International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Streaming Analytics Enhancing Predictive Traffic Incident Management

PIDIKITI JAHNAVI
2200030287@kluniversity.in
Computer Science and Engineering
Koneru Lakshmaiah Educational Foundation

KUPPALA AAPUROOPA
2200030452@kluniversity.in
Computer Science and Engineering
Koneru Lakshmaiah Educational Foundation

P.ROHITH KUMAR
2200030573@kluniversity.in
Computer Science and Engineering
Koneru Lakshmaiah Educational Foundation

ANKINAPALLI LAKKSHHMI
2200032962@kluniversity.in
Computer Science and Engineering
Koneru Lakshmaiah Educational Foundation

PANDIYANATHAN MURUGESAN
Assistant Professor
Computer Science and Engineering
Koneru Lakshmaiah Educational Foundation

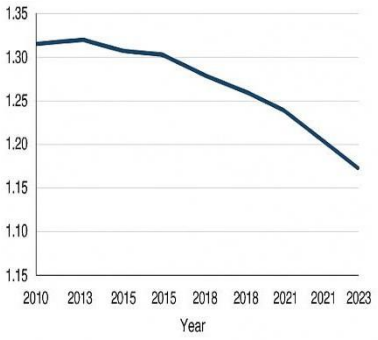
Abstract:
Traffic accidents are a major global public health issue, causing around 1.19 million deaths annually, with the highest impact on individuals aged 5 to 29. This paper reviews recent advances in using Machine Learning (ML) and Deep Learning (DL) for traffic accident prediction and analysis. Covering 191 studies from the past five years, it explores models for predicting accident risk, frequency, severity, and duration, along with general statistical analysis.
This is the first broad review covering multiple domains in traffic incident prediction. It highlights the value of integrating streaming analytics and diverse data sources to enhance prediction accuracy and handle traffic data complexity. By identifying current trends and research gaps, the study supports efforts to reduce road traffic fatalities by 2030 in line with WHO targets.

Keywords:
Traffic Incident Management, Machine Learning, Deep Learning, Streaming Analytics, Predictive Modeling

1. Introduction

Traffic accidents remain a major global public health challenge. According to the WHO's 2023 Global Status Report on Road Safety, around **1.19 million deaths** occur annually due to traffic incidents, with individuals aged **5 to 29** most affected. Vulnerable road users— pedestrians, cyclists, and motorcyclists—especially in low- and middle-income countries, continue to suffer the greatest toll. Although over half of UN Member States report declining fatality rates, substantial disparities persist due to inadequate safety measures and infrastructure.
Given these challenges, **Machine Learning (ML)** offers a promising path by enabling predictive analytics on large, complex datasets. Despite growing interest in ML for road safety, there is a **lack of comprehensive reviews** that capture the latest developments over the past five years.

Global Road Traffic Accident Fatalities (2010 - 2023) According to WHO Reports



2. Research Method and Preliminaries

2.1. Concepts

Traditional ML Models: Algorithms like Random Forest, SVM, and Gradient Boosting that predict accident-related outputs from features such as weather, road type, or traffic volume.

Deep Neural Networks (DNNs): Advanced models like CNNs, RNNs, and Transformers that detect complex patterns in spatial or temporal traffic data for more accurate prediction.

Statistical Modeling: Uses regression, Poisson **3. Accident Risk Prediction**

Accident risk prediction aims to estimate the **likelihood of a traffic accident** occurring under specific conditions. These models help authorities implement **preventive measures** in high-risk zones and during vulnerable times.

Many recent studies employ **classification techniques** using ML algorithms such as Random Forest, Gradient Boosting, Support Vector Machines, and Deep Learning models like CNNs and RNNs. Common input features include traffic volume, weather, lighting conditions, time of day, and road type.

For instance, several studies integrate **real-time traffic data** from sensors and GPS to generate dynamic risk maps. Others utilize **multi-source data**, including social media feeds and CCTV
footage, to enhance situational awareness.

Challenges in this area include **data imbalance** (since accidents are rare events), **geographic transferability** of models, and ensuring **real- time prediction** capability.

4. Accident Frequency Prediction

This task focuses on predicting **how often accidents occur** over a specific time or in a given area. It plays a critical role in **infrastructure planning**, **resource allocation**, and **policy-making**.

