الجمهوريـة الجزائـريـة الديمقراطيـة الشعبيـة

People's Democratic Republic of Algeria

وزارة التعليــم العالـي والبحـث العلمـي

Ministry of Higher Education and Scientific Research

No Ref:………………..

Abdelhafid Boussouf Mila University Centre

Institute of Mathematics and Computer Science

Department of Computer Science

Specialty: Artificial Intelligence and its Applications.

Master's Thesis

Application of Machine Learning and Deep Learning Techniques in Road Safety

Presented by:

- Benghalia Khawla
- Khelil Omima

Supported by the jury:

President:	Mr.	DIB Abderrahim	Rank: M.A.A
Examiner:	Mrs.	DEFFAS Zineb	Rank: M.A.A
Supervisor:	Mrs.	ZEKIOUK Mounira	Rank: M.A.A

Academic year: 2024/2025

بسم الله الرحمن الرحیم

إهداء

الحمد لله على لذة الإنجاز، والحمد لله عند البدء، وعند الختام...

إلى والدي العزيز **رياض**، الذي أضاء دروبي، وكان قدوتي في كل خطوة أخطوها.

وإلى أمي الحنونة **نعمة**، حضني الدافئ وسمائي التي لم تتركني يومًا، ولا يكتمل يومي بدونها.

إلى أخواتي الغاليات **آية**، **رونق**، و**مريم**، من كنّ لي السند والدعم في كل مراحل تعليمي.

إلى صديقة عمري **شرين**، رفيقة الأيام الجميلة والمواقف الصعبة، التي كانت دائمًا بجانبي.

وإلى **أميمة**، رفيقة دراستي في كل تفصيلة، من تشاركنا مشوار العلم بكل ما فيه من تعب وسهر وإنجاز. كنا دائمًا معًا في كل مرحلة، نكمل بعضنا وندعم بعضنا. صداقتها كانت نعمة، ورفقتها من أجمل ما مرّ في هذه الرحلة.

وإلى صديقاتي العزيزات **فاتن** و**أسماء**، من كانت صحبتُهن زادًا في درب الإنجاز، وأثرًا طيبًا لا يُنسى.

وإلى كل من دعمني بكلمة أو دعوة أو حضور، خالص امتناني ومحبة صادقة.

ولا يفوتني أن أتقدم بالشكر لأستاذتنا العزيزة **الأستاذة منيرة**، على دعمها وتشجيعها الدائم.

أهديكم جميعًا هذا العمل المتواضع، ثمرة جهدي وتعب الأيام، راجية من الله أن يجعله خالصًا لوجهه الكريم، وبداية موفقة لما هو آت.

ختامًا، أحمد الله على التمام، وأشكره على كل عون، وأسأله الإخلاص والتوفيق لما هو قادم.

خولة

إهداء

﴿وَآخِرُ دَعْوَاهُمْ أَنِ الْـحَمْدُ للهِّ رَبِّ الْعَالَمِينَ ﴾

قال النبي ﷺ: "من لا يشكر الناس لا يشكر الله" ,أرفع هذا التخرج عربون امتنان وعرفان، وأهديه بكل فخر وحب ووفاء

إلى أبي الغالي **محمد الطاهر**، يا من كنت لي الأمان حين خفت، والسند حين تعبت، والنور حين أظلمت الدروب.

إلى أمي الحنونة **مليكة**، نبع الحنان، وسرّ الصبر، وصوت الدعاء الذي لم يغِب عن أيّ لحظة في طريقي.

إلى شقيقتيَّ الحبيبتين: **شيماء** و**أنفال**، وإلى **ندى، الأخت التي لم تلدها أمي**، أنتنّ النبض الدافئ في قلبي، والضوء الذي كان ينير عتمتي، ورفيقات الدرب في الحياة قبل الدراسة. وجودكن نعمة أحمد الله عليها كل يوم.

إلى إخوتي الأحبة: **زكرياء، وياسر، وأبو بكر**، أنتم السند الحقيقي، ورفاق الروح، والضحكة التي لا تنطفئ مهما اشتدّ التعب.

إلى **خولة**، شريكة البحث والسهر والتفاصيل، شكراً لكِ من القلب على صبرك وصدقك، على دعمك في الأوقات الحرجة، وعلى كل لحظة كنتِ فيها القوة عندما خارت قواي. وجودك في هذه الرحلة كان نعمة لا تُقدّر.

إلى **جدي الحبيب – رحمه الله** ، الذي انتظر هذا اليوم بقلبٍ فخور، وروحٍ مملوءة بالدعاء, فكان حنانه وذكراه رفيق دربي في الغياب.

وفي ختام هذا الإهداء، لا يسعني إلا أن أرفع كفّ الشكر والامتنان لكلّ من كان نورًا في دربي.
فلكلّ من زرع في قلبي أملاً، أو مسح عني تعبًا، أو صدح باسمي فخرًا، أقول إن كنتم جزءًا من رحلتي، فأنتم بلا شك جزءٌ من هذا الإنجاز.

فالحمد لله أولاً وآخرًا، ظاهرًا وباطنًا، الذي بنعمته تتمّ الصالحات.

أميمة

Abstract

Road traffic accidents are a growing global concern, causing around 1.3 million deaths and over 50 million injuries annually. This crisis highlights the urgent need for intelligent systems that can enhance road safety and reduce both the severity and frequency of accidents.

This project investigates the use of ML and DL to analyze traffic accident data and support decision-making. It focuses on building predictive models that learn from many factors and past patterns to help prevent future accidents.

Two key tasks are studied in this project: ***predicting accident severity*** using datasets such as US Accidents (2016–2023) and Road Traffic Accidents, and ***predicting accident frequency*** using Thailand Fatal Road Accidents and Great Britain Accident datasets. Each task used data suited to its objective.

The project follows a specialized modeling process for machine learning and deep learning, known as CRISP-ML. A comprehensive data preparation steps were applied to clean and structure the data for effective training. By leveraging ML and DL techniques, several models are built and evaluated to demonstrate the effectiveness of these new approaches in enhancing road safety.

Keywords: Road Traffic Accidents, Machine Learning, Deep Learning, Accident Severity Prediction, Accident Frequency Prediction, Road Safety, CRISP-ML.

RÉSUMÉ

Les accidents de la route constituent une préoccupation mondiale croissante, causant environ 1,3 million de décès et plus de 50 millions de blessés chaque année. Cette crise souligne le besoin urgent de systèmes intelligents pour améliorer la sécurité routière et réduire la gravité ainsi que la fréquence des accidents.

Ce projet explore l'utilisation de l'apprentissage automatique (ML) et de l'apprentissage profond (DL) pour analyser les données d'accidents de la route et soutenir la prise de décision. Il se concentre sur la construction de modèles prédictifs capables d'apprendre à partir de multiples facteurs et de schémas passés afin de prévenir les accidents futurs.

Deux tâches principales sont étudiées : la prédiction de la gravité des accidents à l'aide de jeux de données tels que US Accidents (2016–2023) et Road Traffic Accidents, et la prédiction de la fréquence des accidents en utilisant les données de Thailand Fatal Road Accidents et Great Britain Accident. Chaque tâche utilise des données adaptées à ses objectifs.

Le projet suit un processus de modélisation spécialisé pour le ML et le DL, connu sous le nom de CRISP-ML. Des étapes complètes de préparation des données ont été appliquées pour nettoyer et structurer les données en vue d'un entraînement efficace. En exploitant les techniques de ML et de DL, plusieurs modèles ont été construits et évalués pour démontrer l'efficacité de ces approches dans l'amélioration de la sécurité routière.

Mots-clés : Accidents de la route, Apprentissage automatique, Apprentissage profond, Prédiction de la gravité des accidents, Prédiction de la fréquence des accidents, Sécurité routière, CRISP-ML.

الملخص

تُعد حوادث المرور مشكلة عالمية متزايدة، تتسبب في حوالي 1.3 مليون وفاة وأكثر من 50 مليون إصابة سنويًا. تعكس هذه الأزمة الحاجة الملحّة إلى أنظمة ذكية تعزز السلامة على الطرق وتقلل من شدة الحوادث وتكرارها.

يتناول هذا المشروع استخدام تقنيات التعلم الآلي (ML) والتعلم العميق (DL) لتحليل بيانات الحوادث ودعم اتخاذ القرار. يركز المشروع على بناء نماذج تنبؤية تتعلم من عوامل متعددة وأنماط سابقة للمساعدة في الوقاية من الحوادث المستقبلية.

تمت دراسة مهمتين أساسيتين في هذا المشروع: التنبؤ بشدة الحوادث باستخدام مجموعات بيانات مثل حوادث الولايات المتحدة (2016–2023) وحوادث المرور الأخرى، والتنبؤ بتكرار الحوادث باستخدام بيانات حوادث الطرق المميتة في تايلاند وحوادث بريطانيا الكبرى. تم استخدام بيانات مناسبة لكل مهمة حسب هدفها.

يعتمد المشروع منهجية متخصصة في النمذجة للتعلم الآلي والعميق تُعرف باسم CRISP-ML. وتم تطبيق خطوات شاملة لتحضير البيانات بهدف تنظيفها وتنظيمها لتدريب النماذج بفعالية. من خلال توظيف تقنيات ML وDL، تم بناء نماذج متعددة وتقييمها لإثبات فعالية هذه المقاربات في تحسين السلامة المرورية.

الكلمات المفتاحية: حوادث المرور، التعلم الآلي، التعلم العميق، التنبؤ بشدة الحوادث، التنبؤ بتكرار الحوادث، السلامة على الطرق، CRISP-ML.

Table Of Contents

List Of Figures

List Of Tables

List Of Abbreviations

ABS Anti-lock Braking System.

ESC Electronic Stability Control.

LSTM Long Short-Term Memory.

GRU Gated Recurrent Units.

SVM Support Vector Machines.

ML Machine Learning.

CRISP-ML Cross-Industry Standard Process for Machine Learning.

RTA Road Traffic Accidents.

NZTA New Zealand Transport Agency.

CAS Crash Analysis System.

ADF Augmented Dickey-Fuller.

DL Deep Learning.

IQR Inter Quartile Range.

SMOTE Synthetic Minority Over-sampling Technique.

AI Artificial Intelligence.

TP True Positives.

TN True Negatives.

FP False Positives.

FN False Negatives.

IDS Intrusion Detection System.

DT Decision Tree.

RF Random Forest.

SHAP Shapley Additive exPlanations.

MSLE Mean Squared Logarithmic Error.

RNN Recurrent Neural Network.

GPU Graphics Processing Units.

General Introduction

Road safety is a critical aspect of public health and urban planning, as it directly impacts the well-being of individuals and communities. Ensuring safe roads helps reduce the number of accidents, injuries, and fatalities, which can have devastating effects on families and society as a whole. Each year, road traffic accidents claim countless lives and result in serious injuries, placing significant emotional and financial burdens on victims and their loved ones [1].

Improving road safety requires teamwork and different strategies that work well together. Key parts include public awareness campaigns that inform drivers and pedestrians about risks and safe behavior on the roads[2]. Regular road maintenance is essential, as it keeps roads safe by fixing problems that could cause accidents. Additionally, connected devices like cameras and sensors provide real-time traffic data for better monitoring. This data is supported by analysis software that helps authorities understand traffic patterns and make smart decisions. Moreover, machine learning and deep learning techniques enhance traffic management by addressing predictive issues, which is the focus of our project[3].

Machine learning and deep learning play a crucial role in enhancing road safety, particularly through predictive modeling. These technologies analyze vast amounts of data from various sources, such as traffic cameras, sensors, and historical accident records, to identify patterns and predict potential problems. By leveraging advanced algorithms, they can forecast traffic conditions, assess the likelihood of accidents, and optimize traffic flow in real-time. This predictive capability allows authorities to proactively address safety issues, implement timely interventions, and make informed decisions about infrastructure improvements.

In this project, we aim to develop predictive models that leverage machine learning and deep learning techniques to mitigate the risks and severity of traffic accidents. We will concentrate on two primary areas: **Accident Severity Prediction and Accident Frequency Prediction**. Our objective is not to produce a final product for real-world application but to conduct an in-depth study of the various steps involved in a machine learning and deep learning project, from data selection to model evaluation.

The project is guided by the CRISP-ML framework as a development process and supported by various deep learning models along with a series of preprocessing steps.

This thesis is organized into four chapters as follows:

- **Chapter 1: Road Safety: A New Field for Machine and Deep Learning Applications:** This chapter introduces the concept of road safety and explores how machine learning and deep learning can be applied to enhance safety measures and reduce traffic accidents.

- **Chapter 2: Project Presentation: Objectives, Boundaries, and Development Process:** This chapter outlines the project's objectives, defines its functional boundaries, and describes the development process employed for modeling.

- **Chapter 3: Business and Data Understanding:** This chapter focuses on understanding the business context and the datasets related to the studied problems (Accident Severity Prediction and Accident Frequency Prediction), highlighting the importance of data in driving effective solutions.

- **Chapter 4: Data Preparation, Model Building, and Performance Evaluation:** This chapter delves into the steps of data preparation, the construction of predictive models, and the evaluation of their performance to ensure effectiveness in real-world applications.

Chapter 01

Road safety: A new field for machine and deep Learning applications

1.Introduction

Road safety is a crucial issue for public health and urban development, significantly affecting the well-being of individuals and communities. This chapter seeks to provide a thorough understanding of road safety, starting with a clear definition of the concept. We will investigate the various factors that influence road safety, such as human behavior, vehicle design, and essential infrastructure components [1,2,3]. Furthermore, we will present a range of strategies and techniques that can be employed to enhance road safety, with a specific focus on the applications of machine learning and deep learning techniques. These advanced technologies present innovative solutions for accident prediction and prevention.

2.Road Safety

In this section, we will first define road safety and then address the various factors that influence it.

2.1.Definition

Road safety refers to methods and measures aimed at reducing the risk of individuals using the road network from being killed or seriously injured. This includes all road users. It encompasses a collection of policies, strategies, and tools designed to prevent traffic accidents and reduce their severity [4].

2.2. Factors affecting road safety

Various elements influence the safety of road users, making it crucial to analyze these factors. In the following sections, the most important factors are summarized for a clearer understanding of their impact on road safety.

2.2.1.Human factors

Human factors are the primary contributors to road traffic accidents, accounting for over 90% of incidents [5]. Among the most critical issues are:

- **Speeding:** Even increasing small speed, such as 1%, can significantly increase the risk of accidents of death by about 4%. Data related to speed collected through tracking devices can be used as a valuable input for automatic learning models.
- **Driving Under the Influence of Alcohol or Drugs:** The consumption of alcohol or drugs changes judgment time and reaction time, significantly increasing the risk of collision [6].
- **Driver Distraction:** Distractions, including mobile phone use, increase the likelihood of downhill tracks. Surveillance cameras can help recognize signs of distraction [7].
- **Fatigue and Drowsiness:** Reducing vigilance and reaction time due to fatigue makes drivers more sensitive to accidents [8].
- **Demographic Characteristics:** Age, gender, and driving experience affect driving behavior; young drivers often participate in risky activities.

2.2.2.Vehicle factors

The role of vehicles in road safety extends beyond human behavior, encompassing design, technologies, and maintenance practices:

- **Safety Technologies:** Advanced systems such as anti-lock brakes (ABS), electronic stability control (ESC), and airbags play an important role in minimizing injury risks. Integrating data on the presence and use of these technologies can enhance predictive models [9].
- **Vehicle Design:** The structure and design of a car, especially from the front, significantly affect the safety of drivers, with softer concepts that help reduce the severity of injury [9].
- **Vehicle Maintenance:** Mechanical failure, including braking dysfunction, continues to be an important factor contributing to a traffic accident [9].

2.2.3.Road and infrastructure factors

Road design and infrastructure play a crucial role in influencing traffic safety [10]. The key contributing factors include:

- **Road engineering and maintenance:** Safety risk increased with poorly designed characteristics, such as sharp curves and inappropriate design.
- **Infrastructure components:** The implementation of specialized rails for pedestrians and cyclists contributes to reducing accidents.
- **Traffic management:** Various control measures, such as traffic lights, loops, and soothing traffic techniques, such as narrowed routes, have proven effective in reducing collision frequency.

