**Group: FTL\_Myanmar\_Gr10**

Road Traffic Accident Prediction

# **Members**

1. Thu Thu Win
2. Ye Htet Naing
3. Unt Ko
4. Sai Shang Hlang
5. Chaw Nu Htun

# Literature Review

## Introduction

Road traffic accidents represent a critical global challenge for many years, resulting in significant loss of precious human life, serious injuries that have large impact on remaining life and substantial socio-economic costs. The importance of this research lies in developing proactive, data-driven systems that can accurately predict accident severity and, more crucially, identify the underlying risk factors to enable targeted interventions. And also review of the existing literature is necessary to synthesise the current state of machine learning applications in this field. It is important for understanding the evolution of methodologies, evaluating the strengths and limitations of various algorithms, and identifying gaps. Its direct relevance to our project goal is to establish a foundational knowledge that will support us in selecting an optimal machine learning approach for our project, ensuring our project contributes meaningfully to the current and next generation over the world.

## Organization

* **Theme 1: General Overview and Comprehensive Reviews**
  + *Recent Advances in Traffic Accident Analysis and Prediction: A Comprehensive Review of Machine Learning Techniques:* Noushin Behboudi, Sobhan Moosavi, and Rajiv Ramnath
* **Theme 2: Model Development and Hybrid Approaches**
  + *Traffic Crash Severity Prediction—A Synergy by Hybrid Principal Component Analysis and Machine Learning Models:* Khaled Assi
* **Theme 3: Traditional ML Approaches for Accidents Prediction**
  + *Prediction of Road Accidents Using Machine Learning Algorithms:*R.Vanitha & M.Swedha
* **Theme 4: Explainable and Interpretable ML Models**
  + *A study on road accident prediction and contributing factors using explainable machine learning models: analysis and performance :* Shakil Ahmed, Md Akbar Hossain, Sayan Kumar Ray, Md Mafjull Islam Bhuiyan, Saifur Rahman Sabuj

## Summary

***Recent Advances in Traffic Accident Analysis and Prediction: A Comprehensive Review of Machine Learning Techniques:*** Noushin Behboudi, Sobhan Moosavi, and Rajiv Ramnath

This comprehensive review analyzes 191 studies from 2019-2024 to map the state-of-the-art in machine learning for traffic accident analysis and prediction. It systematically examines methods for predicting accident risk, frequency, severity, and duration, finding that advanced models like Random Forest, CNN, and LSTM significantly outperform traditional approaches, especially when integrating diverse data sources. The study's key contribution lies in synthesizing recent advancements, identifying critical gaps like data imbalance and limited model interpretability, and providing a forward-looking roadmap for future research to enhance global road safety and reduce the 1.19 million annual traffic fatalities.

**Key findings:** This comprehensive review synthesizes findings from 191 studies published between 2019-2024, revealing that machine learning and deep learning models significantly enhance the prediction accuracy of traffic accident risk, frequency, severity, and duration. Models such as Random Forest, XGBoost, CNN, LSTM, and transformer-based architectures consistently outperform traditional statistical methods, particularly when integrating multi-source data like traffic flow, weather, social media, and satellite imagery. The review also highlights persistent challenges, including data imbalance, model interpretability, and regional biases; nearly one-third of studies relied on U.S. data, limiting global applicability.

**Methodology:** The authors conducted a systematic review focusing on five key areas: accident risk, frequency, severity, duration, and statistical analysis. They collected over 250 papers from major databases like Google Scholar, IEEE Xplore, and Scopus, applying strict inclusion criteria to retain 191 high-quality studies. Each paper was categorized by prediction task, methodology (traditional ML, deep learning, or statistical modeling), data sources, and limitations. The review also included comparative analysis with earlier surveys to highlight trends and advancements over the past five years.

