

Real-Time Traffic Flow Prediction and Anomaly Detection Using Big Data

Bhavana H S, Dr. Geetha M

P.G. Student, Department of Master of Computer Applications, BIET, Davangere-577004.

bhavanah65@gmail.com, Davangere, Karnataka, India

Associate Professor of Master of Computer Applications, BIET, Davangere-577004.

bmgeethamallad@gmail.com, Davangere, Karnataka, India

ABSTRACT

Efficient traffic regulation in smart cities requires accurate prediction of road traffic flow and timely detection of abnormal events such as congestion spikes, accidents, or sensor failures. The rapid development of heterogeneous traffic data produced from loop detectors, GPS devices, cameras, and connected vehicles has enabled the usage of big data analytics and machine learning techniques for smart transportation systems. This paper offers a real-time traffic flow prediction and anomaly detection system that integrates big data processing with spatio-temporal learning Patterns. The framework utilizes historical and streaming traffic data to forecast short-term traffic conditions and identify anomalies using prediction residual analysis and reconstruction-based validation. Experimental results display that deep learning models, particularly recurrent and graph-based neural networks, significantly outperform traditional statistical approaches in forecasting accuracy. The proposed method achieves reliable anomaly detection while maintaining low latency suitable for real-time deployment.

Keywords:

Traffic Flow Prediction, Anomaly Identification, Big Data Analytics, Smart Transportation Systems, Spatio-Temporal Modeling, Graph Neural Networks

I. INTRODUCTION

Rapid urbanization and rising vehicle density have intensified traffic congestion problems worldwide. Intelligent Transportation Systems aim to enhance

mobility efficiency through predictive analytics and automated decision-making. Accurate traffic flow prediction enables proactive congestion control, route optimization, and signal timing adjustments. Simultaneously, anomaly detection plays a crucial role in identifying incidents, unexpected congestion, and infrastructure faults in real time.

The availability of large-scale traffic data from road sensors, connected vehicles, and mobile devices has created chances to apply big data analytics and advanced machine learning techniques. However, traffic data are inherently nonlinear, dynamic, and spatially correlated, which makes modeling challenging. This research presents a comprehensive structure combining big data processing with spatio-temporal learning models for accurate prediction and anomaly detection in real time.

II. LITERATURE REVIEW

Earlier studies relied on statistical time-series models such as Autoregressive Integrated Moving Average and Kalman filters for traffic forecasting. While these patterns perform adequately under stationary conditions, they struggle with nonlinear and complex traffic dynamics.

Machine learning methods such as Support Vector Regression, Random Forest, and Gradient Boosting improved predictive capabilities by modelling nonlinear relationships. More lately, deep learning techniques including LSTM networks and Gated Recurrent Units have demonstrated strong performance in capturing temporal dependencies. CNN have been applied to

spatially structured traffic data, while Graph Neural Networks effectively represent road networks as graph structures to model spatial interactions between sensors.

Anomaly detection research has explored clustering methods, statistical thresholding, and autoencoder-based reconstruction techniques. Prediction-based anomaly detection, where abnormality is identified through large deviations between predicted and observed values, has gained popularity due to its simplicity and effectiveness in real-time systems. Integration with big data streaming platforms has further enabled scalable deployment in urban environments.

III. METHODOLOGY

The proposed methodology integrates big data ingestion, preprocessing, predictive modeling, and anomaly detection into a unified framework.

The dataset used in this research consists of real-world traffic sensor data such as the PeMS dataset, which includes traffic flow, speed, occupancy, timestamp, and station identifiers collected at regular intervals. Additional contextual features such as weather conditions, day-of-week indicators, and holiday flags are incorporated to enhance model robustness.

Data preprocessing requires handling missing values through interpolation, normalization of numerical features, and aggregation into consistent time intervals. Spatial relationships among traffic sensors are represented using an adjacency matrix constructed from road network topology or correlation analysis. Feature engineering includes generation of lag features, moving averages, and periodic time indicators to capture daily and weekly traffic patterns.

