IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 February 2024, accepted 17 April 2024, date of publication 26 April 2024, date of current version 10 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3393981

RESEARCH ARTICLE

Enhancing Urban Traffic Management Through Real-Time Anomaly Detection and Load Balancing

MY DRISS LAANAOUI[1], MOHAMED LACHGAR[2], HANINE MOHAMED[2], HRIMECH HAMID[3], SANTOS GRACIA VILLAR[4,5,6], AND IMRAN ASHRAF[7]

[1]Ecole Normale Superieure, Department of Computer Science, Cadi Ayyad University, Marrakech 40000, Morocco
[2]LTI Laboratory, ENSA, Chouaib Doukkali University, El Jadida 24000, Morocco
[3]LAMSAD Laboratory, ENSA, Hassan First University, Berrechid 26000, Morocco
[4]Higher Polytechnic School, Universidad Europea del Atlántico, 39011 Santander, Spain
[5]Universidad Internacional Iberoamericana, Campeche 24560, Mexico
[6]Universidade Internacional do Cuanza, Cuito, Bié, Angola
[7]Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, Republic of Korea

Corresponding authors: Imran Ashraf (ashrafimran@live.com) and My Driss Laanaoui (d.laanaoui@uca.ma)

This work was supported by the European University of Atlantics.

ABSTRACT Efficient traffic management has become a major concern within the framework of smart city projects. However, the increasing complexity of data exchanges and the growing importance of big data makes this task more challenging. Vehicular ad hoc networks (VANETs) face various challenges, including the management of massive data generated by different entities in their environment. In this context, a proposal is put forth for a real-time anomaly detection system with parallel data processing, thereby speeding up data processing. This approach accurately computes vehicle density for each section at any given time, enabling precise traffic management and the provision of information to vehicles regarding traffic density and the safest route to their destination. Furthermore, a machine learning-based prediction system has been developed to mitigate congestion problems and reduce accident risks. Simulations demonstrate that the proposed solution effectively addresses transportation issues while maintaining low latency and high precision.

INDEX TERMS Urban traffic management, real-time anomaly detection, intelligent transportation systems, traffic density prediction.

I. INTRODUCTION

Intelligent transport systems (ITS) have long stood as the backbone of efficient road traffic management, enabling optimization and seamless information exchange [1]. They are pivotal not only in elevating user quality of experiences but also in ensuring the seamless operation of the entire transportation network. However, as the landscape of transportation undergoes continuous transformation, a significant challenge has emerged, one rooted in data exchange [2]. The surge of big data and the ever-growing spectrum of stakeholders within the ITS ecosystem have brought to the forefront capacity limitations that obstruct the ability to effectively address road and highway congestion [3].

The convergence of diverse vehicle types and the escalating number of vehicles on roads further complicates this intricate web of communication [4]. Challenges in communication are amplified by the sheer diversity and volume of data in play, all while real-time systems become increasingly indispensable for enhancing the efficiency of ITS [5]. These multifaceted issues cast a shadow over the broader context, rendering it more susceptible to traffic congestion, inadequately designed intersections, transportation incidents, and other complications [6], [7], [8]. The collective impact of these challenges resonates profoundly, hampering the smooth operation of road infrastructure and exacerbating congestion-related issues.

The associate editor coordinating the review of this manuscript and approving it for publication was Yanli Xu.

In response to these multifaceted challenges, the emergence of effective traffic management takes center stage as a pivotal solution. Predicting traffic density on highways and roads stands out as one such solution. This predictive approach offers valuable insights into potential trouble spots, paving the way for proactive issue resolution. Within this context, vehicular ad hoc networks (VANETs) have risen to prominence, evolving into a highly sought-after paradigm for facilitating communication, both vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) [9]. Over time, numerous research projects and applications have reaffirmed the appropriateness of VANETs for ITS implementation [10].

Key advancements include the proliferation of on-board units (OBUs), significantly boosting the efficiency of mobile node communication [11]. Additionally, roadside units (RSUs) equipped with dedicated short-range communications (DSRC) have demonstrated their effectiveness in facilitating communication between stationary and non-stationary nodes [12]. Furthermore, RSUs customized for the management of VANET applications and the coordination of actions have further extended the capabilities of this technology [13]. A growing consensus acknowledges the advantages of centralizing these applications to streamline management and deployment, thereby enhancing the overall system's efficiency.

This study embarks on an exploration of the diverse units employed in both ITS and VANET networks. It also introduces an innovative approach to predicting high-risk areas, thus contributing to the ongoing dialogue on enhancing the efficiency and safety of transportation systems. As the transportation landscape continues to evolve, the pursuit of innovative solutions becomes increasingly crucial for the well-being of roadways and the individuals who depend on them.

Additionally, the study harnesses the power of the Lambda architecture, a dynamic framework that enables real-time adaptation by adjusting predictive models through continuous batch processing. This architectural approach allows for the seamless integration of new data, ensuring the system's ability to respond to changing traffic conditions promptly. It equips the system with the capability to continuously monitor traffic, evaluating both historical and real-time data to remain well-informed and promptly address shifts in the traffic landscape. Notably, the Lambda architecture is designed to maintain operational continuity, even in the event of a component failure. For example, if the real-time processing layer encounters an issue, the batch processing layer can seamlessly take over, ensuring uninterrupted system functionality. This architectural innovation further underscores the robustness and reliability of the approach, making it a valuable asset in the quest to enhance the efficiency and safety of transportation systems, ultimately benefiting both the infrastructure and the individuals who rely on it.

This article is structured into several sections, each addressing specific aspects of road traffic management in the context of VANETs and presenting an innovative solution based on big data and machine learning technologies. Section II examines prior research related to road traffic management, VANETs, and the utilization of big data and machine learning in this domain. Section III provides a detailed overview of the proposed traffic management approach, emphasizing the creation of a real-time database and the use of machine learning for traffic anomaly detection and accident prevention. Section IV showcases simulations and outcomes obtained through the proposed solution, demonstrating its effectiveness in congestion reduction and accident prevention, analyzes the implications of the findings, and explores possibilities for further refinement and extension of the proposed approach. Section V presents some identified limitations of the study. Lastly, Section VI summarizes the key contributions of this study and outlines future research prospects in road traffic management within VANETs.

II. RELATED WORKS

Before delving into the detailed presentation of these research papers, it is essential to underscore the magnitude of the collective effort exerted in the field of transportation management and big data analytics within VANETs. These works represent a vast array of innovative and interconnected contributions aimed at addressing crucial challenges in the realm of intelligent urban mobility. By combining big data techniques, machine learning, real-time processing, and communication security, these researchers have pushed the boundaries of understanding and managing road traffic while ensuring more reliable and secure communication networks for connected vehicles. The examination of these works will shed light on significant advancements and innovative solutions emerging to tackle challenges concerning road traffic, vehicle safety, and the efficiency of communication networks within smart cities.

The study [14] provides a survey on the topic of group data communication within connected vehicles. The authors review and compare existing group data communication strategies in VANETs with a focus on routing strategies and quality of service (QoS) awareness. Security parameters are also analyzed. Similarly, a comprehensive overview of emergency communication networks is provided in [15]. It surveys the classification, characteristics, and applications of emergency communication networks. It discusses multiple network technologies, including satellite, wireless mesh network (WMN), mobile ad hoc network (MANET), VANET, flying ad hoc network (FANET), SANET, cellular networks, and wireless private networks. The paper also summarizes existing emergency communication schemes, their challenges, and potential directions. It emphasizes the integration of various networking technologies and the shift towards broadband multimedia networks, promoting advanced emergency communication capabilities.

In [16], the authors provide a comprehensive survey of the intersection between vehicular cloud computing and big

data. This survey likely explores the current state of research and developments in these areas. The authors introduce vehicular cloud computing (VCC), which is a promising solution that combines VANETs and cloud computing. VCC offers several advantages, including cost reduction, reduced traffic congestion, and enhanced safety. However, the paper discusses challenges within the vehicular cloud network, particularly in managing big data. It covers VCC architecture, types, characteristics, applications, and advantages when it is combined with big data and discusses the challenges within the vehicular cloud network, particularly in managing big data.

The authors present a two-phase clustering method for identifying traffic accident black spots in [17]. The study proposes a novel black spot identification model, integrating geographic information system (GIS)-based processing with hierarchical density-based spatial clustering of applications with noise. Parameters are optimized using a clustering validation index to minimize subjectivity. Validated with 3536 accident data in Hangzhou, China, it identifies 39 black spots. Results reveal that these spots account for 75% of accidents but only 23.26% of the total road network length.

The study [18] discusses a method for selecting reliable relays for routing emergency messages in intermittently connected VANETs. This research likely explores strategies to ensure the dependable and timely transmission of critical emergency messages in challenging network conditions. A relay selection scheme, EMR-ICN, is presented for emergency message routing in intermittently connected networks. EMR-ICN employs V2I and V2V communication, uses position prediction and mobility metrics for stable relay selection, and minimizes channel contention by adapting beacon intervals. By predicting relative vehicle positions and choosing reliable relays, it outperforms existing protocols, showing improved message delivery, reduced latency, and fewer hop counts, making it suitable for both dense and sparse network environments.

The focus of [19] is to predict vehicle acceleration using machine learning models and an analysis of driving behavior. The research explores how machine learning can be applied to anticipate vehicle acceleration patterns, considering various driving behaviors and factors. In addition, the authors introduce a hidden Markov model (MHMM) for analyzing personalized driving behaviors in car-following scenarios. The MHMM effectively segments driving data into behavioral segments and maintains reasonable durations for each segment. Using the MHMM analysis results, similar drivers are grouped together. The study then uses a controlled experiment to predict driver acceleration, comparing the performance of gated recurrent unit (GRU) and long short-term memory (LSTM) models. Results confirm the MHMM's effectiveness in personalized driving behavior analysis and indicate that GRU outperforms LSTM.

The authors discuss the use of big data in the context of smart mobility [20]. They provide insights into the approaches, applications, and challenges of employing large datasets to enhance and optimize smart mobility solutions. This paper discusses the growing influence of the Internet of Things (IoT), artificial intelligence (AI), and big data in the realm of smart mobility. It highlights the positive impact on convenience, safety, and transportation efficiency. However, challenges like job displacement and privacy concerns exist. The rapid technological advancements require a cultural shift, and issues like personal data security, autonomous vehicle security, and traffic optimization need to be addressed. Big data and AI are poised to enhance smart mobility, offering convenience, safety, and improved traffic management through data processing.

The study [21] focuses on a real-time big data analytics system for proactive traffic safety management. It also explores how big data analytics are used to monitor and enhance traffic safety in real-time, along with a visualization system to support these efforts. The authors present a web-based proactive traffic safety management (PATM) and real-time big data visualization tool, which won recognition from the US Department of Transportation (USDOT). It utilizes real-time data, including traffic, weather, and CCTV video streams, to develop modules for crash prediction, expedited detection, PATM recommendations, data sharing, and report generation. The system effectively visualizes real-time data, identifies hidden patterns, and demonstrates excellent crash prediction performance. The paper also discusses its current and future implementation.