2.2.4.Environmental factors

Environmental factors play a significant role in road safety:

- **Weather conditions:** Fog, rain, and snow impair visibility and increase the risk of slipping, necessitating weather data prediction models for analysis [11].
- **Light:** Road illumination is often inadequate, especially at night or in rural areas, raising the likelihood of accidents [11].
- **Road surface conditions:** Wet or icy surfaces make vehicle control more difficult and extend braking distances [11].
- **Surrounding environment:** Dense advertisements along the roadside distract drivers, contributing to an increase in distraction-related accidents [11].

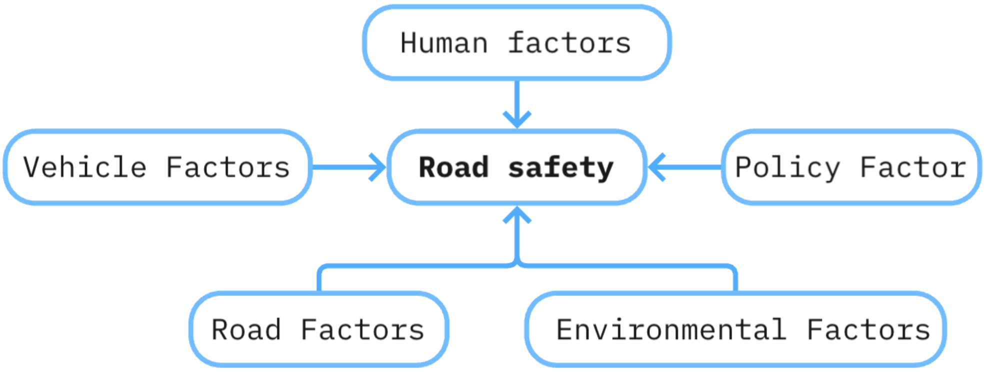
2.2.5.Legal and policy factors

The legal and policy factors play a central role in achieving road safety:

- **Traffic Legislation:** Laws such as speed limits and seat belt requirements help reduce accident severity and fatalities [12].
- **Enforcement Mechanisms:** Effective enforcement, like random breath tests and speed cameras, helps deter risky behavior [12].

- **License and training:** Strict licensing procedures and driver training programs improve driving skills and reduce accidents, especially among new drivers [12].

Figure 1.1 Factors Affecting Road Safety.



2.3. Strategies and tools for road safety enhancement:

Improving road safety is a challenging task that requires teamwork. Different efforts and tools need to come together to create effective strategies for managing traffic and preventing accidents. By combining real actions, public awareness campaigns, new technologies, and data-driven predictive solutions, we can make the roads safer for everyone. Here are some important elements that help promote road safety:

2.3.1.Real interventions by law enforcement

Law enforcement agencies are vital in enforcing traffic laws and ensuring compliance. Regular patrols and checkpoints can discourage reckless behavior and encourage safer driving practices [13].

2.3.2.Public awareness campaigns

Raising public awareness about road safety is crucial for driving behavioral change. Campaigns can inform drivers and pedestrians about the risks and responsibilities involved in road use.

2.3.3.Regular road maintenance

Regular road maintenance is essential for ensuring the safety and efficiency of transportation systems. It involves activities such as repairing potholes, updating signage, and cleaning road surfaces to prevent hazards.

2.3.4.Integration of connected materials

Connected devices such as cameras and sensors significantly improve road monitoring. These technologies offer real-time data on traffic conditions and help identify potential hazards.

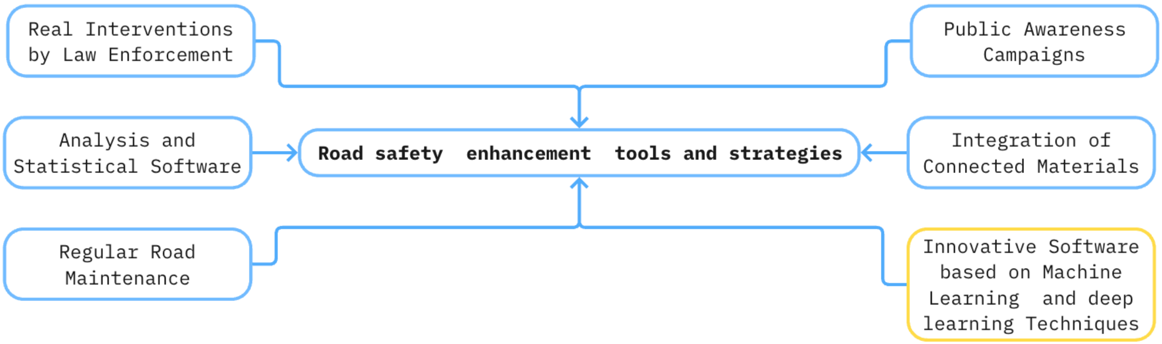
2.3.5.Analysis and statistical software

Analysis and statistical software play a crucial role in traffic management by providing valuable insights into traffic patterns and behaviors. These tools enable authorities to collect, process, and analyze data from various sources, such as traffic cameras and sensors. By utilizing statistical methods, agencies can identify trends, assess the effectiveness of current traffic measures, and make informed decisions to enhance safety and efficiency on the roads [14].

2.3.6.Innovative software based on machine learning and deep learning techniques

Developing software applications that utilize machine and deep learning can significantly enhance traffic management. These tools analyze large datasets to optimize traffic flow and improve safety measures. By employing advanced algorithms for predictive analysis of traffic patterns, authorities can proactively address safety issues and implement effective solutions. In our project, we focus only on the use of machine and deep learning tools to enhance road safety [15].

Figure 1.2 Key Tools and Strategies for Enhancing Road Safety.


Real Interventions by Law Enforcement
Public Awareness Campaigns
Analysis and Statistical Software
Road safety enhancement tools and strategies
Integration of Connected Materials
Regular Road Maintenance
Innovative Software based on Machine Learning and deep learning Techniques

3.Applications of Machine Learning and Deep Learning in Road Safety

Machine learning and deep learning techniques are widely used in road safety in various forms, such as automated surveillance, smart traffic signals, road condition assessment, and predictive models. In our study, we specifically focus on predictive models.

3.1. Accident risk prediction

Accident risk prediction models are designed to carry out binary classification based on multiple factors to assess the likelihood of an accident occurring. These models are commonly employed in connected cars to provide real-time notifications and alert drivers. Most developed models to solve this problem are based on decision trees, support vector machines (SVM), Gradient Boosting, and ensemble methods, which allow for improved accuracy by combining predictions from multiple trees [16]. The following table summarizes some models developed in the field of accident risk prediction:

Table 1.1 Summary of Studies on Accident Risk Prediction.

Study	Year	Main Approach	High-level Summary	Data Source
Mor et al.	2020	Machine Learning	Develop an Accident Prediction Model for Haryana, India using linear regression, showing high accuracy in predictions.	Accident data from 1996-2016 in Haryana (India)
Venkat et al	2020	Machine Learning	Evaluated ensemble ML models for predicting road accident severity in NZ, highlighting Random Forest’s effectiveness using SHAP analysis.	Datasets from Data.gov.uk, US-Accidents (2.25 million records) and UK police accidents (1.6 million records), 2000-2019.
Yeole et al	2022	Deep Learning	Developed an ANN-based model to predict traffic accidents, outperforming multiple linear regression, using real-world data.	Data from 887 accidents over six years in Pune, India, analyzed with weather, traffic, and road conditions.

3.2. Accident frequency prediction

Accident frequency prediction involves using statistical and machine learning models to estimate how often accidents are likely to occur in a given area by analyzing historical data. The models dedicated to frequency prediction are based on Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRU), which are particularly effective in handling time-series data [16]. *Table 1.2* below summarizes several models developed for accident Frequency prediction:

Table 1.2 Summary of Studies on Accident Frequency Prediction.

Study	Year	Main Approach	High-level Summary	Data Source
Hebert et al	2019	machine Learning	Developed high-resolution accident frequency prediction models for Montreal using big data analytics and the Balanced Random Forest algorithm.	Data from Montreal city, the Canadian government, and historical weather records.
Wu et al	2021	Deep Learning	Developed a CNN-GRU fusion deep learning model to predict crash driver frequency, outperforming other approaches.	Traffic Police Corps, Taoyuan Police Department database.
Feng et al	2020	Deep/machine learning	Developed a big data analytics platform using ML and DL for UK traffic accident analysis and prediction.	Compiled road accident data from Albanian online news using web scraping.

3.3. Accident severity prediction

Accident severity prediction refers to the process of estimating the potential seriousness of injuries or damages resulting from a traffic accident before it occurs. Severity prediction is viewed as a multiclass classification problem that can be tackled using various algorithms, such as Logistic Regression, Decision Trees, and Random Forest. Support Vector Machines (SVM) and Gradient Boosting [16]**.** The following *Table 1.3* summarizes some models developed in the field of accident severity prediction:

Table 1.3 Summary of Studies on Accident Severity Prediction.

Study	**Year**	**Main Approach**	**High-level Summary**	**Data Source**
Kashyap et al	2000	Machine Learning	Developed a framework using ML models like Random Forest and Logistic Regression, achieving 87% accuracy in predicting traffic accident severity.	Data from UK traffic accident records (2005-2014), accident and vehicle information databases.
Koramati et al	2022	Deep Learning	Developed ANN models for urban crash prediction in India using Hyderabad police data, highlighting sensitive crash factor	Road crash database from police records: 7464 total (2015-2019), 1714 total (2019) crashes analyzed
Niyogisubizo et al.	2021	Machine Learning	Utilized RF, MNB, KC, KNN for predicting road crash severity, highlighting RF's superior accuracy and feature importance analysis	Used Victoria, Australia's road accident data from 2015-2020, provided by the Department of Transport

3.4. Accident duration prediction

Accident duration prediction refers to the estimation of how long an accident will disrupt traffic flow following an incident. Accurately predicting the duration of road disruptions is crucial for effective traffic management and minimizing the impact on commuters [16]. *Table 1.4* below outlines several models created to predict accident duration:

Table 1.4 Summary of Studies on Accident Duration Prediction.

Study	Year	Main Approach	High-level Summary	Data Source
Hamad et .al	2020	Machine Learning	Explores machine learning models, notably SVM and GPR, for accurate traffic incident duration prediction.	Data from the Transtar operators' database including 146,573 incidents with 74 attributes, modified for prediction analysis.
Li et al	2020	Machine Learning	Introduced a deep fusion model using RBMs to predict traffic accident durations considering spatial-temporal correlations.	Data from Highway Safety Information System (HSIS)

4.Conclusion

In conclusion, understanding road safety is essential for fostering healthier and more sustainable communities. By examining the key factors that contribute to road safety, including human behavior, vehicle design, and infrastructure, we can identify critical areas for intervention. The integration of machine learning and deep learning techniques into road safety strategies offers promising opportunities to enhance accident prediction and prevention. In the next chapter, we will outline the issues that we will focus on in our studies.

Chapter 02

Project Presentation: Objectives, Boundaries, and Development Process

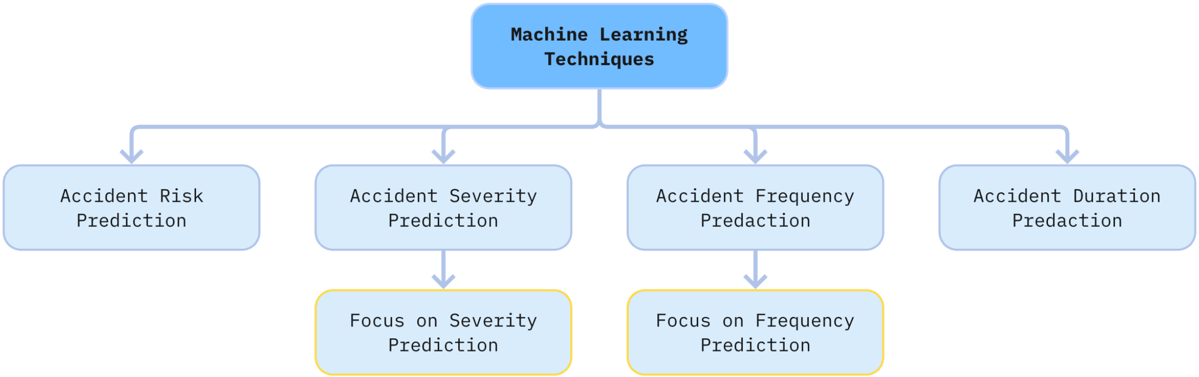
1.Introduction

In this chapter, we will present the main objective and functional boundaries of our project, which is designed to offer a specific form of decision support aimed at enhancing road safety. We will begin by outlining the primary goals and defining the functional boundaries of the project, clarifying what is included and excluded in our scope to ensure a focused approach. Following this, we will delve into the adopted development process (CRISP-ML), highlighting the methodologies that will guide our work.

2.Main Objective and Functional Boundaries of the Project

The primary objective of this thesis is to develop predictive models using machine learning and deep learning to mitigate the risks and severity of traffic accidents. The study addresses two sub-problems: **Accident Severity Prediction** and **Accident Frequency Prediction**, each with specific goals aimed at providing practical, data-driven insights for planners and safety authorities. Our aim is not to create a comprehensive application with backend and frontend components; rather, we will concentrate solely on the processes of developing machine learning and deep learning models for the two issues. These models may be integrated into various other applications, extending beyond the scope of our study.

Figure 2.1 Machine Learning Techniques.


Machine Learning Techniques
Accident Risk Prediction
Accident Severity Prediction
Accident Frequency Predaction
Accident Duration Predaction
Focus on Severity Prediction
Focus on Frequency Prediction

3.Adopted Development Process

Like all computer science projects, machine learning and deep learning projects require a framework. However, traditional software engineering methodologies and processes do not apply well to data science. To guide ML practitioners through the development lifecycle, the Cross-Industry Standard Process for the Development of Machine Learning Applications (CRISP-ML) methodology was recently proposed [17].

3.1.CRISP-ML presentation

CRISP-ML is a structured process for developing machine learning models that builds on CRISP-DM (Cross-Industry Standard Process for Data Mining) [17]. It aims to standardize the various stages of the ML project lifecycle, from problem understanding to production deployment and maintenance.

It is a methodology that allows for:

- **A Structured approach:** It provides a framework for addressing ML projects methodically, ensuring that business objectives and technical requirements are considered [17,18].
- **Quality assurance:** It emphasizes quality at every stage of the process, from data understanding to production deployment and model maintenance [17,18].
- **Adaptation to different contexts:** It can be applied to various ML projects, regardless of the algorithm or application domain [17,18].

3.2. CRISP-ML steps

The CRISP-ML framework consists of several essential steps that guide the development of machine learning models. Each step plays a critical role in ensuring that the final model is effective and aligned with business objectives [19]. Below are the key stages involved in the CRISP-ML process:

3.2.1.Business and data understanding

The first step is to thoroughly understand the business elements and the issues that the project aims to solve. Also, this phase aims to accurately determine and select the datasets relevant to the problem, as well as to establish the connection between the model and possible use cases [19].

3.2.2.Data engineering (Preparation)

This phase encompasses activities related to preparing the dataset for the next step. It includes data cleaning and, most importantly, encoding the data to make it compatible with the algorithms that will be used [19].

3.2.3.Model building

This step involves selecting, parameterizing, and training different algorithms, as well as chaining them together to form a model [19].

3.2.4.Model evaluation

The evaluation aims to verify and compare the obtained models to ensure they meet the objectives established at the beginning of the process. It also contributes to the decision to deploy the best model based on several parameters [19].

3.2.5.Deployment

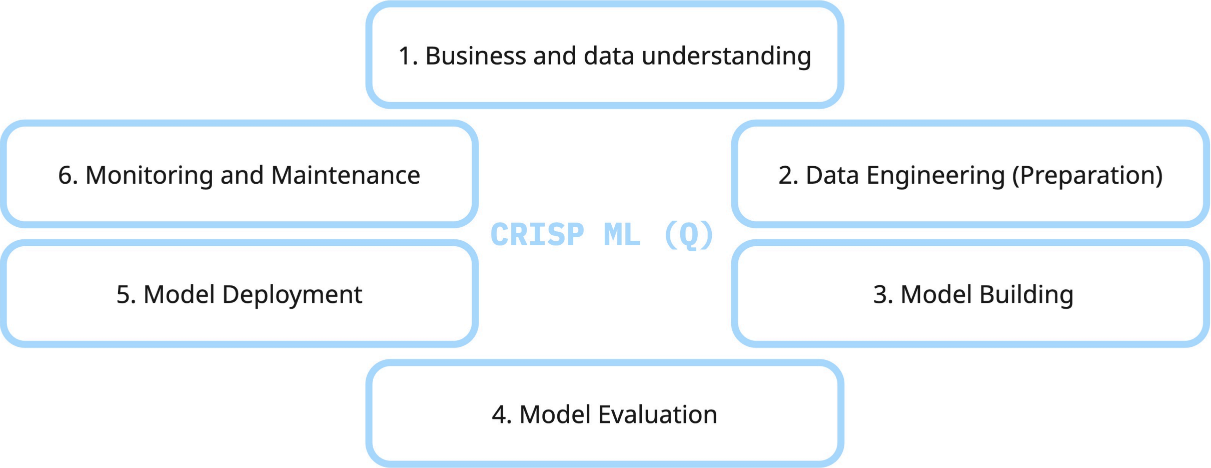
This step involves putting the obtained models into production for end users [19].