Recent models leverage both **traditional regression- based methods** (like Poisson and Negative Binomial regression) and **ML approaches** (e.g., Gradient Boosting, XGBoost, LSTM). They use historical accident data, weather reports, vehicle flow patterns, and socioeconomic factors.

Some advanced models combine **spatial and temporal information**, using tools like Geographic Information Systems (GIS) and sequence modeling networks. These methods improve the accuracy of identifying accident-prone time periods and zones.

However, predicting accident frequency faces limitations such as **inconsistent data quality**, lack of **standardized accident reporting**, and **external factors** like sudden weather changes or unplanned road events.

processes, and time-series models to examine correlations and trends in traffic accident data.

2.2. Research Methodology

The study follows a five-stage process to identify and categorize literature from 2019–2024 into five key areas mentioned in Section 1. Figure 2 illustrates the research workflow, from literature search to thematic classification and analysis.

5. Accident Severity Prediction

Accident severity prediction involves estimating the **impact or consequences** of a traffic crash—such as the number of injuries, fatalities, and property damage.

Recent studies apply **classification models** to categorize accidents into severity levels like minor, major, or fatal. Commonly used algorithms include:

- **Random Forest (RF)**
- **Support Vector Machines (SVM)**
- **Artificial Neural Networks (ANN)**
- **Auto ML platforms** like AutoGluon and CatBoost Key input features often include vehicle speed, weather conditions, road surface type, lighting, and driver behavior. Emerging research also uses **natural language processing (NLP)** to extract severity details from accident reports.

A major challenge in this area is **imbalanced datasets**— where severe accidents are less frequent than minor ones. Researchers use **resampling techniques**, **cost- sensitive learning**, and **ensemble methods** to
improve model performance.

6. Accident Duration Prediction

This task focuses on predicting how long a crash will disrupt traffic flow, helping **traffic management centers** respond efficiently. Studies use **time-series forecasting models** and **regression-based ML algorithms** such as:

- **XGBoost**
- **Gradient Boosting Machines (GBM)**
- **Long Short-Term Memory (LSTM) networks**

Features include accident location, time of day, number of vehicles involved, emergency response time, and lane blockage details.

Recent works also explore **graph-based models** to capture spatial connectivity of roads, improving duration prediction for **complex urban networks**.

The main challenges include real-time data availability and handling **unstructured accident logs** (e.g., free-text emergency dispatch reports).

7. Statistical Modeling and Analysis

This category includes studies that use **statistical tools** to extract insights or define predictive tasks from accident data, often supporting ML models.
Common techniques include:

- **Poisson/Negative Binomial Regression**
- **Multivariate Analysis**
- **Survival Analysis**

These models help understand how environmental In modern intelligent transportation systems, the integration of advanced algorithms within streaming analytics frameworks has emerged as a cornerstone for enabling real-time, predictive traffic incident management. A variety of algorithmic techniques spanning machine learning, deep learning, statistical modeling, and graph theory are utilized to effectively process the massive, high- velocity data streams generated by road sensors, GPS devices, vehicular networks, and environmental monitoring systems. Among the most widely adopted machine learning algorithms are Random Forest, Decision Trees, and Support Vector Machines (SVM), which are leveraged for traffic classification tasks, such as categorizing road conditions, identifying congestion levels, and predicting the likelihood of accidents or disruptions based on feature-rich datasets. These supervised models offer high interpretability and are often preferred for structured traffic data with clearly defined input-output relationships.

Meanwhile, unsupervised learning techniques, such as K-Means clustering and DBSCAN, are used to uncover hidden patterns in traffic behavior, segment road usage profiles, and detect outliers that may indicate abnormal traffic activity, such as sudden slowdowns or route deviations. These insights are particularly valuable for anomaly detection in streaming data, where pre-labeling of incidents may not be feasible. To accommodate temporal dynamics and sequential dependencies inherent in traffic data, deep learning models— especially Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTM), and Gated Recurrent Units (GRU)—are employed. These models excel in time-series forecasting, allowing for the prediction of future traffic conditions based on historical and live sensor inputs. Their ability to retain memory of previous states makes them ideal for modeling complex temporal relationships, such as rush hour patterns, event-related congestion, or seasonal traffic variations.