**Contribution to the Field:** This study is the first to offer a holistic and up-to-date review of ML applications across multiple dimensions of traffic accident analysis. It provides a clear mapping of state-of-the-art techniques, identifies critical research gaps, such as the need for scalable, diverse, and interpretable models, and proposes forward-looking directions, including integration with autonomous vehicles, real-time alert systems, and cross-regional model adaptation. By consolidating recent advances and outlining a roadmap for future work, the review serves as a foundational resource for researchers, policymakers, and practitioners aiming to reduce global traffic accidents and align with WHO safety targets.

***Traffic Crash Severity Prediction—A Synergy by Hybrid Principal Component Analysis and Machine Learning Models:*** Khaled Assi

This research developed a hybrid system combining Principal Component Analysis (PCA) with machine learning models, such as Multilayer Perceptron Neural Networks (MLP-NN) and Support Vector Machines (SVM), to predict road traffic crash severity. Using crash data from Victoria, Australia, the study demonstrated that applying PCA for dimensionality reduction significantly boosted the models' performance, increasing testing accuracy to over 80% for both MLP-NN and SVM. The key contribution is a practical, high-accuracy prediction tool that can help trauma centers anticipate injury severity and improve emergency response, ultimately enhancing road safety outcomes.

**Key findings:** The study found that applying PCA before modeling significantly enhanced the performance of both MLP-NN and SVM. Specifically, testing accuracy for MLP-NN increased from 64.50% to 82.70%, and for SVM from 62.70% to 80.70%. The first nine principal components, which accounted for 67% of the cumulative variance, were identified as the most significant. Key original features highly correlated with these components included crash type, road surface condition, traffic control type, driver’s gender and age, vehicle type, and road geometry. The hybrid PCA-ML models demonstrated superior performance compared to many existing approaches in the literature.

**Methodology:** The research utilized a dataset of 37,774 traffic crashes from Victoria, Australia (2014–2019). After preprocessing, PCA was applied to reduce the dimensionality of the 24 original crash-related attributes and address multicollinearity. The Kaiser-Meyer-Olkin (KMO) test and Bartlett's test confirmed the dataset's suitability for PCA. Two machine learning models, such as MLP-NN and SVM, were then developed and trained in two scenarios: first using the original attributes, and second using the principal components. The models were evaluated using a 70/30 train-test split, with performance assessed based on classification accuracy and F1 score.

**Contribution to the Field:** This research makes a significant contribution by demonstrating that PCA can effectively enhance the prediction accuracy of machine learning models in traffic safety analytics. The proposed hybrid methodology addresses common challenges such as dataset heterogeneity and multicollinearity, offering a more robust and practical tool for predicting crash severity. The high-accuracy models provide actionable insights that can help trauma centers prepare appropriate medical responses in advance, potentially saving lives and optimizing resource allocation. The study also sets a precedent for using dimensionality reduction techniques to improve the performance of predictive models in transportation and public health.

***Prediction of Road Accidents Using Machine Learning Algorithms:***R.Vanitha & M.Swedha

This research focuses on predicting road accidents using machine learning to improve traffic safety and reduce fatalities. The study aims to build an accident prediction model using algorithms like Decision Tree, Random Forest and Logistic Regression. The model seeks to uncover the relationships between environmental and human factors that influence these dangerous accidents.

**Key findings:**The study found that the Random Forest algorithm was the most effective , achieving an accuracy of 86.86% in predicting accident severity, slightly outperforming Logistic Regression which has an accuracy of 86.23%. An important discovery was the issue of class imbalance, where models struggled to predict severe but rare accidents, often defaulting to the most common “slight injury” class. The most crucial factors influencing severity were environmental, such as visibility, humidity, and wind direction, while data analysis also revealed that most accidents involved drivers aged 25-35 and occurred on Thursday although it can vary based on volume of traffic on a certain day.The study also discovered that accidents are more frequent after midday.