3.2.6.Monitoring and maintenance

This step refers to the ongoing processes that ensure the model continues to perform effectively after deployment [19].

As part of our project, we focus only on the first four phases, as the last two steps require a broader framework and proactive models for monitoring.

Figure 2.2 Development Process: CRISP ML.


1. Business and data understanding
6. Monitoring and Maintenance
2. Data Engineering (Preparation)
CRISP ML (Q)
5. Model Deployment
3. Model Building
4. Model Evaluation

4.Conclusion

In conclusion, this chapter has established a clear framework for our project aimed at enhancing road safety through effective decision support. By outlining the main objectives and defining the functional boundaries, we have ensured a focused approach that addresses specific needs while clarifying the scope of our work. Additionally, the introduction of the CRISP-ML development process underscores our commitment to a structured methodology.

Chapter 03

Business and Data Understanding

1.Introduction

In this chapter, we will focus on the first step of business and data understanding within the CRISP-ML process. This phase is essential for laying the groundwork for our project, as it enables us to thoroughly analyze the studied problems and identify different use cases for the targeted models, which helps in selecting the relevant datasets.

2.Business Understanding

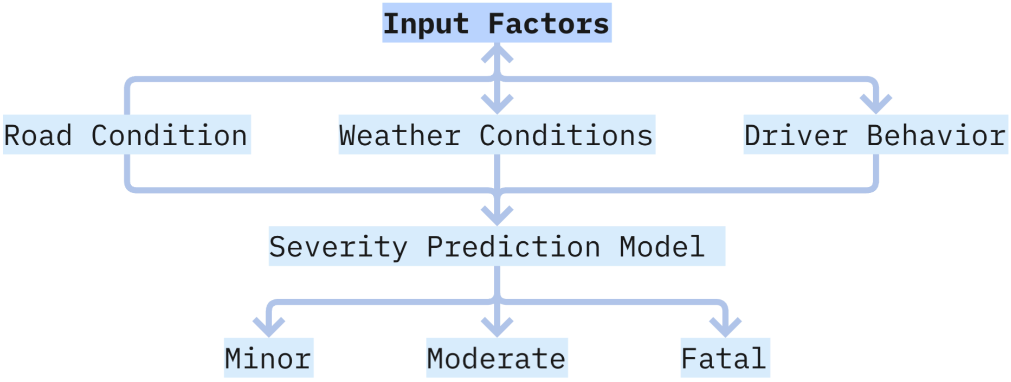
In the following subsections, we will define the inputs and outputs of our two future models, along with the possible use cases for each model to clarify their functional limits.

2.1.Model for accident severity prediction

- **Inputs and outputs of the model:**

  The primary objective is to build a predictive model capable of determining the severity of a traffic accident (minor, moderate, fatal, etc.) before it occurs, based on various input factors such as road conditions, weather, and driver behavior. Our model will focus on using ***tabular data*** [20].

Figure 3.1 Output and Input of the Severity Prediction Mode.



- **Possible use cases of the model:**

A severity prediction model can serve various purposes:

- The model can be integrated into intelligent transport systems and connected cars to provide real-time safety alerts.

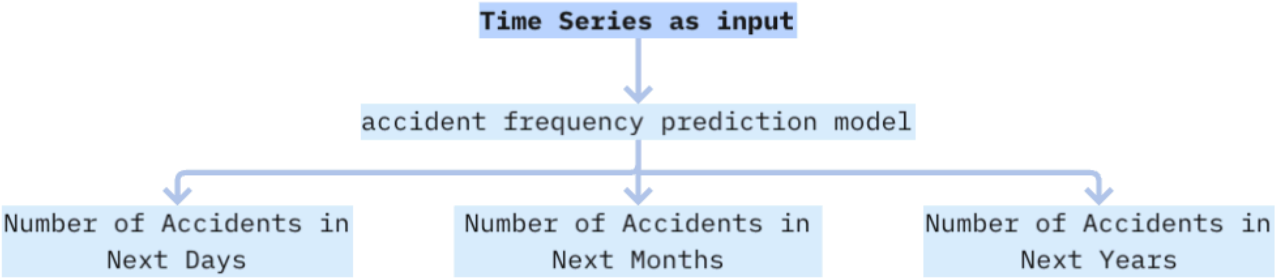
- The model can be used to simulate the severity of accidents under different changes in impact factors.
- The model can be used in combination with other algorithms to optimize intervention resources.
- The model can help in the installation of warning dynamic signs

2.2.Model for accident frequency prediction

- **Inputs and outputs of the model :**

The main goal of our accident frequency prediction model is to estimate the number of accidents in future periods. In contrast to accident severity prediction, which relies on analyzing various impact factors (such as weather and road conditions), accident frequency prediction uses historical accident count data, known ***as time series*** to project the number of accidents in future periods (e.g., the number of accidents in the coming days, months, or years) [21].

Figure 3.2 Output and Input of the Frequency Prediction Model.


Time Series as input
accident frequency prediction model
Number of Accidents in Next Days
Number of Accidents in Next Months
Number of Accidents in Next Years

- **Possible use cases of the model:**

Generally, accident frequency prediction models are used for long-term planning projects:

- The model can assist city planners, transportation agencies, and policymakers in making decisions about huge infrastructure development, road redesigns, road safety measures, and resource allocation in the future.
- Understanding cyclical trends in the number of accidents, such as the increase in accidents during festival periods or summer.

- The model can serve to make a deep analysis in combination with other algorithms to understand the real causes of the predicted number of accidents.

3.Data Understanding

Data understanding and dataset selection are the most important aspects of any machine learning and deep learning project, regardless of the domain. In the following subsection, we will select the appropriate datasets for training our two models, considering the inputs and outputs specified in the previous section [22].

Given the lack and near absence of accident datasets in Algerian public databases, we have decided to choose from the available public datasets of other countries.

3.1.Dataset selection for accident severity prediction

3.1.1.Candidate datasets

To support accident severity prediction, we explored multiple publicly available datasets from diverse geographical regions and periods. These datasets vary in size, feature richness, and structure. *Table 3.1* summarizes the key details of each dataset, including its source, region of coverage, and data collection period.

Table 3.1 Lists the datasets for Accident Severity Prediction.

Dataset Name	Source	Region	Time Period	Number of Rows	Number of Features
US Accidents (2016 - 2023)	Kaggle (Moosavi, 2023)	USA	2016–2023	7.000.000	46
UK Road Accident Dataset	Kaggle (Odariya, 2022)	UK	2005-2014	1.000.000	33
Road Traffic Accidents	Kaggle (Shahane, 2022)	Addis Ababa, Ethiopia	2017-2020	12.000	32
Road Accident Dataset	Kaggle (Berge, 2023)	Global	2020-22022	40.000	23

Dataset Name	Source	Region	Time Period	Number of Rows	Number of Features
Global Road Accidents Dataset	Kaggle (Panday, 2023)	Global (USA, UK, India…)	2000-2024	130.000	30
Road Accident in Montreal	Données Québec (CC-BY 4.0, accessed 20 Dec 2023)	Quebec, Montreal	2011-2022	1.000.000	25

3.1.2.Selected and excluded datasets

Selecting the right datasets is a critical step in ensuring that the predictive model is accurate. After thoroughly evaluating all the explored datasets, two were selected based on their structure and feature richness (the US Accidents dataset and the Road Traffic Accidents), to train our severity prediction model. In the following subsections, for reasons of space, we will only present the structure of the two datasets selected for training. For the excluded datasets, we will only mention their weaknesses.

1.US Accidents (2016 - 2023):

This dataset was chosen because of its large volume and comprehensive feature set, especially about environmental and temporal conditions such as weather, road conditions, and timestamps. It provides detailed contextual data before accidents occur, which is valuable for severity prediction.

- The **Severity** is classified into **4 levels**:
  1 (Low), 2 (Moderate), 3 (Serious), and 4 (High Severity).

Table 3.2 Key Features of the US Accidents Dataset.

Feature Group	Name Feature	Feature Type	Accident	Number Classes
Identifier	ID	Categorical	After	7728394
	Source			3
Target	Severity	Numerical	After	4
	Start_Time		Before	6131796

Feature Group	Name Feature	Feature Type	Accident	Number Classes
Temporal	End_Time	Categorical	After	6705355
	Timezone		Before	4
Location	Start_Lat	Numerical	Before	2428358
	Start_Lng			2482533
	End_Lat		After	1568172
	End_Lng			1605789
	Distance(mi)			22382
	Description	Categorical		3761578
	Street		Before	336306
	City			13678
	County			1871
Location	State	Categorical	Before	49
	Zipcode			825094
	Country			1
	Airport_Code			2045
Weather	Weather_Timestamp	Categorical	After	941331
	Temperature(F)	Numerical	Before	860
	Wind_Chill(F)			1001
	Humidity(%)			100
	Pressure(in)			1144
Weather	Visibility(mi)	Numerical	Before	92
	Wind_Direction	Categorical		24
	Wind_Speed(mph)	Numerical		184
	Precipitation(in)			299
	Weather_Condition	Categorical		144
	Amenity			
	Bump			
	Crossing			

Feature Group	Name Feature	Feature Type	Accident	Number Classes
Road	Give_Way	Boolean	Before	2
	Junction			
	No_Exit			
	Railway			
	Roundabout			
	Station			
	Stop			
	Traffic_Calming			
	Traffic_Signal			
	Turning_Loop			1
Twilight Conditions	Sunrise_Sunset	Categorical	Before	2
	Civil_Twilight			
Twilight Conditions	Nautical_Twilight	Categorical	Before	2
	Astronomical_Twilight			

2.Road Traffic Accidents (RTA):

Selected for its focus on driver-specific and behavioral attributes not present in the US dataset, offering a complementary perspective. It includes details such as driver age, experience, vehicle condition, and pedestrian information, which are crucial for a well-rounded severity model.

- The **Severity** is classified into **3 levels**:

Slight Injury(1), Serious Injury(2), and Fatal Injury(3).

Table 3.3 Key Features of the Road Traffic Accidents.

Feature Group	Name Feature	Feature Type	Accident	Number Classes
Temporal	Time	Categorical	Before	1074
	Day_of_week			7
Driver Demographics	Age_band_of_driver	Categorical	Before	5
	Sex_of_driver			3
	Educational_level			7
Driver-Vehicle Info	Vehicle_driver_relation	Categorical	Before	4
	Driving_experience			7
Vehicle Info	Type_of_vehicle	Categorical	After	17
	Owner_of_vehicle			4
	Service_year_of_vehicle			6
	Defect_of_vehicle			3
Road Environment	Area_accident_occured	Categorical	After	14
Road Environment	Lanes_or_Medians	Categorical	Before	7
	Road_allignment			9
Road Environment	Types_of_Junction	Categorical	Before	8
	Road_surface_type			5
Road Conditions	Road_surface_conditions	Categorical	Before	4
	Light_conditions			4
	Weather_conditions			9
Accident Characteristics	Type_of_collision		Before	10
Accident Characteristics	Number_of_vehicles_involved	Numerical	After	6
	Number_of_casualties	Numerical	After	8
	Vehicle_movement	Categorical		13
	Cause_of_accident		Before	20
Casualty Info	Casualty_class	Categorical	After	4
	Sex_of_casualty			3

Feature Group	Name Feature	Feature Type	Accident	Number Classes
Casualty Info	Age_band_of_casualty	Categorical	After	6
	Casualty_severity			4
	Work_of_casuality			7
	Fitness_of_casuality			5
Pedestrian Info	Pedestrian_movement	Categorical	After	9
Target	Accident_severity	Categorical	After	3

3.UK Road Accident:

The UK dataset is similar to the US Accidents dataset in structure and weather-related data, but it is less detailed and contains fewer records, which reduces its statistical significance. Its features are also more limited compared to the US dataset (which includes traffic flow, visibility, and wind direction) or the RTA dataset (which covers driver behavior, vehicle type, and driver condition), making it less suitable for developing a comprehensive and generalizable model.

4.Road Accident Dataset:

This dataset shares similarities with both the **US Accidents dataset** (in terms of weather-related data) and the **RTA dataset** (in terms of driver and road conditions). However, it is **incomplete,** many important features are missing, and the ones that are present are **not sufficiently representative** of real-world conditions.

5.Global Road Accidents Dataset:

Although it is large, the **Global Road Accidents Dataset** suffers from **low variability;** many of its features have uniform values across all entries, which is indicative of poor data quality. This raises concerns that the dataset may be **synthetically generated** or excessively cleaned, making it less reliable for training predictive models.

6.Road Accident in Montreal:

This dataset offers **68 features**, but only **25 are publicly available**, while the rest are either internal or require government approval to use. This restriction prevents **full utilization** of the dataset and limits the ability of the model to leverage all available data.

3.2.Dataset selection for accident frequency prediction

3.2.1.Candidate datasets

For accident frequency prediction, we will focus on four well-known datasets. These datasets contain regular and daily accident data collected with few or no interruptions, making them suitable for frequency prediction.

Table 3.4 Key Details of Each Dataset.

Dataset Name	Data Source	Time Coverage	Records Count	Temporal Variables
Road Accident (UK)	Kaggle ,Devansodariya) (2023	2015–2005	+1,500,000	(Time ,Date)
Thailand Fatal Road Accident	Kaggle Thaweewatboy) (2023 ,	2022–2011	200,000	Date
Great Britain Road Accidents	Kaggle ,Nichaoku) (2022	2016–2005	+1,900,000	(Time ,Date)
million UK 1.6 traffic accidents	Kaggle (2021 ,Hickey)	2016–2005	1,600,000	Date

3.2.2.Selected and excluded datasets

Accident frequency prediction does not rely directly on datasets that contain information about accidents; instead, it requires the transformation of these datasets into what we call a time series. This transformation is called time series creation. Therefore, selecting the best dataset for frequency prediction comes down to choosing the best time series.

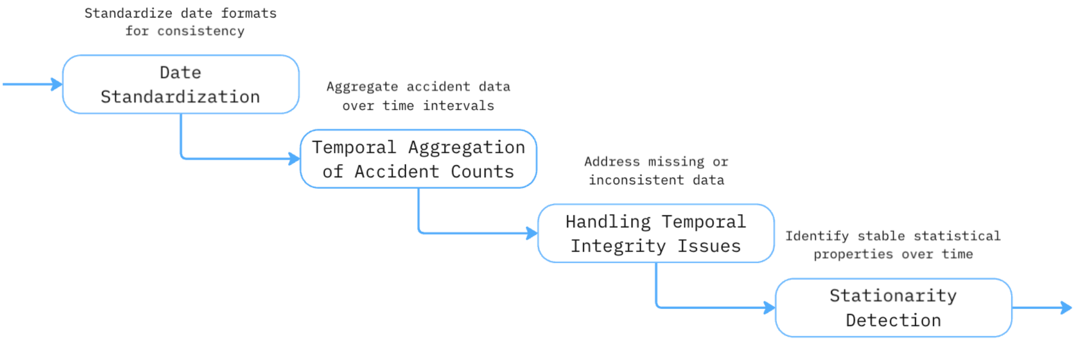
A time series of accident frequency is a sequence of data points representing the number of accidents recorded over specific time intervals (every day, every week, or every year).

Table 3.5 Example of a Time Series for Accident Frequency.

Day	Accidents number
3-11-2024	10
4-11-2024	4
5-11-2024	6
6-11-2024	9
7-11-2024	4
8-11-2024	3
9-11-2024	5

To create time series and select the best datasets, we follow a four-step approach:

Figure 3.3 Time Series Creation and Data Set Selection Steps.


