flow prediction: Deep belief networks with multitask learning," IEEE Transactions on Intelligent Transportation Systems, vol. 15, no. 5, pp. 2191-2201, Oct. 2014, [Online]. Available: https://doi.org/10.1109/TITS.2014.2311123.

[9] 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," Transportation Research Part C, vol. 54, pp. 187-197, 2015.

[10] Z. Cui, R. Ke, and Y. Wang, "Deep Bidirectional and Unidirectional LSTM Recurrent Neural Network for Network-wide Traffic Speed Prediction," 2019, [Online]. Available: https://arxiv.org/abs/1801.02143.

[11] X. Shi, 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 Advances in Neural Information Processing Systems (NeurIPS), 2015.

[12] J. Sun and K. W. Axhausen, "Understanding spatiotemporal travel patterns with deep learning: A review," Transportation Research Part C, vol. 117, p. 102668, 2020.

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

[14] J. Xu, H. Zheng, J. Cao, and Z. Li, "Informer: Efficient Transformer for Long Sequence Time-Series Forecasting," IEEE Trans. Knowl. Data Eng., vol. 34, no. 10, pp. 5029-5040, Oct. 2022.

[15] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), pp. 5998-6008, 2017.

[16] Z. Zhao, W. Chen, X. Wu, P. C. Chen, and J. Liu, "Traffic-BERT: A Pretrained Model to Understand Traffic Context," IEEE Internet Things J., vol. 8, no. 12, pp. 9876-9887, Jun. 2021.

[17] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A Comprehensive Survey on Graph Neural Networks," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 1, pp. 4-24, Jan. 2021, [Online]. Available: https://doi.org/10.1109/TNNLS.2020.2978386.

[18] H. H. Saleh and S. T. Hasson, "Improving Communication Reliability in Vehicular Networks Using Diversity Techniques," Journal of Computational and Theoretical Nanoscience, vol. 16, no. 3, pp. 838-844, Mar. 2019, [Online]. Available: https://doi.org/10.1166/jctn.2019.7963.

[19] Y. Zhang and A. Haghani, "A gradient boosting method to improve travel time prediction," Transportation Research Part C: Emerging Technologies, vol. 58, pp. 308-324, 2015.

[20] Y. Jiang, S. Yu, X. Wang, and Q. Chen, "Spatio-Temporal Multi-Graph Attention Networks for Urban Traffic Forecasting," IEEE Trans. Intell. Transp. Syst., vol. 24, no. 5, pp. 5312-5324, May 2023.