LLM-DRIVEN TRAFFIC RISK PREDICTION IN CLOUD–EDGE NETWORKS

Author's: M. Nandhini[1], G. Thilipkumar[2], K. Sujith[3], N. Vanjulavalli[4]

Abstract

The rapid growth of intelligent transportation systems (ITS) and connected vehicular infrastructure has generated massive volumes of heterogeneous traffic data across distributed cloud–edge environments. Efficiently leveraging this data for real-time traffic risk prediction and accident severity analysis remains a critical challenge due to latency constraints, scalability requirements, and the presence of both structured sensor data and unstructured incident reports. This research proposes a Cloud–Edge collaborative framework that integrates Large Language Models (LLMs) with machine learning techniques to enhance traffic risk assessment and accident severity prediction.In the proposed architecture, edge nodes collect real-time traffic sensor data, including vehicle speed, traffic density, weather conditions, visibility, road type, and junction information. To enrich predictive capabilities, unstructured textual incident reports generated by traffic operators, emergency responders, or connected vehicles are processed using LLM-based semantic feature extraction. The LLM module transforms textual descriptions into structured risk indicators such as fog presence, overspeeding behavior, congestion, collision type, and environmental hazards. These extracted semantic features are combined with structured sensor inputs to form a unified feature representation.The cloud layer performs centralized model training using supervised learning algorithms for two predictive tasks: (1) binary traffic risk prediction to identify high-risk traffic conditions and (2) multi-class accident severity analysis to classify incidents into minor, moderate, or severe categories. Advanced ensemble models such as Random Forest and logistic regression are utilized to capture nonlinear relationships between traffic parameters and risk outcomes. The trained models are periodically deployed to edge nodes to support low-latency inference and real-time alert generation.