FDRE TECHNICAL AND VOCATIONAL TRAINING INSTITUTE

SCHOOL OF GRADUATE STUDIES

DIVISION OF ELECTRICAL, ELECTRONICS AND ICT

DEPARTMENT OF INFORMATION AND COMMUNICATION TECHNOLOGY

A MACHINE LEARNING MODEL FOR ROAD TRAFFIC ACCIDENT FACTOR CLASSIFICATION: THE CASE OF KAMBA WOREDA, GAMO ZONE, SOUTHERN ETHIOPIA

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Technology

BY

AMANUEL DULEBO SHURGA

ID No: TTMR029/16

PRINCIPAL ADVISOR: PROFESSOR REVANDIRA BABU

ADDIS ABABA, ETHIOPIA

JUNE 2026

DECLARATION

I, Amanuel Dulebo Shurga, the undersigned, hereby declare that this thesis entitled "A Machine Learning Model for Road Traffic Accident Factor Classification: The Case of Kamba Woreda, Gamo Zone, Southern Ethiopia," is my original work. The research was conducted independently under the guidance and supervision of my research advisor.

I further declare that this thesis has not been submitted earlier in whole or in part, for any degree, diploma, or other academic qualification at this or any other institution. All sources of information and materials used in this study have been properly acknowledged and cited.

Declared by:

Name: Amanuel Dulebo Shurga

Signature:

Department: ICT

Date: June 3 2026

This thesis has been submitted for examination with my approval as the candidate's advisor.

Principal Advisor: _____

Signature: _____

Date: _____

Co-Advisor (if any): _____

Signature: _____

Date: _____

APPROVAL PAGE

This is to certify that the thesis entitled "A Machine Learning Model for Road Traffic Accident Factor Classification: The Case of Kamba Woreda, Gamo Zone, Southern Ethiopia", submitted by Amanuel Dulebo Shurga (ID No. TTMR029/16) in partial fulfillment of the requirements for the Degree of Master of Science in Information Technology, has been examined and approved by the Board of Examiners of the Department of Information and Communication Technology, FDRE Technical and Vocational Training Institute.

Examination Role	Name	Signature	Date
Principal Advisor			
Co-Advisor (if any)			
Internal Examiner			
External Examiner			
Head, Department of ICT			
Dean, School of Graduate Studies			

Place: Addis Ababa, Ethiopia

Date of Final Approval: _____

ACKNOWLEDGEMENTS

First and foremost, I offer my gratitude to the Almighty God for the countless blessings throughout my life and during this research.

I am deeply grateful to my advisor, Professor Revandira Babu, for his invaluable guidance, constructive criticism, and continuous feedback from the inception of this study to its completion. His mentorship played a significant role in shaping both the direction and the quality of this research.

My sincere appreciation goes to the instructors and staff of the Department of Information and Communication Technology for their support, encouragement, and academic contributions throughout my studies.

I would also like to extend my heartfelt thanks to the Gamo Zone Kamba Woreda Traffic Police officers and the accident data record office staff for providing the necessary data and offering dedicated assistance during the research process.

Finally, I express my deepest gratitude to my family and friends for their unwavering support, patience, encouragement, and motivation, which sustained me throughout this journey.

ABSTRACT

Road traffic accidents are events that occur on roads open to public traffic and result in injury, loss of life, vehicle damage, and substantial economic costs. As the number of vehicles increases worldwide, including in Ethiopia, road traffic accidents have risen accordingly, yet it remains difficult to determine the specific conditions that lead to them. Previous studies have largely focused on classifying accident severity levels or predicting whether an accident will occur, often using a limited set of features. This study addresses a different problem: classifying the contributing factors of road traffic accidents into four categories — human, vehicle, road, and environmental factors — using supervised machine learning. The dataset was obtained from the Gamo Zone Kamba Woreda Traffic Police Office and covers seven years of accident records (2011–2017 E.C.). After preprocessing, the dataset comprised 5,250 instances described by 19 attributes, including the target class. An experimental research approach was adopted to identify the best-performing model. The implementation was carried out in the Python programming language using established machine learning libraries. Four classification algorithms were trained and compared: Random Forest, Decision Tree, Multi-Layer Perceptron, and Extreme Gradient Boosting (XGBoost). Models were evaluated using a 70/30 percentage split together with stratified k-fold cross-validation, and performance was assessed using accuracy, precision, recall, F1-score, ROC-AUC, and confusion-matrix analysis. The results show that the XGBoost classifier outperformed the other models, achieving a test-set accuracy of approximately 91%. Based on this result, a prototype system capable of classifying the contributing factors of road traffic accidents was developed and deployed using the Streamlit framework.

Keywords: Road traffic accidents; machine learning; classification; factor classification; data preprocessing; feature selection; Random Forest; Decision Tree; XGBoost; Multi-Layer Perceptron; model evaluation metrics; Gamo Zone Kamba.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

Abbreviation	Full Form
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under the Curve
CNN	Convolutional Neural Network
CSV	Comma-Separated Values
CV	Cross-Validation
DT	Decision Tree
E.C.	Ethiopian Calendar
FTVETI	Federal Technical and Vocational Education and Training Institute
GDP	Gross Domestic Product
IQR	Interquartile Range
IRB	Institutional Review Board
KDD	Knowledge Discovery in Databases
ML	Machine Learning
MLP	Multi-Layer Perceptron
MoU	Memorandum of Understanding
PCA	Principal Component Analysis
RF	Random Forest
RFE	Recursive Feature Elimination
ROC	Receiver Operating Characteristic
RTA	Road Traffic Accident
RTAF	Road Traffic Accident Factor
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
WHO	World Health Organization
XGBoost	Extreme Gradient Boosting

CHAPTER ONE: INTRODUCTION

1.1 Background of the Study

Road traffic accidents are a leading cause of death and injury worldwide, and developing countries bear the greatest share of this burden. A road traffic accident is an event occurring on a road open to public traffic that results in injury, loss of life, or property damage. Such accidents take several forms, including collisions between vehicles, between a vehicle and a pedestrian, between a vehicle and an animal, and between a vehicle and a fixed physical object. Road users include pedestrians, cyclists, motorcyclists, motorists, commuters, and passengers of public transport.

According to the World Health Organization (WHO), road traffic accidents claim approximately 1.35 million lives globally each year, and around 92% of these deaths occur in low- and middle-income countries (WHO, 2021). Vulnerable road users — pedestrians, cyclists, and motorcyclists — account for more than half of all traffic deaths. In Ethiopia, the situation is particularly difficult. Reported figures indicate that road traffic accident deaths reached an estimated 31,564, equivalent to about 5.6% of total deaths in the country in 2020 (World Life Expectancy, 2020). The corresponding age-adjusted death rate places Ethiopia among the countries with the highest road traffic fatality burden relative to its vehicle fleet.

In 2021, the Federal Police Commission recorded 15,034 road accidents during the fiscal year, in which 4,161 people lost their lives (Federal Police Commission, 2021). Studies on traffic safety in Ethiopia estimate that an average of 13 people die every day as a result of traffic accidents, amounting to more than 4,700 deaths annually. Although asphalt roads account for only about 14% of the national road network, a large proportion of recorded accidents occur on these roads. Beyond the loss of life, road traffic injuries impose a heavy economic and emotional burden on victims and their families through medical costs and lost income, and they have been estimated to cost countries up to 3% of their annual gross domestic product.

In southern Ethiopia, the incidence of road traffic accidents has risen at an alarming rate, with neighboring zones such as Wolaita reporting high accident rates (Kurika et al., 2020). Within the Gamo Zone, the Kamba Woreda has similarly experienced a high frequency of traffic accidents

attributable to a range of contributing factors. Most existing research on road traffic accidents focuses on predicting accident severity levels or on predicting whether an accident will occur, and these studies typically use a limited number of features and cover limited geographical areas. As a result, there is a recognized gap in models that classify the underlying contributing factors of accidents. This motivated the present study, which applies appropriate machine learning algorithms to a dataset of real road traffic accidents collected from the Kamba Woreda of the Gamo Zone. The study explores the development of a machine learning model for road traffic accident factor classification, to strengthen the capacity to understand accident causation, reduce accidents, and save lives.

1.2 Statement of the Problem

In the Gamo Zone, and particularly in the Kamba Woreda, road traffic accidents continue to rise despite the use of conventional analytical methods. Traditional statistical approaches have shown limited effectiveness in accurately identifying the complex and interacting factors that contribute to accidents. This limitation has resulted in inadequate insights and has hindered the design and implementation of effective preventive strategies. Road traffic accidents are among the leading causes of property damage, injury, and death in Ethiopia, and, with the rapid increase in the number of vehicles, ensuring road safety has become a significant national concern.

Although many factors contribute to road traffic accidents, some are location-specific and may occur only on particular roads or under particular conditions. This makes the identification and classification of contributing factors both essential and challenging. Despite extensive research on road traffic accidents globally and within Ethiopia, most studies have focused on predicting accident severity levels or the likelihood of accident occurrence, often using a limited number of features and limited geographical coverage. Few studies have addressed the classification of accident factors, and fewer still have done so in the context of southern Ethiopia and the Gamo Zone Kamba Woreda.

The volume and complexity of traffic accident data require advanced analytical methods, because conventional statistical techniques struggle to handle large datasets and to capture interactions among multiple features. Machine learning offers the capability to process large, multidimensional datasets and to uncover hidden patterns in accident causation. By training

appropriate machine learning models on well-prepared datasets, it becomes possible to classify the major contributing factors of road traffic accidents with measurable accuracy. Traditional statistical approaches such as logistic regression often perform poorly on nonlinear relationships and multi-class classification problems; prior studies report logistic-regression accuracies of roughly 65–75% in multi-class traffic accident modeling, whereas ensemble machine learning models have demonstrated accuracies exceeding 85% on comparable transportation datasets. This study therefore investigates whether advanced machine learning algorithms can significantly improve accident factor classification performance over conventional methods, and it develops a classification model that identifies the key contributing factors of road traffic accidents in the Kamba Woreda. The outcomes are intended to support traffic police and other stakeholders in making data-driven decisions and in implementing targeted and effective road safety interventions.

1.3 Research Objectives

1.3.1 General Objective

The general objective of this study is to develop a machine learning model capable of accurately classifying the factors that contribute to road traffic accidents in the Kamba Woreda of the Gamo Zone, southern Ethiopia.

1.3.2 Specific Objectives

To achieve the general objective, the study pursues the following specific objectives:

- ✓ To collect, clean, and preprocess road traffic accident records obtained from the Kamba Woreda Traffic Police Office into a dataset suitable for machine learning.
- ✓ To develop and evaluate supervised machine learning algorithms — namely Decision Tree, Random Forest, Multi-Layer Perceptron, and XGBoost — for road traffic accident factor classification.
- ✓ To optimize the selected models through hyperparameter tuning and feature selection to improve classification performance.
- ✓ To validate and comparatively analyze model performance using standard evaluation metrics, including accuracy, precision, recall, F1-score, and ROC-AUC.

- ✓ To develop a prototype system, based on the best-performing model, that classifies the contributing factors of road traffic accidents and supports data-driven decision-making for local authorities and policymakers.

1.4 Research Questions

The study is guided by the following research questions, which correspond to the specific objectives stated above:

1. What are the primary factors contributing to road traffic accidents in the Kamba Woreda of the Gamo Zone?
2. How do different supervised machine learning algorithms perform in classifying road traffic accident factors?
3. Which machine learning algorithm achieves the highest predictive performance in classifying road traffic accident factors?
4. How effectively can the developed model provide actionable insights to support road safety measures in the study area?

1.5 Scope of the Study

The study is geographically limited to the Kamba Woreda of the Gamo Zone in southern Ethiopia. It focuses on applying supervised machine learning classification algorithms to identify and classify the factors that contribute to road traffic accidents. The classification is performed using a dataset limited to the period from 2011 to 2017 E.C. The study does not cover other regions of Ethiopia or accident records beyond 2017 E.C.; consequently, its results cannot be generalized to the country as a whole.

1.6 Limitations of the Study

Despite its contributions, the study has several limitations. First, it is geographically restricted to the Kamba Woreda and does not include other regions of Ethiopia, so its findings cannot be generalized to the entire country. Second, the dataset is limited to accident records from 2011 to 2017 E.C.; records beyond this period were not included, which means that recent trends and emerging traffic patterns are not captured. Third, the study relies solely on structured historical

data obtained from institutional records, which may contain missing or incomplete entries. Finally, the study does not incorporate real-time data sources or unstructured data such as images or video recordings, which could have provided additional analytical insight.

1.7 Significance of the Study

This study contributes both scientifically and practically to the application of machine learning in road traffic accident analysis within the Ethiopian context. Scientifically, it introduces a factor-classification approach tailored to the local context of the Kamba Woreda, demonstrates the effectiveness of optimized machine learning algorithms relative to traditional statistical methods, and provides a replicable analytical framework that can guide future road traffic accident research in similar developing regions.

Practically, the classification model is relevant to traffic police offices, transport authorities, and road safety stakeholders. It improves understanding of accident patterns and supports the development of data-driven policies and strategies. By enabling targeted interventions based on key contributing factors — such as road infrastructure, driver behavior, vehicle condition, and environmental conditions — the study contributes to improved road safety. The automated classification system also increases operational efficiency by reducing the time and effort required for manual analysis, and it promotes evidence-based decision-making in road safety management.

1.8 Purpose of the Study

The purpose of this study is to develop a robust, data-driven model that classifies road traffic accident factors into human, vehicle, road, and environmental categories using supervised machine learning algorithms. The study analyses accident records collected between 2011 and 2017 E.C. from the Kamba Woreda of the Gamo Zone. It aims to generate reliable insights into the dominant causes of road traffic accidents in order to support evidence-based policymaking and to contribute to the improvement of road safety interventions in the Ethiopian context.

1.9 Motivation of the Study

The motivation for this study arises from the urgent need to address the increasing rate of road traffic accidents in the Gamo Zone, particularly in the Kamba Woreda. Despite existing road

safety measures, road traffic accidents remain a persistent challenge that significantly affects public health and economic stability. The limited application of advanced data-driven analytical techniques in the local context has created a gap in the understanding of accident causation. The adoption of advanced machine learning approaches is therefore considered a promising means of improving accident factor classification, enhancing analytical accuracy, and supporting the development of more effective preventive strategies and interventions.

CHAPTER TWO: REVIEW OF RELATED LITERATURE

2.1 Theoretical Review

This chapter reviews the literature relevant to the objectives of the study in order to clarify the research problem and to provide a technical foundation for the work. It begins with an overview of machine learning and machine learning tasks, followed by a review of the major factors and causes of road traffic accidents. The theoretical foundation of the study is grounded in machine learning and in supervised learning in particular. Supervised learning algorithms are well suited to the problem of classifying road traffic accident factors because they use labeled datasets to train models that predict categorical outcomes. This review justifies the application of machine learning techniques to traffic safety challenges and to data-driven accident analysis.

2.2 Overview of Machine Learning

Machine learning has been widely adopted across domains such as education, business, healthcare, agriculture, and transportation for the purpose of extracting hidden patterns and knowledge that support data-driven decision-making. Machine learning algorithms are capable of processing large datasets with many classification parameters and of producing useful patterns from them. Machine learning tasks are commonly grouped into supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning (Nasteski, 2017; Sharma, Sharma, & Sharma, 2021).

2.2.1 Supervised Learning

Supervised learning involves training models on labeled datasets so that input attributes are mapped to a desired output. Classification is a primary task in supervised learning and aims to assign data instances to predefined categories. Common classification algorithms include logistic regression, decision trees, random forests, support vector machines, k-nearest neighbors, and naïve Bayes (El Morr, Jammal, Ali-Hassan, & El-Hallak, 2022). Each algorithm has its own strengths and weaknesses; for example, logistic regression is simple and interpretable but limited to linear decision boundaries, whereas random forests provide robust performance by aggregating multiple decision trees but are less interpretable (Sharma et al., 2021).

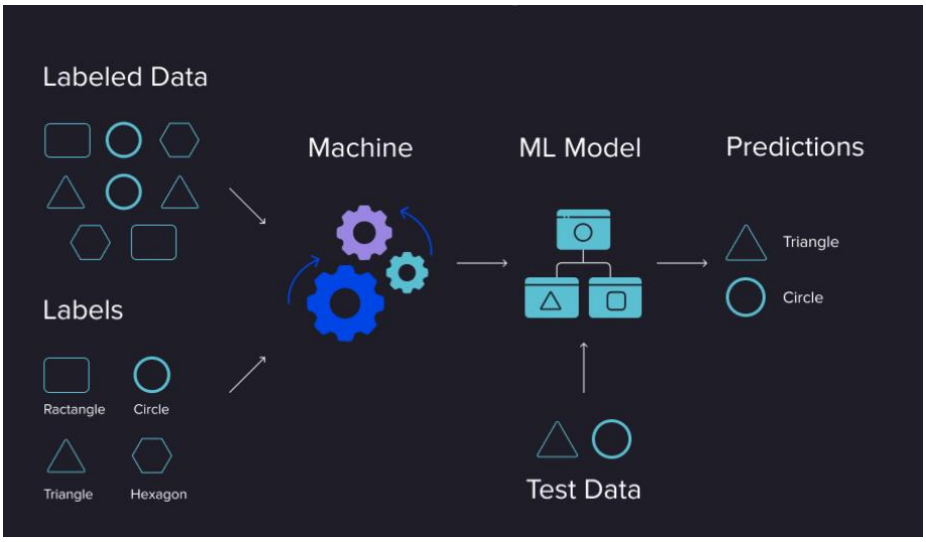


Figure 2.1:An overview of how supervised learning works

2.2.2 Unsupervised Learning

Unsupervised learning operates on unlabeled data and identifies hidden structure through techniques such as clustering. Methods such as k-means clustering and Principal Component Analysis are widely used to explore accident datasets by grouping similar cases or by reducing feature dimensionality for visualization (Aslanyan, 2023).

2.2.3 Semi-Supervised Learning

Semi-supervised learning combines a small amount of labeled data with a larger amount of unlabeled data to construct an appropriate classifier. It is useful when labeling data is expensive or time-consuming.