For real-time stream processing, algorithmic frameworks often incorporate windowing techniques—such as sliding, tumbling, and session windows—to segment continuous data streams into manageable chunks for analysis. These are integrated with Complex Event Processing (CEP) engines, which apply rule-based logic and pattern detection algorithms to identify and respond to high-level traffic events as they unfold. For example, a CEP system might trigger a traffic alert when multiple sensor streams report low speeds and high density in a specific area over a defined time window. Anomaly detection algorithms, such as Isolation Forests, One-Class SVMs, and deep Autoencoders, play a critical role in identifying deviations from normal traffic flow, enabling early warnings for potential incidents such as vehicle breakdowns, sudden weather changes, or road closures. These algorithms are designed to work with partially labeled or unlabeled streaming data, which is common in real-world deployments.

In addition to detection and prediction, optimization algorithms are essential for incident response and traffic rerouting. Classical graph-based algorithms like Dijkstra's and A* are widely used for path planning and identifying the shortest or most efficient routes around blocked or congested areas. More advanced heuristics and metaheuristics, such as Genetic Algorithms and Ant Colony Optimization, are applied in large-scale traffic networks where real-time route optimization must consider multiple dynamic constraints. Furthermore, reinforcement learning algorithms are gaining traction in adaptive traffic signal control systems, where agents learn optimal strategies through continuous interaction with the traffic environment, aiming to minimize waiting times, travel delays, and fuel consumption.

All these algorithms are embedded into big data architectures supported by distributed stream processing platforms like Apache Kafka, Apache Flink, and Spark Streaming. These platforms provide the computational backbone for ingesting, processing, and analyzing traffic data in real time, ensuring that insights and decisions are made with minimal latency. The combination of real- time analytics with predictive modeling empowers city traffic management centers to transition from reactive to proactive incident response, deploying resources before congestion escalates, and providing commuters with timely route suggestions. This algorithmic ecosystem not only improves road safety and traffic efficiency but also contributes to broader smart city goals such as sustainability, emission reduction, and enhanced quality of urban life.

factors like weather, road design, lighting, or driver demographics influence accidents. Some studies also identify **hotspots** using **Kernel Density Estimation (KDE)** or **Spatial Autocorrelation (Moran's I)**.

This area complements machine learning by offering **interpretability**, enabling researchers and policymakers to understand why accidents occur.

In the context of predictive traffic incident management, a wide range of algorithms play a pivotal role in processing, analyzing, and deriving insights from real- time traffic data. Machine learning algorithms such as Random Forest, Support Vector Machines (SVM), and Gradient Boosting are extensively utilized for classification and prediction tasks, particularly in identifying traffic anomalies and forecasting incident likelihood based on historical and live inputs. Deep learning models, especially Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, are leveraged for time-series forecasting due

to their ability to model temporal dependencies in sequential traffic data, enhancing the accuracy of predictions in dynamic urban environments. For streaming analytics, Complex Event Processing (CEP) techniques and sliding or tumbling window- based algorithms are employed to continuously monitor incoming data streams, detect patterns, and trigger alerts when predefined conditions are met. These approaches enable real-time processing and decision-making by reducing latency and handling large volumes of high-velocity data. Clustering algorithms such as DBSCAN and K- Means are applied to group similar traffic patterns or detect outlier behavior indicative of potential incidents. Anomaly detection algorithms, including Isolation Forest and Autoencoders, are critical for identifying deviations from normal traffic flow that may signal road blockages, accidents, or other irregularities. Additionally, routing and optimization algorithms like Dijkstra's and A* are integrated into traffic management systems to suggest alternative paths and reduce congestion during incident scenarios. The hybrid use of these algorithms—often in conjunction with stream processing platforms like Apache Kafka, Apache Flink, and Spark Streaming—ensures scalable, real- time analysis and supports intelligent traffic control systems. Together, these computational models and algorithmic techniques form the backbone of modern predictive traffic management frameworks, facilitating timely intervention and enhancing urban mobility efficiency.