The authors of [22] provide a comprehensive review of techniques for predicting urban traffic flow. The study provides a review of intelligent techniques for predicting urban traffic flow. Various models, including deep learning and hybrid algorithms, are examined for their predictive precision. The study finds that models based on deep learning and hybrid algorithms, particularly graph convolutional networks (GCN), demonstrate promising results, outperforming traditional models like autoregressive integrated moving average (ARIMA), support vector regressor (SVR), K nearest neighbor (KNN), and support vector machine (SVM). Challenges remain in addressing spatial-temporal external data and varying spatial dependencies in different regions to enhance urban traffic predictions.

The study [23] presents various aspects of trust establishment and maintenance in VANETs, addressing critical security and reliability issues within these networks. This work explores trust management in VANETs. It emphasizes the need for balancing privacy, security, and quality of service. The paper provides an overview of ITS and VANETs, discussing their importance and security challenges. Various trust management schemes are surveyed, and a taxonomy is proposed based on Artificial Intelligence and emerging technologies. The work also suggests research directions related to Federated Learning, clustering, energy consumption, and the impact of emerging technologies on trust models.

The authors make the application of a Levenberg-Marquardt artificial neural network model for predicting vehicular traffic flow in [24]. The study is carried out in the context of the Italy road transportation system, using survey-derived parameters. The key findings include the model's effectiveness in predicting traffic flow, the significance of factors like vehicle speed, traffic volume, time, and the number of vehicles, and the potential use of this research for advanced traveler information systems and traffic flow forecasts. The study also highlights the need for further research on factors like traffic accidents and weather conditions. Limitations include the study's specific focus and the omission of factors like weather conditions and road accidents.

The study [25] discusses the temporal prediction of traffic characteristics in real-world road scenarios in Amman. It focuses on predicting traffic congestion for drivers by utilizing big data and machine learning regression techniques on road scenarios in Amman, Jordan. Three regression methods were tested, resulting in accurate predictions of traffic characteristics such as vehicle positions, speed, and locations. The best regression model was then used to forecast future traffic congestion levels, providing real-time information to drivers for route planning and optimization.

The authors of [28] introduce a system for predicting the estimated time of arrival in real-time by leveraging historical surveillance data. They establish a comprehensive infrastructure for real-time arrival time estimations of all flights heading to Malpensa-Milan airport, solely based on aircraft surveillance data. Data warehousing and distribution solutions were developed to make the information accessible for machine learning model training. The model, trained on almost three years of surveillance data using a state-of-the-art platform, demonstrated improved accuracy compared to current services. Future work may explore different machine learning algorithms, including artificial neural networks, and expand data sources, such as weather information, to enhance feature vectors.

A real-time GIS in the context of smart cities is discussed in [29]. The authors touch on a vast topic, raising more questions than it answers. It encourages readers to pursue their own research in this field. The cities of the future will be significantly different from today, driven by factors like connected and autonomous vehicles, online retail, digital technology, climate change, global demographics, and healthcare advancements. These changes will create numerous opportunities for GIS specialists.

The study [26] explores the development of a protocol called, traffic prediction protocol (TDPP), for predicting real-time traffic distribution in vehicular networks. The introduced TDPP estimates traffic characteristics on downtown and highway road scenarios. It evaluates and predicts traffic speed, density, and estimated travel time based on basic traffic data. The TDPP protocol proves efficient, consuming low bandwidth and providing accurate predictions for various road scenarios. The limits of this work are related to detecting obstacles, emergency vehicles, and exit/entrance ramps by analyzing vehicle speed distribution over the area of interest.

In [30], a comprehensive investigation into the use of data mining and machine learning techniques is provided regarding ITS and control systems. The main idea is to enhance understanding of traffic management technologies, particularly through data mining and machine learning. The study categorizes existing research into various approaches: real-time traffic parameter measurement, detection of moving objects, routing identification, analysis of driver and pedestrian behaviors, and a focus on traffic light signals. Future work involves proposing a new traffic management approach to further advance the field.

The authors of [27] present the Vanet-TSMA approach, which focuses on traffic safety management in VANETs for smart road transportation. The study analyzes the trend of road traffic accidents in India. It introduces a VANET-based traffic safety management approach, VANET-TSMA, aimed at reducing accidents. This approach efficiently utilizes message distribution, traffic management, and congestion control, relying on multiple information and communication technology (ICT)-based safety mechanisms. It addresses collision avoidance, detection, and traffic congestion to enhance the safety of future connected vehicles on the road.

A new method known as SES-AD (Space-Embedding Strategy for anomaly detection) is proposed for detecting anomalies in multivariate time collection [31]. The technique involves projecting the uncooked sequences into a decrease-dimensional area, where widespread abrupt change points may be effortlessly captured from the dissimilarity vector. A statistical strategy is then used to decide the capability anomalies. The effectiveness of the SES-AD approach was validated by way of applying it to a large quantity of multivariate time collection. The experimental consequences tested that SES-AD is greater green than 5 current tactics. Overall, the SES-AD model is appropriate for anomaly detection in excessive-dimensional datasets and ensures computational efficiency and accuracy. The authors in [32] introduce the LRRDS (Local Recurrence Rate based Discord Search), a unique computational framework for detecting discords in multivariate time collection (MTS) records. The proposed technique utilizes a recurrence plot derived from the unique time series facts to correctly identify discords. To improve efficiency, a strategy is employed for pairwise distance assessment among subsequences. Extensive simulations on various MTS datasets display the effectiveness and performance of LRRDS. A contrast with existing techniques, which include GDS, exhibits that LRRDS outperforms them in terms of performance. Furthermore, LRRDS correctly addresses the adaptability trouble of discord sequences in multi-dimensional areas, ensuring computational effectiveness and performance.

Despite the results reported in the above-discussed studies and their reported efficacy, these approaches share a common

TABLE 1. Performance evaluation based on key metrics.

Approach	Scalability	Real time	Density	Throughput
Novel Black Spot Model [17]	Average	Low	Average	×
Vehicular Cloud Computing (VCC) [16]	High	Low	High	×
Emergency Message Routing in Intermittently Connected Networks (EMR-ICN) [18]	×	×	Average	High
MHMM (Hidden Markov Model) [19]	High	Average	Low	Low
Proactive traffic safety management (PATM) [21]	Low	High	Low	High
Levenberg-Marquardt artificial neural network model for predicting vehicular traffic flow [24]	Low	Average	High	Average
A protocol called TDPP for predicting real-time traffic distribution in vehicular networks [26]	Average	Low	Average	Low
The Vanet-TSMA approach [27]	Average	High	Average	Low

shortcoming, i.e., the absence of effective real-time. Table 1 presents the limitations of the main works presented based on the following criteria:

- **Scalability**: In larger VANETs, the ability to scale the network and maintain efficient communication becomes crucial. Metrics related to scalability evaluate the network's performance as it expands.
- **Real time**: The real-time factor (RTF) is a commonly used metric to measure the speed of an automatic speech recognition system during the decoding phase or at runtime. It can be applied to scenarios involving real-time processing.
- **Density**: Density describes the concentration or volume of vehicles on a road or in a specific area at a given time, affecting traffic flow and congestion levels.
- **Throughput**: Throughput refers to the rate at which a system, process, or network can process, transfer, or deliver data or items within a specific timeframe, indicating its efficiency or capacity.

All these approaches share a common shortcoming: the absence of effective real-time road traffic management and a comprehensive overview of the ongoing road conditions. It is imperative to maintain an up-to-date database that continuously tracks vehicle densities on every road, facilitating the equitable distribution of traffic loads among different routes. The implementation of load-balancing techniques holds the promise of lighter traffic conditions, subsequently reducing the inherent risk of accidents. The pressing need of the hour is a method capable of dynamically managing the ever-changing state of road traffic in real-time.

III. MATERIALS AND METHODS

A. OVERVIEW

The methodology proposed in this article is divided into several stages, which are as follows:

i) **Data Collection**: Data is gathered from various sensors installed on the roads, such as surveillance cameras, speed sensors, vehicle density sensors, etc.
ii) **Data Preprocessing**: The collected data is cleaned, filtered, and preprocessed to remove unnecessary data and correct errors.
iii) **Data Storage**: The preprocessed data is stored in a real-time database for later analysis.
iv) **Data Analysis**: The stored data is analyzed in real-time using big data processing techniques such as Hadoop MapReduce and Apache Storm.
v) **High-Risk Zone Prediction**: High-risk zones are predicted using the collected and analyzed data.
vi) **Anomaly Detection**: Anomalies are detected in real-time using machine learning techniques such as classification and regression.
vii) **Load Management**: The load is balanced in real-time using vehicle density data for each section of the road.

Figure 1 provides a visual representation of the stages followed in the adopted methodology. It illustrates the flow of data and the sequential order of these essential steps.

In the following subsections, the key stages of the proposed methodology are outlined.

B. ANOMALY DETECTION PROCESS

Currently, urban traffic management predominantly relies on computer systems, particularly in densely populated areas. These systems prove effective under normal traffic conditions but tend to struggle when traffic reaches a state of saturation. Saturation typically occurs in areas with uneven supply and demand, often at intersections. Two distinct scenarios can be discerned: one involving gradual degradation of service quality as demand steadily rises (referred to as a loaded situation), and the other marked by an abrupt drop in average service levels when saturation suddenly occurs.

To counteract saturation, preventive measures involve increasing network capacity and implementing regulations to ensure that the supply of roads consistently exceeds the demand from vehicles, particularly on frequently traveled routes. Remedial measures come into play when instantaneous demand surpasses supply. In such instances, the excess demand must be redirected to designated areas where saturation can be intentionally induced and controlled. This technique, commonly used in various traffic management systems, includes the establishment of retention areas.