2.2.4 Reinforcement Learning

Reinforcement learning focuses on sequential decision-making, in which an agent learns optimal strategies through trial-and-error interactions with an environment. Although it is less commonly used for accident classification, reinforcement learning has applications in traffic signal optimization and autonomous driving systems (Puppala, 2025).

2.3 Machine Learning Tasks

All machine learning algorithms take data as input, but their objectives differ according to the problem being addressed. Depending on the analytical objective, machine learning tasks can be broadly classified into predictive and descriptive categories. Predictive tasks aim to forecast or

classify outcomes from data and include classification, regression, time-series analysis, and prediction. Descriptive tasks focus on discovering patterns within data and include clustering, summarization, sequence discovery, and association-rule mining. Predictive approaches are typically used when labeled data are available, whereas descriptive approaches are applied to explore hidden structure in unlabeled data. The classification of machine learning tasks is illustrated in Figure 2.2.

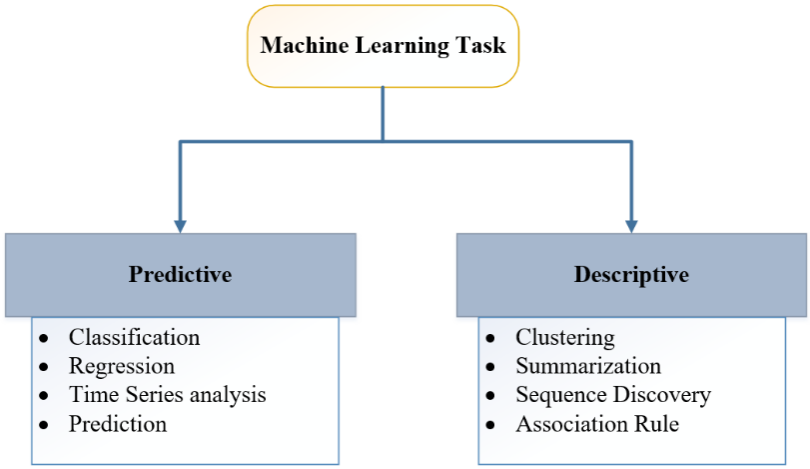


Figure 2.2:Classification of machine learning tasks

2.4 Overview of Road Traffic Accidents in Ethiopia

Road traffic accidents in Ethiopia constitute a major problem with multiple causes, compounded by gaps in road safety management and policy. According to WHO reports and records of the Ethiopian Federal Police Commission, large numbers of road users are killed or injured each year, with pedestrians accounting for an estimated 45–55% of casualties, passengers for 25–35%, and drivers for up to 25%, alongside substantial economic loss and material damage (WHO, 2023). WHO data indicate that the road traffic death rate in Ethiopia is among the highest in the region, with associated economic damage estimated in billions of birr annually. Within the country, regions such as Amhara and Oromia consistently record the highest fatality counts; the southern region reports comparatively lower absolute fatality figures but a higher proportion of pedestrian and passenger casualties, which is associated with limited road infrastructure.

2.5 Conceptual Review

Various research articles, related works, and journal papers on road traffic accidents and machine learning were reviewed to develop a conceptual understanding of the problem, to identify research gaps, and to formulate the problem statement and research questions (Asnake et al., 2019; National Highway Traffic Safety Administration [NHTSA], 2015). Conceptually, the study treats road traffic accidents as a multifactorial phenomenon and classifies their contributing factors into four major categories using data-driven methods.

Road traffic accidents are unintended events occurring on public roads that result in injury, death, or property damage. They are widely recognized as a multifactorial phenomenon, meaning that accidents rarely arise from a single cause but rather from the interaction of several contributing factors. In this study, road traffic accidents are conceptualized as outcomes influenced by four broad categories of factors: human, vehicle, road, and environmental.

2.5.1 Road Traffic Accident Factors

Understanding the factors that contribute to road traffic accidents is essential for developing preventive strategies and for building data-driven models for accident classification. Road traffic accidents are generally caused by the interaction of human behavior, vehicle condition, road infrastructure, and environmental conditions.

2.5.1.1 Human Factors

Human behavior is recognized as the dominant contributor to road traffic accidents, and drivers, pedestrians, and passengers all play significant roles in road safety. Driver-related factors include over-speeding, drink-driving, failure to use safety equipment such as seat belts and helmets, fatigue, disregard of traffic signals, failure to understand road signs, and violation of traffic rules. Over-speeding reduces a driver's ability to react to unexpected situations and increases stopping distance; drink-driving impairs cognitive and motor skills, leading to poor decision-making and delayed responses; and fatigue and disregard of signals and signs contribute to human error. Pedestrians are vulnerable road users who may contribute to accidents through unsafe behavior such as crossing roads at inappropriate locations, jaywalking, and ignoring pedestrian crossings, often as a result of limited awareness of traffic rules. Passenger behavior can also indirectly

influence road safety; actions such as distracting the driver, projecting body parts outside the vehicle, or boarding and alighting from unsafe sides increase accident risk.

2.5.1.2 Vehicle Factors

Vehicle condition significantly affects road safety and operational reliability. Mechanical failures such as brake malfunction and tyre bursts are frequent contributors to accidents. Poorly maintained vehicles with inadequate headlights reduce visibility during night driving and adverse weather conditions, and overloading vehicles beyond their rated capacity affects stability and braking efficiency. Regular vehicle maintenance and adherence to safety standards are therefore essential for accident prevention.

2.5.1.3 Road Infrastructure Factors

Road design and condition are critical determinants of traffic safety. Poorly constructed roads with potholes, uneven surfaces, and inadequate signage increase accident risk, and wet or icy surfaces reduce tyre grip, leading to skidding and loss of control. Inappropriate road curvature, inadequate speed control measures, and construction zones without proper warnings further increase hazards. Improving road infrastructure and applying safety-oriented design principles can significantly reduce accidents.

2.5.1.4 Environmental and Weather Factors

Environmental conditions play an important role in road safety. Weather phenomena such as fog, heavy rainfall, and strong winds reduce visibility and vehicle stability, while wet and slippery roads increase braking distance and the likelihood of skidding. Dust, mud, and extreme temperatures may also affect road surface conditions and vehicle performance. Drivers must adapt their behavior to prevailing environmental conditions to minimize risk.

2.6 Types of Road Traffic Crashes

Road traffic crashes frequently result in injury, disability, death, and property damage, as well as significant economic loss to individuals and society. Classifying crash types is essential for understanding accident patterns and for developing effective prevention strategies. Crashes can be categorized according to the entities involved in the collision.

2.6.1 Vehicle-to-Pedestrian Crashes

These crashes occur when a motor vehicle collides with a pedestrian. They are among the most severe crash types because pedestrians are unprotected road users, and they often result in serious injury or death, particularly in urban areas with high pedestrian traffic. Contributing factors include driver negligence, pedestrian rule violations, and poor road design, such as the absence of pedestrian crossings.

2.6.2 Vehicle-to-Structure Crashes

These crashes occur when a vehicle collides with roadside structures such as guardrails, traffic poles, trees, barriers, bridges, or buildings. They commonly result from loss of vehicle control, over-speeding, mechanical failure, or slippery road conditions. The severity of such crashes depends on the speed of the vehicle and the type of structure involved.

2.6.3 Vehicle-to-Vehicle Crashes

These crashes occur when one motor vehicle collides with another and are among the most common types of road traffic accident. They include head-on collisions, rear-end collisions, side-impact crashes, and multi-vehicle pile-ups. Contributing factors include over-speeding, improper overtaking, failure to maintain a safe following distance, distracted driving, and traffic rule violations.

2.6.4 Vehicle-to-Other-Object Crashes

These crashes involve collisions with animals, road debris, fallen objects, or other stationary obstructions not classified as road structures. They are more common in rural areas where animal crossings are frequent,and road monitoring is limited. Although they may cause less severe human injury than other crash types, they can still result in significant property damage and traffic disruption.

2.7 Generalized Factors of Road Traffic Accidents

Building on the discussion above, this study generalizes the causes of road traffic accidents into four main categories: human-related factors, vehicle-related factors, road-related factors, and environmental factors. These four categories form the target classes for the classification model developed in the study.

2.8 Machine Learning in Accident Analysis

In this study, machine learning is conceptualized as a computational approach that enables a system to learn patterns from historical accident data and to classify new cases into factor categories. The variables used in the analysis are organized into four groups corresponding to the four accident-factor categories.

2.8.1 Human-Related Variables

Human-related variables capture characteristics of drivers and behavioral factors that may influence the occurrence and severity of road traffic accidents. These include driver age, which influences reaction time and risk perception; driver experience, which affects decision-making ability; the behavioral cause of the accident, such as speeding, distraction, alcohol use, fatigue, or rule violation; driver sex, which may capture behavioral differences in driving patterns; and driver educational level, which may influence awareness of traffic regulations and safety practices. Priority-rule variables, such as whether the driver gave priority to pedestrians or to other vehicles, are also included.

2.8.2 Vehicle-Related Variables

Vehicle-related variables describe the mechanical and operational characteristics of the vehicle involved in the accident. These include vehicle type (for example, car, truck, bus, motorcycle, or minibus), vehicle condition, and vehicle service age, expressed as the number of years the vehicle has been in service. Different vehicle types are associated with different risk and crash-severity patterns, and older vehicles may carry a higher risk of mechanical failure.

2.8.3 Road-Related Variables

Road-related variables describe the physical characteristics of the road at the time of the accident. They include road geometry (straight, curve, intersection, or slope), road type (asphalt, gravel, cobblestone, or dirt), accident area (urban, rural, or highway), road surface state (dry, wet, slippery, or icy), and overall road condition (good, damaged, under construction, or affected by potholes).

2.8.4 Environment-Related Variables

Environmental variables describe external conditions at the time of the accident that may influence visibility, vehicle control, and driver behavior. They include weather conditions, time of day, and lighting conditions. Adverse weather such as rain and fog reduces visibility and road friction; accident patterns vary by time of day owing to differences in traffic volume and driver alertness, and poor lighting increases the likelihood of pedestrian and vehicle collisions.

2.9 Conceptual Framework of the Study

The conceptual framework of the study integrates accident-causation theory with machine learning techniques. It is based on the following assumptions: that road traffic accidents result from the interaction of human, vehicle, road, and environmental factors; that historical accident data can be transformed into structured features; and that machine learning models can classify accident factors with measurable accuracy. This framework guides the research design, data preprocessing, model development, and evaluation. The processing layer of the framework comprises data preprocessing — including data cleaning, encoding of categorical variables, and normalization of numerical features — followed by the application of four supervised classification algorithms: Decision Tree, Random Forest, XGBoost, and Multi-Layer Perceptron. These algorithms were selected for their effectiveness in handling structured data and classification tasks.

Conceptually, the study shifts the focus of analysis from severity prediction to accident-factor classification, thereby providing deeper insight into the root causes of accidents. Its conceptual contribution lies in supporting policymakers in designing targeted interventions, demonstrating the application of machine learning to traffic safety research, and providing an evidence-based basis for accident-reduction strategies.

2.10 Empirical Review

2.10.1 Summary of Related Work

Previous studies on road traffic accidents have predominantly focused on predicting accident severity rather than on classifying causative factors. Although a variety of machine learning models — including Random Forest, support vector machines, XGBoost, and Multi-Layer

Perceptron — have been explored, analytical comparisons and thematic interpretations of their strengths and weaknesses remain limited. This section synthesises selected studies in order to establish a clear research gap and to justify the methodology adopted in the present study. A review of prior work reveals several recurring patterns: an emphasis on severity prediction rather than factor classification; limited geographical representation, with much research originating from data-rich, high-income contexts; an algorithm-centric focus on accuracy that overlooks interpretability and implementation feasibility; and reliance on small datasets and limited feature sets, which can lead to overfitting and weak generalization.

2.10.2 Review of Empirical Studies

Several studies have applied machine learning techniques to the analysis of road traffic accidents, primarily focusing on accident-severity prediction. Bharti et al. (2016) used support vector machines and a Multi-Layer Perceptron to predict accident severity, achieving 94% accuracy using only two features — alcohol level and vehicle speed — although the very narrow feature set limits the generality of the result. Rajkumar et al. (2020) tested logistic regression, naïve Bayes, and XGBoost on a small dataset to classify accident severity; logistic regression performed best with 85% accuracy, but the model lacked regional specificity. Yuan et al. (2018) proposed a ConvLSTM model that predicts accidents using spatio-temporal features but excludes human and environmental factors. Ramasamy et al. (2020) applied ID3 and Random Forest in WEKA and RapidMiner to predict severity levels, with a scope limited to severity rather than root causes, and Wenqi and Dongyu (2017) used convolutional neural networks for accident prediction, achieving 78.5% accuracy while focusing only on occurrence prediction.

In the Ethiopian context, Assefa et al. (2020) applied J48, naïve Bayes, and Random Forest to 1,611 accident records from the Wolaita Zone; although Random Forest achieved up to 95% accuracy, the study focused only on driver-related features and did not classify accidents by cause. Bokaba et al. (2022) compared multiple classifiers — naïve Bayes, logistic regression, k-nearest neighbours, AdaBoost, support vector machines, and Random Forest — on South African traffic data, integrating advanced preprocessing techniques such as Principal Component Analysis, Linear Discriminant Analysis, and Multiple Imputation by Chained Equations, and concluded that Random Forest combined with imputation was the most accurate. Kurika and Sundado (2020), working in the Wolaita Zone, applied decision-tree classifiers (J48, Random

Forest, and RepTree) and Bayesian classifiers (naïve Bayes and Bayesian network) under the Knowledge Discovery in Databases process model; J48 and RepTree achieved comparable F-measures of approximately 97.9% and 97.8%,respectively, while Random Forest achieved approximately 90.9%. The study identified the J48 tree as the best-performing model but was restricted to a small, geographically limited dataset.

Table 2.1 summarizes a selection of related studies, highlighting their methods, contributions, and limitations.

Table 2.1:Comparison of related studies on machine learning for road traffic accident analysis

No.	Author(s) and Year	Title and Country	Methods and Contribution	Critical Remarks
1	Ramasamy et al. (2020)	Predicting severity level of road traffic accidents in the Oromia East Shewa Zone, Ethiopia	Used ID3, naïve Bayes, Random Forest, and decision-tree classifiers in WEKA and RapidMiner; ID3 was identified as the best model for severity prediction.	Limited to severity prediction; the dataset comprised only 2,141 records.
2	Ramya, Reshma et al. (2019)	Accident severity prediction using data-mining methods, India	Used Random Forest, decision tree, naïve Bayes, and SVM; Random Forest achieved the highest accuracy (88%).	The study does not report dataset size or train/test split, and excludes real-time, GPS, and sensor data.
3	Gatarić et al. (2023)	Predicting road traffic accidents using artificial neural network models (Bosnia and Herzegovina, Serbia, Switzerland)	Developed an ANN with multiple hidden layers and compared it with regression-based models; the ANN achieved higher accuracy and robustness.	ANN models are difficult to interpret; they require large, high-quality datasets that are often unavailable in developing countries.
4	Rajkumar et al. (2020)	Prediction of road accident severity using machine learning algorithms	Compared logistic regression, naïve Bayes, and XGBoost; logistic regression performed best (85% accuracy).	Based on only two years of data (about 5,000 records) and lacks regional specificity.
5	Kurika and Sundado (2020)	Predicting factors of vehicular accidents using machine learning algorithms (Wolaita Zone, Ethiopia)	Used J48, Random Forest, RepTree, naïve Bayes, and Bayesian networks under the KDD model; J48 achieved the best F-measure (about 97.9%).	Used a small dataset (1,611 records) restricted to the Wolaita Zone; only a limited set of models was evaluated.
6	Prasad and Reddy (2023)	Prediction of road accident severity using	Used decision tree, Random Forest, SVM, and naïve	Dataset is regionally limited and excludes real-time

No.	Author(s) and Year	Title and Country	Methods and Contribution	Critical Remarks
		machine learning algorithms, India	Bayes; Random Forest and XGBoost performed best.	traffic, GPS, and CCTV data; feature importance is not analysed in depth.
7	Kalaivani et al. (2025)	Road accident severity prediction using machine learning, India	Applied logistic regression, Random Forest, SVM, k-NN, XGBoost, and MLP to a large dataset; XGBoost and Random Forest performed best.	Dataset size and class distribution are not clearly reported; no class-imbalance handling is applied, and explainability is not addressed.

Taken together, these studies confirm the value of ensemble methods and robust preprocessing, but they also reveal that the literature is dominated by severity prediction, that Ethiopian and low-resource contexts are underrepresented, and that few studies translate model outputs into actionable insight for traffic authorities.

2.11 Technical Critique of Existing Studies

This section critically examines empirical studies that have applied machine learning and statistical methods to road traffic accident analysis. The majority of these works emphasise severity prediction — fatal, serious, or slight — rather than the classification of contributing factors. Both global and Ethiopian studies are considered.

2.11.1 Global Studies

Several international studies have applied machine learning and statistical techniques to road traffic accident analysis. Research using the Philippine National Police dataset applied naïve Bayes, decision trees, and rule-induction techniques and identified time of day and day of the week as critical determinants of accident severity (Das et al., 2009); however, the study focused on severity outcomes rather than structured factor classification. Yu and Abdel-Aty (2014) evaluated J48, a Multi-Layer Perceptron, and Bayesian-network classifiers on a relatively small dataset of approximately 150 instances, with the Multi-Layer Perceptron achieving about 85% accuracy, although the limited dataset size raises concerns about generalisability. Behnood and Mannering (2015), using Omani accident data, applied boosted-tree regression and multiplicative statistical models and found that human factors accounted for approximately 91% of accidents,

while non-human factors contributed only 9%. Analysis of the United States Fatality Analysis Reporting System dataset employed association-rule mining, naïve Bayes, and k-means clustering to identify patterns linked to fatal accidents (De Oña et al., 2013), and recent Asian studies have used deep-learning approaches such as ConvLSTM and Fast R-CNN to detect dangerous crossings and accident-prone sites (Yuan et al., 2018), although these approaches often exclude detailed behavioral and environmental variables.