Multiple algorithms are analyzed for traffic prediction. A baseline ARIMA pattern is implemented for comparison. ML models such as Random Forest and Gradient Boosting are useful to capture nonlinear patterns. Deep Learning models comprising LSTM and GRU are employed to model temporal dependencies. A Graph CNN combined with a recurrent layer is implemented to catch both spatial and temporal dependencies within the traffic network.

The training task involves splitting the dataset into training and testing sets. Model parameters are optimized using cross-validation, and performance

is calculated using Mean Absolute Error and RMS Error.

For anomaly detection, a residual-based method is adopted. The prediction model generates expected traffic values in real time. The residual between predicted and real values is calculated.

If the residual exceeds a dynamically computed statistical threshold, the event is classified as anomalous. An autoencoder model is also used to measure reconstruction error, providing an additional verification mechanism for anomaly detection.

The system workflow follows a continuous pipeline structure:

Data Collection from Traffic Sensors

Data Cleaning and Normalization Feature Engineering and Graph Construction Model Training and Validation Real-Time Traffic Prediction Residual Computation Anomaly Detection Alert Generation and Visualization This pipeline ensures continuous monitoring and scalable real-time analytics.

IV. RESULTS AND DISCUSSION

Experimental evaluation establishes that deep learning models exceed traditional statistical approaches in short-term traffic forecasting. LSTM and Graph Convolutional Network models achieve lower prediction errors compared to ARIMA and Random Forest models due to their capability to capture nonlinear spatio-temporal relationships.

The anomaly detection module successfully identifies abnormal congestion events and sudden traffic drops. Evaluation metrics such as precision, recall, and F1-score indicate improved detection performance when combining residual-based detection with autoencoder validation.

Latency analysis shows that the system can operate efficiently in near real-time environments when grouped with big data streaming platforms. The combined prediction and detection framework provides practical benefits for traffic management centers by supporting early warning and proactive control strategies.

V. EXPERIMENTAL RESULTS



VI. CONCLUSION AND FUTURE WORK

This paper represented a real-time traffic flow prediction and anomaly detection framework leveraging big data analytics and advanced ML models. The combination of spatio-temporal deep learning techniques with streaming anomaly detection provides accurate forecasting and reliable abnormal event identification.

Future work includes incorporating probabilistic forecasting to quantify uncertainty, applying transfer learning for cross-city model adaptation, integrating multimodal transportation data, and deploying lightweight compressed models for edge computing environments. Further research may also explore reinforcement learning-based traffic control strategies using predicted traffic states and detected anomalies.

REFERENCES

- [1] J. Lv et al., "Traffic Flow Prediction With Big Data: A Deep Learning Approach," IEEE Transactions on Intelligent Transportation Systems, 2015.
- [2] Y. Li et al., "Diffusion Convolutional Recurrent Neural Network for Traffic Forecasting," International Conference on Learning Representations, 2018.
- [3] B. Yu et al., "Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting," IJCAI, 2018.
- [4] W. Ma et al., "LSTM Neural Network for Traffic Speed Prediction Using Remote Microwave Sensor Data," Transportation Research Part C, 2015.
- [5] Z. Cui et al., "Traffic Graph Convolutional Recurrent Neural Network," IEEE Access, 2019.
- [6] X. Jia et al., "Prediction-Based Traffic Anomaly Detection Using Big Data Analytics," IEEE Access, 2020.
- [7] H. Zheng et al., "Urban Traffic Flow Prediction: A Deep Learning Approach," IEEE Transactions on Intelligent Transportation Systems, 2019.
- [8] Y. Wu et al., "Graph WaveNet for Deep Spatial-Temporal Graph Modeling," IJCAI, 2019.
- [9] L. Sun et al., "Using Autoencoders for Traffic Anomaly Detection," IEEE Intelligent Transportation Systems Conference, 2017.
- [10] M. Zaharia et al., "Apache Spark: A Unified Engine for Big Data Processing," Communications of the ACM, 2016.
- [11] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, 1997.
- [12] T. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," ICLR, 2017.