One of the key challenges in urban traffic regulation is the need to acknowledge that users are not automated machines; they exhibit individual behaviors. Encouraging

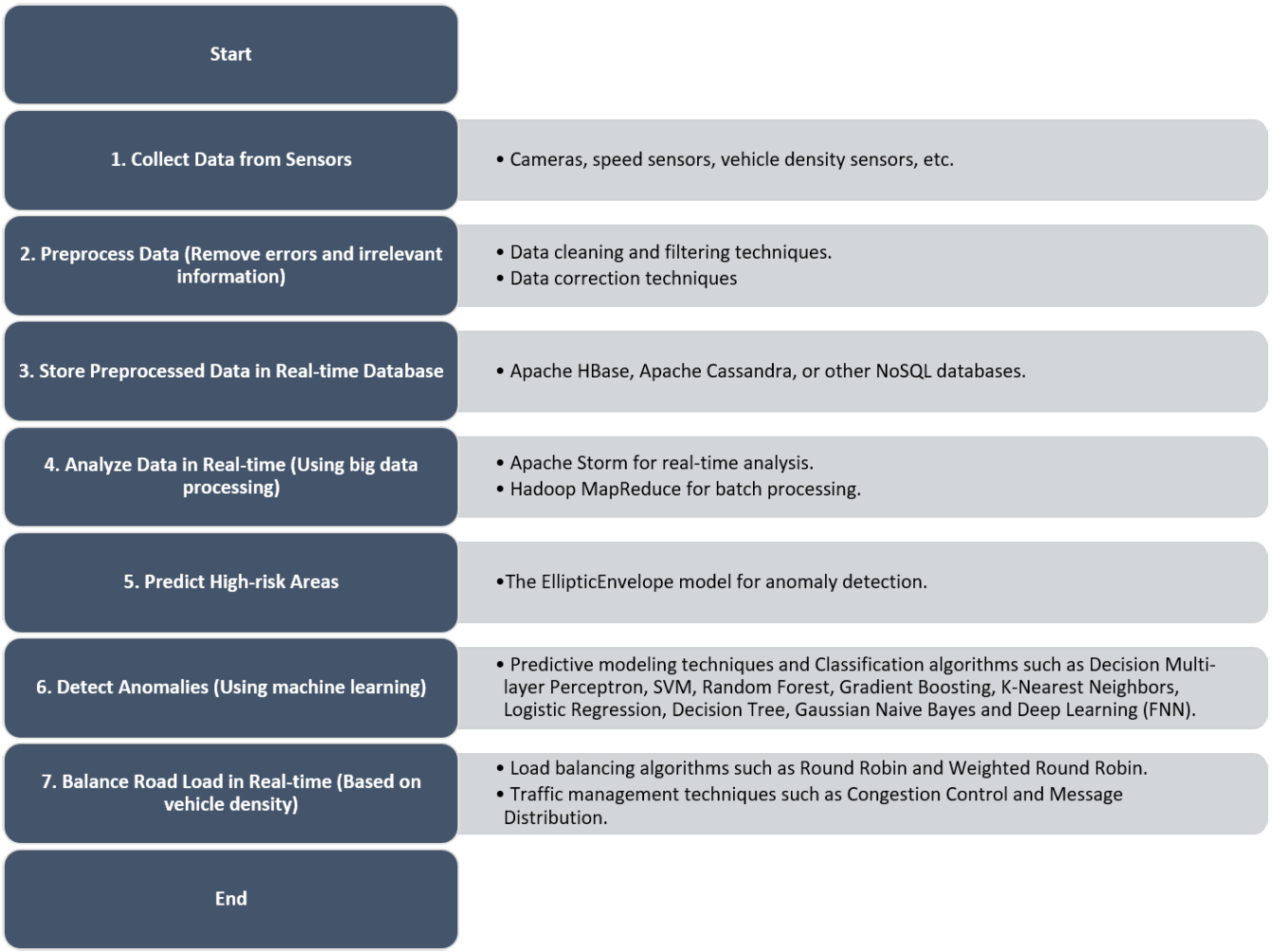


FIGURE 1. Steps followed in the proposed methodology.

them to adhere to traffic regulations, including signals, traffic lights, and road markings, is crucial for effective traffic management. To be effective, these regulatory constraints must align reasonably with natural human behavior. Anomalies can be categorized into two types based on their behavior and potential treatment: chronic anomalies and accidental anomalies.

i) **Chronic anomalies** stem from inherent network under-capacity and are frequently examined in studies on saturation. They exhibit a predictable pattern, often occurring on a daily or weekly basis.

ii) **Accidental anomalies**, on the other hand, result from traffic incidents such as deliveries, vehicle breakdowns, improperly parked vehicles, and roadwork. These anomalies are random in nature.

C. DATA COLLECTION

To enhance traffic management and alleviate congestion, a solution has been devised involving the creation of a real-time database containing road density information indexed by city and time. This database will facilitate improved routing for vehicles, promoting better load distribution within cities (urban traffic) and on highways.

The routing mechanism will enable various stakeholders in the system to circumvent congested roads and select alternative routes. Each vehicle will record data when it traverses a road section between its two endpoints, transmitting this information, including its unique identifier (ID), to an RSU at the road's edge. Subsequently, the system will calculate the vehicle density for each road section, providing real-time density data stored in the database.

The database, functioning as an RSU layer, will capture information transmitted by vehicles at each step of their journey, including road entry, speed, direction, and more, all identifiable by a unique ID. This process generates a wealth of real-time data about each vehicle's road crossings. The sequence diagram to construct the database is illustrated in Figure 2.

Regarding routing, the database, which has already received vehicle information, will serve as the input source to generate routes on demand for specific vehicles. Vehicles will

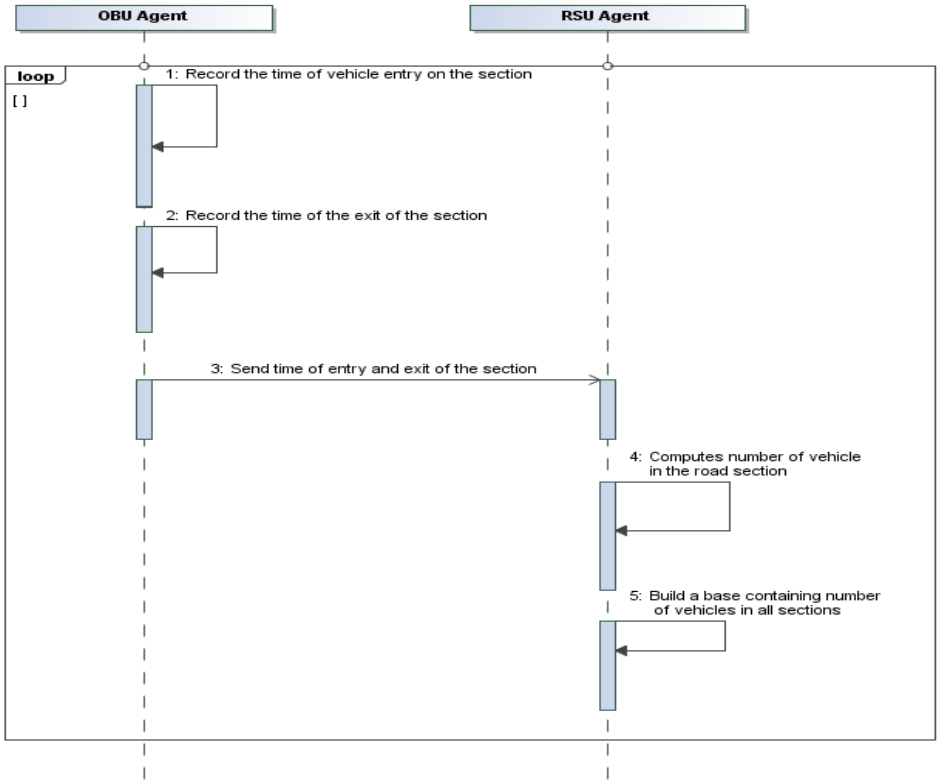


FIGURE 2. Sequence diagram to construct the database.

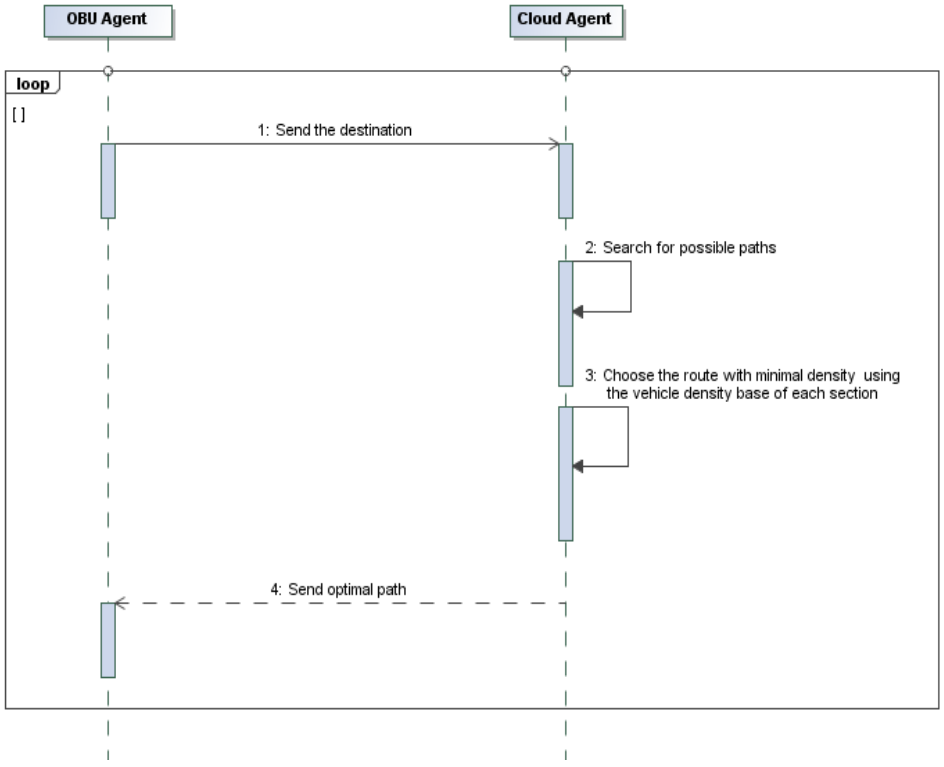


FIGURE 3. Sequence diagram of routing.

submit requests with specific details such as destination, vehicle weight, and type. Taking into account road density and congestion avoidance (scheduling), the system will respond with a comprehensive route plan, including estimated travel times and road information (e.g., stop signs, traffic signals, etc.), which can be integrated into future human-machine