2.11.2 African and Ethiopian Studies

In the African and Ethiopian context, machine learning applications in traffic safety research are growing but remain limited in scope. Tibebe et al. (2011) applied classification and regression trees, Random Forest, and related ensemble methods to Ethiopian accident data and identified driver behaviour, speeding, and alcohol consumption as dominant contributors, demonstrating the potential of data-driven approaches in the Ethiopian setting. Abellán et al. (2013) combined genetic algorithms with fuzzy classifiers to classify injury types and showed that feature selection significantly improved predictive performance, although the focus remained on injury severity rather than comprehensive factor categorization. Kurika et al. (2020), in the Wolaita Zone, implemented J48, Random Forest, RepTree, naïve Bayes, and Bayesian networks under the Knowledge Discovery in Databases framework; J48 and RepTree achieved the highest F-measures, and the generation of decision rules enhanced interpretability, although the dataset was geographically restricted and relatively small.

2.11.3 Predominant Use of Black-Box Models

Many prior studies employ complex machine learning algorithms such as neural networks, support vector machines, and ensemble methods. Although these models frequently achieve high predictive accuracy, they are inherently complex and lack transparency, and are therefore commonly referred to as black-box models because their internal decision processes are difficult to interpret. In safety-critical domains such as traffic safety analysis, limited explainability reduces stakeholder trust and constrains the practical adoption of predictive systems by policymakers and traffic authorities.

2.11.4 Limited Use of Interpretable Models and Explainability Techniques

Interpretable models such as decision trees and logistic regression provide transparent, rule-based explanations that are easier to understand and communicate, but they are relatively underused in road traffic accident research or are applied only as baseline comparisons. Explainability techniques such as feature-importance analysis, SHAP, and LIME can improve model interpretability, yet they remain underutilized in existing research. The present study addresses this gap by selecting models with differing levels of interpretability: the Decision Tree provides clear rule-based explanations, Random Forest and XGBoost offer feature-importance measures that identify dominant accident factors, and the Multi-Layer Perceptron is included for performance comparison despite being less interpretable.

2.11.5 Accuracy–Interpretability Trade-offs and Practical Deployment

There is often a trade-off between model accuracy and interpretability. Highly accurate models such as ensemble methods and neural networks tend to be complex and difficult to interpret, whereas simpler models such as decision trees provide clearer explanations but may achieve lower predictive performance. Because traffic authorities and policymakers require clear and understandable insights in order to design effective prevention strategies, models that combine reliable prediction with interpretable results are more suitable for practical deployment. The literature therefore suggests a need to integrate explainability techniques into machine learning models for road traffic accident analysis in order to improve transparency and to strengthen stakeholder trust.

2.12 Research Gaps

Despite the growing body of literature on road traffic accidents and machine learning, several important gaps remain. First, most existing studies emphasize the prediction of accident severity rather than the classification of underlying causative factors, which limits the ability to design targeted interventions. Second, Ethiopian contexts are underrepresented, and localized accident dynamics — particularly in specific zones such as the Gamo Zone — remain insufficiently explored. Third, previous studies often rely on a narrow set of attributes and overlook variables such as driver experience, road geometry, vehicle condition, and environmental conditions, which reduces predictive power. Fourth, several studies rely mainly on accuracy and do not

incorporate complementary metrics such as precision, recall, F1-score, ROC-AUC, and confusion matrices, which are essential for imbalanced datasets. Finally, many studies do not translate model findings into actionable insights for traffic authorities and policymakers. This study addresses these gaps by developing a multi-class accident-factor classification model, using a locally collected dataset, integrating a comprehensive set of features, applying a full range of evaluation metrics, and producing a deployable prototype.

2.13 Chapter Summary

This chapter reviewed the theoretical, conceptual, and empirical literature relevant to the study. It established machine learning and supervised learning in particular, as the theoretical foundation of the work; conceptualized road traffic accidents as a multifactorial phenomenon with human, vehicle, road, and environmental factors; and critically examined prior empirical studies. The review identified clear research gaps — most notably the dominance of severity prediction over factor classification and the underrepresentation of localized Ethiopian contexts — which the present study is designed to address.

CHAPTER THREE: RESEARCH METHODOLOGY

This chapter presents the research design and methodological procedures adopted to achieve the objectives of the study. The main objective is to develop a machine learning model for classifying the factors that contribute to road traffic accidents. The chapter describes the study area, the research design and approach, the data sources and collection methods, the data preprocessing techniques, the machine learning algorithms employed, and the methods used to evaluate model performance.

3.1 Description of the Study Area

The study was conducted in the Gamo Zone, specifically in the Kamba Woreda, in southern Ethiopia. The administrative center of the Gamo Zone is Arba Minch, which lies to the south of Addis Ababa and is geographically characterized by its proximity to Lake Abaya and Lake Chamo and its location on the edge of Nechisar National Park. The name Arba Minch, which translates as "Forty Springs," reflects the abundance of natural springs in the area. The Kamba Woreda, the specific focus of this study, is located approximately 100 kilometers from Arba Minch.

According to available demographic data, the 2007 Ethiopian national census reported approximately 1,104,360 individuals identifying as Gamo, representing about 1.56% of the national population at that time. Despite the growing population and socio-economic activity in the study area, road infrastructure remains limited; in particular, there is a shortage of pedestrian walkways, traffic signals, and adequate road safety signage. These infrastructural deficiencies contribute significantly to the occurrence of road traffic accidents in the area.

3.2 Research Design and Approach

The study adopts a mixed research approach that integrates quantitative and qualitative methods. The quantitative component involves experimental research, in which machine learning techniques and statistical analysis are used to classify road traffic accident factors. The qualitative component involves an extensive literature review used to explore existing research, identify knowledge gaps, and support the formulation of the research questions (Hadi, 2019). Four supervised machine learning algorithms — Random Forest, Decision Tree, XGBoost, and

Multi-Layer Perceptron — are trained and compared in order to identify the model with the best predictive performance for accident-factor classification.

The dataset was collected from traffic police records in the Kamba Woreda of the Gamo Zone. Data preprocessing procedures include handling missing values, detecting and treating outliers, encoding categorical variables, performing feature selection, and applying normalization where necessary. The processed dataset is stored in CSV format for modeling. All models are trained and evaluated under consistent experimental conditions, using a stratified train–test split and appropriate hyperparameter tuning to ensure a fair and reliable comparison.

The implementation is carried out in the Python programming language using the Jupyter Notebook and Spyder environments. The principal libraries used are pandas for data manipulation, scikit-learn for model development, the XGBoost library for gradient-boosting implementation, and Matplotlib and Seaborn for visualisation. The final optimised model is deployed using Streamlit as a web-based decision-support interface. Model performance is evaluated using standard classification metrics — accuracy, precision, recall, F1-score, ROC-AUC, and confusion-matrix analysis — and the study concludes with a comprehensive evaluation of the findings and practical recommendations for road safety improvement. The research design adopted for the study is illustrated in Figure 3.1.

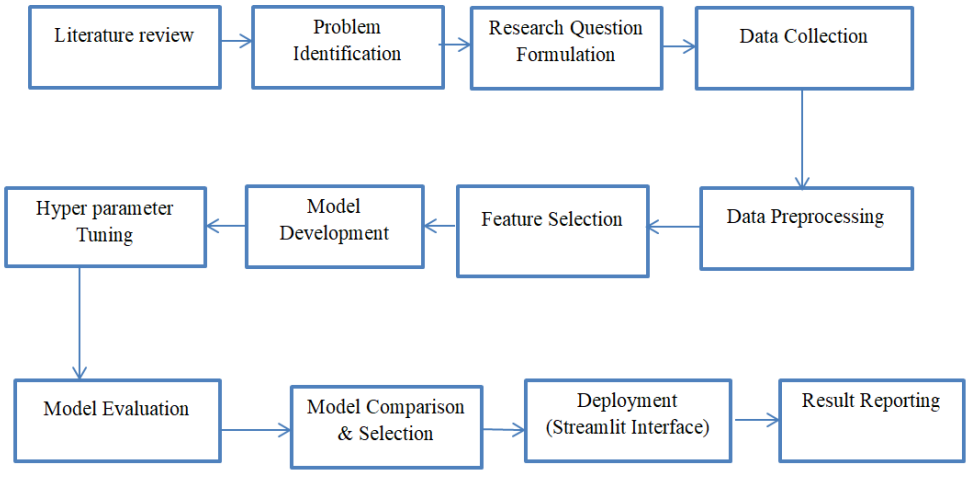


Figure 3.1:Research design adopted for this study

3.3 Data Sources and Collection

The dataset for this study was collected from the Gamo Zone Kamba Woreda Traffic Police Department. A combination of data collection techniques was used to obtain both structured accident records and qualitative contextual information. Qualitative data were gathered primarily through structured interviews with traffic police officers, accident-record personnel, and experienced drivers. This qualitative information played an important role in refining the feature-engineering process and in ensuring that the variables used in the model reflect real-world accident causes that may not be fully captured in structured records. For example, interview responses helped clarify how factors such as driver fatigue, poor signage, road visibility, and time-of-day risk are represented in accident records, and they guided the grouping of accident causes into broader human-related and environmental categories.

The primary data source consisted of interviews conducted with traffic police officers, the head of the traffic police office, and the data recorder, using questions designed to develop a deep understanding of the study domain. The secondary data source consisted of accident records and related documents previously prepared by the traffic office, together with books, reports, and other written materials. The methods of data collection are summarised in Figure 3.2.

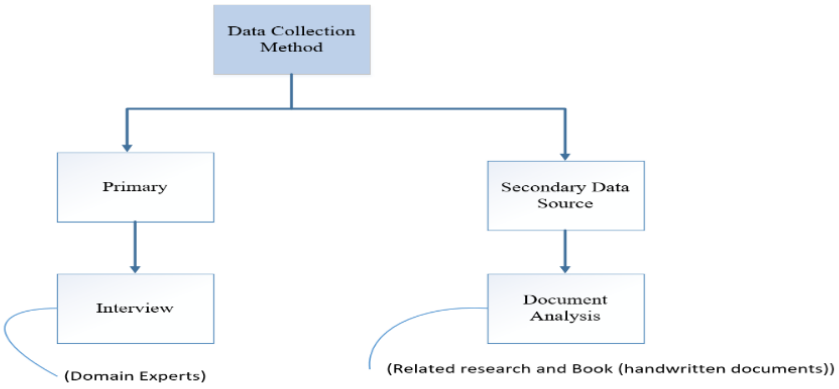


Figure 3.2:Methods of data collection

Accident records were originally registered manually in handwritten and word-processed documents. Considerable effort was therefore required to convert these unstructured records into a structured electronic dataset. Table 3.1 summarizes the number of attributes and records collected from the traffic police office for each year of the study period.

Table 3.1:No of attributes and records collected from the Kamba Woreda Traffic Police Office

Year (E.C.)	Number of Attributes Recorded	Number of Records
2011	21	775
2012	21	804
2013	21	775
2014	21	875
2015	21	789
2016	21	677
2017	21	555
Total	—	5,250

As recorded by the traffic police office, each accident was initially described using 21 attributes (fields)**.** After data preprocessing and feature selection (Section 3.8), 2 irrelevant or redundant attributes were removed**,** resulting in a final dataset of **19 attributes** used for model development. These consist of **18 predictor variables and 1 target variable (Accident Factor)**.

3.4 Dataset and Sampling

The dataset comprises 5,250 road traffic accident records collected from the Kamba Woreda Traffic Police Office over the seven-year period from 2011 to 2017 E.C. The dataset is considered representative of the traffic and accident patterns of the Kamba Woreda, because it covers a sustained multi-year period and includes accidents of varied type, location, and severity. To support reliable model development and unbiased evaluation, the data were divided into training and testing subsets using a 70/30 ratio with stratified sampling, which preserves the proportional distribution of the four accident-factor classes across both subsets.

3.5 Data Description

The dataset consists of road traffic accident records obtained from the Kamba Woreda Traffic Police Office covering the period from 2011 to 2017 E.C. Each record describes an accident event using attributes such as the date and time of occurrence, the accident area, weather and road conditions, the type of vehicle involved, driver demographic information, and the contributing accident factor. Together, these variables provide a comprehensive representation of

accident circumstances and serve as input features for developing and evaluating the classification models. The attributes of the dataset are described in Table 3.2.

Table 3.2:Description of dataset attributes

No.	Attribute	Data Type	Description
1	Accident area	Nominal	Whether the accident occurred in an urban or a rural area
2	Lighting condition	Nominal	Day; night
3	Driver sex	Nominal	Sex of the driver involved in the accident
4	Driver age	Numeric	Age of the driver involved in the accident
5	Driver educational level	Nominal	Educational level of the driver
6	Driver experience	Numeric	Driving experience of the driver, in years
7	Vehicle service	Numeric	Number of years the vehicle has been in service
8	Vehicle type	Nominal	Type of vehicle involved in the accident
9	Road type	Nominal	Type of road infrastructure (asphalt, cobblestone, gravel, etc.)
10	Road geometry	Nominal	Geometric structure of the road (straight, curve, slope, etc.)
11	Weather condition	Nominal	Weather condition at the time of the accident
12	Accident type	Nominal	Type of accident (vehicle-to-pedestrian, vehicle-to-vehicle, etc.)
13	Casualty sex	Nominal	Sex of the casualty
14	Casualty age	Numeric	Age of the casualty
15	Cause of accident	Nominal	Reported cause of the accident
16	Time	Nominal	Time of the accident (morning, afternoon, evening, or night)
17	Road surface	Nominal	State of the road surface at the time of the accident
18	Road condition	Nominal	Overall condition of the road (damaged or not damaged)
19	Accident factor (target)	Nominal	Contributing factor category: human, vehicle, road, or environmental

3.6 The Proposed Road Traffic Accident Factor Classification Model

The proposed model classifies the contributing factors of road traffic accidents in the Kamba Woreda into four categories: human, vehicle, road, and environmental factors. The overall structure of the proposed solution is illustrated in Figure 3.3. After accident data are collected from official traffic police records, the dataset undergoes preprocessing, including missing-value

handling, outlier detection, encoding of categorical variables, feature selection, and treatment of class imbalance using the Synthetic Minority Oversampling Technique (SMOTE). These steps ensure data quality, consistency, and suitability for machine learning modelling.

The four supervised machine learning algorithms — Random Forest, Decision Tree, XGBoost, and Multi-Layer Perceptron — are then trained on the training portion of the dataset and evaluated on the testing portion using accuracy, precision, recall, F1-score, and ROC-AUC. The comparative evaluation identifies the best-performing classifier, which constitutes the final predictive solution for road traffic accident factor classification. Detailed experimental results and performance comparisons are presented in Chapter Four.

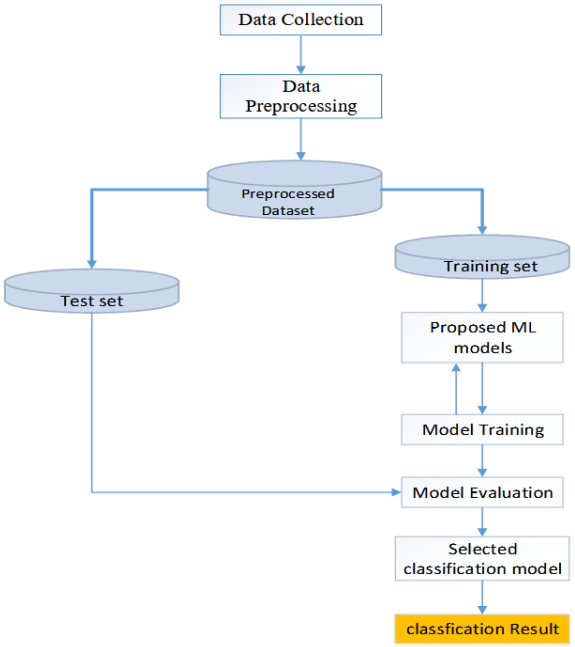


Figure 3.3:The proposed road traffic accident factor classification model

3.7 Data Preprocessing

Data preprocessing prepares the collected data for machine learning modeling. Because the accident records were originally registered in handwritten and word-processed documents, the researcher first transferred the data into Microsoft Excel and, after careful organization and verification, exported it to CSV format, which is suitable for the Python machine learning libraries. The final dataset consisted of 5,250 accident instances described by 19 attributes,

including the target variable. Several preprocessing steps were then carried out before model development, as summarized in Table 3.3.

Table 3.3:Data preparation and validation strategy used before model development

Step	Method Applied
Train–test split	70:30 ratio with stratified sampling (70% training, 30% testing)
Missing-value handling	Numerical variables imputed with the mean; categorical variables imputed with the mode
Outlier detection	Interquartile Range (IQR) method, supported by Z-score thresholds
Class-imbalance handling	Synthetic Minority Oversampling Technique (SMOTE), applied to the training set only
Data validation	Verification of encoded labels and consistency checks

3.7.1 Data Cleaning and Missing-Value Handling

Data cleaning is the first step of preprocessing and is used to identify missing values, smooth noisy data, detect outliers, and correct inconsistencies; uncorrected errors of this kind can degrade model performance and lead to unreliable output. In this study, data cleaning removed duplicate records, corrected typographical errors, and standardized inconsistent categorical labels — for example, by merging variants such as “High Way” and “Highway” into a single category. Outliers in numerical variables, such as driver age, were detected using the Interquartile Range method; obvious data-entry errors were corrected or removed, while legitimate extreme values were retained to preserve real-world variability. The proportion of missing values for each affected attribute is reported in Table 3.4.

Table 3.4:Missing-value statistics for the selected attributes

Attribute	Missing Values (%)
Casualty age	0.59%
Driver education	0.17%
Accident type	0.76%
Driver experience	0.44%
Vehicle type	2.27%
Vehicle service	0.82%
Time	0.27%
Driver age	0.29%

Common strategies for handling missing values include dropping the affected attribute, filling values manually, and statistical imputation. Dropping an attribute is appropriate only when the proportion of missing values is very high, and the attribute is not essential, while manual filling is impractical for large datasets. Because the proportion of missing values was low for all affected attributes, statistical imputation was adopted: mean imputation was used for numerical attributes, and mode imputation — replacing missing values with the most frequent category — was used for categorical attributes, which preserves the distribution of each variable.

3.7.2 Handling Categorical Variables

Most machine learning algorithms operate on numerical inputs, whereas the dataset used in this study contains many categorical variables. These categorical variables must therefore be converted into a numerical representation before model development. Several encoding techniques exist, including label encoding and one-hot encoding. In this study, label encoding was applied because it is efficient and suitable for the structured, predominantly categorical dataset; an illustrative example of the encoding process is shown in Figure 3.4.