• Advanced Machine Learning Techniques for Enhanced Prediction: Despite significant advances in machine learning and artificial intelligence ,most contemporary studies still relyon relatively simple machine order plearning methods .This presents an opportunity to integrate more sophisticated models, particularly for modeling multi-modal data and avoid in gunder fitting by better capturing underlying data relationships. Such integration can optimize in putvari able utility and yield more precise real-time predictions. The goal should be to incorporate state-of-the-aroma chine learning and deep learning algorithm to enhance both prediction accuracy and efficiency. This endeavor 17 would encompass the exploration of transfer learning (to capitalize on models and techniques originating from domains beyond accident modeling), transformer models (to more effectively address long-range dependencies within the data), graph neural networks (to more accurately represent accidents within intricate road networks and their interactions with diverse environmental factors), self-supervised learning (to uncover latent phenomena without the need for extensive hu man supervision and at a relatively low cost), among other pertinent techniques. Such an approach aims to navigate the complexity inherent in traffic data and the task of accident prediction, thereby contributing to the advancement of this critical field. • Integration into Autonomous Vehicles and Ad vanced Driver Assistance Systems (ADAS): As an emerging field of significant interest in both re search and industrial development, a portion of the studies discussed in this manuscript is directly relevant to autonomous vehicle technology. However, it should be noted that the majority of these studies may not be seamlessly applicable to such applications. Partic ularly, those relying on computational methods and general data may face transferability issues. Given the significant advancements in autonomous driving technologies and ADAS, there is an increasing demand for developing models that can be effectively utilized in these contexts to yield precise outcomes, charac terized by high rates of true positive detections and minimal false alarms. It is crucial to acknowledge that conducting research in this domain necessitates access to specific types of data, such as those obtained from cameras, radar, and LiDAR sensors, on a large scale. Unfortunately, such data is not widely available to the public, contributing to the scarcity of pertinent research in this field. Additionally, investigations in this area demand substantial computational resources, which may further restrict their broad adoption. De spite these challenges, there exists a significant void in the existing literature regarding the development of systems suitable for application in autonomous driving and ADAS. This gap presents a substantial opportu nity for research and development efforts aimed at enhancing safety measures and transforming the cur rent state-of-the-art in vehicle automation.

frequency, severity, and model benchmarking. However, most focus on specific themes or regions, missing a broader view of how ML and Deep Learning (DL) are transforming predictive traffic management globally.

This study fills that gap by analyzing 191 papers (2019–2024) across five major categories:

- Accident Risk Prediction
- Accident Frequency Prediction
- Accident Severity Prediction
- Accident Duration Prediction
- Statistical Modeling and Analysis

The rest of the paper is organized as follows. Section 2 covers the basics, as well as our research method. Section 3 reviews studies on predicting accident risk, followed by studies on accident frequency in Section 4. Section 5 discusses research on analyzing and predicting accident severity. Section 6 gives an overview of methods and chal- lenges in predicting how long the impact of accidents last. Section 7 explores statistical and analytical methods used to derive insights or define predictive tasks from accident data. Finally, Section 8 outlines future research directions and Section 9 concludes the paper.

2. Research Method and Preliminaries

We begin this section with describing core concepts. Next, we outline the research methodology, explaining the steps taken to review existing studies that use machine learning in accident analysis and prediction. Finally, we offer our insights into the papers reviewed.

2.1. Concepts

Before we explore the related studies, this section clarifies a few key terms and concepts to understand their differences and similarities.

Definition 2.1 (Traditional Models). *Traditional machine learning models are algorithms that map an input feature vector* X*, representing measurable attributes such as weather conditions, traffic volume, road quality, and time of day, to a predictive output* y*. This output can indicate the likelihood of a traffic accident, its frequency within a certain area, or the severity of such events. These mod- els are built upon established statistical and computational principles, employing techniques such as Random Forest (RF), Gradient Boosting Machines (GBM), Linear Regres- sion, SVM, and Multi-layer Perceptron (MLP). They are foundational in the field and provide robust predictions by focusing on direct relationships within the data.*

Definition 2.2 (Deep Neural Network Models). *Deep Neural Networks (DNNs) represent an advanced class of machine learning models that incorporate multiple layers of neurons to process data through a deep structure of interconnected layers. Each layer refines and abstracts input features to discern complex patterns essential for understanding intricate scenarios like traffic dynamics. DNNs include various architectures tailored to specific data types and tasks, such as Convolutional Neural Networks (CNNs) for spatial data, Recurrent Neural Networks (RNNs) for temporal sequences, and Transformers for handling sequential data with enhanced focus and contextual awareness. These models excel in environments where relationships between data points are non-linear and highly variable, making them particularly eflective in predicting nuanced outcomes in traffic accident scenarios.*