Standardize date formats for consistency
Date Standardization
Aggregate accident data over time intervals
Temporal Aggregation of Accident Counts
Address missing or inconsistent data
Handling Temporal Integrity Issues
Identify stable statistical properties over time
Stationarity Detection

- **Date standardization**

In time series analysis, maintaining temporal consistency across multiple datasets is essential for accurate modeling and comparison. However, real-world data often comes from diverse sources with varying date formats and naming conventions. In this step, we parse inconsistent date strings and correct format errors.

- **Temporal aggregation of accident counts**

Temporal aggregation of accident counts refers to the process of summarizing raw event data (typically timestamped per incident) into fixed time intervals such as daily, weekly, or monthly totals [23]. In this project, daily aggregation was selected.

Table 3.6 Temporal Aggregation Results.

Dataset Name	Raw Records	Start Data	End Data	Total Days
Road Accident (UK)	1,917,274	01-01-2005	2016-12-31	4,383
Thailand Fatal Road Accident	1,780,653	2005-01-01	31-12-2015	4,017
Great Britain Road Accidents	1,504,150	2005-01-01	31-12-2014	3,653
1.6 million UK traffic accidents	1,504,150	2005-01-01	31-12-2014	3,653

- **Handling temporal integrity issues**

Temporal integrity was preserved by correcting missing dates through dataset-specific strategies [24].

For example, in Thailand, the Fatal Road Accidents dataset exhibited 88 missing days. A complete daily calendar was created and merged with the original data. Missing values were handled using a two-step imputation strategy:

- Forward-fill (ffill) for short gaps.
- 7-day rolling mean for residual missing values.
- Result: A continuous, imputed time series ready for modeling.

Figure 3.4 Daily Accident Count Time Series with Missing Days.

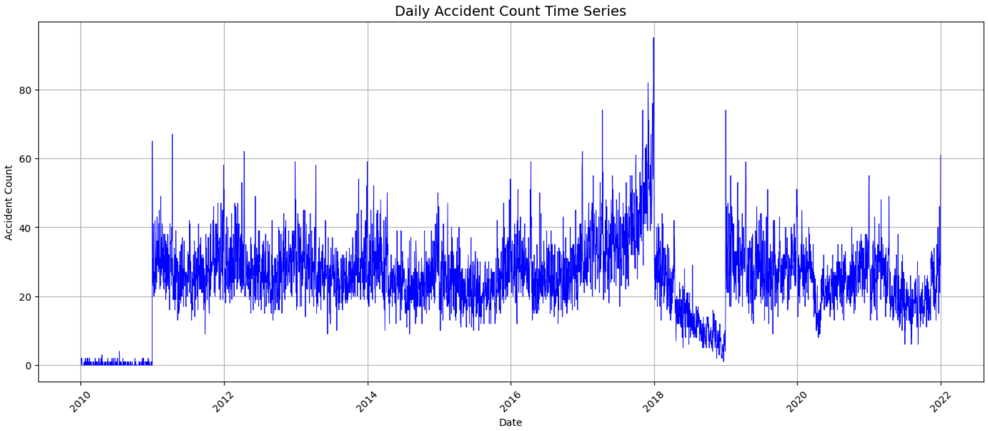
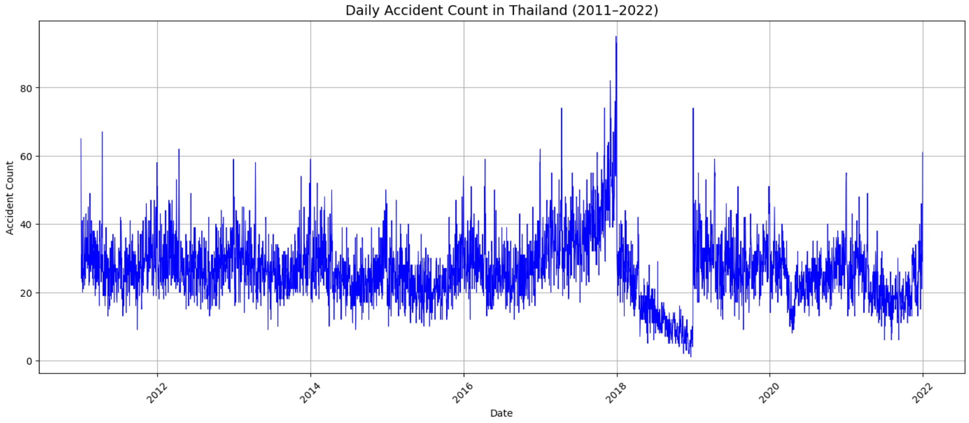


Figure 3.5 Daily Accident Count Time Series without Missing Days.



The "1.6 Million UK Traffic Accidents" and "Road Accidents UK" datasets lacked all records for 2008. To preserve the continuity of the time series without introducing artificial values, data from 2009 onward was shifted back by one year (e.g., 2008→ 2009). This displacement maintained trend integrity and seasonal patterns essential for modeling.

Figure 3.6 Daily Accident Count Time Series with Missing Days.

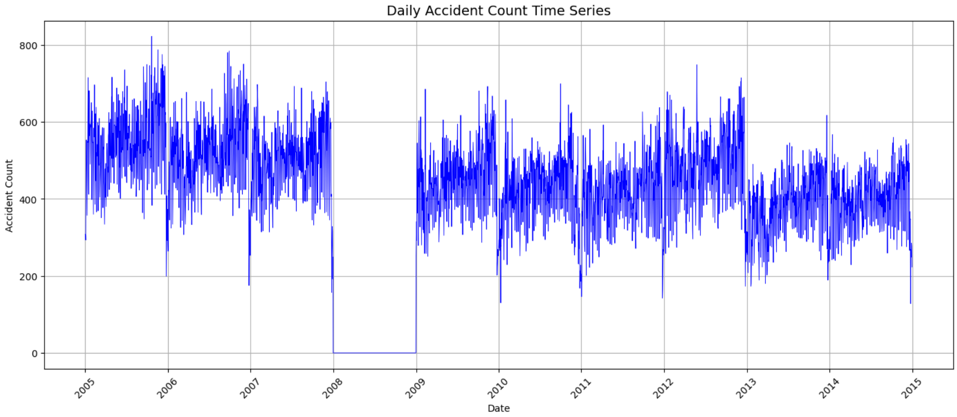
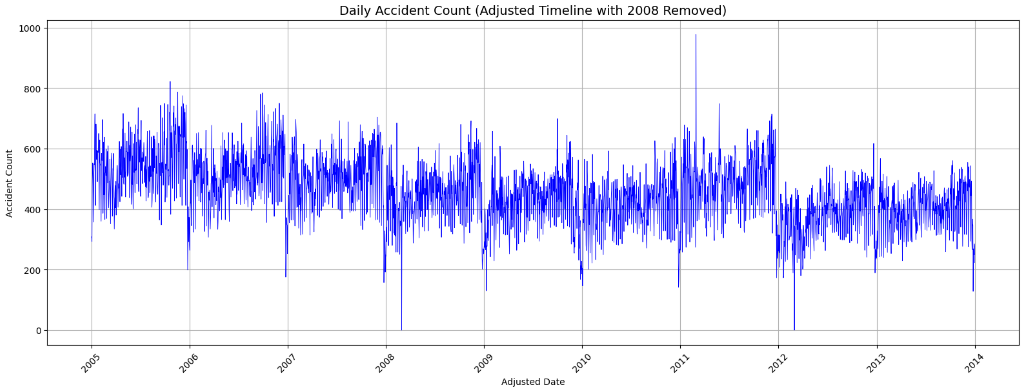


Figure 3.7 Daily Accident Count Time without Missing Days.



❋ Stationarity detection

Employing a stationarity test is essential for selecting the most suitable time series to train our model. A stationary time series maintains a consistent mean and variance over time, enabling more reliable predictions. To identify stationarity, we utilize the Augmented Dickey-Fuller (ADF) test, which is commonly applied in time series analysis. In our work, we use this test without delving into complex mathematical formulas [25].

The **ADF Test** was applied to each dataset's daily accident count series:

- **p-value ≤ 0.05 ➔** Stationary.
- **p-value > 0.05 ➔** Non-stationary.

Table 3.7 ADF Test Results.

Dataset	p-value ADF	Stationarity Status
UK Road Accident	0.060234	Non-Stationary
Thailand Fatal Road Accident	0.010689	Stationary
Great Britain Road Accidents	0.000030	Stationary
1.6M UK Traffic Accidents	0.057034	Non-Stationary

The results indicate that only two time series are stationary: Great Britain Road Accidents and Thailand Fatal Road Accidents. The other two time series, Road Accidents (UK) and 1.6 million UK traffic accidents, are not stationary and have significant gaps of missing days, as noted in the previous section. Therefore, we will select the first two datasets for our training.

4.Conclusion

In summary, choosing the right datasets is essential for effective road safety research and for building reliable models that predict accident severity and frequency. This chapter examined various datasets for severity prediction, highlighting the **US Accidents** and **Road Traffic Accident** Dataset due to its extensive features and significant volume. For frequency prediction, datasets such as the **Thailand Fatal Road Accident** and **Great Britain Road** were selected for their temporal continuity and stationarity. The next chapter will delve into these datasets, utilizing advanced machine learning and deep learning techniques to build accurate models.

Chapter 04

Data Preparation, Model Building and Performance Evaluation

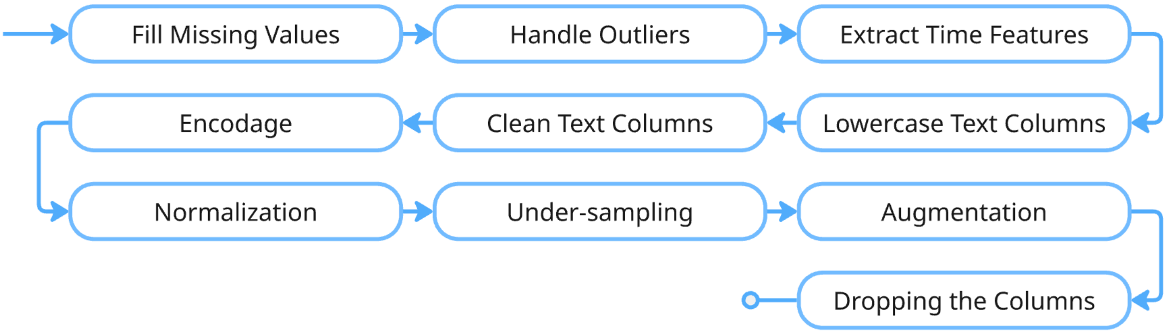
1. Introduction

In this chapter, we will explore the critical steps of data preparation, model building, and performance evaluation within the modeling process CRISP-ML. These steps are essential for ensuring that our models are both accurate and reliable, as they involve cleaning and transforming raw data into a usable format, constructing effective models, and rigorously assessing their performance.

2. Data Preparation for Severity Prediction

This section describes the key preprocessing steps used to prepare road traffic accident datasets for the model-building step. It focuses on cleaning and structuring the data by addressing missing values, outliers, and format inconsistencies. Additionally, it discusses undersampling and augmentation techniques to handle imbalanced data.

Figure 4.1 Data Preparation Steps for Severity Prediction.



2.1. Fill Missing Values

Missing values were addressed to ensure completeness and improve data quality before the modeling process. Initially, the proportion of missing values in each column was examined individually. If the percentage of missing values in a column exceeded 50%, it was completely deleted as it was unreliable and negatively impacted the accuracy of the analysis [26].

For columns with less than 50% missing values, those values were filled in depending on the data type:

- In the case of numeric data (such as temperature or humidity), the median was used instead of the mean, in order to avoid the influence of outliers on the filling process.
- In the case of categorical data (such as road type or weather condition), the most frequent value (Mode) was used to fill in the blanks, as it represents the most common category.

Table 4.1 Example of Missing Value Handling.

Column Name	Data Type	Missing Value Percentage	Column Name
Sunset_Sunrise	Categorical	58%	Dropped (Too many missing)
Humidity(%)	Numeric	12%	Filled with Median
Type_of_vehicle	Categorical	8%	Filled with Mode
Visibility(mi)	Numeric	3%	Filled with Median
Sex_of_driver	Categorical	1%	Filled with Mode

2.2.Handle Outliers

Outliers are data points that differ significantly from other observations in a dataset. In our project, outliers were addressed using the interquartile range (IQR) method. This involved calculating the range between the first quartile (Q1) and the third quartile (Q3) and setting the lower and upper limits as 1.5 times the Q1 average, positioned below Q1 and above Q3, respectively. Any data points that fell outside these limits were replaced with the median value of the corresponding attribute. This approach helps mitigate the impact of outliers on model performance, enhances model stability, and reduces bias while preserving the overall integrity of the dataset [27].

Table 4.2 Example of Outlier Handling Using IQR Method.

Feature Name	Q1	Q2	Q3	Lower Bound	Upper Bound	Outlier Treatment
Vehicle_movement	40	70	30	40 − 1.5×30 = −5	70 + 1.5×30 = 115	Replaced with Median
Temperature(F)	10	25	15	10 − 1.5×15 = −12.5	25 + 1.5×15 = 47.5	Replaced with Median

2.3.Extract Time Features

Temporal features were engineered from the timestamp columns in both datasets. In the **US dataset**, the **Start_Time** column *(e.g., 2016-02-11 07:18:39)* was used to extract the year, month, day of the week, hour, minute, and time of day (categorized into morning, afternoon, evening, or night).

Table 4.3 Extracted Time Features from Start_Time Column.

Start_Time	Extracted Feature	Value
2016-02-11 07:18:39	Year	2016
	Month	2
	Day_of_Week	Thursday
	Hour	7
	Minute	18
	Time_of_Day	Morning

2.4.Lowercase Text Columns

Text columns containing categorical data, such as “Weather_Condition”, have been standardized by converting all text entries to lowercase. This normalization step ensures consistency across the dataset by treating entries such as “Light Rain” and “light rain” as identical values [26].

Table 4.4 Standardizing Text Columns to Lowercase.

Column Name	Original Value	Standardized Value
Direction_Wind	Light Rain	light rain
Weather_Condition	HEAVY SNOW	heavy snow
Type_of_collision	No junction	no junction
Vehicle_movement	Going Straight	going straight

2.5.Clean Text Columns

Text columns have been cleaned by removing special characters and extra spaces from entries. This preprocessing improves the uniformity and quality of the text data [26].

Table 4.5 Cleaning Text Columns.

Column Name	Original Value	Cleaned Value
Light_conditions	Darkness - lights lit	Darkness lights lit
Wind_Direction	S/SW	ssw

2.6.Encoding

Ordinal Encoding was applied to convert categorical values into integers based on their order. All columns with categorical data types were identified and encoded to ensure compatibility with numerical model input requirements [28].

Table 4.6 Ordinal Encoding of Categorical Features.

Column Name	Original Categories (Sample)	Encoded Values (Sample)
Weather_Condition	Rain, light rain, Overcast	0, 1, 2
Wind_Direction	Calm, SW, SSW, WSW	0, 1, 2, 3
Sex_of_driver	Male, Unknown, Female	0, 1, 2
Day_of_week	Monday, Sunday, Friday	0, 1, 2

2.7.Normalization

Normalization is a key preprocessing step used to ensure that all input features contribute equally to machine learning models by bringing them into a common scale. Among the various normalization methods, **Min-Max Normalization** was employed in this study. This technique linearly rescales feature values to a fixed range, typically between 0 and 1, preserving the relative distances between data points [29].

$$X_{\text{normalized}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

Table 4.7 Feature Normalization Using Min-Max Scaling.

Feature Name	Sample Value	Original Min	Original Max	Normalized Value	Formula Used
Visibility	5	0	10	0.50	(5 − 0) / (10 − 0)
Wind_Speed	10	0	50	0.20	(10 − 0) / (50 − 0)

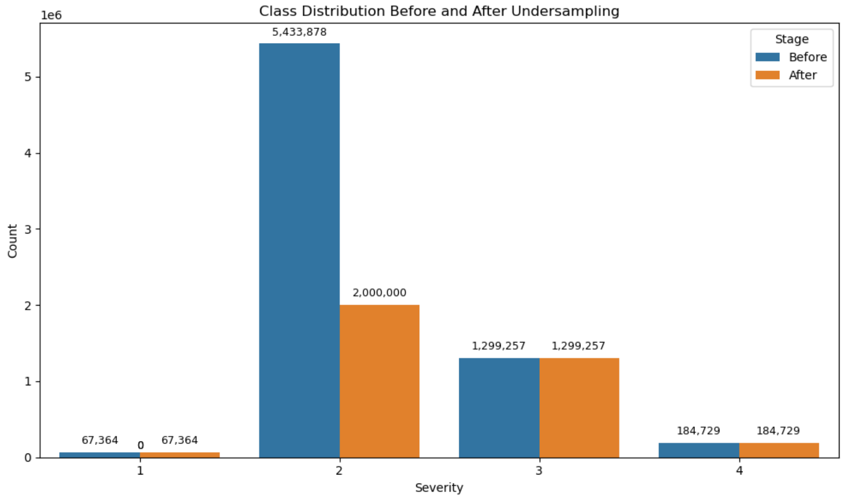
2.8.Undersampling

To address class imbalance within the US dataset, undersampling was applied by reducing the number of samples in the overrepresented class *(severity level 2)* to a fixed target count [30].