	Accident_area	Sex_of_Casualty	Age_of_Casualty	Driver_education	Driver_Sex	Cause_of_Accident	Accide_type	Road_type	Road_Geometry	Road_Conditi
0	1	0	45	3	1	5	3	0	3	
1	1	0	23	3	1	3	3	0	3	
2	0	0	36	3	1	5	3	0	3	
3	1	0	28	3	1	2	1	0	3	
4	0	1	24	3	1	2	1	0	3	

Figure 3.4:Example of categorical-value encoding (label encoding

3.7.3 Class-Imbalance Handling

Traffic accident datasets often exhibit class imbalance, in which some accident-factor categories — such as human-related causes — dominate the dataset while others, such as environmental or vehicle-related factors, occur much less frequently. Such imbalance can bias a model towards the majority class and lead to poor predictive performance for minority classes, which is particularly problematic in public safety studies where even infrequent factors may have serious consequences. Several techniques exist to address class imbalance, including undersampling of majority classes, oversampling of minority classes, and algorithm-level approaches such as weighted loss functions.

In this study, the Synthetic Minority Oversampling Technique (SMOTE) was selected as the primary method of handling class imbalance. Rather than duplicating existing observations, SMOTE generates synthetic minority-class instances by interpolating between neighbouring samples in feature space. SMOTE was applied only to the training set, after the train–test split, in order to prevent data leakage and to ensure that evaluation on the testing set reflects the real-world class distribution. By balancing the class distribution, SMOTE improves the model's ability to learn meaningful decision boundaries for minority categories and thereby enhances recall and F1-score for underrepresented accident factors.

3.7.4 Data Transformation and Normalization

Data transformation converts data from one format into another to make it suitable for machine learning. Normalization, a key transformation step, rescales numerical values onto a common scale and is required when attributes have widely differing ranges or units of measurement (Jo, 2019; Jayalakshmi & Santhakumaran, 2011). In this dataset, casualty age ranges from 0 to 94, driver age ranges from 18 to 70, and vehicle service age ranges from 1 to 30 years; these attributes were therefore normalized so that their differing scales would not bias the models. The min–max normalization technique was used, as defined in Equation 3.1.

$$X' = (X - \min(X)) / (\max(X) - \min(X)) \qquad (3.1)$$

In Equation 3.1, X is the original value, X′ is the normalized value, and min(X) and max(X) are the minimum and maximum values of the attribute, respectively. Normalization improves model performance and helps to reduce overfitting.

3.8 Feature Selection

Feature selection is the process of identifying and retaining the subset of features most relevant to model construction, based on domain knowledge and statistical relevance. In this study, candidate features were drawn from the human, vehicle, road, and environmental categories. Domain knowledge and correlation analysis were first used to identify potentially relevant predictors, after which statistical tests — the Chi-square test for categorical variables and the ANOVA F-test for numerical variables — were applied to evaluate the significance of each feature with respect to the target accident-factor categories. Principal Component Analysis was explored to examine the underlying feature structure and reduce redundancy, and Recursive

Feature Elimination was applied with selected classifiers to iteratively remove the least important attributes. These combined techniques ensured that only informative, non-redundant features were retained, thereby improving classification accuracy, reducing overfitting, and enhancing interpretability. The 19 attributes retained for model development are listed in Table 3.4.

Table 3.5:Features selected for model development

No.	Attribute	Data Type	Possible Values
1	Accident area	Categorical	Urban; rural
2	Casualty sex	Categorical	Female; male
3	Casualty age	Numeric	Integer
4	Driver age	Numeric	Integer
5	Driver sex	Categorical	Female; male
6	Driver education	Categorical	Illiterate; elementary; secondary; above secondary
7	Driver experience	Numeric	Continuous (years)
8	Vehicle type	Categorical	Truck; bus; minibus; taxi; private car; Bajaj; etc.
9	Vehicle service	Numeric	Integer (years)
10	Cause of accident	Categorical	Over-speeding; failure to give priority; brake failure; sleep; etc.
11	Accident type	Categorical	Vehicle-to-pedestrian; vehicle-to-vehicle; vehicle-to-other
12	Road condition	Categorical	Good; not good
13	Road geometry	Categorical	Straight; curve; uphill; downhill
14	Road type	Categorical	Asphalt; cobblestone; gravel; pista
15	Road surface	Categorical	Dry; wet/damp; snow; frost/ice
16	Lighting condition	Categorical	Day; night
17	Weather condition	Categorical	Cold; good; hot
18	Time of accident	Categorical	Morning; afternoon
19	Accident factor (target)	Categorical	Human; environmental; road; vehicle

3.9 Machine Learning Models

This study develops a road traffic accident factor classification model using supervised machine learning applied to a labeled accident dataset. Four classification algorithms — Random Forest, XGBoost, Decision Tree, and Multi-Layer Perceptron — were selected on the basis of their established performance in accident-classification research and their suitability for structured

tabular data (Ramya et al., 2019; Sharma et al., 2016). The algorithms were also chosen to balance interpretability, predictive performance, and robustness. All models were trained on the preprocessed dataset, and their hyperparameters were tuned to obtain optimal configurations.

3.9.1 Decision Tree

A Decision Tree is one of the simplest and most widely used supervised learning algorithms and can be applied to both classification and regression problems. It has a flowchart-like structure in which each internal node represents an attribute, each branch represents a decision rule, and each leaf node represents an outcome; the tree learns to partition the data on the basis of attribute values (Chauhan, 2020). The Decision Tree is also the building block of the Random Forest algorithm. Several parameters influence its performance and help to control overfitting, including the maximum depth of the tree (max_depth), the minimum number of samples required to split an internal node (min_samples_split), the splitting strategy (splitter), the function used to measure split quality (criterion), and the maximum number of features considered at each split (max_features). The architecture of the Decision Tree classifier is shown in Figure 3.5.

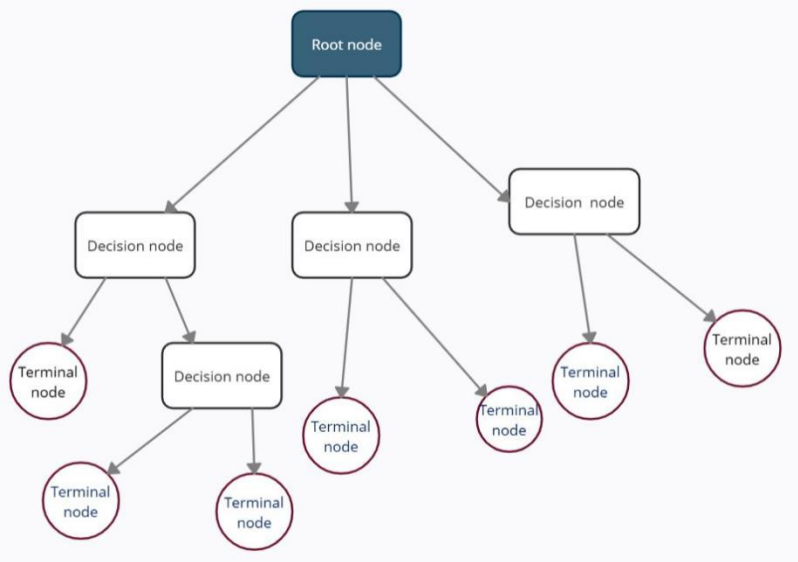


Figure 3.5:Architecture of the Decision Tree classifier

3.9.2 Random Forest

Random Forest is a popular ensemble learning method used for classification and regression. It constructs many decision trees from randomly selected subsets of the data and features, obtains a prediction from each tree, and combines these predictions — typically by majority vote — to

produce the final output. Random Forest is able to handle large, high-dimensional datasets, improves predictive accuracy, and reduces overfitting; in general, the more trees the forest contains, the more robust it is. The algorithm also provides a useful measure of feature importance, automatically computing a relevance score for each feature during training. Its principal parameters include the number of trees in the forest (n_estimators), the maximum number of features considered at a split (max_features), the maximum depth of each tree (max_depth), the minimum number of samples in a leaf node (min_samples_leaf), and the split-quality criterion (criterion). The architecture of the Random Forest classifier is shown in Figure 3.6.

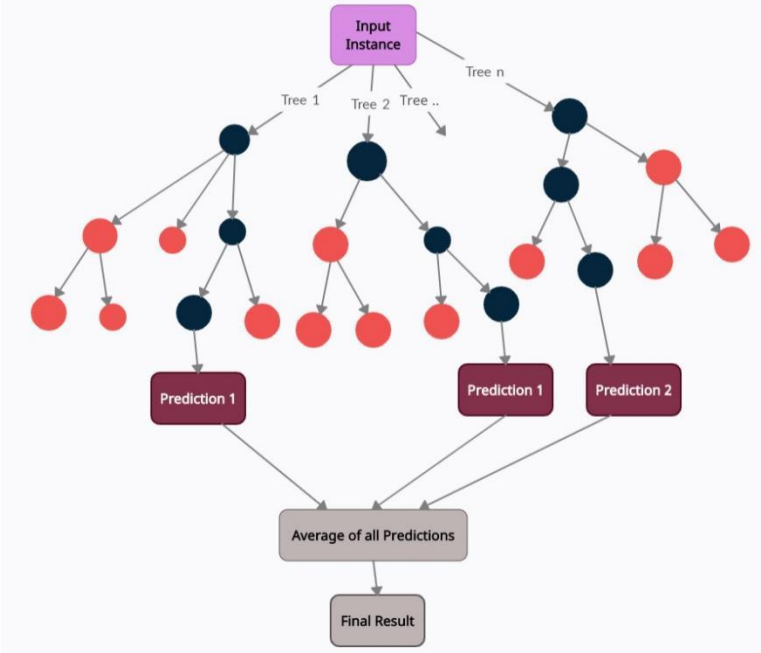


Figure 3.6:Architecture of the Random Forest classifier

3.9.3 XGBoost Classifier

Extreme Gradient Boosting (XGBoost) is a supervised learning algorithm based on gradient-boosted decision trees that has become a dominant method for structured and tabular data. It builds an ensemble of trees sequentially, with each new tree correcting the errors of its predecessors, and it is optimised for both speed and predictive performance (Chen & Guestrin, 2016). XGBoost can be used for both classification and regression and is widely valued for its computational efficiency and strong performance on difficult machine learning tasks. Its principal hyperparameters include the number of estimators (n_estimators), the learning rate, the

maximum tree depth (max_depth), the subsample ratio, and the column-sampling ratio (colsample_bytree).

3.9.4 Multi-Layer Perceptron Classifier

The Multi-Layer Perceptron (MLP) is a feed-forward artificial neural network capable of achieving high predictive performance and of modeling complex, nonlinear relationships, although it is sometimes criticized for having a large number of free parameters (Windeatt, 2008). An MLP consists of at least three layers: an input layer, one or more hidden layers, and an output layer. Its principal hyperparameters include the hidden-layer configuration (hidden_layer_sizes), which specifies the number of layers and the number of nodes in each layer; the maximum number of training iterations or epochs (max_iter); the activation function for the hidden layers; and the solver used for weight optimization. The general architecture of the MLP is shown in Figure 3.7.

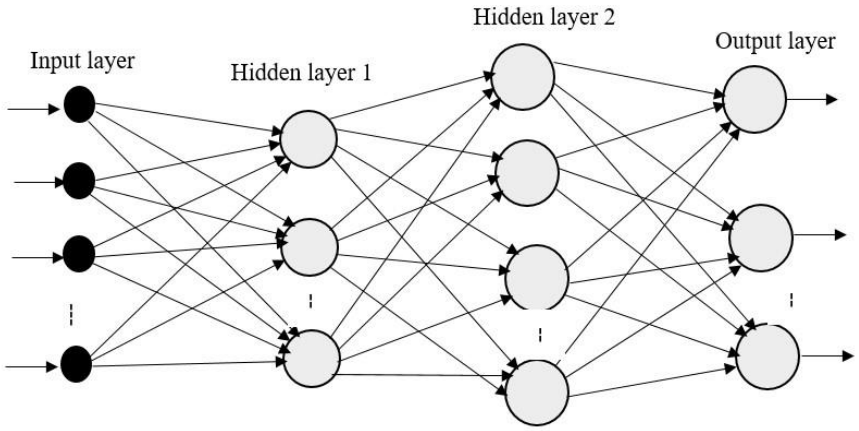


Figure 3.7:General architecture of the Multi-Layer Perceptron

3.10 Hyperparameter Tuning

Hyperparameter tuning is the process of selecting the parameter values that optimize model performance, and it is an important step in building an effective machine learning model (Jain, 2021). Three tuning strategies were used in this study, each matched to the complexity of the model concerned. Grid Search, implemented through the GridSearchCV function of scikit-learn, exhaustively evaluates every combination of values in a predefined grid and was applied to models with relatively small parameter spaces, such as the Decision Tree and Random Forest. Randomized Search samples random combinations from specified parameter distributions and

was used in exploratory phases to identify promising parameter ranges efficiently. Bayesian Optimization, which uses a probabilistic model to learn from previous trials and to select new parameter combinations intelligently, was used for the more complex models, namely XGBoost and the Multi-Layer Perceptron. For each model, the hyperparameters not included in the search were left at their default values.

3.11 Cross-Validation

Cross-validation was used to ensure the robustness and generalizability of the models. Specifically, stratified k-fold cross-validation was adopted because it preserves the original class distribution within each fold, thereby reducing bias in imbalanced classification problems. In this approach, the dataset is partitioned into k subsets — ten in this study — and the model is trained on k−1 folds and validated on the remaining fold; this process is repeated k times, and the performance metrics are averaged to provide a reliable estimate of model accuracy and stability. Stratification was particularly important in this study because the four accident-factor categories are not evenly distributed.

3.12 Performance Evaluation Metrics

Several standard metrics were used to evaluate the performance of the machine learning models, including accuracy, precision, recall, F1-score, ROC-AUC, and the confusion matrix. ROC-AUC was used to assess each model's ability to distinguish between classes, and confusion matrices were used to visualize correct and incorrect predictions and to identify misclassification patterns. Together, these metrics provide a rigorous evaluation framework. The metrics are defined in terms of four fundamental quantities:

- True Positive (TP): the predicted and actual values are both positive.
- True Negative (TN): the predicted and actual values are both negative.
- False Positive (FP): the predicted value is positive, but the actual value is negative.
- False Negative (FN): the predicted value is negative, but the actual value is positive.

In a multi-class setting, the confusion matrix shows the number of instances correctly classified (the diagonal cells) and incorrectly classified (the off-diagonal cells) for each class (Tekie &

Beshah, 2020). The structure of the confusion matrix for the four accident-factor classes is shown in Table 3.6.

Table 3.6:Confusion matrix structure for the four accident-factor classes

Actual \ Predicted	Environmental	Human	Road	Vehicle
Environmental	True Environmental	False Human	False Road	False Vehicle
Human	False Environmental	True Human	False Road	False Vehicle
Road	False Environmental	False Human	True Road	False Vehicle
Vehicle	False Environmental	False Human	False Road	True Vehicle

Accuracy measures how often the model is correct overall and is calculated as the ratio of correctly classified instances to the total number of instances (Equation 3.2). Precision measures the proportion of instances predicted as belonging to a class that actually belong to that class (Equation 3.3), and recall, also known as sensitivity, measures the proportion of actual members of a class that are correctly identified (Equation 3.4). The F1-score is the harmonic mean of precision and recall, with a best value of 1 and a worst value of 0 (Equation 3.5).

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (3.2)$$

$$\text{Precision} = TP / (TP + FP) \quad (3.3)$$

$$\text{Recall} = TP / (TP + FN) \quad (3.4)$$

$$\text{F1-score} = (2 \times \text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (3.5)$$

Because the four accident-factor classes are imbalanced, macro-averaged precision, recall, and F1-score were also computed. Macro-averaging calculates each metric separately for every class and then takes the weighted mean, thereby giving equal importance to all classes regardless of their frequency.

3.13 Hardware and Software Requirements

To support the development and deployment of the prototype Streamlit application, the hardware and software specifications recommended in Table 3.7 were used.

Table 3.7:Hardware and software requirements

No.	Component	Minimum Requirement	Recommended
1	Processor	Dual-core CPU (2 GHz)	Quad-core (Intel Core i5/i7 or AMD Ryzen)
2	Memory (RAM)	4 GB	8 GB or more
3	Storage	500 MB free space	1 GB or more
4	Operating system	Windows 10/11 (64-bit)	Windows 10/11 (64-bit)
5	Software	Python 3.9+, Streamlit 0.84+, scikit-learn, pandas, NumPy, Matplotlib	Latest stable versions of the listed software

3.14 Users of the System

The prototype road traffic accident factor classification system is designed to support two primary groups of users. The first group consists of traffic police analysts, who can use the system to identify the main contributing factors of accidents under specific conditions, to support risk profiling and targeted interventions, and to benefit from consistent, explainable outputs based on real accident data. The second group consists of transport safety researchers and policymakers, who can use the classified accident factors to analyse accident trends, to generate evidence-based reports, to evaluate existing road safety policies, and to develop more effective accident-prevention strategies.

3.15 Ethics and Data Governance

The study adhered strictly to national and institutional research ethics guidelines, and all data were handled confidentially and used solely for academic research. Formal permission to access road traffic accident records was obtained through a signed Memorandum of Understanding between the researcher and the Gamo Zone Kamba Woreda Traffic Police Department, and ethical clearance was sought from the Institutional Review Board of the Federal Technical and Vocational Training Institute. For the qualitative component of the study, all interview participants provided informed consent, both verbal and written, after being briefed on the purpose and scope of the research.

To protect privacy, all personal identifiers — including names, licence-plate numbers, and contact details — were removed or masked before analysis, and unique codes were assigned to records to maintain confidentiality while preserving data integrity. The digitised dataset is stored

on an encrypted, access-restricted drive managed by the Federal Technical and Vocational Training Institute and, in line with institutional policy, will be retained for a defined period for audit or re-analysis purposes before being permanently deleted.

3.16 Risks and Mitigation

Several risks were anticipated across the stages of the research — data collection, preprocessing, feature selection, model building, evaluation, and deployment. Table 3.8 outlines these risks together with the strategies adopted to mitigate them.