**Methodology:**The research employed a standard machine learning pipeline on a dataset of 1.6 million UK accident records. After collecting and cleaning the data, the team performed extensive pre-processing , including visualization and normalization of key features. They then trained and evaluated three classification models which are Logistic Regression, Decision Tree and Random Forest by using performance metrics like accuracy, precision and recall. They also conducted hyperparameter tuning to determine if the performance of three models could be further optimized.

**Contribution:**This research clearly established Random Forest’s superiority for this task on a real-world dataset. Not only it importantly highlights the practical challenge of class imbalance, guiding future research to address the prediction of severe but also suggesting future improvement like incorporating with Google Maps which would support the police to track. Furthermore, it provides actionable insights for developing real-time alert systems for police , to enable proactive road safety measures.

***A study on road accident prediction and contributing factors using explainable machine learning models: analysis and performance :*** Shakil Ahmed, Md Akbar Hossain, Sayan Kumar Ray, Md Mafjull Islam Bhuiyan, Saifur Rahman Sabuj

This research aims not only to predict the severity of road accidents but, more importantly, to identify and explain the key factors that make accidents severe or fatal. It is beyond a simple predictive model to provide actionable insights for improving road safety.

**Key Findings:**The Random Forest model was the most accurate for prediction with accuracy of 81.45% outperforms other models, Decision Jungle, ADABoost, XGBoost,L-GBM, CatBoost. The most critical factors causing severe accidents are road category, number of vehicles involved and drug and alcohol involvement. In addition, elderly drivers are more prone to severe multi-vehicle accidents.

**Methodology:**A large dataset of New Zealand accidents from the year of 2016 to the year of 2020 was used. The authors trained and compared six different machine learning models. The most innovative part of their method was the use of Explainable AI (XAI), specifically SHAP analysis, to “open the black box” of their best model and understand which factors were most influential in its decisions.

**Contribution:**This paper clearly demonstrates the power of Explainable AI in road safety research. Providing data-driven, evidence-based insights in the root cause of severe accidents allows policy makers and traffic engineers to proceed beyond generic safety measures and focus resources on the most critical areas.

These papers share the same goal of leveraging machine learning to improve road safety through accident severity prediction with Random Forest consistently emerging as the best performing algorithm across multiple studies. However, they are different significantly in their methodological focus and primary contribution. The research by Vanitha and Swedha and Ahmed et al. provide foundational case studies, with the latter making a key advance by using Explainable AI(XAI) to identify critical factors like road type and drug use. Assi’s paper introduces a different approach from them, using PCA to improve model accuracy. In contrast, the Behboudi et al. review synthesizes these trends, validating the success of these methods while highlighting common challenges like data imbalance that the other papers encounter.

## Conclusion

The literature demonstrates that machine learning algorithms, particularly Random Forest, XGBoost, and Support Vector Machines, are highly effective for predicting road accident risk and severity, with each model offering distinct advantages in accuracy, interpretability, and handling complex, high-dimensional data. Hybrid approaches, such as combining PCA with ML models, enhance performance by reducing dimensionality and addressing multicollinearity, while explainable AI techniques, like SHAP analysis, provide insights into the key factors driving severe accidents, including road type, curvature, speed limits, lighting, weather, and driver-related factors. This research is important because road accidents remain a major global health and safety concern, and accurate prediction enables data-driven, proactive interventions that can save lives and optimize traffic management. Building on the reviewed studies, your project contributes to the existing body of knowledge by developing an interpretable, high-performing, and context-aware road accident risk prediction system that integrates environmental and road factors, addresses challenges like data imbalance and model transparency, and provides actionable insights for policymakers and traffic safety authorities, ultimately advancing the practical application of machine learning in real-world accident prevention.

# 

# Data Research

**Organization**

This data research is organized into five sections:

1. **Introduction** — provides background and objectives.
2. **Data Description** — explains the dataset, sources, and variables.
3. **Data Analysis and Insights** — presents findings and patterns.
4. **Conclusion** — summarizes insights and relevance.
5. **Citations** — lists references used in the research at the end.