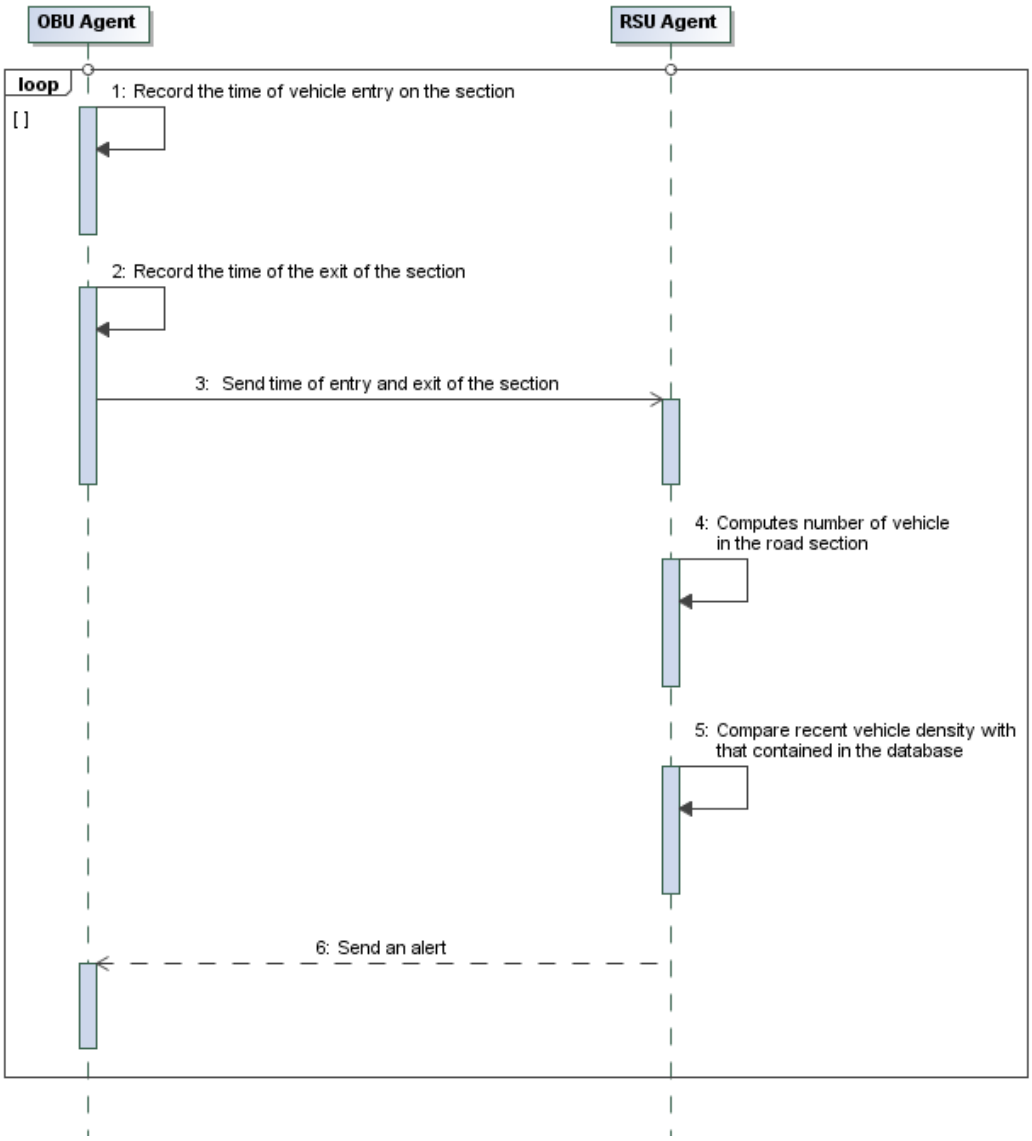


FIGURE 4. Anomaly detection sequence diagram.

interfaces. Figure 3 shows the sequence diagram for routing scheme.

The proposed architecture also possesses the capability to detect anomalies in real-time, distinguishing between chronic and accidental anomalies through a gateway that compares current traffic conditions with historical data of recurring jams or anomalies in specific city sections or roads.

The system continuously receives updated vehicle density information and compares it with previously recorded values. When the difference between current and historical density for a section exceeds a predefined threshold, the system updates the vehicle density database. Vehicles intending to overtake or change their route declare their destination, and the system calculates the route that minimizes vehicle density across various sections. The anomaly detection sequence is presented in Figure 4.

Initially, this change is stored in the database and relayed as an alert to vehicles routed through the affected section, enhancing the efficiency of their decision-making, be it a route change or a stop. In cases where the section returns to normal density, new routes may be selected, or alerts about the anomaly can be removed. Utilizing the database's vehicle density information to reach their destination, vehicles opt for paths with minimal vehicle density. This entire methodology is summarized in Figure 5.

The proposed approach offers drivers a way to avoid hazardous situations, thus reducing the risk of accidents. Prior to setting off, each vehicle transmits a notification containing section identification, destination, and departure time. The system, comparing this information with the vehicle's current route, decides whether to maintain the existing route or suggest a change. Given the frequent fluctuations in traffic

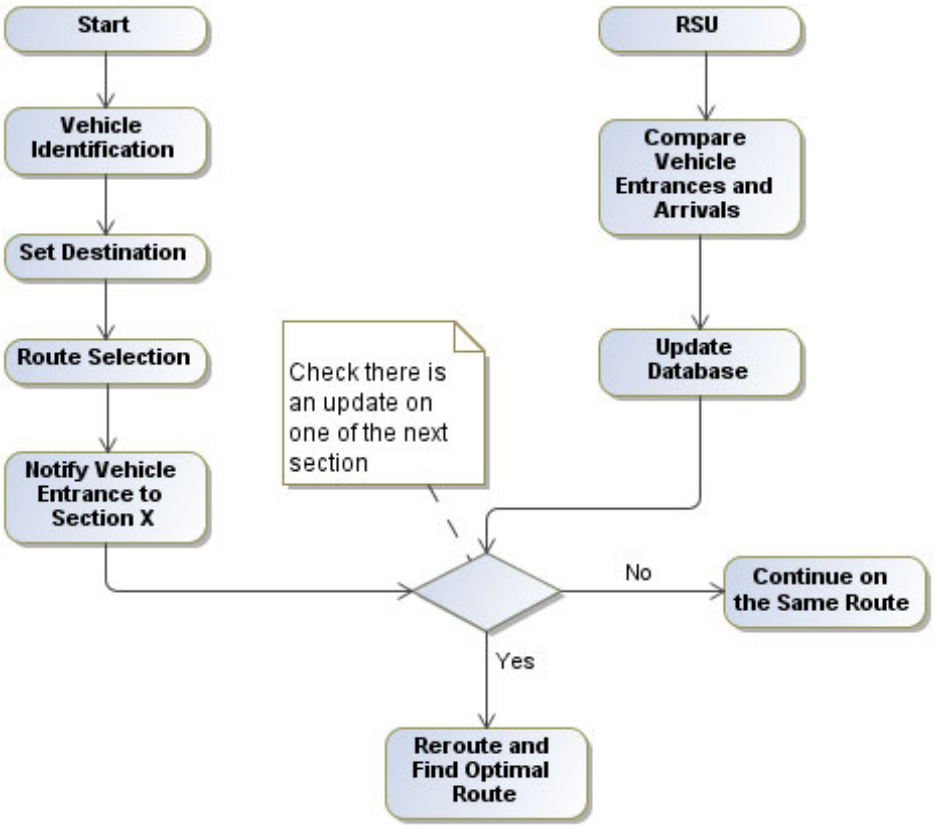


FIGURE 5. Data handling and predictive framework.

density within VANET networks, the trajectory is evaluated each time a vehicle enters a new section.

D. PREDICTION OF ANOMALIES

Traffic management has long been a substantial area of study aimed at ensuring the safety of all participants in the transportation system, including vehicles and pedestrians. Nevertheless, unpredictable human behavior means that anomalies, such as accidents, cannot be entirely prevented.

Therefore, proactive measures to minimize human losses are of paramount importance to both businesses and governments. Currently, there is no system that can accurately detect anomalies in real-time on both urban roads and highways. This is where predictive anomaly pattern recognition emerges as a significant advancement in road safety. Despite the challenge of obtaining training data due to data scarcity, the system, aided by the aforementioned database, can create a more comprehensive dataset with labeled inputs for historical analysis using machine learning (ML) frameworks.

The system is adeptly designed for classifying various data patterns, including meteorological and accident datasets, as well as intricate holiday traffic patterns. This is facilitated by leveraging a rich array of data sourced from sensors. The focus is on employing a suite of supervised anomaly detection techniques, utilizing a diverse set of machine learning models. These models are effectively executed using libraries compatible with Apache Spark. To further enhance the robustness and generalizability of our model's performance evaluation, the system now incorporates additional benchmarking datasets in its simulation experiments. This integration expands the database to include not just holiday-related information specific to regions and countries but also other critical datasets, ensuring a comprehensive approach to anomaly detection. The integration of additional benchmarking datasets significantly fortifies the system's capabilities. This enhancement not only augments the robustness of the model but also substantially elevates its generalizability. This strategic inclusion ensures that the model's performance evaluation is not only comprehensive but also widely applicable across a variety of urban contexts. Such an expansion in the model's applicability and reliability marks a significant step forward in the field of urban traffic management and anomaly detection.

Several ML methods can be employed for prediction. Some studies opt for Naive Bayes (NB) [33], [34], while others favor discriminant random forest (DRF) [35], [36]. However, conducting a comparative study between these approaches is necessary, as the choice between them depends not on their inherent capabilities but rather on specific requirements.

Naive Bayesian classifier implementations fall within the family of linear classifiers, while DRF represents a more complex ML algorithm. DRF, which combines random subspaces and bagging concepts [37], operates through multiple decision trees generated from slightly different data subsets. These trees are calculated based on decision tree learning. Breiman's [38] improved method addresses issues like the sensitivity of individual trees to predictor order. It involves calculating a set of partially independent trees while considering congestion levels.

A decision tree (DT) serves as a decision support tool, visually resembling a tree with various possible decisions represented as branches or leaves. Decisions are made at each stage by following the path from the root to the corresponding leaf.

Further expanding the spectrum, Logistic Regression is a popular choice for binary classifications, such as distinguishing between normal and abnormal traffic conditions, due to its effectiveness in analyzing linear relationships [39]. For high-dimensional data, typical in traffic systems, Random Forest is a preferred choice. Its ability to handle various attributes like vehicle speed and flow through an ensemble of decision trees makes it robust for complex scenarios [40].

Gradient Boosting, another powerful model, stands out for its sequential error correction mechanism. This feature is particularly useful in traffic data where reducing bias and variance in predictions is crucial for accuracy [41], [42]. Support Vector Machine (SVM) thrives in scenarios with non-linear decision boundaries, common in complex traffic patterns. Its capability to handle high-dimensional data is essential for accurate anomaly detection [43].

K-Nearest Neighbors (KNN), with its instance-based learning approach, is valuable for identifying anomalies based on the proximity to historical traffic patterns. Lastly, neural network models like the Multi-layer Perceptron and Feedforward Neural Networks are excellent for their complex pattern recognition capabilities, a necessity in traffic systems where anomalies might be subtle and deeply embedded in high-volume data [44].

Each model offers unique strengths and can be tailored to specific aspects of traffic anomaly detection, depending on the data and characteristics of the traffic system under analysis.

In academic literature, the advantages of well-established anomaly detection models are commonly highlighted as follows:

- High performance:
  The Isolation Forest model [45] is a well-established model that uses decision trees for anomaly detection. In a comparative study conducted by [46], the Isolation Forest outperformed other models in anomaly detection across various datasets. Numerical results: The Isolation Forest achieved an average precision of 95% and an average recall of 93% on the datasets used in the study [46].
- Proven methodologies:
  The One-Class SVM model [47] is a well-established model used for anomaly detection. It is based on solid concepts in machine learning and has been extensively studied in the literature.
- Interpretability:
  Rule-based models, such as the association rule-based anomaly detection algorithm [48], provide easy interpretation of results. The identified association rules can be directly examined to understand specific patterns leading to anomaly detection.