Table 3.8:Risks and mitigation strategies

Risk	Likelihood	Impact	Mitigation Strategy
Incomplete or low-quality data	High	High	Validate records with domain experts; apply data cleaning and imputation
Limited computing or internet infrastructure	Medium	Medium	Use lightweight, offline-capable tools such as Streamlit
Difficulty engaging domain experts	Medium	Medium	Schedule early consultations and build rapport with stakeholders
Low model accuracy or poor generalisation	Low	High	Apply cross-validation, feature engineering, and hyperparameter tuning
User resistance to new technology	Medium	High	Conduct demonstrations and train users with an interactive prototype

CHAPTER FOUR: EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

This chapter presents and discusses the experimental results obtained from the machine learning models developed for road traffic accident factor classification. A series of experiments was conducted using four supervised learning algorithms with the widely adopted 70/30 percentage-split option, in which 70% of the dataset was used for training and the remaining 30% for testing. The experiments were performed using the Random Forest (RF), Decision Tree (DT), Extreme Gradient Boosting (XGBoost), and Multi-Layer Perceptron (MLP) classifiers. For every classifier, performance was assessed using a consistent set of metrics — accuracy, precision, recall, F1-score, ROC-AUC, and the confusion matrix — all implemented with the help of Python's scikit-learn library. The results are presented in the subsequent sections.

4.2 Prepared Dataset for the Model

After the road traffic accident data had been collected, detailed data analysis and experimental preparation were carried out using Python. The final prepared dataset comprises 5,250 instances and 19 features, including the target class. The target variable, accident factor, contains 630 instances of "Environmental factor (0)", 3,747 instances of "Human factor (1)", 442 instances of "Road factor (2)", and 431 instances of "Vehicle factor (3)". These four categories constitute the target classes into which accident factors are classified on the basis of the labelled historical data. The distribution of the classes in the dataset is shown in Figure 4.1.

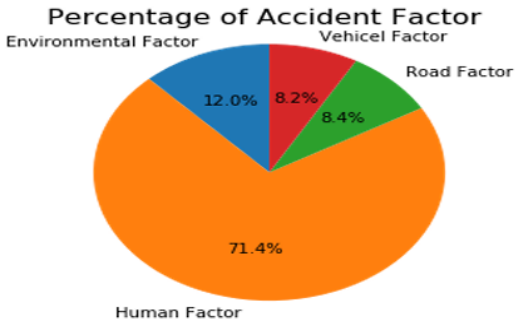


Figure 4.1:Percentage distribution of accident-factor classes in the dataset

As Figure 4.1 shows, the dataset is highly imbalanced: the human-factor class dominates the data, while the environmental, road, and vehicle classes are comparatively underrepresented. To address this imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied to the training partition only, as described in Chapter Three, so that model evaluation on the testing set continued to reflect the original real-world class distribution.

4.3 Dataset Splitting

Several data-splitting strategies are available for training and testing machine learning models. In this study, an experimental comparison of different percentage splits — including 70/30, 80/20, and others — was carried out using a trial-and-error procedure. The best result was obtained with a 70/30 split, in which 70% of the accident dataset (3,675 instances) was allocated for training and 30% (1,575 instances) for testing on previously unseen data. The 70/30 split is also one of the most widely used data-partitioning techniques in machine learning research and has been adopted by many researchers (Vrigazova, 2021). The experiments reported in this chapter were therefore carried out using a 70/30 percentage split.

4.4 Machine Learning Model Selection

Many algorithms can be used for classification, including the decision tree, random forest, neural networks, support vector machines, k-nearest neighbours (KNN), the multi-layer perceptron, naïve Bayes, and XGBoost. In this study, an extensive experimental approach was adopted in which more than ten classification algorithms — among them naïve Bayes, KNN, SVM, Decision Tree, Random Forest, MLP, AdaBoost, and XGBoost — were trained and evaluated on the traffic accident records. The Decision Tree, Random Forest, MLP, and XGBoost classifiers outperformed the others in terms of accuracy, recall, F1-score, ROC-AUC, and precision. Consequently, these four classifiers were selected for further experimentation and detailed evaluation.

4.5 Hyperparameter Tuning

Hyperparameter tuning, also known as hyperparameter optimization is an important process in machine learning that improves model performance and yields more accurate classification results. It identifies the optimal combination of parameters for building an effective model (Jain,

2021). In this study, the Grid Search technique was applied to systematically evaluate different hyperparameter combinations and to select the best-performing configuration for each algorithm.

Grid search is one of the fundamental methods for identifying the best parameters for a model. It builds a model for every possible combination of the supplied hyperparameter values, evaluates each model, and selects the parameter set that produces the highest accuracy (Chauhan & Singh, 2021). The technique is not limited to a single model but can be applied to any machine learning algorithm. The implementation used the GridSearchCV() method from the scikit-learn library. The hyperparameters and search ranges considered for each algorithm are listed in Table 4.1; all remaining hyperparameters were left at their default values.

Table 4.1:Hyperparameter search space used for the Grid Search method

Model	Parameter	Values
RF	n_estimators	100, 200
	min_samples_split	2, 5
	min_samples_leaf	1, 2
	max_depth	10, 20, 30
	max_features	'sqrt', 'log2'
DT	min_samples_split	2, 5, 10
	max_depth	5, 10, 15, 20
	criterion	gini, entropy
	min_samples_leaf	1, 2, 4
XGBoost	n_estimators	100, 200
	learning_rate	0.01, 0.05, 0.1
	max_depth	5, 10
	subsample	0.8, 1.0
MLP	hidden_layer_sizes	(50), (100), (100, 50)
	max_iter	500
	activation	tanh, relu
	learning_rate_init	0.001, 0.01

4.6 Experimental Results and Analysis

This section discusses the experimental results of the selected classifiers, obtained using Python machine learning tools. After preparation and preprocessing, the dataset was ready for experimentation. The road traffic accident data contains a target variable with four accident-factor classes: human factors, road factors, environmental factors, and vehicle factors. Experiments were conducted for the four selected classification algorithms — Random Forest, Decision Tree, XGBoost, and MLP — using the 70/30 percentage split, so that 3,675 instances were used for training and 1,575 instances for testing.

4.6.1 Random Forest (RF) Experimental Results

In this experiment, a model was built using the Random Forest classifier. The experiment was conducted with different parameter combinations in order to determine the hyperparameters for the proposed model, all of which were obtained through grid search during model training to achieve optimal results. The selected optimal hyperparameters are shown in Table 4.2.

Table 4.2:Selected hyperparameters for the RF model

Best RF parameter	n_estimators	min_samples_split	max_depth	min_samples_leaf	max_features
Values	100	2	20	1	sqrt

The experiment was carried out using the selected parameters above. Table 4.3 shows the percentage of correctly and incorrectly classified instances produced by the Random Forest classifier under the 70/30 split with grid search.

Table 4.3:Classification training result using the Random Forest algorithm

Random Forest classifier	Correctly classified	Incorrectly classified
Result	87.4%	12.6%

As Table 4.3 shows, the Random Forest classifier correctly classified 87.4% of instances and incorrectly classified 12.6% under the 70/30 percentage split. The summary of the experimental results, expressed through the standard evaluation metrics, is presented in Table 4.4.

Table 4.4:Experimental results using the Random Forest algorithm

Class	Precision	Recall	F1-score
Environmental (0)	0.55	0.76	0.64
Human (1)	0.95	0.89	0.92

Class	Precision	Recall	F1-score
Road (2)	0.86	0.85	0.86
Vehicle (3)	0.96	0.94	0.95
Macro avg	0.83	0.86	0.84
Weighted avg	0.89	0.87	0.88
ROC-AUC score	0.97		

As Table 4.4 indicates, the experiment was evaluated using precision, recall, and F1-score. The recall (true-positive rate) was 76% for environmental factors, 89% for human factors, 85% for road factors, and 94% for vehicle factors. The precision, recall, and F1-score were respectively 55%, 76%, and 64% for environmental factors; 95%, 89%, and 92% for human factors; 86%, 85%, and 86% for road factors; and 96%, 94%, and 95% for vehicle factors. The macro averages of precision, recall, and F1-score were 83%, 86%, and 84% and the weighted averages were 89%, 87%, and 88%, respectively.

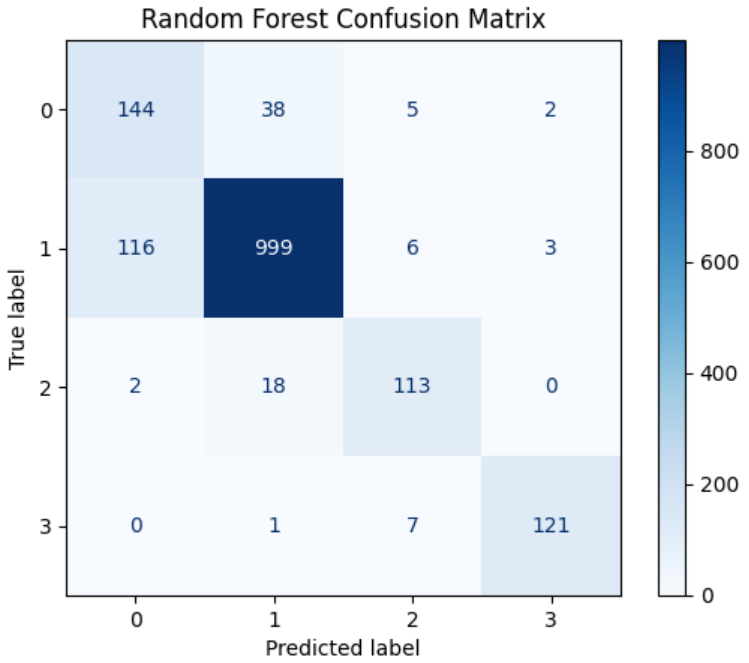


Figure 4.2:Confusion matrix of the Random Forest algorithm

Figure 4.2 shows the confusion matrix of the Random Forest classifier. Out of the 1,575 test records, 1,377 were correctly classified, giving an overall accuracy of 87.4%. For the individual classes, 144 environmental-factor, 999 human-factor, 113 road-factor, and 121 vehicle-factor records were correctly classified. The classifier misclassified 38 environmental-factor instances

as human factors, 5 as road factors, and 2 as vehicle factors, and misclassified 116 human-factor instances as environmental factors, 6 as road factors, and 3 as vehicle factors.

4.6.2 Decision Tree (DT) Experimental Results

The Decision Tree was used as an alternative classifier in the search for the most suitable machine learning model for road traffic accident factor classification. As with the other models, the experiment was conducted with different parameter combinations, all obtained through grid search during model training, in order to achieve optimal results. The selected optimal hyperparameters for the Decision Tree model are shown in Table 4.5.

Table 4.5:Selected hyperparameters for the DT model

DT best parameters	max_depth	criterion	min_samples_leaf	min_samples_split
Values	20	gini	1	2

The experiment was conducted using the Decision Tree classifier with the hyperparameters listed above; the result is shown in Table 4.6.

Table 4.6:Classification training result using the Decision Tree algorithm

Decision Tree classifier	Correctly classified	Incorrectly classified
Result	86%	14%

As Table 4.6 shows, the accuracy of the Decision Tree algorithm was evaluated on 1,575 instances: 86% of the instances were correctly classified, and the remaining 14% were incorrectly classified.

Table 4.7:Experimental results of the Decision Tree classifier

Class	Precision	Recall	F1-score
Environmental (0)	0.58	0.70	0.64
Human (1)	0.94	0.89	0.91
Road (2)	0.71	0.82	0.76
Vehicle (3)	0.95	0.89	0.92
Macro avg	0.80	0.83	0.81
Weighted avg	0.88	0.86	0.87
ROC-AUC score	0.88		

As Table 4.7 indicates, the experiment used the 70/30 percentage split and was evaluated using precision, recall, ROC-AUC, and F1-score. The precision, recall, and F1-score were respectively

58%, 70%, and 64% for environmental factors; 94%, 89%, and 91% for human factors; 71%, 82%, and 76% for road factors; and 95%, 89%, and 92% for vehicle factors. The macro averages of precision, recall, and F1-score were 80%, 83%, and 81%, the weighted averages were 88%, 86%, and 87%; and the ROC-AUC score was 88%.

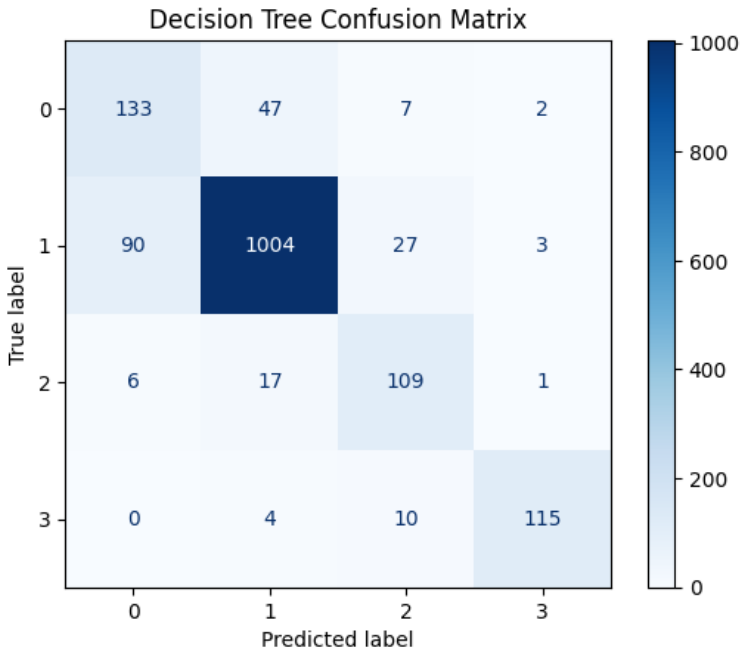


Figure 4.3:Confusion matrix of the Decision Tree classifier

As Figure 4.3 shows, the Decision Tree classifier correctly classified 1,361 of the 1,575 test records. The confusion matrix indicates that 133 environmental-factor, 1,004 human-factor, 109 road-factor, and 115 vehicle-factor records were classified correctly.

4.6.3 XGBoost Experimental Results

XGBoost is a supervised learning algorithm that has recently emerged as a leading approach for structured or tabular datasets. It is an implementation of gradient-boosted decision trees optimized for speed and performance. The classifier has many hyperparameters that influence model performance, and these were tuned with the grid search method to obtain the optimal model. The best parameters identified through this process are shown in Table 4.8.

Table 4.8:Selected hyperparameters for the XGBoost model

XGBoost best parameters	max_depth	subsample	n_estimators	colsample_bytree	learning_rate
Values	8	0.8	200	0.8	0.1

Using the best hyperparameters identified through grid search, the XGBoost model was built to obtain optimal results. The experimental results are shown in Table 4.9.

Table 4.9:Classification training result using the XGBoost classifier

XGBoost classifier	Correctly classified	Incorrectly classified
Result	90%	10%

As Table 4.9 shows, the XGBoost classifier correctly classified approximately 90% of the instances and incorrectly classified about 10%. The model achieved a test accuracy of 91.52% on the 1,575-instance testing dataset, while Table 4.15 reports a training accuracy of 100%. The difference occurs because the training accuracy reflects the model's performance on the training data, whereas the test accuracy measures its performance on unseen data and therefore provides a more realistic evaluation of the model's generalization capability.

Table 4.10:Experimental results of the XGBoost classifier

Class	Precision	Recall	F1-score
Environmental (0)	0.65	0.74	0.69
Human (1)	0.94	0.93	0.94
Road (2)	0.89	0.83	0.86
Vehicle (3)	0.96	0.95	0.96
Macro avg	0.86	0.86	0.86
Weighted avg	0.91	0.90	0.90
ROC-AUC score	0.974		

As Table 4.10 indicates, the experiment used a 70/30 split and was evaluated with accuracy, precision, recall, and F1-score. The XGBoost classifier registered the best performance, with an accuracy of 90%. The precision, recall, and F1-score were respectively 65%, 74%, and 69% for environmental factors; 94%, 93%, and 94% for human factors; 89%, 83%, and 86% for road factors; and 96%, 95%, and 96% for vehicle factors. The macro averages of precision, recall, and F1-score were 86%, 86%, and 86% with an ROC-AUC score of 97.4%, and the weighted averages of precision, recall, and F1-score were 91%, 90%, and 90%, respectively.

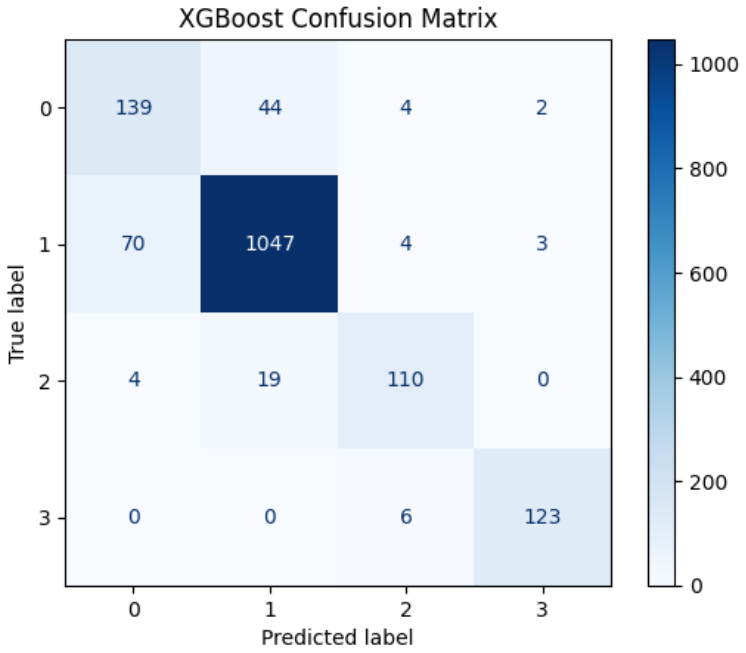
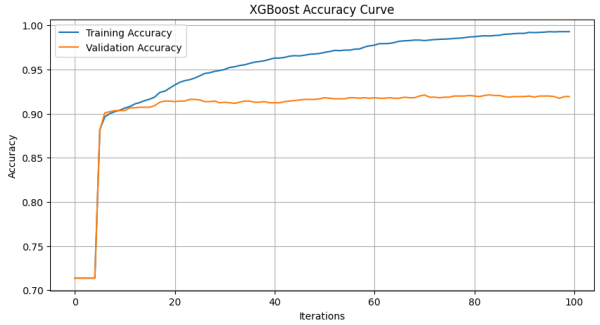


Figure 4.4:Confusion matrix of the XGBoost algorithm

As Figure 4.4 shows, the XGBoost classifier correctly classified 1,419 of the 1,575 test records under the 70/30 split. For the individual classes, 139 environmental-factor, 1,047 human-factor, 110 road-factor, and 123 vehicle-factor records were classified correctly.