Table 4.8 Class Distribution of Severity Levels Before and After Undersampling.

Severity	Before under-sampling	After under-sampling
1	67364	67364
2	5433878	2000000
3	1299257	1299257
4	184729	184729

Figure 4.2 Distribution of Severity Classes Before and After Under-sampling.



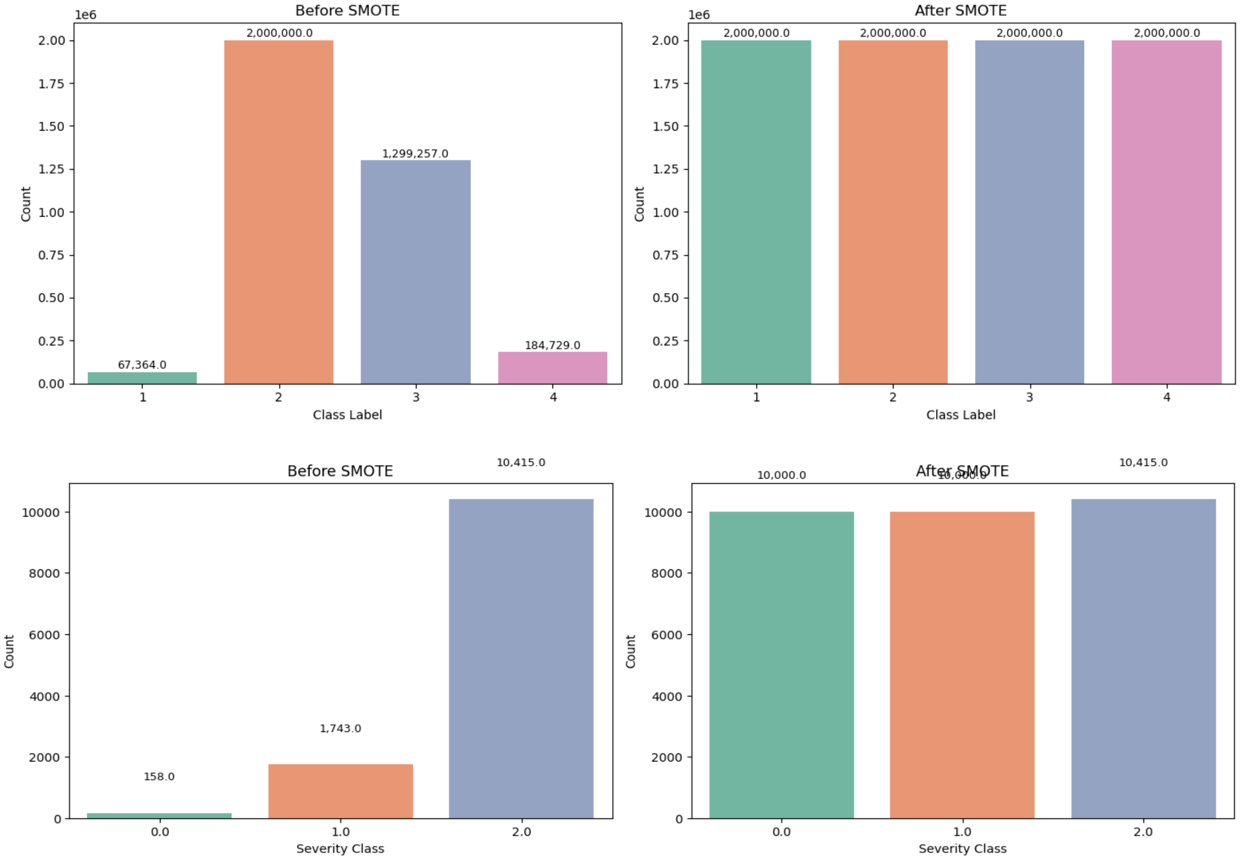
2.9.Augmentation

Following the application of undersampling to reduce the dominance of severity class 2, several augmentation techniques were explored to enhance the representation of minority classes *(severity levels 1, 3, and 4)* and achieve a more balanced class distribution. The tested methods included **SMOTE**, **SMOTE-Tomek**, **Random Oversampling**, **SMOTE-ENN**, and **SMOTE-NC**. Each method aimed to improve class balance and model generalization by generating or duplicating minority class samples. Among these, **SMOTE** was ultimately applied, as it effectively generates synthetic samples by interpolating between existing instances, enhancing minority class representation while maintaining data consistency [31].

Table 4.9 Summary of Data Augmentation Methods for Addressing Class Imbalance.

Data	Severity	Before augmentation	After augmentation
US	1	67364	2000000
	2	2000000	2000000
	3	1299257	2000000
	4	184729	2000000
RTA	0	158	10000
	1	1743	10000
	2	10415	10415

Figure 4.3 Severity class Before and After SMOTE.



2.10.Dropping unused features

In this step, irrelevant features were removed from the dataset to improve model efficiency and interpretability. These included post-accident details and identifiers that do not contribute to predicting accident severity. The removal process ensured the model focuses on relevant predictors only [32].

Table 4.10 Dropped Columns and Justification.

Column Name	Reason for Removal
ID	Unique identifier; carries no predictive information.
Distance(mi)	Post-accident outcome; not available at prediction time.
Casualty_severity	Post-accident outcome; not available at prediction time.

3.Data preparation for frequency prediction

Since our frequency prediction model relies only on simple time series, the only necessary data preparation step is normalization.

3.1.Time Series Normalization

After aggregating accident counts daily, **Min-Max normalization** is applied to scale the values into the [0, 1] range [33].

This transformation is essential for deep learning models such as **LSTM** and **GRU** [34].

Table 4.11 Examples Feature Normalization Using Min-Max Scaling.

Original Value	Normalized Value
2	0.68085106
308	0.26988636

4.Development Environment and Tools Used

Kaggle is an online platform designed for data science and machine learning (ML) projects. It allows users to work directly in the cloud without needing to install anything on their personal computers. In our project, we used Kaggle to load, analyze, and train models using Python. We didn't need to download the datasets to our local machines, as everything was done online using built-in tools.

One of **Kaggle's** most useful features is its free access to GPUs, which speeds up model training. However, GPU usage is limited to a few hours per week. We also benefited from the pre-installed libraries such as pandas, NumPy, matplotlib, TensorFlow, and scikit-learn, which made it easier and faster to start working. Overall, Kaggle allowed us to manage our code, test models, and visualize results in a simple, efficient, and well-organized environment.

5. Model Training and Evaluation

This section outlines the training and evaluation of models used to predict accident severity and frequency. It includes data splitting, model selection, training, and performance assessment using standard metrics.

5.1. Training and evaluation for severity prediction

This section describes how models were trained and evaluated for severity prediction.

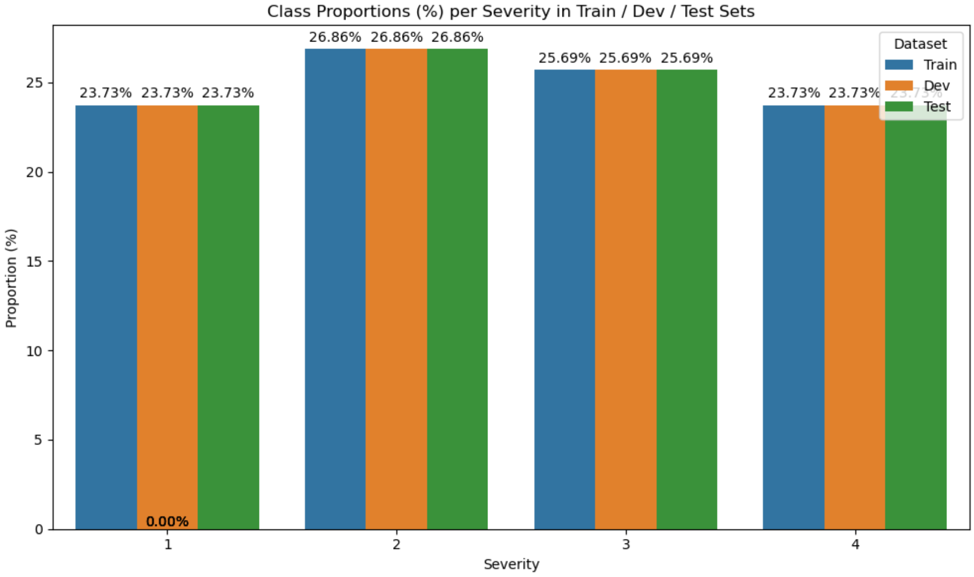
5.1.1. Data splitting

Before selecting and training models, the datasets were split into training (70%), validation (15%), and test (15%) subsets to support effective model training. Stratified sampling was used to maintain the original distribution of accident severity classes across all subsets, ensuring balanced representation [35]. This method improves the reliability and accuracy of performance assessment. The class distribution for each subset is shown in *Table 4.12* and *Figure 4.4*.

Table 4.12 Stratified Distribution of Classes Across Training, Validation, and Test Sets.

Split Data	Classes	Size	Proportions(%)		Class Counts
Trainig	1	3540408	70%	23.73%	840000
	2			26.86%	950928
	3			25.69%	909480
	4			23.73%	840000
Dev	1	758659	15%	23.73%	180000
	2			26.86%	203771
	3			25.69%	194888
	4			23.73%	180000
Test	1	758659	15%	23.73%	180000
	2			26.86%	203770
	3			25.69%	194889
	4			23.73%	180000

Figure 4.4 Class Proportions (%) per Severity in Train / Dev / Test Sets.



5.1.2.Optimizing Memory Usage

Memory optimization was applied by converting data types to more efficient formats, reducing memory usage while preserving data accuracy. Numerical and categorical variables were adjusted to minimize storage requirements, resulting in faster processing and improved system performance, especially with large datasets. This step contributed to better computational efficiency and resource management [36]. The impact of this optimization is summarized in *Table 4.13*, which compares memory usage before and after the adjustments.

Table 4.13 Memory Usage of Dataset Splits Before and After Optimization (MB).

Dataset	Memory Usage Before Optimization (MB)	Memory Usage After Optimization (MB)
X_train	1053.436	540.223
X_dev	225.736	115.762
X_test	225.736	115.762

5.1.3.Validation metrics of ML & DL

To comprehend model strengths, detect weaknesses, and enable comparisons, we utilize confusion matrices and evaluation metrics. This section will clearly explain these measures and their practical application.

Confusion matrix

A confusion matrix is a tabular tool used to measure classification accuracy. It tracks model performance by summarizing predicted outcomes against actual results [37]. For binary classification, the confusion matrix *Table 4.14* includes four components:

- TP (True Positives): correctly predicted positive cases.
- TN (True Negatives): correctly predicted negative cases.
- FP (False Positives): instances where the model incorrectly labels a true negative as positive.
- FN (False Negatives): instances where the model incorrectly labels a true positive as negative.

Table 4.14 Confusion Matrix.

		Predicted Class	
		+	-
Class	+	**TP**	**FN**
	-	**FP**	**TN**

❋ Performance evaluation metrics

This section outlines key performance metrics used to evaluate model effectiveness, all derived from the components of the confusion matrix: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

(a) Accuracy represents the proportion of correct predictions, both positive (TP) and negative (TN), out of all predictions made by the model. It is most reliable with balanced datasets, but can be misleading with imbalanced data.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

(b) Precision measures the accuracy of positive predictions. It is calculated by dividing the number of true positives by the total number of predicted positives. A high number of false positives will lower the precision.

$$Precision = \frac{TP}{TP + FP}$$

(c) Recall indicates the model's ability to correctly identify all actual positive cases. It is computed by dividing true positives by the total actual positives. A low recall implies a high number of false negatives.

$$Recall = \frac{TP}{TP + FN}$$

(d) F1-Score provides a balanced measure by combining precision and recall into a single metric using their harmonic mean.

$$F1 - Score = \frac{2 \times \mathrm{Recall} \times \mathrm{Precision}}{\mathrm{Recall} + \mathrm{Precision}}$$

5.1.4.Models training

For accident severity prediction, we train and compare the performance of six models:

1. Decision Tree Model

The Decision Tree (DT) model is a straightforward and interpretable algorithm commonly used in classification tasks due to its low computational cost [38]. In this study, its performance in predicting accident severity was evaluated using two datasets, with results and confusion matrices summarized in *Tables 4.15*, *4.16* and *Figure 4.5*.

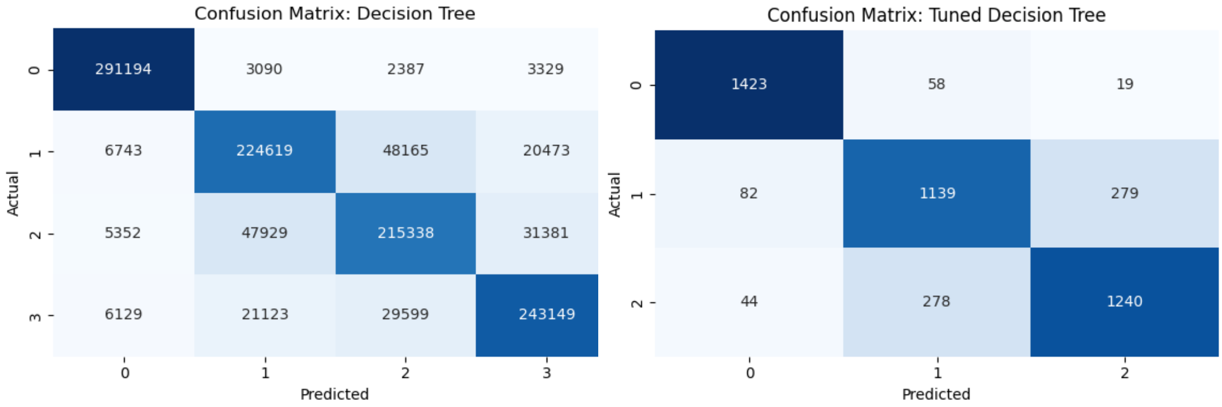
Table 4.15 Results of Decision Tree Models in Multi-class Classification - UK Dataset.

severity	Accuracy	Precision		Recall		F1-score	
1	0.8119	0.9411	0.8104	0.9706	0.8119	0.9556	0.8111
2		0.7569		0.7487		0.7528	
3		0.7288		0.7178		0.7232	
4		0.8150		0.8105		0.8128	

Table 4.16 Results of Decision Tree Models in Multi-class Classification - RTA Dataset.

severity	Accuracy	Precision		Recall		F1-score	
1	0.8334	0.9187	0.8320	0.9487	0.8334	0.9334	0.8326
2		0.7722		0.7593		0.7657	
3		0.8062		0.7939		0.8000	

Figure 4.5 Confusion Matrix for Decision Tree - US and RTA data.



2. Random Forest Model

Random Forest (RF) is a reliable and efficient ensemble learning algorithm that builds multiple decision trees and combines their outputs to enhance accuracy and reduce overfitting. It performs well in classification tasks and handles complex datasets with high dimensionality [39]. In this study, the model's effectiveness in predicting accident severity is demonstrated through results and confusion matrices summarized in *Tables 4.17, 4.18*, and *Figure 4.6*.