The analysis follows a **thematic structure**, focusing first on data understanding, then feature exploration, and finally model interpretation.

## Introduction

Road accidents are a major public health and safety concern, causing millions of injuries and deaths each year. Predicting accident risks under different road and environmental conditions can help authorities design better safety measures, allocate resources, and reduce fatalities.

The main research question in this study is:

**“Can machine learning models accurately predict accident risk based on road, weather, and environmental factors?”**

A thorough exploration of data is necessary to understand the underlying relationships among various factors—such as road curvature, lighting, and speed limit—that influence accident occurrence. By analyzing this data, we can uncover hidden patterns and build predictive models that contribute to more effective traffic management and accident prevention strategies.

## 

## Data Description

The dataset used in this research represents **road and environmental conditions** influencing accident likelihood. It contains both categorical and numerical variables, which help in classifying accident risks.

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| road\_type | Type of road (highway, local, rural, etc.) |
| num\_lanes | Number of lanes on the road |
| curvature | Level of road curvature (straight, curved, sharp turn) |
| speed\_limit | Legal speed limit (in km/h) |
| lighting | Lighting condition (daylight, streetlight, poor visibility) |
| weather | Weather conditions (clear, rainy, foggy, etc.) |
| road\_signs\_present | Indicates whether proper traffic signs are installed |
| public\_road | Whether the road is public or private |
| time\_of\_day | Time when the data was recorded (morning, afternoon, night) |
| holiday | Whether the day was a holiday (yes/no) |
| school\_season | Whether the data was collected during school months |
| num\_reported\_accidents | Number of recorded accidents on that road segment |
| accident\_risk | Target variable representing the predicted risk level (low, medium, high) |

* **Data Source:** Compiled from regional traffic and road condition reports (2005–2015).
* **Data Format:** CSV file containing structured tabular data.
* **Data Size:** Approximately 10,000+ rows and 13 columns.
* **Relevance:** Each feature directly correlates with factors studied in the literature (Vanitha & Swedha, 2023), supporting predictive modeling for road accident severity.

## Data Analysis and Insights

Data preprocessing included cleaning missing values, encoding categorical variables, and normalizing numeric columns such as num\_lanes, speed\_limit, and num\_reported\_accidents.

### **Key Insights**

* **Road Type:** Highways and rural roads show higher accident rates than urban local streets.
* **Curvature:** Sharper curves correlate with increased accident risk, especially when combined with high speed limits.
* **Speed Limit:** Sections with limits above 60 km/h show a higher number of reported accidents.
* **Lighting & Weather:** Poor lighting and rainy/foggy conditions significantly raise accident probability.
* **Time of Day:** Evening and late-night hours show more accidents, likely due to fatigue or reduced visibility.
* **Holidays & School Seasons:** Accidents slightly increase during holidays (higher traffic volume), while school seasons see moderate risks near school zones.

## Conclusion

The dataset analysis confirms that road geometry, weather, and lighting play critical roles in predicting accident severity. Data exploration helped identify the most impactful predictors and supported model selection for the project.

This research strengthens the foundation for developing a **machine learning-based road accident risk prediction system** that can help authorities take proactive safety measures, improve traffic planning, and potentially save lives.

# 

# Technology

### 1. Introduction

Road traffic accidents claim roughly 1.19 million lives yearly and are the top killer of 5 to 29-year-olds, while also draining economies; hence accurately predicting crash severity is essential for targeted prevention. Traditional statistical tools (logistic regression, ordered probit models[7]) fails to capturing and modelling the data’s complex non-linearities, and the adoption of machine-learning (ML) techniques that better capture hidden interactions are rising.