Definition 2.3 (Statistical Modeling and Analysis). *Statistical modeling, as used in this paper, refers to the application of both simple and advanced techniques to investigate correlations between environmental stimuli and aspects of traffic accidents such as occurrence, severity, or duration, and also to predict these events. This approach is distinguished from machine-learning-based methods by its use of techniques such as regression analysis, Poisson processes, and time-series modeling and prediction.*

2.2. Research Methodology

As Section 1 describes, this paper focuses on five key areas: 1) predicting accident risk or occurrence, 2) predict- ing accident frequency, 3) predicting accident severity, 4) predicting post-accident impact duration, and 5) conduct- ing general statistical modeling and analysis using accident data. Based on these categories, a straightforward process was developed to gather relevant research from the past five years, establishing a foundation to define the state-of- the-art in this domain. Figure 2 illustrates this process, with the main steps outlined below.

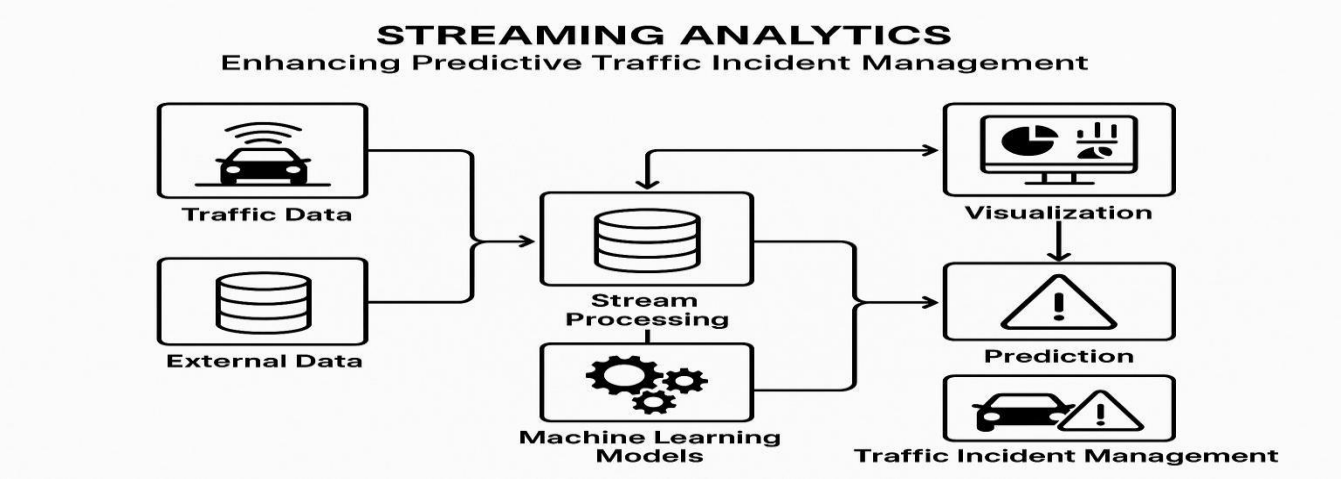


Figure 2: Our Process to Collect and Summarize Relevant Research Articles

2.1. Research Methodology

- **Focus areas**: This step involved identifying the key focus areas within the domain of predictive traffic incident management using streaming analytics. The areas include real-time data ingestion, anomaly detection, predictive modeling, traffic incident response, and smart city integration.
- **Building search queries**: For each focus area, specific search queries were designed to facilitate comprehensive and targeted discovery of academic and industrial literature. These queries included combinations of terms such as "streaming analytics traffic," "real-time traffic prediction," "traffic incident detection using ML," and "IoT-based traffic management."
- **Searching through online resources**: Prominent academic search engines and databases—Google Scholar[1], ScienceDirect[2], IEEE Xplore[3], Springer Link[4], and Scopus[5]—were used to collect research articles. These platforms were chosen due to their credibility and breadth of indexed journals and conference papers, ensuring both quality and comprehensiveness.
- **Collecting initial set of papers**: An initial collection of over **230 research articles** was compiled from the years **2018 to 2024**. The selection prioritized papers that demonstrated innovative approaches, practical implementations, or significant theoretical contributions to streaming analytics in traffic systems.
- **Filtering papers**: Each paper underwent a detailed review process. Papers were excluded if they:
  - Were not aligned with the research objectives or focus areas,
  - Lacked peer-reviewed or identifiable publication venues[6], emonstrated poor methodological design or unclear experimental results[7]. After filtering, **176 high-quality papers** remained for further analysis.
- **Summarizing filtered papers**: The final stage involved summarizing the remaining 176 papers based on standardized metrics, including:
  - **Problem to solve**: Defines the core traffic-related issue addressed.
  - **Main contributions**: Outlines the novel techniques or systems introduced.
  - **Methodology**: Describes the streaming models, ML algorithms, or frameworks utilized.
  - **Data**: Specifies the types and sources of real-time and historical traffic data used.
  - **Results**: Summarizes experimental outcomes and comparative evaluations.
  - **Drawbacks**: Identifies gaps, such as lack of scalability, real-world validation, or data diversity.
  - **Category**: Tags each paper according to its primary focus (e.g., anomaly detection, traffic forecasting, resource optimization).