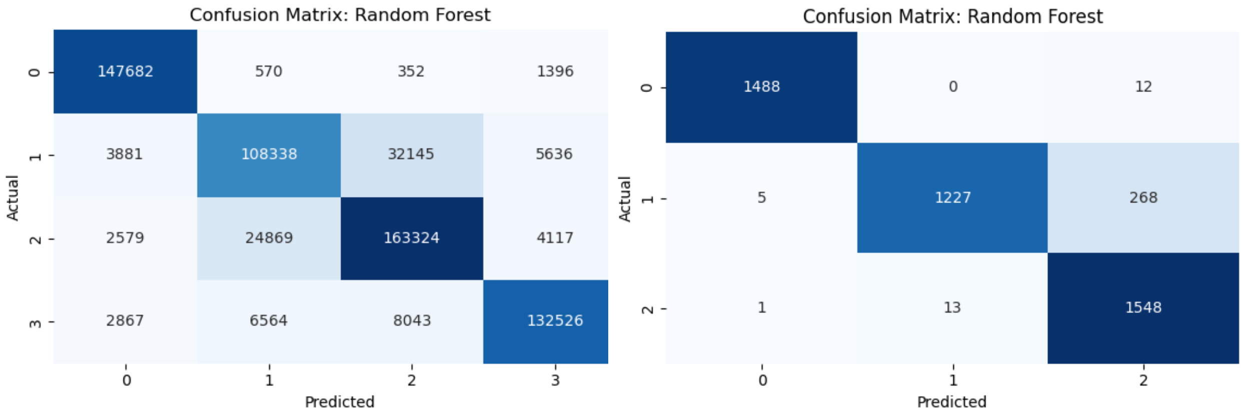
Table 4.17 Results of Random Forest models in multi-class classification - US data.

severity	Accuracy	Precision		Recall		F1-score	
1	0.8752	0.9521	0.8736	0.9915	0.8752	0.9714	0.8742
2		0.8313		0.8036		0.8172	
3		0.8038		0.7893		0.7965	
4		0.9071		0.9162		0.9116	

Table 4.18 Results of Random Forest models in multi-class classification - RTA data.

severity	Accuracy	Precision		Recall		F1-score	
1	0.9345	0.9960	0.9428	0.9920	0.9345	0.9940	0.9340
2		0.9895		0.8180		0.8956	
3		0.8468		0.9910		0.9133	

Figure 4.6 Confusion Matrix for Random Forest - US and RTA data.



Confusion Matrix: Random Forest

Actual \ Predicted	0	1	2	3
0	147682	570	352	1396
1	3881	108338	32145	5636
2	2579	24869	163324	4117
3	2867	6564	8043	132526

Confusion Matrix: Random Forest

Actual \ Predicted	0	1	2
0	1488	0	12
1	5	1227	268
2	1	13	1548

3. LightGBM Model

LightGBM is an efficient gradient boosting framework based on decision tree algorithms, designed for high speed and scalability. It uses a leaf-wise growth strategy with depth constraints, allowing faster training and lower memory usage [40]. In this study, its performance in multi-class accident severity classification is demonstrated using two datasets, with results and confusion matrices presented in *Tables 4.19*, *4.20* and *Figure 4.7*.

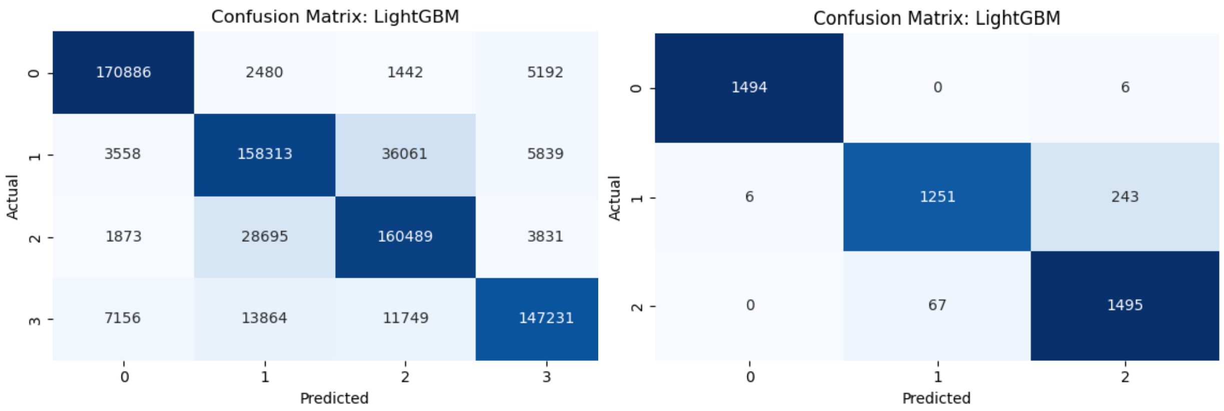
Table 4.19 Results of LightGBM Models in Multi-class Classification - US data.

severity	Accuracy	Precision		Recall		F1-score	
1	0.8467	0.9330	0.8497	0.9526	0.8467	0.9427	0.8472
2		0.7866		0.7888		0.7877	
3		0.7720		0.8312		0.8005	
4		0.9218		0.8230		0.8696	

Table 4.20 Results of LightGBM Models in Multi-class Classification - RTA data.

severity	Accuracy	Precision		Recall		F1-score	
1	0.9294	0.9960	0.9331	0.9960	0.9294	0.9960	0.9291
2		0.9492		0.8340		0.8879	
3		0.8572		0.9571		0.9044	

Figure 4.7 Confusion Matrix for LightGBM - UK and RTA data.



4. AdaBoost Model

AdaBoost is an ensemble learning method that combines multiple weak learners, typically decision trees, into a strong classifier by focusing on misclassified instances [41]. The model's performance in multi-class accident severity classification is reflected in the results and confusion matrices shown in *Tables 4.21*, *4.22*, and *Figure 4.8*.

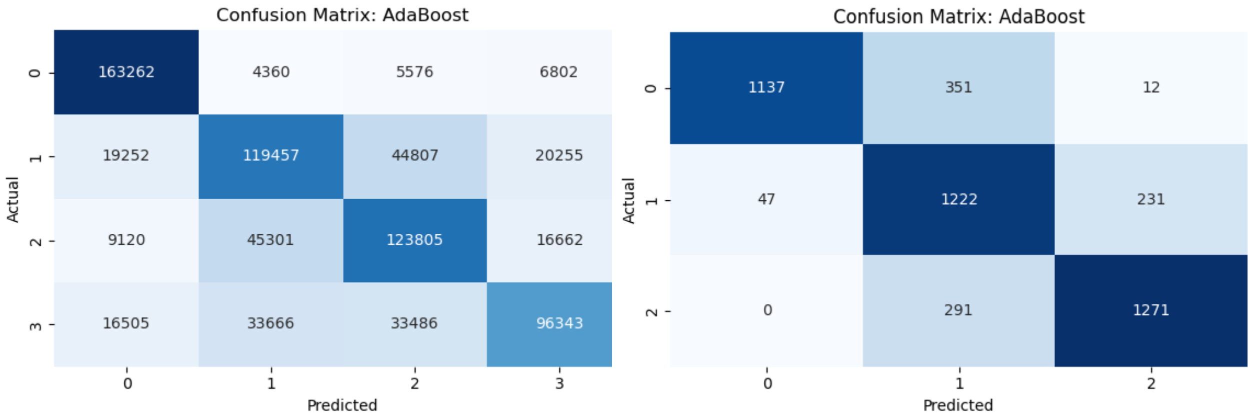
Table 4.21 Results of AdaBoost Models in Multi-class Classification - US data.

severity	Accuracy	Precision		Recall		F1-score	
1	0.6628	0.7844	0.6607	0.9070	0.6628	0.8413	0.6583
2		0.5891		0.5862		0.5877	
3		0.5962		0.6353		0.6151	
4		0.6879		0.5352		0.6020	

Table 4.22 Results of AdaBoost Models in Multi-class Classification - RTA data.

severity	Accuracy	Precision		Recall		F1-score	
1	0.7957	0.9603	0.8187	0.7580	0.7957	0.8472	0.8004
2		0.6556		0.8147		0.7265	
3		0.8395		0.8137		0.8264	

Figure 4.8 Confusion Matrix for AdaBoost - UK and RTA data.



5. CatBoost Model

CatBoost is a gradient boosting algorithm tailored for handling categorical features efficiently. It builds symmetric decision trees and uses techniques like ordered boosting to reduce overfitting [42]. In this study, CatBoost's performance in multi-class accident severity classification is demonstrated through the summarized results and confusion matrices in *Tables 4.23*, *4.24*, and *Figure 4.9*.

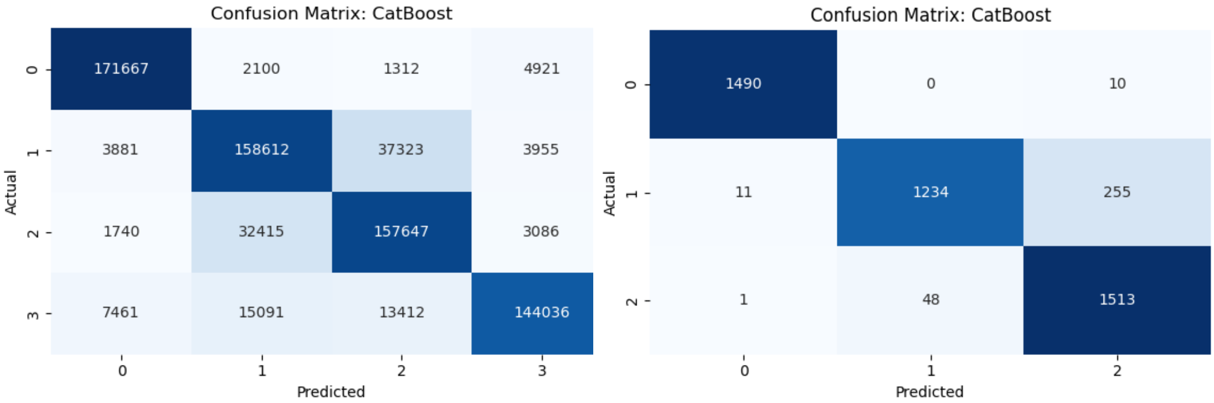
Table 4.23 Results of CatBoost Models in Multi-class Classification - US data.

severity	Accuracy	Precision		Recall		F1-score	
1	0.8330	0.9292	0.8373	0.9537	0.8330	0.9413	0.8338
2		0.7618		0.7784		0.7700	
3	0.8330	0.7518	0.8373	0.8089	0.8330	0.7793	0.8338
4		0.9233		0.8002		0.8574	

Table 4.24 Results of CatBoost Models in Multi-class Classification - RTA data.

severity	Accuracy	Precision		Recall		F1-score	
1	0.9288	0.9920	0.9340	0.9933	0.9288	0.9927	0.9283
2		0.9626		0.8227		0.8871	
3		0.8510		0.9686		0.9060	

Figure 4.9 Confusion Matrix for CatBoost - UK and RTA data.



6. XGBoost Model

XGBoost is a powerful and scalable gradient boosting framework that enhances prediction by iteratively combining weak learners while minimizing error through regularization [43]. It offers fast training, efficient handling of sparse data, and high accuracy. In this study, XGBoost's effectiveness in multi-class accident severity classification is demonstrated through the results and confusion matrices summarized in *Tables 4.25*, *4.26*, and *Figure 4.10.*

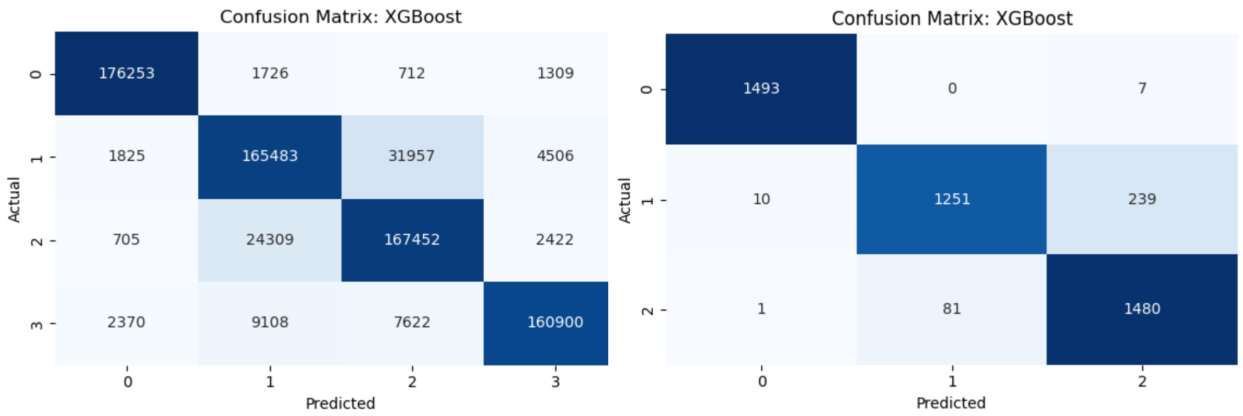
Table 4.25 Results of XGBoost Models in Multi-class Classification - US data.

severity	Accuracy	Precision		Recall		F1-score	
1	0.8833	0.9730	0.8852	0.9792	0.8833	0.9761	0.8838
2		0.8248		0.8121		0.8184	
3		0.8061		0.8592		0.8318	
4		0.9513		0.8939		0.9217	

Table 4.26 Results of XGBoost Models in Multi-class Classification - RTA data.

severity	Accuracy	Precision		Recall		F1-score	
1	0.9259	0.9927	0.9288	0.9953	0.9259	0.9940	0.9256
2		0.9392		0.8340		0.8835	
3		0.8575		0.9475		0.9002	

Figure 4.10 Confusion Matrix for XGBoost - UK and RTA data.



5.1.5.Discussion

A comparative analysis was conducted on several trained models (Decision Tree, Random Forest, LightGBM, AdaBoost, CatBoost, and XGBoost) to evaluate their ability to predict accident severity:

- All models demonstrated improved performance with accuracy and precision values close to 1 (0.88, 0.83, 0.78, etc.).
- The Random Forest and XGBoost achieved the highest accuracy and precision across both datasets, while AdaBoost provided the lowest results among the models.

- In the U.S. dataset, it was observed that the first and last classes were predicted with higher accuracy compared to the two intermediate classes.

As shown in *Table 4.27, Figures 4.11* and *4.12*, these models produced consistent and reliable results, confirming the effectiveness of these methods in handling imbalanced data and supporting their application in real-time road safety initiatives.

Table 4.27 Model Accuracy Comparison: US and RTA data.

Models	Accuracy US	Accuracy RTA
Decision Tree	0.81	0.83
Random Forest	0.86	0.93
LightGBM	0.84	0.93
AdaBoost	0.66	0.80
CatBoost	0.83	0.93
XGBoost	0.88	0.93

Figure 4.11 Model Accuracy Comparison - US data.

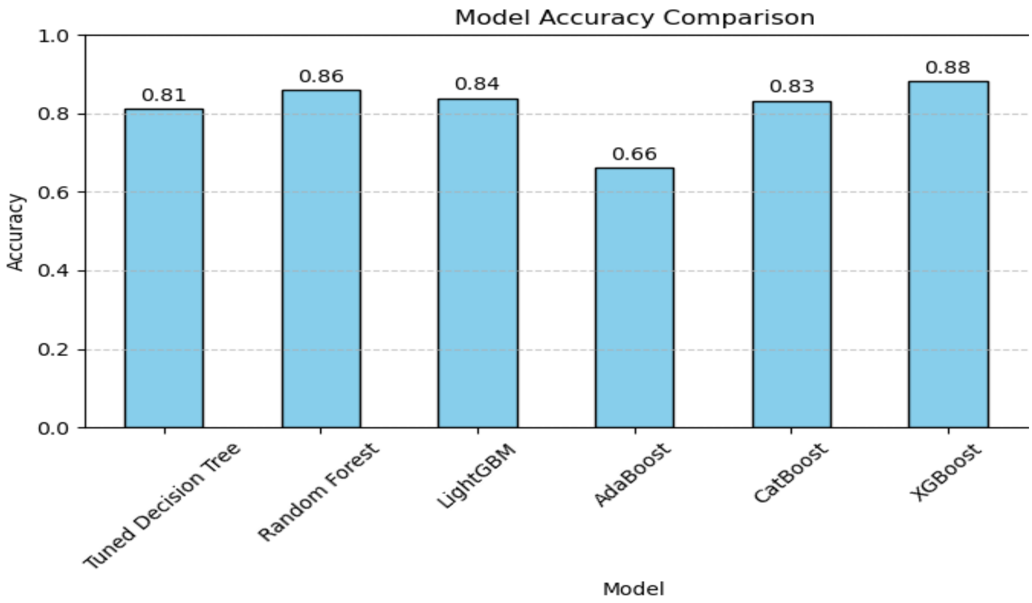
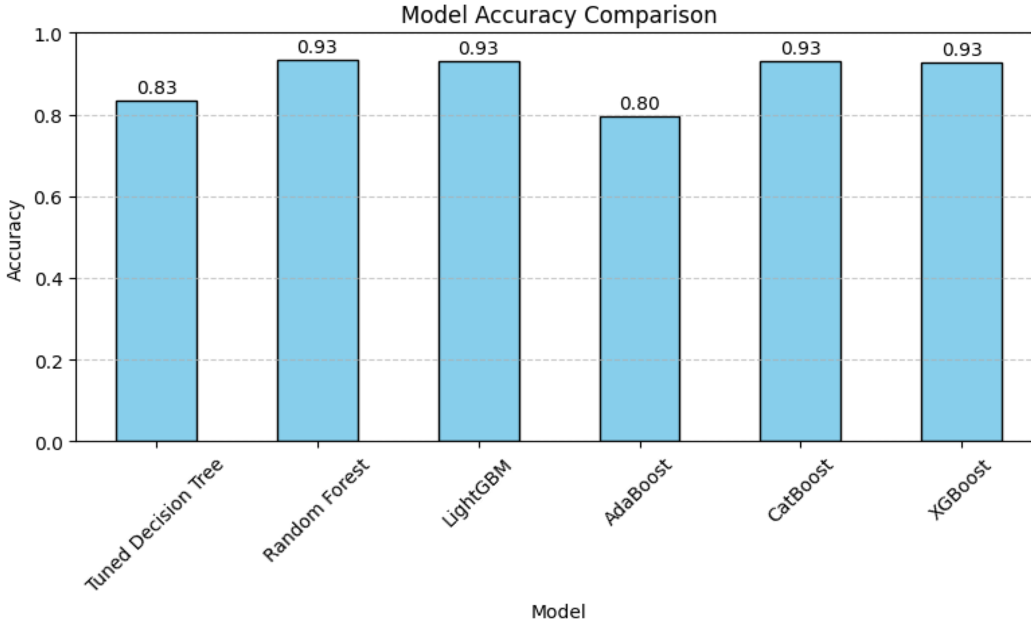


Figure 4.12 Model Accuracy Comparison - RTA data.