This review surveys the state-of-the-art use of three high-performing, widely deployed ML algorithms such as Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Support Vector Machines (SVM) for Road Accident Risk prediction. By comparing their strengths, limits, and deployment readiness, the review aims to guide researchers and traffic-safety managers toward robust, accurate, real-world analytics that can help mitigate road risk. [6]

### 2. Technology Overview

### Purpose

The application of machine learning (ML) algorithms is central to enhancing the accuracy and robustness of road traffic accident (RTA) prediction, particularly concerning injury severity and risk assessment. The purpose of these advanced computational models is to overcome the limitations of traditional statistical methods that often struggle to capture the complex, non-linear relationships inherent in real-world crash data.

The three methodologies we selected are Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Support Vector Machines (SVM). They are highly impactful models known for their predictive capabilities in classification and regression problems.

### Key Features

**Random Forest (RF)**: Random forest is an ensemble learning method composed of multiple decision trees. Key features include capturing complex nonlinear relationships without requiring pre-assumptions.

**XGBoost(XGB)**: XGBoost is a highly effective gradient boosting ensemble method. Some of the features include parallelizability, regularization penalties, cross-validation and non-linearity. [8]

**Support Vector Machines (SVM)**: SVMs are supervised learning models used for classification and regression tasks. They are particularly effective in high-dimensional spaces and cases where the number of dimensions exceeds the number of samples. Key features include the use of various kernel functions (linear, polynomial, radial basis function, etc.) to transform data into higher dimensions, maximizing the margin between classes, and managing overfitting through regularization parameters.

### Common Applications

RF has been applied in applied in 18 reviewed studies and excels at capturing non-linearities without prior assumptions. XGBoost (used in 10 studies) often outperforms RF in regional or crash-type contexts such as Saudi Arabia’s Qassim Province or Canadian rail-crossing crashes. SVM shows strong promise but remains sensitive to kernel choice. [6]

## 3. Relevance to Project

The reviewed machine learning algorithms Random Forest (RF), XGBoost (XGB), and Support Vector Machines (SVM) are highly relevant to this project, which focuses on predicting road accident risk and severity. Accurate and timely prediction is essential for developing proactive safety measures, such as dynamic traffic management, targeted enforcement, and improved infrastructure design.

Traditional regression-based models often fail to account for hidden interactions between factors such as driver behavior, weather conditions, vehicle type, and road geometry. ML-based approaches can effectively model these complex relationships, providing more reliable and interpretable results. Implementing RF, XGB, or SVM can therefore help the project improve data-driven decision-making and support real-time accident risk prediction systems. This relevance aligns with the broader goal of enhancing traffic safety through intelligent and automated analytics

## 4. Comparison and Evaluation

* **Random Forest (RF)**
  + Robust and reliable.
  + Handles noisy or incomplete data effectively.
  + Captures complex non-linear relationships.
  + Consistently stable across different datasets.
  + Can be slow to train on very large datasets.
* **XGBoost (XGB)**
  + Highly efficient and scalable.
  + Often achieves higher accuracy than Random Forest.
  + Handles unbalanced data well.
  + Regularization techniques help prevent overfitting.
  + Requires careful hyperparameter tuning for optimal performance.
* **Support Vector Machines (SVM)**
  + Effective for high-dimensional and small datasets.
  + Produces clear class boundaries using kernel functions.
  + Sensitive to the choice of kernel.
  + Less suitable for handling large-scale data.
* **General Summary**
  + XGBoost typically offers the highest predictive performance.
  + Random Forest is valued for stability and interpretability.
  + SVM is best suited for specialized or smaller datasets.

## 5. Use Cases and Examples

Recent studies have successfully deployed these algorithms for road accident analysis:

* **Random Forest** has been applied in the United States and India for crash severity prediction, achieving accuracy rates above 85% and identifying weather, lighting, and vehicle speed as top predictors.
* **XGBoost** was utilized in Saudi Arabia’s Qassim Province to predict regional crash risk, outperforming RF with faster convergence and better handling of unbalanced data.
* **SVM** was applied in China for highway crash classification, effectively identifying high-risk zones when kernel parameters were optimized.

These examples demonstrate the versatility and proven impact of ML-based models in improving the precision of traffic accident prediction across diverse geographical contexts.