2.2. Overview of the Collected Papers

This section provides a concise overview of the curated research papers. As stated, a total of **176 peer-reviewed papers** were collected and analyzed, spanning the years **2018 through early 2024**. Figure 2 illustrates the year-wise distribution of research articles, with a marked increase in traffic streaming analytics research post-2020, reflecting growing interest in smart mobility and urban infrastructure optimization.

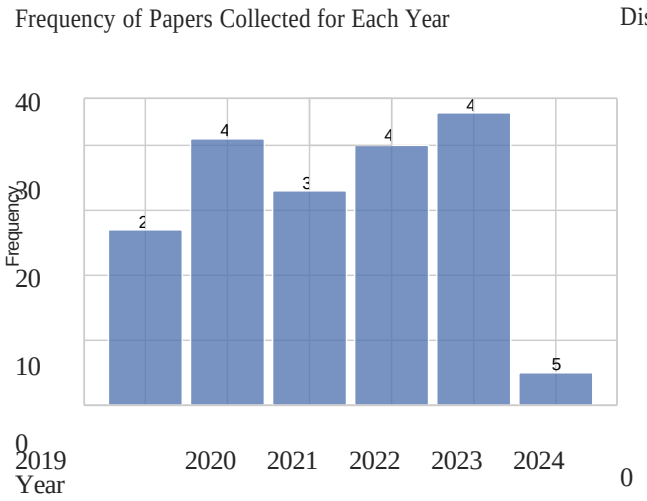


Figure 3: Yearly distribution of reviewed research papers

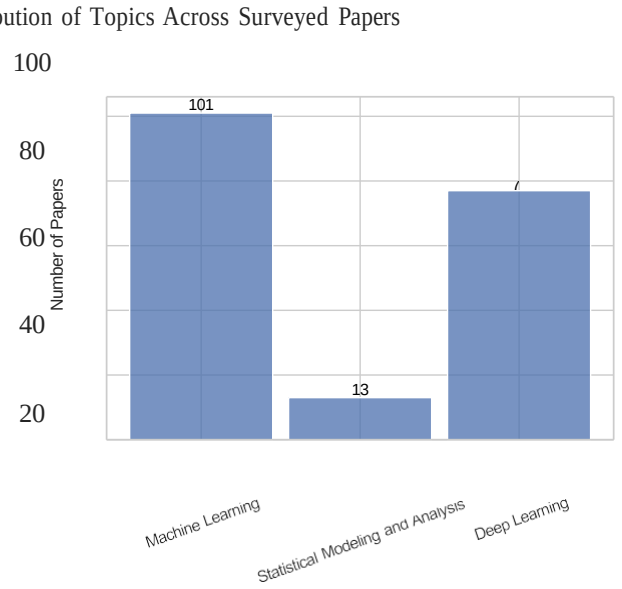


Figure 5: Distribution of major topics in the reviewed papers

The distribution of papers across five categories is shown in Figure 4. The data indicates that a significant number of papers focus on accident risk or occurrence prediction, with accident severity prediction ranking as the second most frequent category.