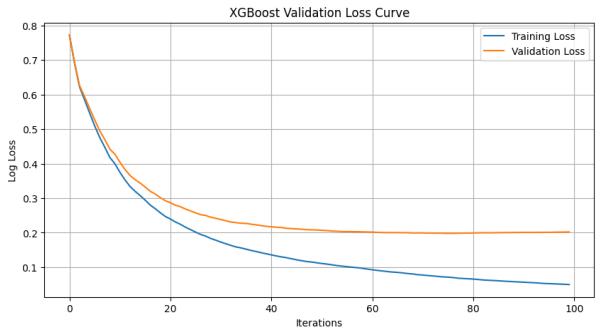


Figure 4.5:XGBoost classifier accuracy and validation-loss curves

The XGBoost accuracy and validation-loss curves demonstrate strong model performance in accident-factor classification. Training accuracy increased steadily to nearly 99%, while validation accuracy stabilized at around 91–92%, indicating good generalization. Both training and validation loss decreased significantly during training, showing effective learning and reduced prediction error. The small gap between the training and validation curves suggests slight overfitting at later iterations. Overall, the results confirm that the XGBoost classifier achieved high predictive performance and is effective for multiclass accident-factor classification.

4.6.4 Multi-Layer Perceptron (MLP) Experimental Results

As with the other classifiers, grid search was used to select the best hyperparameters for the MLP model, which were determined empirically through a series of experiments on the dataset that yielded the highest classification accuracy. The selected hyperparameters for the MLP model are summarized in Table 4.11.

Table 4.11:Selected hyperparameters for the MLP model

Best MLP parameter	hidden_layer_sizes	activation	solver	learning_rate	alpha
Values	(100, 50)	tanh	adam	constant	0.0001

After the best hyperparameters had been selected through grid search, the MLP model was built using those parameters to obtain optimal results. The experimental results are shown in Table 4.12.

Table 4.12:Classification training result using the MLP algorithm

MLP classifier	Correctly classified	Incorrectly classified
Result	85%	15%

Table 4.12 shows the performance of the MLP classifier: 85% of the instances were correctly classified, and 15% were incorrectly classified.

Table 4.13:Experimental results of the MLP classifier

Class	Precision	Recall	F1-score
Environmental (0)	0.52	0.60	0.56
Human (1)	0.92	0.89	0.90

Class	Precision	Recall	F1-score
Road (2)	0.72	0.79	0.76
Vehicle (3)	0.94	0.91	0.92
Macro avg	0.77	0.80	0.78
Weighted avg	0.86	0.85	0.85
ROC-AUC score	0.93		

As Table 4.13 shows, the MLP classifier was evaluated using accuracy, precision, recall, and F1-score. The classifier correctly classified 85% of instances and incorrectly classified 15%. The precision, recall, and F1-score were respectively 52%, 60%, and 56% for environmental factors; 92%, 89%, and 90% for human factors; 72%, 79%, and 76% for road factors; and 94%, 91%, and 92% for vehicle factors. The macro averages of precision, recall, and F1-score were 77%, 80%, and 78% with an ROC-AUC score of 0.93, and the weighted averages were 86%, 85%, and 85%, respectively.

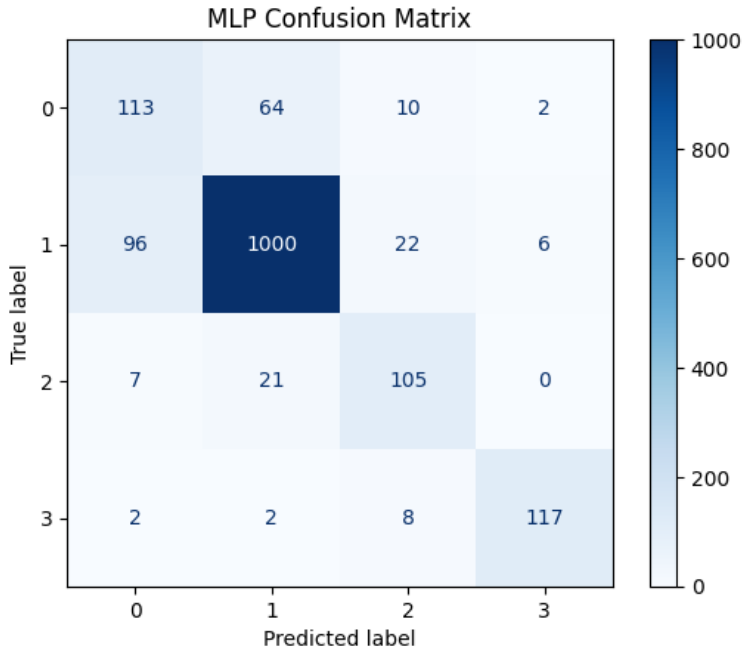


Figure 4.6:Confusion matrix of the MLP classifier

As Figure 4.6 shows, the MLP classifier correctly classified 1,335 of the 1,575 test records, giving an overall accuracy of 85%. The confusion matrix indicates that 113 environmental-factor, 1,000 human-factor, 105 road-factor, and 117 vehicle-factor records were classified correctly.

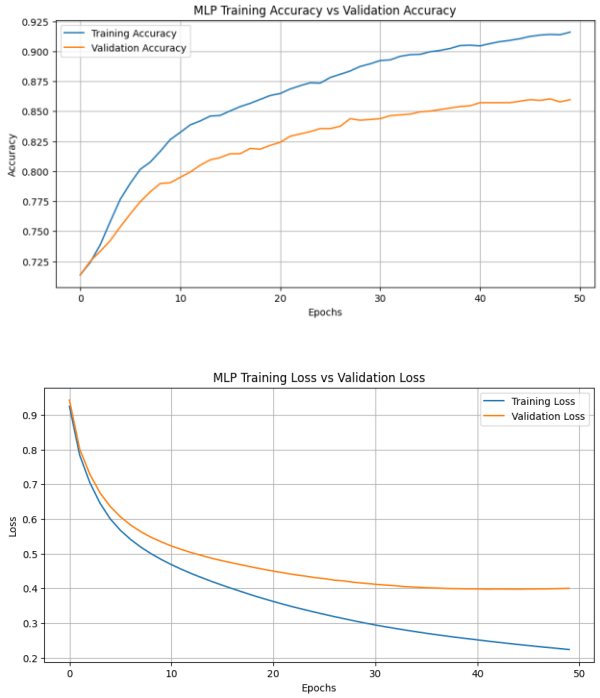


Figure 4.7:MLP classifier accuracy and loss curves

As Figure 4.7 shows, the MLP model does not exhibit significant overfitting. There is no substantial gap between the training and validation accuracy, and the validation loss decreases in line with the training loss.

4.7 Discussion of Experimental Results

The experiments produced four trained models based on the 70/30 percentage split, evaluated with accuracy, recall, precision, ROC-AUC, and F1-score. Table 4.14 summarizes the comparative performance of the four classifiers across these metrics.

Table 4.14:Summary of the comparative evaluation results

Evaluation metric	Random Forest	Decision Tree	XGBoost	MLP
Correctly classified instances	1,377	1,361	1,419	1,335
Incorrectly classified instances	198	214	156	242
Accuracy	0.85	0.86	0.91	0.85
Average precision	0.90	0.89	0.91	0.84
Average recall	0.90	0.88	0.91	0.85
Average F1-score	0.90	0.88	0.91	0.84
ROC-AUC score	0.97	0.88	0.97	0.92

As Table 4.14 shows, the XGBoost classifier achieved the highest accuracy at 0.91, followed by the Decision Tree at 0.86, with the Random Forest and MLP classifiers both at 0.85. The XGBoost algorithm performed best across all evaluation metrics, with an overall accuracy of 0.91. Figure 4.8 presents the comparison graphically.

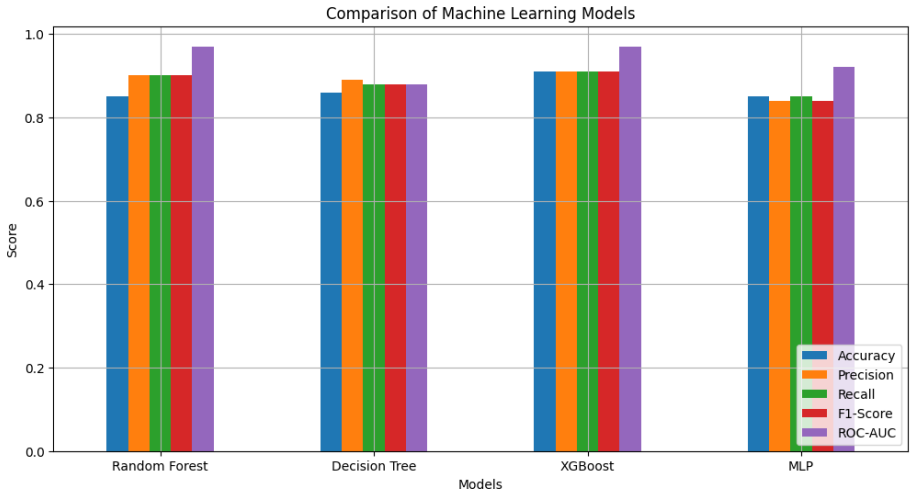


Figure 4.8:Comparative summary of model performance

As Figure 4.8 shows, the XGBoost classifier outperformed the Decision Tree, Random Forest, and MLP classifiers across all evaluation metrics — precision, recall, F1-score, ROC-AUC, and accuracy — with an overall performance of approximately 0.91 on each metric.

4.8 Model Testing and Evaluation

Evaluating model performance on unseen or held-out data is an essential phase of any machine learning study. In this study, both a held-out test set and k-fold cross-validation were used. The cross-validation employed k = 10, partitioning the data into ten equal folds in which nine folds were used for training and the remaining fold for validation; this was repeated ten times, and the results averaged. The cross_val_predict() method with k = 10 was used to implement the evaluation. Table 4.15 reports the training, testing, and cross-validation results for each model.

Table 4.15:Model testing and evaluation results

Algorithm	Metric	Training accuracy	Testing accuracy	Cross-validation	Precision	Recall	F1-score	ROC-AUC
Random Forest	Accuracy	0.9998	0.9171	0.8834	0.9132	0.9171	0.9120	0.9766
Decision Tree	Accuracy	1.0000	0.8952	0.8596	0.8956	0.8952	0.8954	0.8869

Algorithm	Metric	Training accuracy	Testing accuracy	Cross-validation	Precision	Recall	F1-score	ROC-AUC
XGBoost	Accuracy	1.0000	0.9152	0.8842	0.9122	0.9152	0.9131	0.9766
MLP	Accuracy	0.9412	0.8867	0.8594	0.8877	0.8867	0.8871	0.9564

As Table 4.15 shows, when the algorithms are compared in terms of overfitting, the Decision Tree, Random Forest, and XGBoost classifiers performed consistently across the training, testing, and cross-validation stages, whereas the MLP algorithm showed a larger difference between training and validation performance. The XGBoost classifier achieved the strongest results across all three evaluation options — training, testing, and cross-validation — and performed best across all evaluation metrics. On this basis, XGBoost was identified as the most appropriate model for road traffic accident factor classification.

4.8.1 Overall Evaluation Performance of the Selected Models

Figure 4.9 describes the overall performance of the four classifiers — Random Forest, Decision Tree, XGBoost, and MLP — across both the train/test split and the k-fold cross-validation. The test set and the cross-validation results differ only slightly from the training accuracy, indicating that the models generalise well.

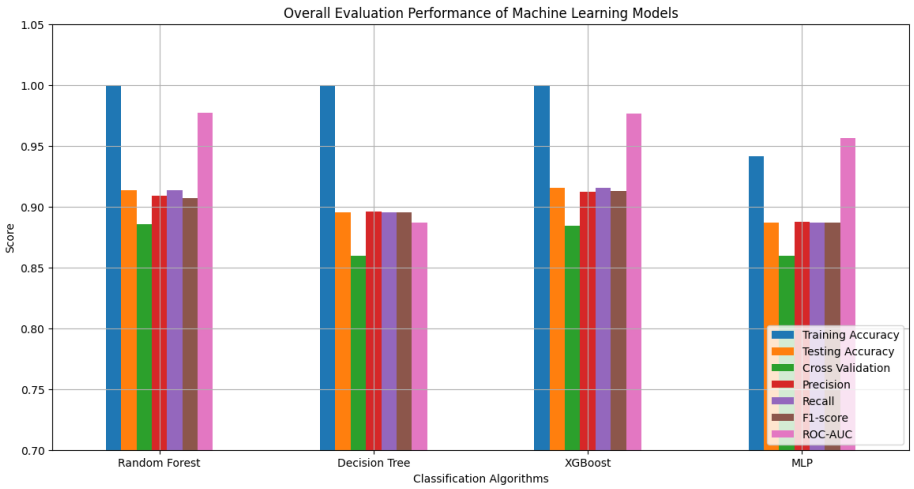


Figure 4.9:Model accuracy in training, testing, and cross-validation

As Figure 4.9 shows, the XGBoost classifier achieved the best overall performance across the three evaluation options. Using this best-performing model, a prototype for classifying the factors contributing to road traffic accidents was developed.

4.9 Prototype of the Classification Model

Following the evaluation, the best-performing model (XGBoost) was deployed as a working prototype using the Streamlit library. Streamlit is used to build lightweight, custom web applications for machine learning models. The prototype accepts the relevant accident attributes as user input and displays the predicted accident-factor class — environmental, human, road, or vehicle. The prototype provides traffic safety officers and other stakeholders with an accessible decision-support tool that demonstrates the practical applicability of the classification model.

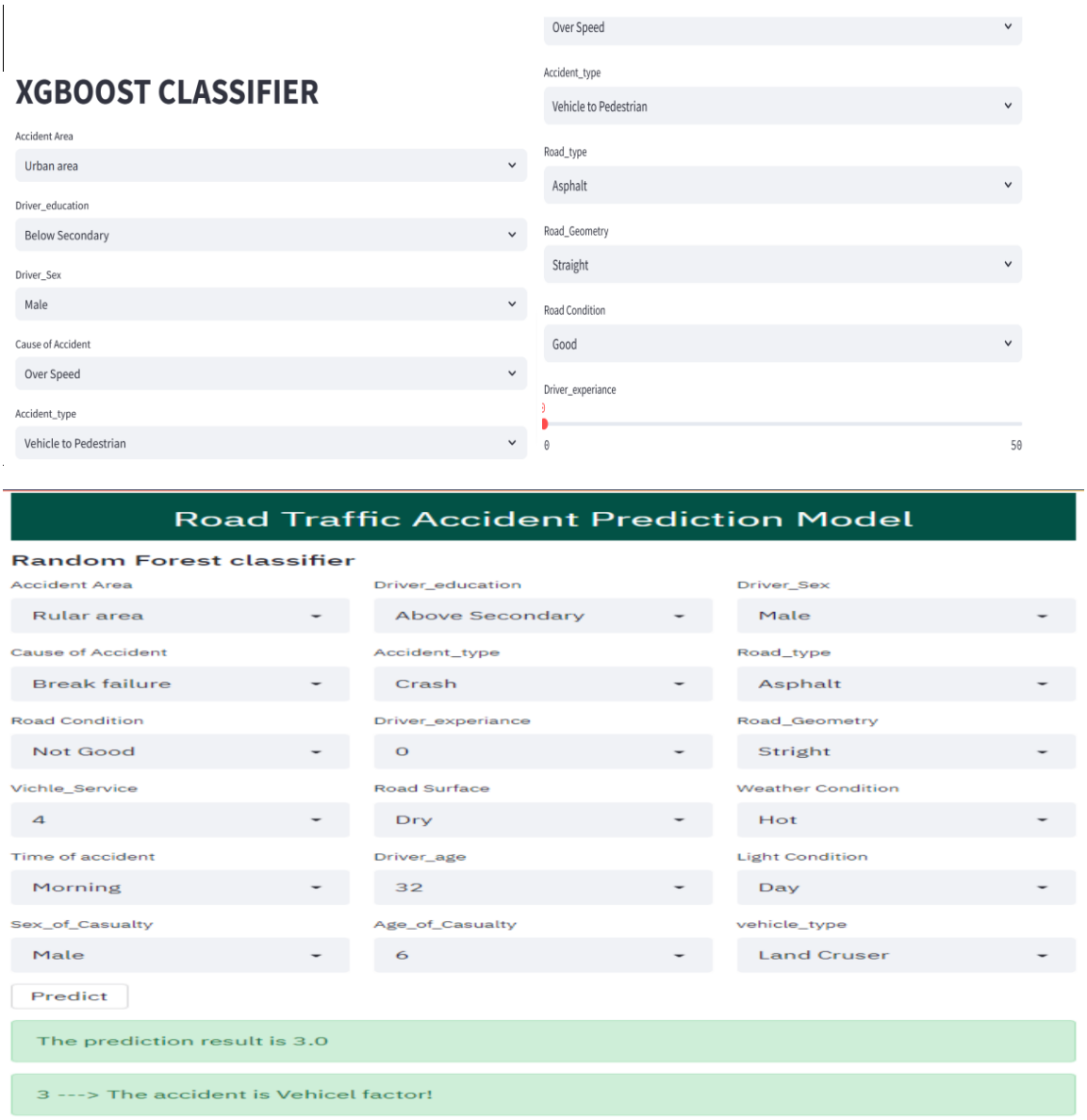


Figure 4.10:Road Accident factor classification model prototype

CHAPTER FIVE: CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

The main objective of this study was to develop a machine learning-based classification model to identify the factors that contribute to road traffic accidents (RTA). More specifically, the study set out to determine the most influential attributes associated with road traffic accidents and to classify accident factors in order to support improved road safety decision-making.