Disadvantages of well-established anomaly detection models:

- Sensitivity to parameters:
  The k-means algorithm [49] is a well-established model used for clustering-based anomaly detection. However, it requires specifying the number of clusters (k) beforehand, which can be challenging to determine in some cases.
- Limited adaptability:
  The Local Outlier Factor (LOF) anomaly detection model [50] is well-established for spatial anomaly detection. However, it may not be suitable for detecting temporal anomalies or other types of data.
- Technological evolution:
  Models based on classical machine learning methods, such as feed-forward neural networks, may be outperformed by newer models based on deep learning architectures, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs) [51].

This study embarked on developing a system for processing real-time data, which was initially acquired in a structured format. The primary step involved segmenting this data, a crucial process that laid the groundwork for effective anomaly detection. The preliminary approach to predicting anomalies was anchored on the implementation of the EllipticEnvelope model, as cited in [52]. This model served as a foundational tool for establishing a normative pattern in traffic behavior, thereby facilitating the identification of anomalies.

Building upon this, the study further delved into the comparative analysis of eight diverse classifiers. These classifiers, each with their unique strengths and applications, included the Multi-layer Perceptron (MLP) as described by [53], the Support Vector Machine (SVM) detailed in [54], the Random Forest (RF) method based on [55], Gradient Boosting (GB) following [41], K-nearest Neighbors (KNN) as per [56], Logistic Regression (LR) outlined in [57], Decision Tree (DT) from [58], and Gaussian Naive Bayes as discussed in [59]. These classifiers were meticulously chosen and implemented to evaluate and compare their performance within the anomaly detection framework.

The first phase of the implementation focused on the application of various anomaly detection models to identify aberrations in traffic data. This phase commenced with

TABLE 2. Performance metrics of various classifiers for three datasets.

	Accident Dataset			Meteorological Dataset			Holiday Traffic Dataset		
Classifier	**Comp. time (s)**	**Accuracy (%)**	**AUC (%)**	**Comp. time (s)**	**Accuracy (%)**	**AUC (%)**	**Comp. time (s)**	**Accuracy (%)**	**AUC (%)**
Deep Learning (FNN)	1600.933	90.56	99.99	1461.951	85.99	99.99	481.728	74.70	99.84
Multi-layer Perceptron	41.265	99.65	99.99	38.677	99.43	99.99	21.005	99.12	99.94
SVM	25.436	98.91	99.99	28.047	98.69	99.99	11.299	98.17	99.93
Random Forest	8.171	99.59	99.99	7.702	99.05	99.98	5.729	98.65	99.89
Gradient Boosting	16.287	99.39	99.98	13.921	98.93	99.96	8.346	97.57	99.54
K-Nearest Neighbors	1.421	98.73	99.81	1.148	97.79	99.57	0.581	97.17	99.24
Logistic Regression	0.106	95.44	99.37	0.126	96.74	99.41	0.087	79.68	89.20
Decision Tree	0.443	99.40	99.45	0.350	98.49	98.71	0.295	97.68	97.55
Gaussian Naive Bayes	0.050	95.68	99.70	0.047	92.52	97.62	0.045	84.45	92.89

the loading and preprocessing of data, selecting pertinent numeric columns such as Timestep, Vehicle_ID, X, Y, Angle, Speed, Position, and Slope, and normalizing them using StandardScaler. A range of models, each chosen for their specific strengths in anomaly detection, were employed. These included the Elliptic Envelope, known for its effectiveness in assuming a Gaussian distribution and identifying outliers; the Isolation Forest, a tree-based model adept at handling multi-dimensional data; the Local Outlier Factor (LOF), an unsupervised algorithm that gauges the local deviation of a data point relative to its neighbors; the One-Class SVM, ideal for outlier detection in single-class datasets; and DBSCAN, a density-based clustering algorithm, augmented with PCA for dimensionality reduction.

The second phase shifted focus towards classification models, aiming to distinguish between normal and anomalous traffic data points. The same numeric columns were maintained, setting the 'EllipticLabel' from the Elliptic Envelope model as the target variable, and balancing the data using Random Under Sampler. In this phase, a spectrum of classifiers was implemented, including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, SVM, KNN, Gaussian Naive Bayes, MLP, and a deep learning model using Keras. Each model, except for the deep learning model, was incorporated into a pipeline with StandardScaler. The deep learning model was specially crafted, featuring dense layers with dropout for regularization. The evaluation of these models was conducted through cross-validation, and their effectiveness was further scrutinized based on ROC-AUC scores. These results were not only plotted in an ROC curve but were also methodically sorted based on AUC scores and systematically documented in an Excel file.

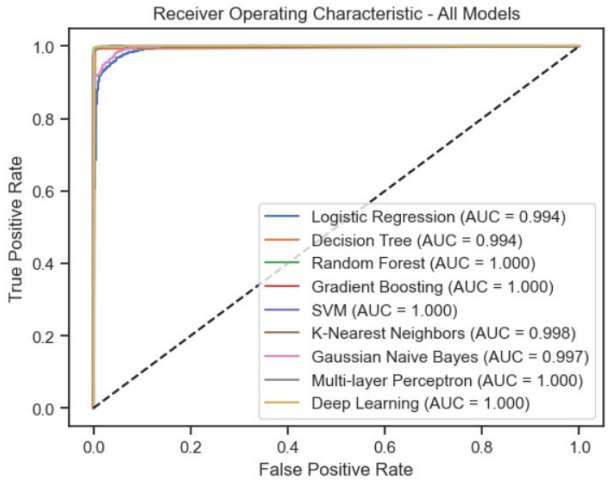
The study's insights and rationale are rooted in several key aspects. Feature Engineering played a pivotal role, as the selected features such as position, speed, angle, and others were critical in understanding traffic patterns and detecting anomalies. The diversity of models used, encompassing tree-based, clustering, SVM, and neural networks, provided a comprehensive and multifaceted approach to the problem. Hyperparameter tuning, particularly the setting of contamination rates and other model-specific parameters, was informed by a combination of experimentation and domain knowledge. Balancing the dataset was a crucial step, particularly important in anomaly detection due to the typically skewed nature of anomalies versus normal data. Lastly, the use of ROC-AUC as a metric was especially suitable for imbalanced datasets, ensuring a robust evaluation of model performance.

Table 2 presents the performance metrics of eight machine learning classifiers across three datasets: Accident Dataset, Meteorological Dataset, and Holiday Traffic Dataset. These metrics include computation time, accuracy, and area under the curve (AUC). Multi-layer Perceptron (MLP) demonstrates the highest accuracy across all datasets, reaching 99.65% for the Accident Dataset, but it also requires significantly longer computation times, for instance, 41.265 seconds for the same dataset. Random Forest (RF) shows impressive predictive capabilities with an AUC of 99.99% in both the Accident and Meteorological Datasets while maintaining relatively lower computation times, such as 8.171 seconds for the Accident Dataset. Gaussian Naive Bayes (GNB) is the fastest, with computation times as low as 0.050 seconds, but it compromises accuracy, particularly in the Holiday Traffic Dataset at 84.45%. Logistic Regression and K-Nearest Neighbors also exhibit notably short computation times, but their accuracy and AUC are comparatively lower. The decision regarding the optimal classifier should therefore consider the specific trade-offs between computational efficiency and predictive accuracy, as different classifiers offer varying balances between these aspects.

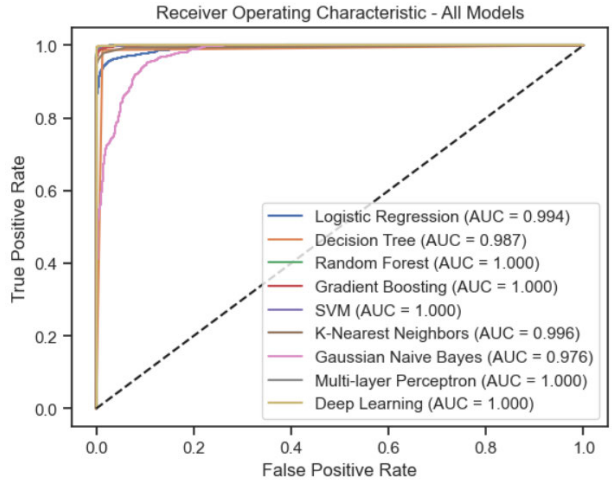
Figure 6 displays the receiver operating characteristic (ROC) curves for all models, providing a visual representation of their performance in terms of true positive rates against false positive rates. The three Receiver Operating Characteristic (ROC) curves provided represent the performance of various classification models on three different datasets: accident, meteorological, and holiday traffic. Here's a brief analysis of each:

i) Accident Dataset ROC Curve:

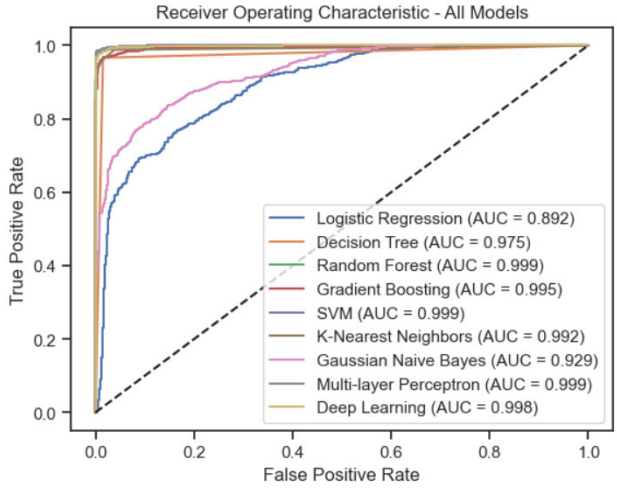
- The models Random Forest, Gradient Boosting, SVM, Multi-layer Perceptron, and Deep Learning all achieve an Area Under Curve (AUC) of 1.000, indicating perfect classification with no overlap between the positive and negative classes.
- The Decision Tree and K-Nearest Neighbors models perform slightly less perfectly, with AUC scores very close to 1.000.
- Gaussian Naive Bayes has the lowest AUC among the models but still performs very well.



(a) accident dataset



(b) meteorological dataset



(c) holiday traffic dataset

FIGURE 6. Receiver operating characteristic (ROC) curves for all models on different datasets.

ii) Meteorological Dataset ROC Curve:

- Once again, most models show an AUC of 1.000, demonstrating excellent classification capabilities.
- Logistic Regression, with an AUC of 0.994, and Gaussian Naive Bayes, with an AUC of 0.976, show slightly lower but still high performance

iii) Holiday Traffic Dataset ROC Curve:

- The Multi-layer Perceptron stands out with the highest AUC at 0.999, closely followed by Random Forest and SVM at 0.999, and K-Nearest Neighbors at 0.992.
- Gaussian Naive Bayes shows a lower AUC of 0.929, indicating that it may not perform as well on this dataset compared to the others.
- Deep Learning, while still high, has a marginally lower AUC of 0.998 compared to its perfect score on the other two datasets.

In all curves, the closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test. The dotted line represents a no-skill classifier; for example, a classifier that predicts the positive class 50% of the time, regardless of the actual class.