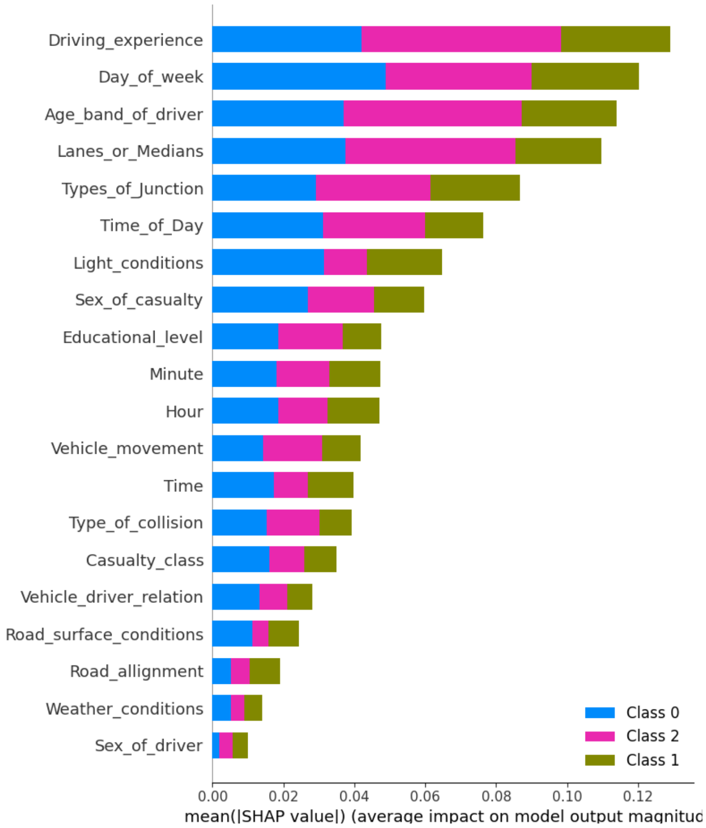
5.1.6.Features importance analysis:

In this section, we used Explainable Machine Learning (SHAP) to identify the most influential features affecting traffic accident severity, offering deeper insight into model decision-making across different severity levels. SHAP (SHapley Additive exPlanations) is an explainable machine learning technique based on game theory that assigns each feature a contribution value to a specific prediction [44]. By leveraging Shapley values, SHAP offers consistent and locally accurate explanations, making it a widely trusted tool for interpreting complex models. It reveals how features impact model predictions.

The overall feature impact on the RTA dataset is visualized in *Figure 4.13*.

Figure 4.13 Global feature importance - RTA data.

The **SHAP** analysis in *Figure 4.13* shows that **Driving_experience**, **Day_of_week**, and **Age_band_of_driver** are the most influential features in predicting accident severity. These variables significantly impact model decisions across all severity classes. In contrast, features like **Sex_of_driver** and **Weather_conditions** have minimal influence. The analysis highlights the importance of driver-related and temporal factors in severity classification.



5.2.Training and evaluation for frequency prediction

This section describes the training and evaluation process for models developed to predict accident frequency. Before starting the training of our models, we must first split our time series and decompose it into sequences. Sequences are the final form of input required by GRU and LSTM models.

5.2.1.Data splitting

The time series was chronologically split into training, validation, and test sets [35].

Table 4.28 Chronological Data Splitting for Time Series Modeling.

Name Data	Split Data	Proportions(%)	Days
Great Britain Road Accidents Dataset	Training	70	3,056
	Validation	15	656
	Test	15	656
Thailand Fatal Road Accidents Dataset	Training	70	3,062
	Validation	15	657
	Test	15	657

5.2.2.Temporal sequence construction for model training

After splitting the dataset into training, validation, and test sets, each segment was independently converted into supervised learning format using a sliding window approach: for a window size of w, each input sequence contained w consecutive normalized values, and the target was the accident count for the next day. Window sizes of 7, 15, and 30 days were tested to assess the impact of sequence length. Result in *Table 4.29.*

Table 4.29 Sequence Generation Summary Using Varying Window Sizes.

Data	Window Size	Before Sequence	After Sequence
Great Britain Road Accidents	7	4383	4376
	15		4368
	30		4353
Thailand Fatal Road Accident	7	4382	4374
	15		4366
	30		4352

5.2.3.Validation metrics

To evaluate the performance of our frequency prediction model, two primary metrics were considered: Mean Squared Logarithmic Error (MSLE) for accuracy, and Per-Run Training Time to assess computational cost during model execution.

- **Mean Squared Logarithmic Error (MSLE)**

MSLE was chosen due to its suitability for **count-based prediction tasks** like accident frequency modeling. It is especially effective at penalizing **underestimations**, which is critical in safety-sensitive applications.

$$\textbf{MSLE} = \frac{1}{n}\sum_{i=1}^{n}\left(\log(1+\hat{y}_i) - \log(1+y_i)\right)^2$$

Where:
y_i= actual value,

$\hat{y}_i$= predicted value,

n = total number of predictions,

The log(1+x) ensures stability for zero values.

MSLE emphasizes **relative error** and reduces the influence of extreme values, making it a robust metric for this task [45].

- **Per-Run Training Time**

For each model (e.g., LSTM, GRU), the **training time in seconds** was recorded during execution. This measures how long the model takes to learn from the training data. Tracking this time helps compare the **computational efficiency** of different models and identify those better suited for real-time or large-scale applications [46].

5.2.4.Models training

To build our severity prediction models, we use two deep learning models: LSTM and GRU.

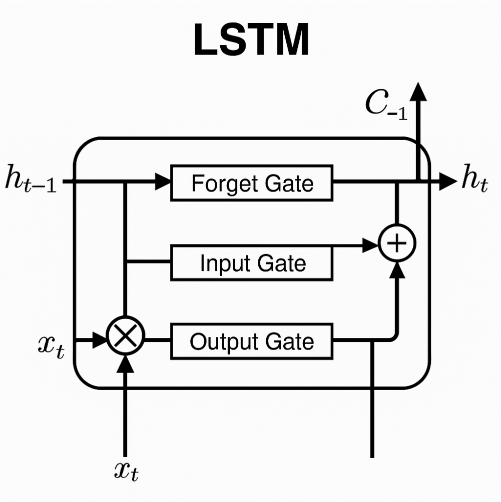
1. **LSTM (Long Short-Term Memory)**

LSTM is a type of recurrent neural network (RNN) designed to learn patterns in sequential data over long periods.

It uses a structure called a memory cell, which allows the network to remember important information and forget irrelevant details.

This makes LSTM especially useful for time series tasks, where past values affect future outcomes, such as predicting daily road accidents [47].

Figure 4.14 LSTM Cell Structure In Hidden Layer.



LSTM architecture

The LSTM models were trained using nine configurations combining three hidden layer depths (0, 2, 4) and three window sizes (7, 15, 30 days).
These configurations were applied separately to two datasets (UK and Thailand).

In addition, several hyperparameters were explored, including the number of epochs, batch size, and different activation functions.

Table 4.30 Comparison Of MSLE On Hidden Layer and Window Size LSTM Variations.

Data	Test	Hidden Layers	Window Size (days)	MSLE	Training Time (s)	The Best Hidden Layers
Great Britain Road Accidents	Test1	0	7	0.018372	19.19	2
		2	7	0.017664	41.81	
		4	7	0.024275	28.76	
	Test2	0	15	0.014538	66.20	0
		2	15	0.031171	20.43	
		4	15	0.032251	39.20	
	Test3	0	30	0.015621	37.51	0
		2	30	0.031901	33.83	
		4	30	0.030895	122.4	
Thailand Fatal Road Accident	Test1	0	7	0.063660	31.65	0
		2	7	0.063960	38.14	
		4	7	0.063723	79.83	
	Test2	0	15	0.063722	24.85	4
		2	15	0.063293	54.21	
		4	15	0.062858	163.03	
	Test3	0	30	0.064213	75.35	0
		2	30	0.066771	35.47	
		4	30	0.064969	247.93	

Table 4.30 shows the performance of the LSTM model using different numbers of hidden layers (0, 2, and 4) and different window sizes (7, 15, and 30 days). The model was tested

on two datasets: road accidents in Great Britain and Thailand. The MSLE value measures prediction accuracy, and training time shows how long the model takes to learn.

We can see that changing the number of layers or the window size affects the results. Some settings gave better accuracy or more stable training, which shows the importance of testing many combinations before choosing the best model.

Figure 4.15 Output Epochs Test Dataset “Great Britain” With LSTM Model.

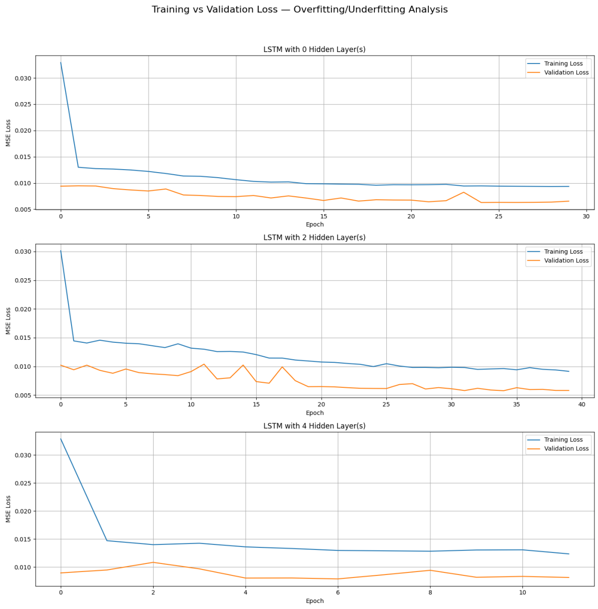
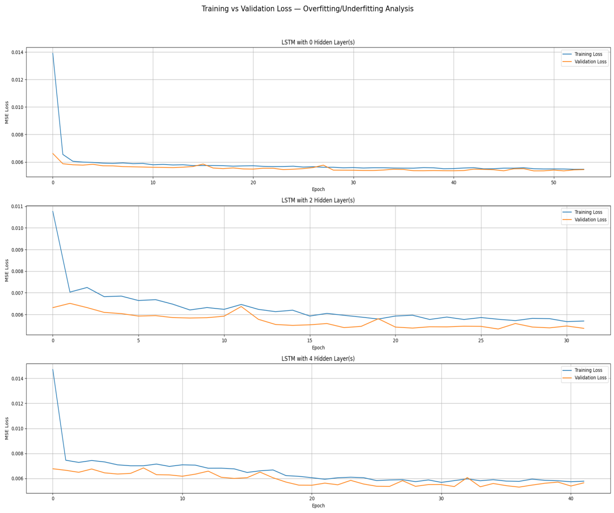


Figure 4.16 Output Epochs Test Dataset “Thailand” With LSTM Model.



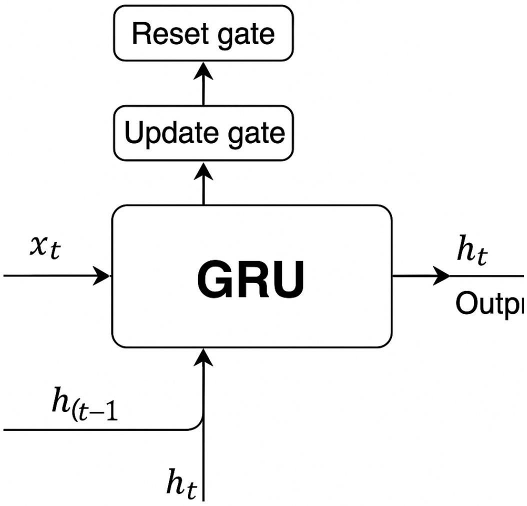
This plot, *Figures 4.15, 4.16,* shows how the loss values changed during training and validation when using the LSTM model. It helps us see if the model is learning well and improving over time. The goal is to make sure the model is not just memorizing the data, but learning in a stable and general way.

From the plot, we can see that simple LSTM models with fewer hidden layers often have smoother and more stable curves. This tells us that a deeper model is not always better, and sometimes a smaller model works just as well

2. GRU (Gated Recurrent Unit)

GRU is a simpler version of LSTM that also handles sequential and time-dependent data. It combines some of the memory functions into fewer gates, making it faster to train and requiring fewer parameters, while still maintaining good performance. GRU is often chosen when computational efficiency is important, without losing much accuracy [48].

Figure 4.17 Internal structure and mechanism of GRU memory cell.



‣ GRU architecture

The GRU model was implemented using three different architectural depths: 0, 2, and 4 hidden layers. These configurations were designed to evaluate how network depth influences the model’s ability to capture temporal dependencies in daily accident data. Each configuration was tested across three window sizes: 7, 15, and 30 days, allowing the model to learn from varying lengths of past observations.

Table 4.31 Comparison Of MSLE On Hidden Layer and Window Size GRU Variations.

Data	Test	Hidden Layers	Window Size (days)	MSLE	Training Time (s)	The Best Hidden Layers
Great Britain Road Accidents	Test1	0	7	0.015643	29.32	0
		2	7	0.017088	49.57	
		4	7	0.015970	81.62	
	Test2	0	15	0.014138	59.69	0
		2	15	0.014548	81.65	
		4	15	0.014281	137.1	
	Test3	0	30	0.013566	107.7	0
		2	30	0.014746	121.4	
		4	30	0.014829	374.8	
Thailand Fatal Road Accident	Test1	0	7	0.063632	19.62	4
		2	7	0.063608	32.31	
		4	7	0.062557	67.40	
	Test2	0	15	0.063683	22.01	0
		2	15	0.063809	52.93	
		4	15	0.064373	68.08	
	Test3	0	30	0.065563	65.16	4
		2	30	0.065140	58.78	
		4	30	0.063318	78.30	

Table 4.31 presents the results of the GRU model under different configurations: 0, 2, and 4 hidden layers, and window sizes of 7, 15, and 30 days. The evaluation was done on both the Great Britain and Thailand datasets. MSLE is used to measure the accuracy, and training time indicates how much time was needed for learning.

From the results, we can observe that some configurations work better than others. The goal of testing many setups is to better understand how the model behaves in different situations, instead of relying on a single structure.

Figure 4.18 Output Epochs Test Dataset “Great Britain” With GRU Model.