To achieve this objective, it was first necessary to understand the nature and characteristics of road traffic accidents. The dataset was collected from the Kamba Woreda Traffic Police Office in the Gamo Zone and transformed from a manual format into an electronic dataset, after which it was cleaned and prepared for analysis. The dataset originally contained 21 attributes; however, after preprocessing and feature selection, 19 attributes (18 predictor attributes and 1 target attribute) were retained for model development. These attributes include the date of accident, accident area, age and sex of casualty, driver sex, driver experience, driver education, cause of accident, accident type, road surface, lighting condition, road geometry, time of accident, vehicle service, and the target attribute accident factor.

Several procedures were carried out to identify the most relevant features influencing accident-factor classification. The most significant attributes were found to include accident area, sex and age of casualty, driver sex and age, driver experience, driver education, cause of accident, accident type, road surface, weather condition, lighting condition, road type, road geometry, time of accident, vehicle type, and vehicle service.

The study analyzed 5,250 road traffic accident records collected from Kamba Woreda, Gamo Zone over a seven-year period. The dataset was analysed using Python and a range of machine learning libraries. Data preprocessing techniques were applied to ensure data quality and suitability for modeling. Missing values were handled using mean imputation for numerical attributes and mode imputation for categorical attributes, and categorical variables were converted into numerical form using label encoding. The target variable, accident factor, was encoded as 0 for environmental factors, 1 for human factors, 2 for road factors, and 3 for vehicle factors.

Four machine learning algorithms — XGBoost, Decision Tree, Multi-Layer Perceptron, and Random Forest — were implemented and trained. The dataset was split into training and testing sets using a 70/30 ratio, and model performance was evaluated using accuracy, precision, recall, F1-score, ROC-AUC, and confusion-matrix analysis.

The experimental results on the test data showed that XGBoost achieved an accuracy of approximately 91%,Random Forest 87.4%, the Decision Tree 86%, and the MLP 85%. The XGBoost classifier demonstrated the best overall performance and was therefore selected as the most appropriate model for classifying the factors contributing to road traffic accidents. Finally, a prototype classification model was developed and tested using the XGBoost algorithm, demonstrating that the model is effective and reliable for identifying and classifying road traffic accident factors.

5.2 Recommendations

On the basis of the findings of this study, the following recommendations are made:

- Obtaining well-organized data was one of the most difficult and time-consuming aspects of this study, as the records had to be converted from a manual format into an electronic one. Traffic offices should therefore record accident history data digitally, so that domain experts and other stakeholders can access the data easily when needed.
- Although a real traffic accident dataset was used, the number of instances may be limited compared with the volume of accident data available at a national scale. Future models should be developed using larger datasets drawn from multiple sources.
- While the accuracy achieved in this study is promising, other modeling approaches may yield more accurate results. Additional models should be applied to this data and their results compared with those reported here.
- Although most of the relevant accident-factor features were considered, some potentially important features may have been excluded. Future work should incorporate additional features that may influence accident-factor classification.
- The classification model could be further improved by testing additional machine learning methods — including unsupervised learning techniques — and, where possible, increasing the sample size, then comparing the results with those of this study.

5.3 Future Work

This study can be extended in several directions. Future research could enlarge the dataset by collecting accident records from additional woredas and zones, enabling the model to generalize across a wider geographical area. Deep learning architectures and ensemble strategies beyond those examined here could be explored and compared. The prototype could be developed into a fully integrated decision-support system for traffic police offices, with multilingual support and the capacity to update the model as new accident data becomes available. Finally, incorporating explainability techniques would help stakeholders understand and trust the predictions produced by the model, supporting its adoption in real-world road safety planning.

REFERENCES

The following list consolidates the references cited throughout this thesis. Duplicate entries present in the original draft have been merged, and the references are presented in alphabetical order.

Alasadi, S. A., & Bhaya, W. S. (2017). Review of data preprocessing techniques in data mining. Journal of Engineering and Applied Sciences, 12(16), 4102–4107.

Anusuya, R., & Rajalakshmi, P. (2020). Predicting severity level of road traffic accidents in Oromiya East Shewa Zone. Journal of Data Mining and Knowledge Engineering, 12(1), 31–37.

Aslanyan, T. (2023, October 24). Machine learning fundamentals handbook — key concepts, algorithms, and Python code examples. freeCodeCamp. https://www.freecodecamp.org/news/machine-learning-handbook/

Assefa, G., Gizachew, M., & Deribe, K. (2020). Predicting factors of vehicular accidents using machine learning algorithms. International Journal of Emerging Trends in Engineering Research, 8(9), 26–35.

Behnood, A., & Mannering, F. (2015). The temporal stability of factors affecting driver-injury severities in single-vehicle crashes: A random parameters approach. Analytic Methods in Accident Research, 8, 7–32. https://doi.org/10.1016/j.amar.2015.01.001

Bharti, S., Garg, N., & Mishra, B. B. (2016). Traffic accident prediction model using support vector machines. Procedia Computer Science, 89, 472–478. https://doi.org/10.1016/j.procs.2016.06.082

Bokaba, T., Doorsamy, W., & Paul, B. S. (2022). Comparative study of machine learning classifiers for modelling road traffic accidents. Applied Sciences, 12(2), 828. https://doi.org/10.3390/app12020828

Chauhan, N. S. (2020, January 14). Decision tree algorithm, explained. KDnuggets. https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Das, A., Abdel-Aty, M., & Pande, A. (2009). Using conditional probability to identify risky drivers and predict crash probability. Accident Analysis & Prevention, 41(1), 142–152. https://doi.org/10.1016/j.aap.2008.10.010

De Oña, J., López, G., & Abellán, J. (2013). Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. Accident Analysis & Prevention, 50, 67–77. https://doi.org/10.1016/j.aap.2012.09.004

De Soto, B. G. (2018). Predicting road traffic accidents using artificial neural network models. Infrastructure Asset Management, 5(4), 132–144.

Dires, D. M., & Dadi, Y. A. (2022). Predicting factors of vehicle traffic accidents using machine learning algorithms: The case of the Wolaita Zone. International Journal of Engineering and Advanced Technology, 11(4), 17–23.

El Morr, C., Jammal, M., Ali-Hassan, H., & El-Hallak, W. (2022). Overview of machine learning algorithms. In Machine learning for practical decision making (pp. 61–115). Springer. https://doi.org/10.1007/978-3-031-16990-8_3

Endalew, M. M., Abebe, S. M., & Tadesse, T. (2024). Road traffic accidents and the contributing factors among drivers of public transportation in Mizan Aman town, Ethiopia. Frontiers in Public Health, 12, 1456789. https://doi.org/10.3389/fpubh.2024.1456789

Federal Police Commission. (2020). Road traffic accident report, July 2019–December 2020. Addis Ababa, Ethiopia.

Ferreira, A. J., & Figueiredo, M. A. T. (2012). Efficient feature selection filters for high-dimensional data. Pattern Recognition Letters, 33(13), 1794–1804.

GeeksforGeeks. (2019, May 6). History of Python. https://www.geeksforgeeks.org/history-of-python/

Getachew, E., Alemu, T., & Bekele, H. (2024). Socioeconomic and behavioral factors of road traffic accidents among drivers in Ethiopia: A systematic review and meta-analysis. BMC Public Health, 24(1), 1123. https://doi.org/10.1186/s12889-024-1123-y

Grandini, M., & Visani, G. (2021). Metrics for multi-class classification: An overview. arXiv preprint arXiv:2008.05756.

Hadi, S. (2019). Book review: Creswell, J. W. (2014). Research design: Qualitative, quantitative, and mixed methods approaches (4th ed.). ResearchGate.

Jayalakshmi, T., & Santhakumaran, A. (2011). Statistical normalization and back propagation for classification. International Journal of Computer Theory and Engineering, 3(1), 89–93.

Jo, J.-M. (2019). Effectiveness of normalization pre-processing of big data to the machine learning performance. The Journal of the Korea Institute of Electronic Communication Sciences, 14(3), 547–552.

Khalak, A. A., Tesfaye, B., & Mohammed, Y. (2024). Road accidents and safety measures: A diverse study of Ethiopia. International Journal of Research Publication and Reviews, 5(7), 234–242.

Kurika, A. E., Ganie, I. A., Kadir, Y., Cerna, P. D., & Desei, F. L. (2020). Predicting factors of vehicular accidents using machine learning algorithms. International Journal of Emerging Trends in Engineering Research, 8(9), 5171–5176. https://doi.org/10.30534/ijeter/2020/46892020

Kurika, A. E., & Sundado, T. S. (2020). Predicting factors of vehicle traffic accidents using machine learning algorithms: The case of the Wolaita Zone. International Journal of Scientific Research in Computer Science and Engineering, 8(4), 1–7.

Nasteski, V. (2017). An overview of the supervised machine learning methods. Horizons B, 4, 51–62.

Nishida, T. (2010). Data transformation and normalization. Rinsho Byori (The Japanese Journal of Clinical Pathology), 58(10), 990–997.

Puppala, A. (2025). A comprehensive overview of machine learning models: Techniques, applications, and future directions. International Journal of Future Multidisciplinary Research, 3(1), 45–60.

Rajkumar, A. R., Prabhakar, S., & Priyadharsini, A. M. (2020). *Prediction of road accident severity using machine learning algorithm*. *International Journal of Advanced Science and Technology, 29*(6), 116–120.

Rajkumar, S., Ramesh, P., & Anandhakrishnan, R. (2020). Prediction of road accident severity using machine learning algorithms. International Journal of Advanced Trends in Computer Science and Engineering, 9(2), 2047–2052. https://doi.org/10.30534/ijatcse/2020/48922020

Ramasamy, A., Dechasa, S., & Mulugeta, A. (2020). Predicting severity level of road traffic accidents in Oromiya East Shewa Zone using Iterative Dichotomiser 3. International Journal of Recent Technology and Engineering, 9(1), 1–7.

Ramya, S., Reshma, S., Manogna, V. D., Saroja, Y. S., & Gandhi, G. S. (2019). Accident severity prediction using data mining methods. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 5(2), 528–536.

Samya, R. (2017). Attribute selection using machine learning technique. International Journal of Computational Intelligence and Informatics, 7(2), 122–132.

Sharma, A., Sharma, S., & Sharma, A. (2021). Machine learning: Algorithms, real-world applications and research directions. SN Computer Science, 2(3), 1–20. https://doi.org/10.1007/s42979-021-00592-x

Sharma, B., Katiyar, V. K., & Kumar, K. (2016). Traffic accident prediction model using support vector machines with Gaussian kernel. In Proceedings of the Fifth International Conference on Soft Computing for Problem Solving. Springer.

Shweta, Yadav, J., Batra, K., & Goel, A. K. (2021). A framework for analyzing road accidents using machine learning paradigms. Journal of Physics: Conference Series.

Tekie, H., & Beshah, T. (2020). Moisture content prediction model for pharmaceutical granules using machine learning techniques. Authorea. https://doi.org/10.22541/au.160313597.69423867/v1

Tibebe, B., Ejigu, D., & Abraham, A. (2011). Pattern recognition and knowledge discovery from road traffic accident data in Ethiopia: Implications for improving road safety. In 2011 World Congress on Information and Communication Technologies (pp. 1241–1246). IEEE. https://doi.org/10.1109/WICT.2011.6141393

Tiwari, V., Kumar, A., & Jain, S. (2020). Road accident severity predictions using machine learning algorithms. International Journal of Scientific & Technology Research, 9(3), 1104–1108.

United Nations Economic Commission for Africa & United Nations Economic Commission for Europe. (2020). Road safety performance review: Ethiopia. UNECE.

Vrigazova, B. (2021). The proportion for splitting data into training and test set for the bootstrap in classification problems. Business Systems Research, 12(1), 228–242.

Wenqi, L., & Dongyu, L. (2017). A model of traffic accident prediction based on convolutional neural network. In 2017 2nd IEEE International Conference on Intelligent Transportation Engineering (pp. 198–202). IEEE. https://doi.org/10.1109/ICITE.2017.8056908

Wikipedia contributors. (n.d.). Road traffic accidents in Ethiopia. Wikipedia. Retrieved June 12, 2025, from https://en.wikipedia.org/wiki/Road_traffic_accidents_in_Ethiopia

Windeatt, T. (2008). Ensemble MLP classifier design. In Computational intelligence paradigms (pp. 133–147). Springer.

World Health Organization. (2021). Road traffic injuries. WHO. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

World Health Organization. (2023). Road safety country profile: Ethiopia. WHO. https://www.who.int/publications/m/item/road-safety-eth-2023-country-profile

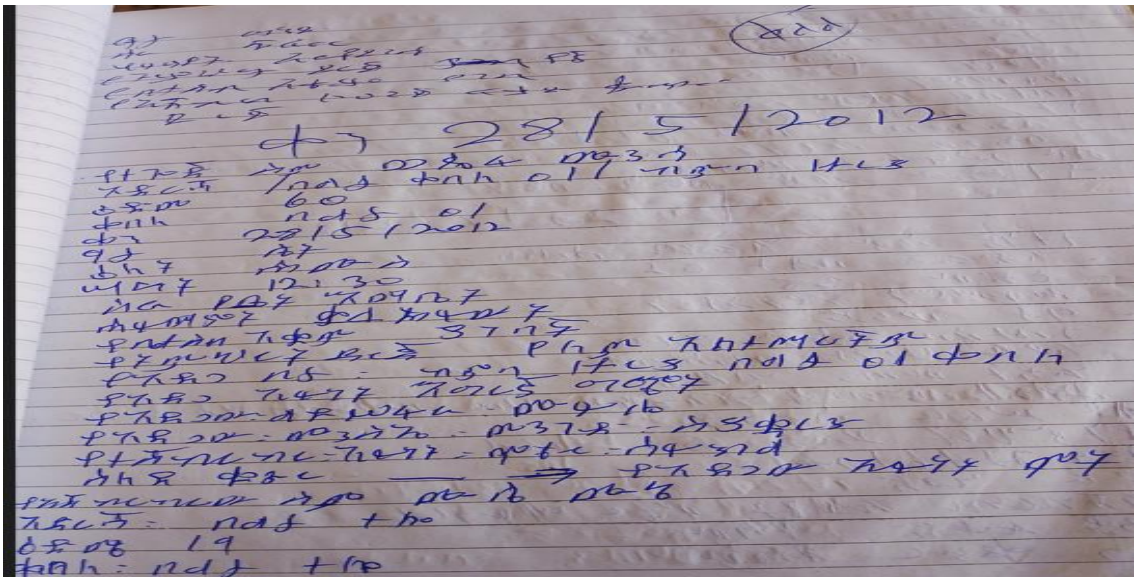
World Life Expectancy. (2020). Road traffic accidents in Ethiopia. https://www.worldlifeexpectancy.com/ethiopia-road-traffic-accidents

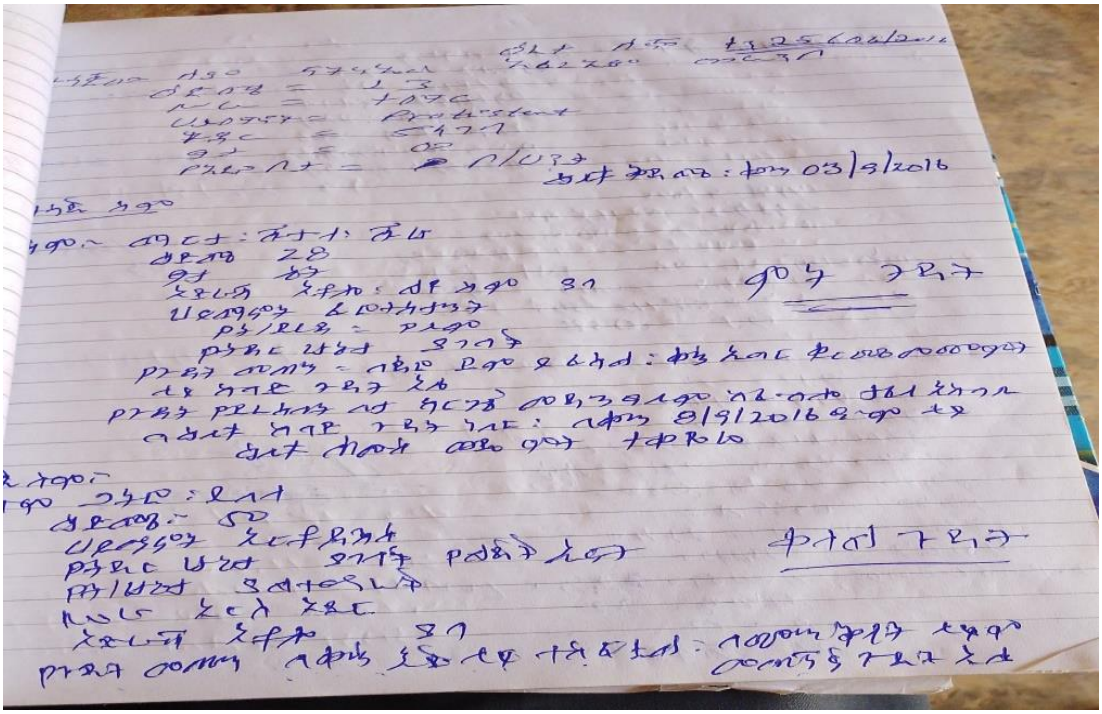
Yu, R., & Abdel-Aty, M. (2014). Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data. Safety Science, 63, 50–56. https://doi.org/10.1016/j.ssci.2013.11.014