The AUC provides a single measure of overall accuracy that is independent of a specific classification threshold and is often used to compare different classifiers. An AUC of 1.0 represents a perfect model that makes no false positive or false negative predictions. As the AUC decreases towards 0.5, the model's performance is no better than random chance. All models here perform significantly better than chance, indicating strong predictive capabilities across the datasets.

IV. RESULTS AND DISCUSSION

This section delves deeper into the simulation results, highlighting their critical role in validating the proposed real-time anomaly detection and load-balancing methods for urban traffic management. These simulations, meticulously crafted to mirror the intricate and representative traffic conditions of Casablanca, particularly highlight the city's notable traffic density and unique urban layout, embodying the typical challenges faced in metropolitan areas.

Utilizing the flexible and open-source SUMO (Simulation of Urban MObility) tool, the simulations adeptly recreated complex urban traffic scenarios, encapsulating the dynamic interactions among different elements such as vehicles, pedestrians, and urban infrastructure. SUMO's ability to function effectively without extensive calibration, while also offering tailored customizations for specific scenarios, was instrumental in accurately simulating Casablanca's unique traffic conditions. This approach provided valuable insights into the effectiveness and practicality of the proposed traffic management solutions.

To ensure realism in simulation, a traffic flow pattern consistent with peak and off-peak hours, as typically observed in urban environments, was adopted. The simulation model underwent thorough calibration to align with actual traffic

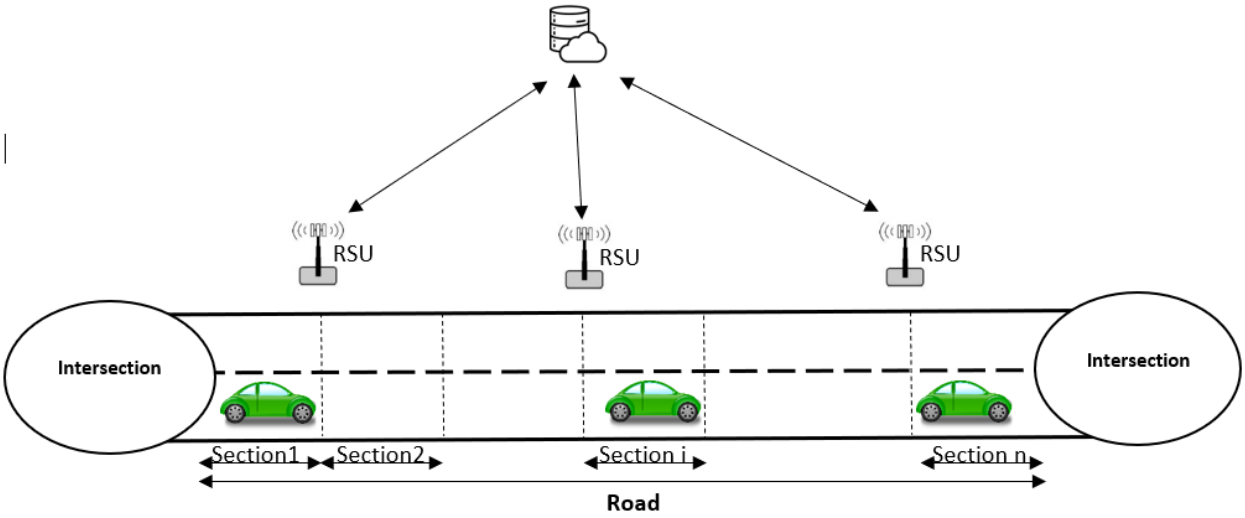


FIGURE 7. VANET elements and road segmentation.

TABLE 3. Vehicle route assigned by the proposed methodology.

Time	Vehicle ID	Source	Path	Duration
14 :00	1	(1,3)	(1,3)-> (1,4) -> (3,4) -> (3,5) -> (5,6) -> (5,7) -> (8,1) -> (9,5)	50 min
14 :02	2	(1,3)	(1,3)-> (1,4) -> (3,4) -> (3,5) -> (5,6) -> (5,7) -> (8,1) -> (9,5)	52 min
14 :18	3	(1,3)	(1,3)-> (1,4) -> (3,7) -> (5,7) -> (6,8) -> (6,9) -> (8,1) -> (9,5)	56 min
14 :21	4	(1,3)	(1,3) -> (1,4) -> (3,7) -> (5,7) -> (6,8) -> (6,9) -> (8,1) -> (9,5)	57 min

data, thus enhancing the reliability and applicability of the results.

To add a layer of realism to the simulations, traffic flow patterns were designed to reflect typical peak and off-peak hours observed in urban environments. The simulation model was meticulously calibrated against actual traffic data, thereby enhancing the reliability and relevance of the results.

To demonstrate the proposed approach, simulations were conducted on a representative urban map. Subsequently, the road network was subdivided into multiple sections, as depicted in Figure 7.

To demonstrate the real-time performance of this approach and its effectiveness in maintaining an up-to-date database, simulations involving four vehicles with identical source and destination points were conducted. The road sections used in the simulation are summarized in Table 3. Vehicle "1," for example, commences its journey on Road "1," Section "1," aiming to reach its destination on Road "9," Section "3." The route taken by each vehicle during its trip is defined by a pair of values (x, y), where "x" represents the road identifier and "y" signifies the section of Road "x."

Table 4 provides insights into various traffic condition variations within the VANET network, reflecting real-time modifications in the database. When Vehicle 2 enters the scenario immediately after Vehicle 1 at 14:02, it aims to reach the same destination as Vehicle 1. As there have been no changes in the database during this short interval, the system assigns the same route to the second vehicle.

However, when Vehicle 3 joins the scenario at 14:18, it is observed that the system assigns a different route compared to Vehicles 1 and 2. This decision is influenced by the elevated traffic density in Section V of Route 3, which could be attributed to factors such as traffic jams or accidents. As a result, the traffic density in (Route 3, Section V) exceeds its previous levels, prompting the system to update the database for improved efficiency. Consequently, Vehicle 3 selects an alternative route distinct from those chosen by Vehicles 1 and 2.

In Table 3, vehicle ID1 starts at 14:00, after 2 minutes, the density increase, since vehicle ID2 starts at 14:02, so the duration is not the same. The use of qualitative indicators such as "low," "medium" and "very high" for density rather than quantitative values depends on the nature of the data. It is a common practice to use qualitative descriptors for simplicity, also qualitative indicators can be more action-oriented and decision-focused. The goal is to guide the system toward appropriate actions which means choosing the way or not. The article suggests that safety is the primary focus, but the implemented traffic management approach incorporates measures that not only enhance safety but also contribute to the efficiency of the transportation system. The efficiency improvements mentioned aim to achieve smoother traffic operations without compromising safety, creating

TABLE 4. An extract of the database containing all road sections density.

(Road, Section)	Vehicle density at 14:00	Vehicle density at 14:10	Vehicle density at 14:20
(1,3)	13	20	27
(1,4)	10	13	11
(1,5)	17	16	20
(2,2)	9	8	10
(2,3)	7	6	7
(3,7)	14	14	16
(4,1)	12	14	15
(3,5)	13	35	30
(7,3)	24	23	21

TABLE 5. Status of different road sections at T and T+ 15min.

time	Road section	Density at T	Density at T+15min
14:00	(1,3)	Low	Medium
14:00	(3,5)	Medium	Very high
14:10	(1,5)	Medium	Low
15:00	(2,5)	Low	Low

a balanced approach to Intelligent Transportation System design.

Additionally, the most recent route can be observed, which is the third entry in the table. This route has a time duration of fifty-six minutes, making it significantly more efficient than the previous routes assigned to vehicle IDs one and two. After an anomaly occurred, the time duration for the first two routes increased to seventy-one minutes. This exemplifies the effectiveness of the proposed mechanism in preventing congestion and reducing the likelihood of accidents.

The proposed mechanism excels in providing users with safe routes to reach their destinations. Unlike methods discussed in the existing literature, this approach does not rely on accident or congestion predictions. Instead, it provides real-time insights into the current traffic conditions based on vehicle density and real-time travel time. When high density or anomalies are detected in specific road sections, the system promptly reroutes subsequent vehicles to less congested roads, ensuring efficient traffic flow.

To maintain load balance, the mechanism intelligently allocates vehicles to various road sections. When a road becomes saturated, the system automatically redirects vehicles to less congested alternatives. This load-balancing effect is demonstrated in Table 5, where road section (1,3) experiences congestion due to an influx of vehicles, but the system avoids using it in subsequent routes, allowing the congestion to dissipate over time. Similarly, road section (3,5) experiences increased vehicle density due to an anomaly, prompting the system to temporarily avoid it until it returns to normal conditions. These continuous updates to the database enhance the system's accuracy in estimating vehicle density and arrival times, guiding vehicles along the most efficient paths.

The proposed architecture consists of centralized batch data storage, processing techniques, and a distributed data storage mechanism for real-time analysis to manage and process vehicle flow at specific locations. Lambda architecture has been leveraged, incorporating Apache Storm for real-time analysis to ensure rapid data processing. Additionally, Hadoop MapReduce is employed to process the massive data used to construct the database, which serves as input to the speed layer, facilitating real-time processing.

The experimental results and analysis not only confirm but emphatically illustrate that the proposed system, with its seamless integration of advanced Big Data technologies, is indeed an optimal solution for near real-time data processing in Intelligent Transportation Systems within a vehicular ad-hoc environment. This system excels in its ability to rapidly process and analyze the vast and complex data generated by urban traffic networks, showcasing a level of efficiency and accuracy that is pivotal for effective traffic management in real-time scenarios.

The core strength of this system lies in its primary focus on road safety, a critical aspect that is often relegated to secondary importance in traditional traffic management strategies. By prioritizing safety over the mere reduction of travel time, this approach fundamentally differentiates itself from previous methodologies. This safety-first perspective is not merely a theoretical stance but is deeply embedded in the system's design and functionality. Through advanced anomaly detection techniques and dynamic load management, the system actively identifies potential hazards and efficiently manages traffic flow to mitigate risks, thereby significantly enhancing the safety of road users.

Furthermore, the emphasis on safety is bolstered by the system's sophisticated use of machine learning algorithms, which work in tandem with Big Data technologies to predict traffic patterns and preemptively address potential congestion and accidents. This proactive approach to traffic management, underpinned by real-time data analysis, ensures that the system is not only reactive but also anticipatory in its strategy, setting a new benchmark in intelligent transportation solutions.

The traffic management improvement approach enhances efficiency. It incorporates a real-time anomaly detection system employing parallel data processing for swift execution. The primary objective of this methodology is to

accurately gauge the time spent crossing each section of all roadways, facilitating efficient traffic control, precise vehicle arrival estimations, and the provision of the safest routes to destinations.