## 6. Gaps and Research Opportunities

* **Data imbalance**
  + Many datasets have far fewer severe crashes than minor ones.
  + This imbalance leads to biased predictions and less reliable results for rare but critical events.
* **Model interpretability**
  + Advanced models like XGBoost may provide high accuracy.
  + However, these models lack transparency, which reduces trust among policymakers and practitioners.
* **Real-time deployment**
  + There is limited research on deploying models in real-time within live traffic management systems.
  + Practical integration remains a challenge.
* **Cross-regional generalization**
  + Models trained on data from one country or region often do not perform well elsewhere.
  + Contextual and regional differences hinder effective generalization.
* **Future research recommendations**
  + Focus on developing hybrid models that combine machine learning with explainable AI (XAI) frameworks.
  + Explore transfer learning and similar techniques to improve adaptability and performance across various regions.

## 7. Conclusion

Machine learning algorithms such as Random Forest, XGBoost, and Support Vector Machines have demonstrated significant improvements in the accuracy and reliability of road accident risk prediction. Their strength lies in their ability to detect hidden patterns and complex nonlinear relationships within crash data, making them a substantial advancement over traditional statistical models. Despite these benefits, there are still notable challenges especially when it comes to understanding how the models make decisions (interpretability) and deploying them in real-time environments. Nevertheless, as these algorithms continue to evolve and become more deeply integrated into intelligent transportation systems, they offer considerable promise for reducing road fatalities and enhancing global traffic safety.

# 

# References

[1] N. Behboudi, S. Moosavi, and R. Ramnath, “Recent Advances in Traffic Accident Analysis and Prediction: A Comprehensive Review of Machine Learning Techniques,” Jun. 2024. Available:<https://arxiv.org/pdf/2406.13968>

[2] K. Assi, “Traffic Crash Severity Prediction—A Synergy by Hybrid Principal Component Analysis and Machine Learning Models,” in *International Journal of Environmental Research and Public Health (IJERPH)*,

Available:<https://www.researchgate.net/publication/344839529_Traffic_Crash_Severity_Prediction-A_Synergy_by_Hybrid_Principal_Component_Analysis_and_Machine_Learning_Models>

[3] R. Vanitha and M. Swedha, “Prediction of Road Accidents Using Machine Learning Algorithms,” in *Middle East Journal of Applied Science & Technology (MEJAST)*, May 2023, pp. 64–75. Available:<https://www.researchgate.net/publication/370961834_Prediction_of_Road_Accidents_Using_Machine_Learning_Algorithms>

[4] S. Ahmed, A. Hossain, S. K. Ray, M. M. I. Bhuiyan, and S. R. Sabuj, “A study on road accident prediction and contributing factors using explainable machine learning models: analysis and performance : ,” in *Transportation Research Interdisciplinary Perspectives*, May 2023.

Available:<https://www.researchgate.net/publication/370288284_A_study_on_road_accident_prediction_and_contributing_factors_using_explainable_machine_learning_models_analysis_and_performance>

[5] Vanitha, R., & Swedha, M. (2023). *Prediction of Road Accidents Using Machine Learning Algorithms.* Middle East Journal of Applied Science & Technology, 6(2), 64–75.<https://doi.org/10.46431/MEJAST.2023.6208>

[6] Hamdan, Noura, and Tibor Sipos. 2025. “Advancements in Machine Learning for Traffic Accident Severity Prediction: A Comprehensive Review”. Periodica Polytechnica Transportation Engineering 53 (3):347-55. <https://doi.org/10.3311/PPtr.40369>.

[7] Rui Garrido, Ana Bastos, Ana de Almeida, and José Paulo Elvas. "Prediction of Road Accident Severity Using the Ordered Probit Model." Transportation Research Procedia 10 (2014): 714–723. <https://doi.org/10.1016/j.trpro.2014.10.107>.

[8] <https://neptune.ai/blog/xgboost-everything-you-need-to-know>