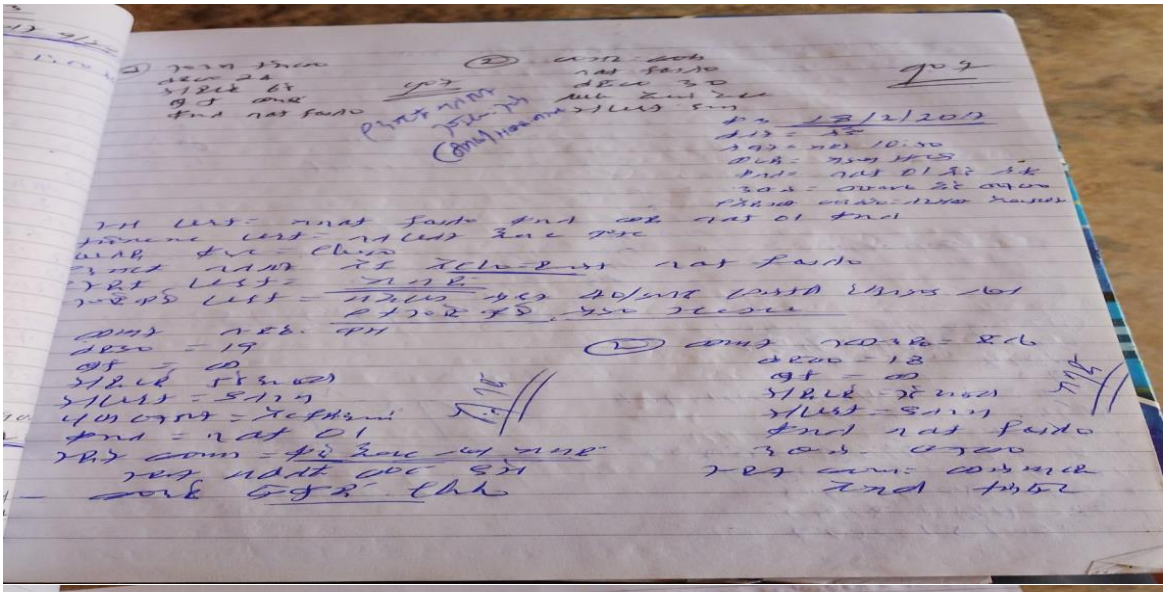
Yuan, Z., Zhou, X., & Yang, T. (2018). Hetero-ConvLSTM: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 984–992). ACM. https://doi.org/10.1145/3219819.3219922

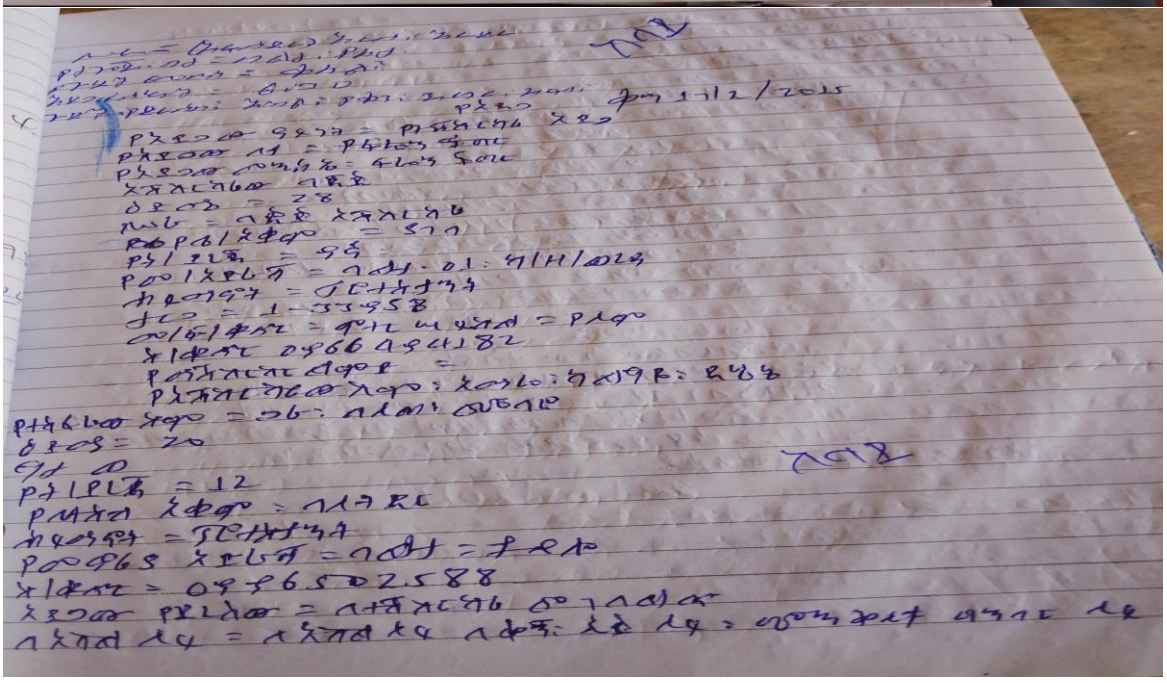
APPENDICES

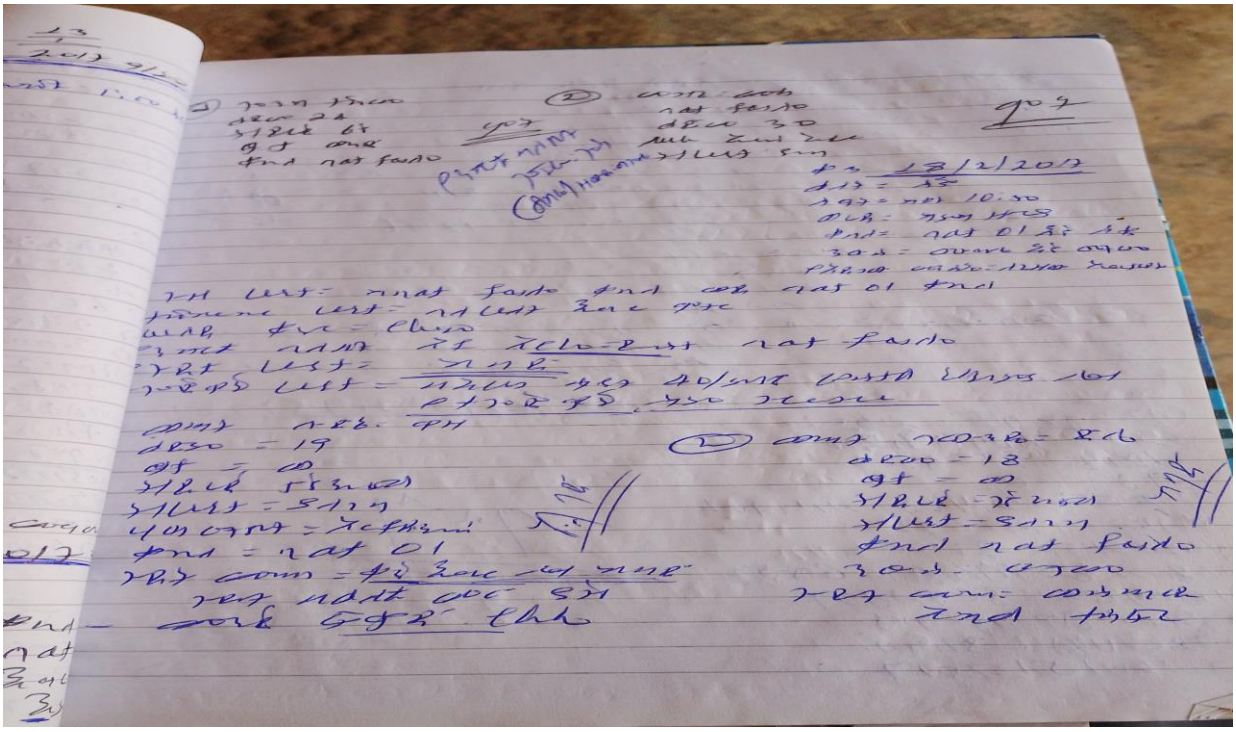
Appendix A: Manual Data before Converting Into Csv File

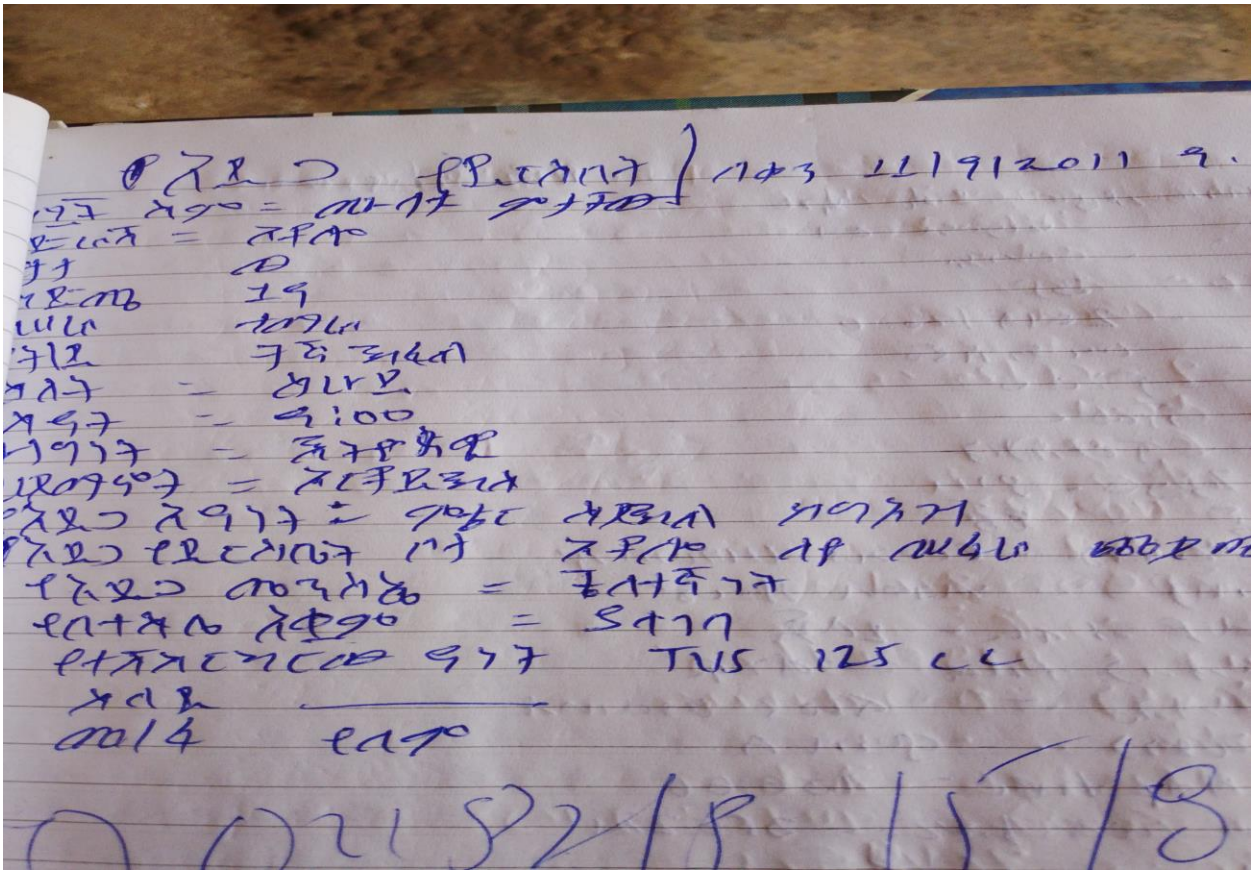












Appendix B: Sample CSV Dataset before Preprocessing

Accident_	Driver_se	Driver_ag	Driver_ed	Driver_ex	Vehicle_se	Vehicle_ty	Road_type	Road_geo	Weather_	Accident_	Casualty_	Casualty_	Cause_of_	Time	Road_surf	Road_con	Lighting_c	Accident_fact
Urban	Male	65	Diploma	8	4	Truck	Earth	Straight	Clear	Single_veh	Female	5	Over_spe	Morning	Dry	Good	Daylight	Human_facto
Rural	Male	46	Diploma	26	1	Truck	Earth	Curve	Foggy	Vehicle_to	Male	28	Drunk_dri	Morning	Dry	Poor	Daylight	Vehicle_facto
Rural	Male	64	Degree	18	2	Motorcycl	Asphalt	Slope	Foggy	Single_veh	Female	74	Brake_fail	Morning	Dry	Good	Night	Human_facto
Urban	Female	35	Degree	21	6	Truck	Gravel	Curve	Rainy	Single_veh	Female	10	Poor_roac	Afternoon	Dry	Good	Night	environment
Urban	Female	67	Primary	8	1	Bus	Gravel	Curve	Clear	Vehicle_to	Female	28	Overloadi	Night	Wet	Good	Night	Human_facto
Rural	Female	55	Degree	12	4	Truck	Earth	Curve	Clear	Vehicle_to	Male	20	Brake_fail	Night	Dry	Poor	Night	environment
Rural	Male	61	Primary	22	9	Bus	Earth	Curve	Clear	Vehicle_to	Female	21	Overloadi	Morning	Wet	Good	Daylight	Vehicle_facto
Urban	Male	41	Secondary	18	9	Car	Earth	Curve	Windy	Vehicle_to	Male	47	Drunk_dri	Afternoon	Dry	Good	Daylight	Road_factor
Urban	Male	26	Degree	18	3	Bus	Earth	Slope	Windy	Vehicle_to	Male	40	Overloadi	Evening	Wet	Poor	Daylight	Human_facto
Rural	Male	55	Secondary	19	4	Car	Asphalt	Slope	Clear	Vehicle_to	Male	5	Drunk_dri	Morning	Dry	Poor	Night	environment
Rural	Male	68	Degree	26	7	Truck	Asphalt	Straight	Windy	Vehicle_to	Female	53	Overloadi	Morning	Dry	Good	Night	Vehicle_facto
Urban	Male	30	Secondary	18	8	Truck	Gravel	Straight	Foggy	Vehicle_to	Male	10	Overloadi	Morning	Dry	Good	Daylight	Human_facto
Rural	Female	31	Degree	29	1	Truck	Gravel	Straight	Windy	Vehicle_to	Female	37	Overloadi	Night	Dry	Good	Night	environment
Urban	Female	21	Primary	19	8	Bicycle	Earth	Straight	Clear	Single_veh	Male	24	Over_spe	Morning	Dry	Poor	Daylight	Human_facto
Urban	Male	44	Diploma	30	5	Truck	Earth	Slope	Foggy	Vehicle_to	Female	51	Brake_fail	Evening	Wet	Poor	Daylight	Vehicle_facto
Urban	Male	52	Secondary	17	5	Truck	Gravel	Straight	Rainy	Vehicle_to	Female	21	Overloadi	Evening	Dry	Poor	Daylight	Vehicle_facto
Urban	Male	35	Diploma	20	4	Bus	Asphalt	Slope	Foggy	Single_veh	Female	33	Over_spe	Morning	Wet	Poor	Daylight	environment
Urban	Female	28	Degree	18	7	Bicycle	Asphalt	Straight	Clear	Single_veh	Male	70	Over_spe	Evening	Dry	Poor	Daylight	Vehicle_facto
Urban	Female	31	Secondary	22	2	Bus	Earth	Curve	Rainy	Vehicle_to	Male	23	Overloadi	Morning	Dry	Poor	Night	Road_factor
Urban	Female	28	Primary	13	1	Motorcycl	Asphalt	Straight	Windy	Vehicle_to	Female	30	Brake_fail	Morning	Dry	Poor	Night	environment
Rural	Female	59	Degree	22	9	Bus	Asphalt	Straight	Foggy	Vehicle_to	Female	5	Over_spe	Night	Wet	Poor	Night	Road_factor

Appendix C: importing packages

```
# ============================
# BASIC LIBRARIES
# ============================
import pandas as pd
import numpy as np

# ============================
# DATA PREPROCESSING
# ============================
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder, StandardScaler

# ============================
# MACHINE LEARNING MODELS
# ============================

# Decision Tree (DT)
```

```
from sklearn.tree import DecisionTreeClassifier

# Random Forest (RF)
from sklearn.ensemble import RandomForestClassifier

# Multi-Layer Perceptron (MLP)
from sklearn.neural_network import MLPClassifier

# XGBoost
from xgboost import XGBClassifier

# ==========================
# MODEL EVALUATION
# ==========================
from sklearn.metrics import (
    accuracy_score,
    precision_score,
    recall_score,
    f1_score,
    confusion_matrix,
    classification_report,
    roc_auc_score
)

# For multiclass ROC AUC
from sklearn.preprocessing import label_binarize

# ==========================
# VISUALIZATION
# ==========================
import matplotlib.pyplot as plt
```

```
import seaborn as sns

# ========================
# OPTIONAL (SAVE/LOAD MODELS)
# ========================
import pickle

# ========================
# IGNORE WARNINGS
# ========================
import warnings
warnings.filterwarnings("ignore")
```

Appendix D: Data Preprocessing Code

```
# ========================
# 1. LOAD DATASET
# ========================
import pandas as pd
import numpy as np

df = pd.read_csv("Traffic_Accident_dataset_final.csv")

# ========================
# 2. DISPLAY BASIC INFO
# ========================
```

```
print(df.head())
print(df.info())
print(df.isnull().sum())

# ============================
# 3. HANDLE MISSING VALUES
# ============================

# Fill numerical missing values with median
for col in df.select_dtypes(include=[np.number]).columns:
    df[col].fillna(df[col].median(), inplace=True)

# Fill categorical missing values with mode
for col in df.select_dtypes(include=["object"]).columns:
    df[col].fillna(df[col].mode()[0], inplace=True)

# ============================
# 4. ENCODING CATEGORICAL VARIABLES
# ============================
from sklearn.preprocessing import LabelEncoder
```

```
label_encoders = {}

for col in df.select_dtypes(include=["object"]).columns:
    le = LabelEncoder()
    df[col] = le.fit_transform(df[col])
    label_encoders[col] = le

# ============================
# 5. SPLIT FEATURES AND TARGET
# ============================

# Target: Accident Factor (Human, Road, Vehicle, Environmental)
target = "Accident_Factor"

X = df.drop(target, axis=1)
y = df[target]

# ============================
# 6. TRAIN-TEST SPLIT
# ============================
from sklearn.model_selection import train_test_split
```

```
X_train, X_test, y_train, y_test = train_test_split(
    X, y,
    test_size=0.2,
    random_state=42,
    stratify=y
)

# ============================
# 7. FEATURE SCALING (IMPORTANT FOR MLP
# ============================
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# ============================
# 8. FINAL OUTPUT SHAPE CHECK
# ============================
```

```
print("Training shape:", X_train.shape)

print("Testing shape:", X_test.shape)
```