In comparison to 'Proactive Traffic Safety Management (PATM)' [21], this approach excels at effectively managing larger systems and ensuring scalability. PATM achieves exceptional real-time performance but is constrained in scalability. The integration of multiple factors into PATM can lead to slower system responses, particularly in high-density scenarios. This limitation arises from the fixed classification approach employed in PATM, making it less adaptable for real-time incident management. Conversely, this approach, founded on probabilistic principles, conveys information to incident management personnel in an intuitive and informative manner. The proposed study presents a comprehensive traffic management approach, leveraging real-time anomaly detection and load balancing for urban traffic improvement. It contrasts with [24]'s method that emphasizes traffic flow prediction through focused predictive modeling. The application of comprehensive, real-time data integration and machine learning techniques signifies an important enhancement in urban traffic management, extending beyond the scope of [26], by offering wider applications for immediate adjustments and infrastructure enhancement.

The results underscore the efficiency of the applied real-time anomaly detection and load balancing techniques, as evidenced by notable improvements in traffic flow and a reduction in congestion within the simulated environments. This efficacy aligns with the initial objectives, reinforcing the potential of these methods to significantly enhance urban traffic management. The results also demonstrate the system's capability to manage large volumes of data generated by various entities within the traffic ecosystem, further underscoring its applicability in contemporary urban settings.

The proposed model offers several advantages from both theoretical and practical perspectives:

- Real-time Anomaly Detection: The model incorporates a real-time anomaly detection system that accurately computes vehicle density for each section at any given time. This enables precise traffic management and provides vehicles with information on traffic density and the safest route to their destination. By detecting anomalies in real-time, the model can quickly identify and address any unexpected events or abnormal traffic patterns, allowing for more efficient traffic management.
- Load Balancing: The model includes load balancing techniques, which help distribute traffic evenly across different sections and routes. By balancing the load, the model can alleviate congestion in heavily congested areas and optimize the overall traffic flow. This leads to reduced travel times, improved efficiency, and a more balanced utilization of road infrastructure.
- Machine Learning-based Prediction: The model incorporates a machine learning-based prediction system to mitigate congestion problems and reduce accident risks. By analyzing historical traffic data and patterns, the model can make accurate predictions about future traffic conditions. This allows for proactive decision-making and the implementation of measures to prevent congestion and accidents before they occur.
- Low Latency and High Precision: The simulations conducted in the study demonstrate that the proposed model effectively addresses transportation issues while maintaining low latency and high precision. This means that the model can process and analyze data quickly, providing real-time insights and actionable information to traffic management systems and individual vehicles promptly. The high precision ensures accurate predictions and reliable decision-making.
- Integration of Big Data: The model takes advantage of big data techniques and technologies to handle the massive amounts of data generated by different entities in the traffic environment. By efficiently processing and analyzing big data, the model can extract valuable insights, identify patterns, and make informed decisions for effective traffic management.

Furthermore, to comprehensively evaluate the computational cost of models, it's crucial to delve into the specifics of the data processing and analysis techniques employed. The models utilize high-dimensional data, requiring intensive computational power for real-time anomaly detection and dynamic load balancing. The implementation of machine learning algorithms, particularly deep learning, significantly increases computational demands due to the need for training on large datasets and the continuous updating of models to adapt to new data. Furthermore, the adoption of the Lambda architecture facilitates the handling of massive data streams efficiently, allowing for both batch and real-time processing. This architecture aids in mitigating computational costs by optimizing data flow and processing tasks, ensuring that the system remains scalable and responsive to the real-time demands of urban traffic management. This detailed approach highlights the intricate balance between achieving high performance in traffic analysis and managing the computational resources effectively to maintain system efficiency and reliability.

V. LIMITATIONS

Despite the advancements and contributions of this study in enhancing urban traffic management through real-time anomaly detection and load balancing, it is essential to recognize certain inherent limitations:

- **Scalability and Generalizability**: The effectiveness of the developed system in diverse urban settings and across various scales remains a subject for further exploration. The models and algorithms may require

adaptation to suit different traffic densities and urban layouts.

- **Data Dependence**: The system's performance heavily relies on the quality and timeliness of the traffic data. Inconsistencies or gaps in data collection could significantly impact the system's accuracy and reliability.
- **Dynamic Traffic Complexities**: The simulations, while crucial, might not fully capture the unpredictability and complexity of real-world traffic scenarios. This includes factors such as human behavior, weather conditions, and unplanned road incidents.
- **Technological Integration**: Future integration of more sophisticated machine learning techniques poses challenges in ensuring seamless compatibility and maintaining the efficiency of the system.
- **Security and Privacy**: Operating within a Vehicle Ad-Hoc Network (VANET) environment raises significant data security and privacy concerns that need to be addressed with robust measures.
- **Future Development**: The plan for future enhancement of the system, especially in terms of integrating advanced machine learning techniques, indicates an ongoing process to address and possibly overcome some of these limitations.

This acknowledgment of limitations is vital for a balanced understanding of the study's scope and for guiding future research directions.

VI. CONCLUSION

In conclusion, this study addresses the challenges posed by the voluminous data generated within the VANET environment, emphasizing the critical role of big data technologies in effectively harnessing valuable insights from this data. Road traffic management, a longstanding field of study aimed at ensuring the safety of both vehicles and pedestrians, benefits significantly from these advancements.

The research has contributed to the field by establishing a real-time anomaly detection system that operates efficiently with parallel data processing. This approach allows for faster execution times and ensures the timely identification of potential traffic anomalies. The core innovation lies in the system's ability to accurately compute vehicle density for each road section within urban and highway networks. This granular density information empowers the system to optimize traffic management by providing vehicles with real-time road conditions and suggesting the safest routes to their destinations.

Furthermore, the work addresses the critical issue of accident prediction and congestion prevention through the integration of a machine learning framework. By leveraging machine learning, the system enhances traffic safety by proactively identifying congestion-prone areas and potential accident risks.

The outcomes of the simulations underscore the effectiveness of the system. It significantly reduces congestion and, consequently, the likelihood of accidents. Notably, the results showcase a remarkable balance between low latency and high precision, making the system a robust tool for traffic management in VANET environments.

Future research plans include enhancing the developed database of vehicle density per road section by integrating advanced machine learning techniques for more precise traffic management solutions. Additionally, optimization strategies such as model pruning, quantization, and the use of efficient data structures, along with leveraging cloud and edge computing resources, will be explored to reduce computational costs while maintaining system efficacy. This approach aims to balance computational resource management with the operational performance of urban traffic management systems, ensuring real-time functionality and economic viability.

REFERENCES

[1] Z. Balaž, "Intelligent transport systems (ITS) for sustainable mobility," UNECE, Geneva, Switzerland, Tech. Rep., p. 123, Feb. 2012.
[2] O. O. Ajayi, A. B. Bagula, H. C. Maluleke, and I. A. Odun-Ayo, "Transport inequalities and the adoption of intelligent transportation systems in Africa: A research landscape," *Sustainability*, vol. 13, no. 22, p. 12891, 2021.
[3] A. Mohandu and M. Kubendiran, "Survey on big data techniques in intelligent transportation system (ITS)," *Mater. Today, Proc.*, vol. 47, pp. 8–17, 2021.
[4] A. Ahmed and B. Aijaz, "A case study on the potential applications of V2V communication for improving road safety in Pakistan," *Eng. Proc.*, vol. 32, no. 1, p. 17, 2023.
[5] S. Kaleem, A. Sohail, M. U. Tariq, and M. Asim, "An improved big data analytics architecture using federated learning for IoT-enabled urban intelligent transportation systems," *Sustainability*, vol. 15, no. 21, p. 15333, Oct. 2023.
[6] Y. Lin, P. Wang, and M. Ma, "Intelligent transportation system(ITS): Concept, challenge and opportunity," in *Proc. IEEE 3rd Int. Conf. Big Data Secur. Cloud (Bigdatasecurity) Int. Conf. High Perform. Smart Comput. (hpsc), IEEE Int. Conf. Intell. Data Secur. (ids)*, May 2017, pp. 167–172.
[7] S. K. John, D. Sivaraj, and R. Mugelan, "Implementation challenges and opportunities of smart city and intelligent transport systems in India," in *Internet of Things and Big Data Analytics for Smart Generation* (Intelligent Systems Reference Library), vol. 154, V. Balas, V. Solanki, R. Kumar, and M. Khari, Eds. Cham, Switzerland: Springer, 2019, pp. 213–235, doi: 10.1007/978-3-030-04203-5_10.
[8] G. Abdelkader, K. Elgazzar, and A. Khamis, "Connected vehicles: Technology review, state of the art, challenges and opportunities," *Sensors*, vol. 21, no. 22, p. 7712, Nov. 2021.
[9] F. D. Da Cunha, A. Boukerche, L. Villas, A. C. Viana, and A. A. Loureiro, "Data communication in VANETs: A survey, challenges and applications," *Ad Hoc Networks*, vol. 44, pp. 90–103, Jul. 2014.
[10] M. Gillani, H. A. Niaz, M. U. Farooq, and A. Ullah, "Data collection protocols for VANETs: A survey," *Complex Intell. Syst.*, vol. 8, no. 3, pp. 2593–2622, Jun. 2022.
[11] S. Mihai, N. Dokuz, M. S. Ali, P. Shah, and R. Trestian, "Security aspects of communications in VANETs," in *Proc. 13th Int. Conf. Commun. (COMM)*, Jun. 2020, pp. 277–282.
[12] M. Arif, G. Wang, O. Geman, V. E. Balas, P. Tao, A. Brezulianu, and J. Chen, "SDN-based VANETs, security attacks, applications, and challenges," *Appl. Sci.*, vol. 10, no. 9, p. 3217, May 2020.
[13] A. Santamaria, M. Tropea, P. Fazio, and F. De Rango, "Managing emergency situations in VANET through heterogeneous technologies cooperation," *Sensors*, vol. 18, no. 5, p. 1461, May 2018.
[14] F. Z. Bousbaa, C. A. Kerrache, N. Lagraa, R. Hussain, M. B. Yagoubi, and A. E. K. Tahari, "Group data communication in connected vehicles: A survey," *Veh. Commun.*, vol. 37, Oct. 2022, Art. no. 100518.
[15] Q. Wang, W. Li, Z. Yu, Q. Abbasi, M. Imran, S. Ansari, Y. Sambo, L. Wu, Q. Li, and T. Zhu, "An overview of emergency communication networks," *Remote Sens.*, vol. 15, no. 6, p. 1595, Mar. 2023.