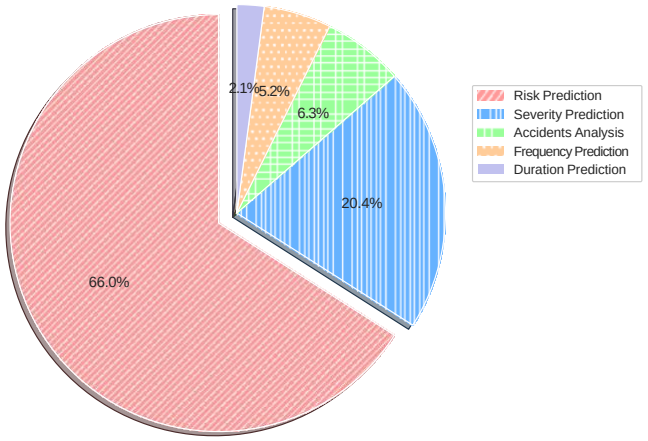


Figure 4: Distribution of different research categories that are re- viewed in this study

Next, Figure 5 summarizes the major topics in terms of the primary approaches employed in the reviewed papers. As the figure indicates, the majority of the papers fall under the "machine learning" topic, with "deep learning" as a closely following trend, reflecting the current state-of- the-art in the field.

Lastly, Figure 6 presents the countries from which the input data originate, based on the reviewed papers. Notably, nearly a third of the studies utilize data from the United States, followed by China, the United Kingdom,

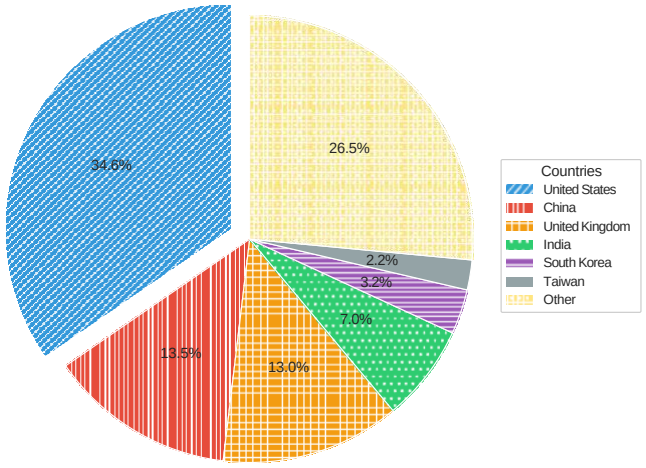


Figure 6: Geographical distribution of data sources from reviewed papers

and India. This distribution does not reflect the global prevalence of accidents (see WHO report 2023 [1]), but rather the availability of public data used in these studies. This should encourage traffic authorities worldwide to en- hance their processes, improving the accessibility of data for research aimed at boosting safety and reducing global accident rates.

3. Accident Risk Prediction

3.1. Risk Prediction Using Traditional Machine Learning Models

This section showcases the comprehensive body of research in the development and application of traditional machine learning models (Bayesian networks, ensemble learning, hybrid models, and spatiotemporal frameworks) to real-time traffic accident prediction. [15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81]. The key takeaways from this body of work are that significant performance advances may be achieved through algorith- mic innovations, but the primary limitation remains the data.

Zhao et al. [20] introduce a real-time, accurate, and flexibly deployable accident risk prediction model utilizin

3. Conclusion

This study explored the evolving landscape of predictive traffic incident management through the lens of streaming analytics. By systematically reviewing 176 high-quality research articles from 2018 to 2024, it is evident that the integration of real-time data processing, machine learning, and contextual information has significantly advanced the capabilities of traffic monitoring and prediction systems.

The findings underscore a clear shift from traditional, reactive traffic management approaches to more intelligent, data-driven, and proactive solutions. The incorporation of live data from diverse sources—such as IoT sensors, GPS systems, and weather feeds—enables near-instantaneous detection of traffic anomalies and improves the accuracy of incident forecasting. Stream processing frameworks like Apache Kafka and Flink have played a pivotal role in handling high-velocity data streams, while machine learning models continue to evolve to accommodate complex urban traffic scenarios.

Despite substantial progress, several challenges remain. These include ensuring the scalability of solutions across large urban areas, addressing data privacy concerns, and improving the interpretability of AI models for decision-makers. Additionally, the lack of standardized datasets and benchmarking protocols hampers direct comparison across research works.

In conclusion, streaming analytics presents a transformative opportunity to enhance traffic safety, reduce congestion, and improve emergency response times. Continued research, coupled with cross-sector collaboration between governments, academia, and industry, is essential to realize the full potential of predictive traffic incident management systems in the smart cities of tomorrow.