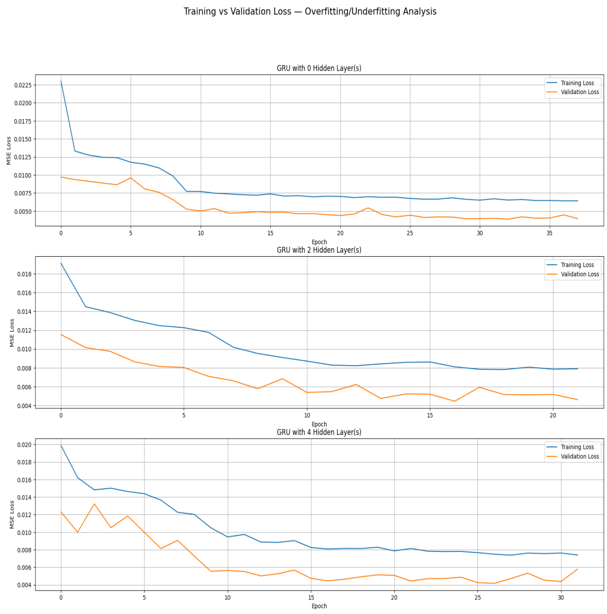
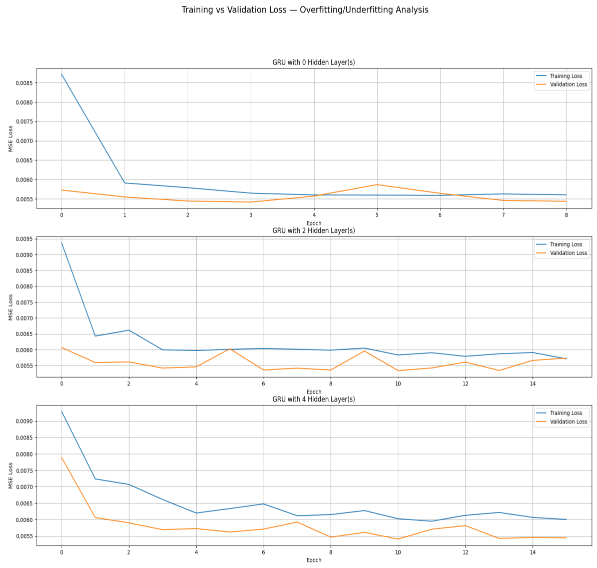


Figure 4.19 Output Epochs Test Dataset “Thailand” With GRU Model.



This plot, *Figures 4.18, 4.19,* shows the training and validation loss when using the GRU model. It is used to check how well the model is learning and to detect any problems like overfitting. The plot gives a clear visual idea of how the model behaves during the learning process.

We noticed that GRU models with fewer layers had more stable learning, while deeper models sometimes showed small changes in the curve. These results helped us better understand which model setups are more reliable.

6. Conclusion

In conclusion, this chapter provided a comprehensive explanation of the practical implementation of ML and DL techniques in the field of road safety. By reviewing various models and evaluating their effectiveness on multiple datasets, our study has proven their ability to achieve excellent results, reflecting the great potential of these techniques in enhancing accident prediction systems and providing more efficient solutions to reduce road traffic risks.

GENERAL CONCLUSION

In conclusion, this project has explored the critical intersection of road safety and advanced technologies such as ML and DL. By focusing on ***Accident Severity Prediction and Accident Frequency Prediction***, we developed and evaluated several predictive models that can serve as a foundation for future research and practical applications.

Through a systematic approach (CRISP-ML), we detailed the entire process, from data selection to model evaluation, ensuring a comprehensive understanding of each stage involved in developing effective predictive models. The results gained from analyzing vast datasets of accident records underscore the potential of ML and DL techniques to enhance road safety significantly.

Like all projects, many obstacles were faced, particularly regarding the quality of the dataset. Inconsistent and incomplete data posed significant challenges, affecting the accuracy of our predictive models. Additionally, the resources required for model training, including computational power and time, were significant.

The completion of this project has significantly deepened our understanding of road safety and ML/DL advanced technologies. By applying the CRISP-ML methodology, we developed valuable skills in data selection, model evaluation, and predictive model development. Analyzing extensive accident datasets enhanced our technical capabilities in machine learning and deep learning.

Looking to the future, there are multiple opportunities for improvement. Subsequent efforts could concentrate on:

- Refining the models by utilizing more diverse datasets.
- Combining the trained models to form a complete predictive architecture.
- Enhancing the thesis by including more comparisons with other research in the same field.
- Leveraging advanced algorithms to improve prediction accuracy, especially with frequency prediction and time series analysis, which require deeper efforts.

- Completing the two final steps of CRISP-ML (deployment and maintenance) will require additional effort.

BIBLOGRAPHY

[1] National Highway Traffic Safety Administration. (2019). *Traffic safety facts: 2019 data*. U.S. Department of Transportation. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812946

[2] World Health Organization. (2013). *Pedestrian safety: A road safety manual for decision-makers and practitioners*. World Health Organization. https://apps.who.int/iris/handle/10665/79753

[3] Campbell, B. J., Zegeer, C. V., Huang, H. H., & Cynecki, M. J. (2004). *A review of pedestrian safety research in the United States and abroad*. Federal Highway Administration. https://www.fhwa.dot.gov/publications/research/safety/pedbike/03042/03042.pdf

[4] United Nations Economic and Social Commission for Western Asia. (n.d.). Road traffic safety. Retrieved June 27, 2025, from https://archive.unescwa.org/road-traffic-safety

[5] Petridou, E., & Moustaki, M. (2000). Human factors in the causation of road traffic crashes.European Journal of Epidemiology,16(9), 819–826. https://doi.org/10.1023/A:1007649804201

[6] Li, G., Brady, J. E., & Chen, Q. (2013). Drug use and fatal motor vehicle crashes: A case-control study. *Accident Analysis & Prevention*, 60, 205–210. https://doi.org/10.1016/j.aap.2013.09.001

[7] Caird, J. K., Johnston, K. A., Willness, C. R., Asbridge, M., & Steel, P. (2014). A meta-analysis of the effects of texting on driving. *Accident Analysis & Prevention*, 71, 311–318.

[8] Philip, P., Taillard, J., Moore, N., Delord, S., Valtat, C., Sagaspe, P., ... & Bioulac, B. (2003). The effects of coffee and napping on nighttime highway driving: A randomized trial. *Annals of Internal Medicine*, 144(11), 785–791.

[9] Christie, R. D., & Smith, J. P. (2023). Vehicle safety factors: Design, technologies, and maintenance in road crash prevention. Journal of Road Safety and Vehicle Engineering,15(2),145–172. https://doi.org/10.1000/jrsve.2023.15.2.145

[10] Awan, B., & Quddus, M. A. (2013). The effect of traffic and road characteristics on road safety: A review and future research directions. Safety Science, 57, 264–275. https://doi.org/10.1016/j.ssci.2013.02.012

[11] Elvik, R., Høye, A., Vaa, T., & Sørensen, M. (2009). *The handbook of road safety measures* (2nd ed.).

[12] Mayhew, D. R., & Simpson, H. M. (2002). The role of driver training and graduated licensing in reducing motor vehicle collisions among young drivers: A review.Accident Analysis & Prevention,34(3), 357–367. https://doi.org/10.1016/S0001-4575(01)00044-2

[13] Lyon, L. et al. (2004). Aggressive traffic enforcement program in Fresno, California: effects on crashes, injuries, and fatalities. *American Journal of Preventive Medicine*. https://pubmed.ncbi.nlm.nih.gov/16688057/

[14] Elvik, R., et al. (2020). A review of data analytic applications in road traffic safety. *International Journal of Transportation Science and Technology*, 9(1), 85–101. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7070501/

[15] Wei, X. (2024). Enhancing road safety in internet of vehicles using deep learning approach for real-time accident prediction and prevention. Internet of Things. https://doi.org/10.1016/j.ijin.2024.05.002

[16] Behboudi, N., Moosavi, S., & Ramnath, R. (2024). Recent advances in traffic accident analysis and prediction: A comprehensive review of machine learning techniques. arXiv preprint arXiv:2406.13705.

[17] Kolyshkina, N., & Simoff, S. (2019). CRISP-ML: A methodology for designing interpretable Machine Learning solutions. *Frontiers in Big Data*. https://doi.org/10.3389/fdata.2021.660206

[18] Studer, S., Bui, T. B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., & Müller, K.-R. (n.d.). Towards CRISP-ML(Q): A machine learning process model with quality assurance methodology.

[19] Guru, S. (2025). CRISP-ML for machine learning products: Managing products that learn, adapt, and occasionally break. Retrieved June 27, 2025.

[20] Mohammed, S., Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Naumann, F., & Harmouch, H. (n.d.). The effects of data quality on machine learning performance on tabular data.

[21] Al-Hasani, G., Khan, A. M., Al-Reesi, H., & Al-Maniri, A. (2019). Diagnostic time series models for road traffic accidents data.International Journal of Applied Statistics and Econometrics,19, 19–41.

[22] Kelleher, J. D., Namee, B. M., & D'Arcy, A. (2015). Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies. MIT Press.

[23] J. Sacks and Y. Zhao, "Temporal aggregation in traffic data: Implications for statistical characteristics and model choice," Transportation Letters: The International Journal of Transportation Research, vol. 3, no. 1, pp. 37–49, Jan. 2011. doi: 10.3328/TL.2011.03.01.37-49.

[24] Abulkhair, A. (2022). Data Imputation Demystified: Time-Series Data. Medium.

[25] Sheiso, D. S., Ware, A. N., & Woyeso, N. M. (2024). Time Series Modelling of Road Traffic Accidents in West Arsi, Ethiopia. Athens Journal of Sciences – Mathematics, 11(2), 15–38.

[26] AI with MTech. (2025). Data Preprocessing in Machine Learning. Retrieved from https://aiwithmtech.com/machine-learning/data-preprocessing-in-machine-learning/

[27] Dastjerdy, B., Saeidi, A., & Heidarzadeh, S. (2023). Review of Applicable Outlier Detection Methods to Treat Geomechanical Data.Geotechnics, 3(2), 375–396. https://doi.org/10.3390/geotechnics3020022

[28] Saxena, A. (2025, April 28). Ordinal Encoding — A Brief Guide. *Applied AI Course*. https://www.appliedaicourse.com/blog/ordinal-encoding/

[29] Number Analytics. (2025). Practical applications of Min–Max scaling in machine learning. https://www.numberanalytics.com/blog/practical-min-max-scaling-machine-learning

[30] Rathi, D. (2021, July 12). Handling imbalanced data — Under-sampling. *Medium*.

[31] Chawla, N. V., Bowyer, K. W., Hall, L.O., & Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

[32] Hwang, J., Park, J., & Shin, K. (2023). Enhancing Traffic Accident Severity Prediction: Feature Identification Using Explainable AI. Data, 7(2), 38.

[33] oPenAI. (2025, March 5). Understanding the role of data normalization in time series forecasting. *GoPenAI Blog*.

[34] Chng, Z. M. (2022, June 20). Using normalization layers to improve deep learning models. *MachineLearningMastery*.

[35] Encord. (2023). Training, Validation, Test Split for Machine Learning Datasets. *Encord Blog*. Retrieved from https://encord.com/blog/train-val-test-split/

[36] Gautamrajotya. (2022, October 20). How to reduce memory usage in Python (Pandas)? *Medium*. Retrieved from https://medium.com/@gautamrajotya/how-to-reduce-memory-usage-in-python-pandas-158427a99001

[37] Ravikumar and Dharshini, "Towards Enhancement of Machine Learning Techniques

Using CSE-CIC-IDS2018 Cybersecurity Dataset," Thesis. Rochester Institute of Technology,2021.

[38] Ibomoiye, D. M., & Jere, N. T. (2023). A Survey of Decision Trees: Concepts, Algorithms, and Applications. *IEEE Access*. https://doi.org/10.1109/ACCESS.2017.DOI

[39] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[40] ProgrammerSought. (2024). LightGBM algorithm overview. Retrieved from https://www.programmersought.com/article/40928374465/

[41] ML Journey. (2025). What is AdaBoost classifier in machine learning? *ML Journey*. Retrieved from https://mljourney.com/what-is-adaboost-classifier-in-machine-learning/

[42] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31. https://arxiv.org/abs/1706.09516

[43] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD*, 785–794. https://arxiv.org/abs/1603.02754

[44] Datacamp. (2023). An introduction to SHAP values and machine learning interpretability. *Datacamp Tutorial*. Retrieved from https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability

[45] Permetrics. (2024). MSLE – Mean Squared Logarithmic Error. Permetrics Documentation. https://permetrics.readthedocs.io/

[46] Ahmad, A. (2024, Sep 27). How to calculate time to train AI Training model? Networking Factors That Impact AI Model Training Time. LinkedIn Pulse.

[47] Saxena, S. (2025, May 1). What is LSTM? Introduction to Long Short-Term Memory. *Analytics Vidhya*.

[48] Patel, D. (2023, May 4). Understanding Gated Recurrent Unit (GRU) in Deep Learning. *Medium*.

ANNEX

Below are summary tables describing the selected features from each dataset used in traffic accident severity prediction. These descriptions help clarify the meaning and context of each variable in **Tables 1** and **2**.

Table 1 Descriptions of Columns in the US Accident.

Feature Name	Description
ID	Unique ID for each accident
Source	Data reporting source.
Severity	Accident seriousness (1–4).
Start_Time	Time the accident started.
End_Time	Time the accident ended.
Start_Lat	Latitude of accident start.
Start_Lng	Longitude of accident start.
End_Lat	Latitude where accident ends.
End_Lng	Longitude where accident ends.
Distance(mi)	Accident length in miles.
Description	Textual accident summary.
Street	Street name of accident.
City	City where accident occurred.
County	County of the accident
State	State where accident happened.
Zipcode	Postal code of the location.
Country	Country of the accident.
Timezone	Local time zone of event.
Airport_Code	Nearest airport code.
Weather_Timestamp	Time of weather data.
Temperature(F)	Temperature in Fahrenheit.
Wind_Chill(F)	Wind-adjusted temperature.

Feature Name	Description
Humidity(%)	Air humidity percentage.
Pressure(in)	Atmospheric pressure.
Visibility(mi)	Visibility in miles.
Wind_Direction	Direction of wind.
Wind_Speed(mph)	Wind speed in mph.
Precipitation(in)	Rain/snow amount.
Weather_Condition	Weather description.
Amenity	Near a public facility.
Bump	Presence of speed bump.
Crossing	At a road crossing.
Give_Way	Near a yield sign.
Junction	At a road junction.
No_Exit	Road with no exit.
Railway	Near railway crossing.
Roundabout	At a roundabout.
Station	Near a station.
Stop	Near a stop sign.
Traffic_Calming	Traffic calming present.
Traffic_Signal	Traffic light present.
Turning_Loop	At a turning loop.
Sunrise_Sunset	Day or night at time.
Civil_Twilight	Light just before sunrise/after sunset.
Nautical_Twilight	Light for navigation.
Astronomical_Twilight	Sky dark for astronomy.

Table 2 Descriptions of Columns in the Road Traffic Accident.

Feature Name	Description
Time	When the accident happened.
Day_of_week	Day of the week of the crash.
Age_band_of_driver	Driver's age group.
Sex_of_driver	Driver’s gender.
Educational_level	Driver’s education level.
Vehicle_driver_relation	Driver’s relation to the vehicle.
Driving_experience	Years of driver’s experience.
Type_of_vehicle	Kind of vehicle involved.
Owner_of_vehicle	Vehicle ownership status.
Service_year_of_vehicle	Vehicle’s age in years.
Defect_of_vehicle	Vehicle mechanical defect.
Area_accident_occured	Location type (urban/rural).
Lanes_or_Medians	Road lane or median type.
Road_allignment	Road shape.
Types_of_Junction	Type of junction.
Road_surface_type	Road material.
Road_surface_conditions	Road condition.
Light_conditions	Lighting at the time of the crash.
Weather_conditions	Weather during the accident.
Type_of_collision	Collision type.
Number_of_vehicles_involved	Count of vehicles.
Number_of_casualties	Count of injured/killed.
Vehicle_movement	Vehicle movement direction.
Casualty_class	Role of injured person.
Sex_of_casualty	Gender of casualty.
Age_band_of_casualty	Age group of the casualty.
Casualty_severity	Injury severity.
Work_of_casuality	The casualty’s occupation.

Feature Name	Description
Fitness_of_casuality	The casualty’s fitness status.
Pedestrian_movement	Pedestrian action.
Cause_of_accident	Main accident cause.
Accident_severity	Overall accident severity.