[16] Y. Soni and S. Jangirala, "Survey on vehicular cloud computing and big data," *EasyChair Preprint*, p. 13, Aug. 2022. [Online]. Available: https://easychair.org/publications/preprint/WSkl
[17] D. Wang, Y. Huang, and Z. Cai, "A two-phase clustering approach for traffic accident black spots identification: Integrated GIS-based processing and HDBSCAN model," *Int. J. Injury Control Saf. Promotion*, vol. 30, no. 2, pp. 270–281, Apr. 2023.
[18] S. Ullah, G. Abbas, M. Waqas, Z. H. Abbas, and A. U. Khan, "RSU assisted reliable relay selection for emergency message routing in intermittently connected VANETs," *Wireless Netw.*, vol. 29, no. 3, pp. 1311–1332, Apr. 2023.
[19] Y. Zou, L. Ding, H. Zhang, T. Zhu, and L. Wu, "Vehicle acceleration prediction based on machine learning models and driving behavior analysis," *Appl. Sci.*, vol. 12, no. 10, p. 5259, May 2022.
[20] D. Lee, D. Camacho, and J. J. Jung, "Smart mobility with big data: Approaches, applications, and challenges," *Appl. Sci.*, vol. 13, no. 12, p. 7244, Jun. 2023.
[21] M. Abdel-Aty, O. Zheng, Y. Wu, A. Abdelraouf, H. Rim, and P. Li, "Real-time big data analytics and proactive traffic safety management visualization system," *J. Transp. Eng., A, Syst.*, vol. 149, no. 8, Aug. 2023, Art. no. 04023064.
[22] B. Medina-Salgado, E. Sánchez-DelaCruz, P. Pozos-Parra, and J. E. Sierra, "Urban traffic flow prediction techniques: A review," *Sustain. Comput., Informat. Syst.*, vol. 35, Sep. 2022, Art. no. 100739.
[23] H. Amari, Z. A. E. Houda, L. Khoukhi, and L. H. Belguith, "Trust management in vehicular ad-hoc networks: Extensive survey," *IEEE Access*, vol. 11, pp. 47659–47680, 2023.
[24] I. O. Olayode, A. Severino, T. Campisi, and L. K. Tartibu, "Prediction of vehicular traffic flow using Levenberg–Marquardt artificial neural network model: Italy road transportation system," *Komunikácie*, vol. 24, no. 2, pp. 74–86, Mar. 2022.
[25] R. Qaddoura and M. B. Younes, "Temporal prediction of traffic characteristics on real road scenarios in amman," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 7, pp. 9751–9766, Jul. 2023.
[26] M. B. Younes, "Real-time traffic distribution prediction protocol (TDPP) for vehicular networks," *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 8, pp. 8507–8518, Aug. 2021.
[27] A. Kaul and I. Altaf, "Vanet-TSMA: A traffic safety management approach for smart road transportation in vehicular ad hoc networks," *Int. J. Commun. Syst.*, vol. 35, no. 9, Jun. 2022.
[28] A. Muñoz Hernández, D. Scarlatti, and P. Costas, "Real-time estimated time of arrival prediction system using historical surveillance data," in *Proc. 45th Euromicro Conf. Softw. Eng. Adv. Appl. (SEAA)*, Aug. 2019, pp. 174–177.
[29] W. Li, M. Batty, and M. F. Goodchild, "Real-time GIS for smart cities," *Int. J. Geographical Inf. Sci.*, vol. 34, no. 2, pp. 311–324, Feb. 2020.
[30] N. O. Alsrehin, A. F. Klaib, and A. Magableh, "Intelligent transportation and control systems using data mining and machine learning techniques: A comprehensive study," *IEEE Access*, vol. 7, pp. 49830–49857, 2019.
[31] Z. Ji, Y. Wang, K. Yan, X. Xie, Y. Xiang, and J. Huang, "A space-embedding strategy for anomaly detection in multivariate time series," *Exp. Syst. Appl.*, vol. 206, Nov. 2022, Art. no. 117892.
[32] M. Hu, X. Feng, Z. Ji, K. Yan, and S. Zhou, "A novel computational approach for discord search with local recurrence rates in multivariate time series," *Inf. Sci.*, vol. 477, pp. 220–233, Mar. 2019.
[33] W. Budiawan, "Traffic accident severity prediction using Naive Bayes algorithm—A case study of semarang toll road," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 598, no. 1, Aug. 2019, Art. no. 012089.
[34] E. B. Anitha, R. Aravinth, S. Deepak, R. Jotheeswari, and G. Karthikeyan, "Prediction of road traffic using Naive Bayes algorithm," *Int. J. Eng. Res. Technol*, vol. 7, no. 1, pp. 1–4, 2019.
[35] A. Kumar and N. Sinha, "Classification of forest cover type using random forests algorithm," in *Proc. ICDIS Adv. Data Inf. Sci.* Singapore: Springer, 2020, pp. 395–402.
[36] V.-H. Dang, N.-D. Hoang, L.-M.-D. Nguyen, D. T. Bui, and P. Samui, "A novel GIS-based random forest machine algorithm for the spatial prediction of shallow landslide susceptibility," *Forests*, vol. 11, no. 1, p. 118, Jan. 2020.
[37] G. Kunapuli, *Ensemble Methods for Machine Learning*. New York, NY, USA: Simon & Schuster, 2023.
[38] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.
[39] O. H. Kwon, W. Rhee, and Y. Yoon, "Application of classification algorithms for analysis of road safety risk factor dependencies," *Accident Anal. Prevention*, vol. 75, pp. 1–15, Feb. 2015.
[40] L. Cheng, X. Chen, J. De Vos, X. Lai, and F. Witlox, "Applying a random forest method approach to model travel mode choice behavior," *Travel Behaviour Soc.*, vol. 14, pp. 1–10, Jan. 2019.
[41] Y. Zhang and A. Haghani, "A gradient boosting method to improve travel time prediction," *Transp. Res. C, Emerg. Technol.*, vol. 58, pp. 308–324, Sep. 2015.
[42] X. Wen, Y. Xie, L. Jiang, Z. Pu, and T. Ge, "Applications of machine learning methods in traffic crash severity modelling: Current status and future directions," *Transp. Rev.*, vol. 41, no. 6, pp. 855–879, Nov. 2021.
[43] P. H. Tran, A. Ahmadi Nadi, T. H. Nguyen, K. D. Tran, and K. P. Tran, "Application of machine learning in statistical process control charts: A survey and perspective," in *Control Charts and Machine Learning for Anomaly Detection in Manufacturing*. Cham, Switzerland: Springer, 2022, pp. 7–42.
[44] M. Abbasi, A. Shahraki, and A. Taherkordi, "Deep learning for network traffic monitoring and analysis (NTMA): A survey," *Comput. Commun.*, vol. 170, pp. 19–41, Mar. 2021.
[45] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Min.*, 2008, pp. 413–422.
[46] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation-based anomaly detection," *ACM Trans. Knowl. Discovery from Data*, vol. 6, no. 1, pp. 1–39, Mar. 2012, doi: 10.1145/2133360.2133363.
[47] B. Schölkopf, R. C. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, "Support vector method for novelty detection," in *Advances in Neural Information Processing Systems*, S. Solla, T. Leen, and K. Müller, Eds., vol. 12. Cambridge, MA, USA: MIT Press, 1999.
[48] C. C. Aggarwal, "High-dimensional outlier detection: the subspace method," in *Outlier Analysis*. Cham, Switzerland: Springer, 2017, pp. 149–184, doi: 10.1007/978-3-319-47578-3_5.
[49] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 129–137, Mar. 1982.
[50] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*. New York, NY, USA: Association for Computing Machinery, May 2000, pp. 93–104, doi: 10.1145/342009.335388.
[51] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[52] M. Belichovski, D. Stavrov, F. Donchevski, and G. Nadzinski, "Unsupervised machine learning approach for anomaly detection in E-coating plant," in *Proc. IEEE 17th Int. Conf. Control Autom. (ICCA)*, Jun. 2022, pp. 992–997.
[53] R. Kruse, S. Mostaghim, C. Borgelt, C. Braune, and M. Steinbrecher, "Multi-layer perceptrons," in *Computational Intelligence: A Methodological Introduction*. Cham, Switzerland: Springer, 2022, pp. 53–124.
[54] M. Pal and P. M. Mather, "Support vector machines for classification in remote sensing," *Int. J. Remote Sens.*, vol. 26, no. 5, pp. 1007–1011, Mar. 2005.
[55] R. Couronné, P. Probst, and A.-L. Boulesteix, "Random forest versus logistic regression: A large-scale benchmark experiment," *BMC Bioinf.*, vol. 19, no. 1, pp. 1–14, Dec. 2018.
[56] X. He, P. Wang, and J. Cheng, "K-nearest neighbors hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 2839–2848.
[57] T. G. Nick and K. M. Campbell, "Logistic regression," in *Topics Biostatistics*, 2021, pp. 273–301.
[58] Y. Y. Song and Y. Lu, "Decision tree methods: Applications for classification and prediction," *Shanghai Arch. Psychiatry*, vol. 27, no. 2, p. 130, Apr. 2015.
[59] M. Ontivero-Ortega, A. Lage-Castellanos, G. Valente, R. Goebel, and M. Valdes-Sosa, "Fast Gaussian Naïve Bayes for searchlight classification analysis," *NeuroImage*, vol. 163, pp. 471–479, Jul. 2017.

MY DRISS LAANAOUI received the Ph.D. degree in computer science from Cadi Ayyad University, Morocco, in 2014. He is currently a Professor of computer science with the Normal Superior School of Marrakech, Cadi Ayyad University. His research interests include routing protocols in MANET and VANET environments, cybersecurity, artificial intelligence, machine learning, and deep learning.

MOHAMED LACHGAR received the engineering degree in computer science from ENSIAS, Université Mohammed V, in 2009, and the Ph.D. degree in computer science from Cadi Ayyad University, in 2017. He is currently an Associate Professor of computer science with the National School of Applied Sciences, Chouaib Doukkali University, El Jadida, Morocco. His research interests include automation tools in embedded software, software modeling and design, metamodel design, model transformation, and methods for model verification and validation, as well as machine learning and deep learning.

HANINE MOHAMED received the Ph.D. degree in computer science, specializing in spatial decision-making from Cadi Ayyad University, Marrakesh, Morocco, in 2017. In 2018, he joined the Department of Telecommunications, Networks, and Computer Science, National School of Applied Sciences, where he educates engineering students in the fields of big data, NoSQL, and business intelligence. He is currently an Associate Professor with the National School of Applied Sciences, Chouaib Doukkali University, El Jadida, Morocco. His research interests include big data, multicriteria decision-making, NoSQL, and business intelligence.

HRIMECH HAMID received the Ph.D. degree in computer science from Art et Métiers ParisTech, France, in 2009. He is currently a Full Professor with the Department of Computer Sciences and Mathematics, ENSA, Hassan First University, Morocco. His research interests include artificial intelligence, collaborative virtual environments, and driving simulation.

SANTOS GRACIA VILLAR received the industrial engineering degree specializing in energy techniques and the Ph.D. degree in industrial engineering from the Polytechnic University of Catalonia. He is currently the Director of the Master's Degree in Design, Management and Project Management. He is an Expert in international cooperation projects.

IMRAN ASHRAF received the M.S. degree (Hons.) in computer science from the Blekinge Institute of Technology, Karlskrona, Sweden, in 2010, and the Ph.D. degree in information and communication engineering from Yeungnam University, Gyeongsan, South Korea, in 2018. He was a Postdoctoral Fellow with Yeungnam University. He is currently an Assistant Professor with the Information and Communication Engineering Department, Yeungnam University. His research interests include positioning using next-generation networks, communication in 5G and beyond, location-based services in wireless communication, smart sensors (LIDAR) for smart cars, and data analytics.