Scientific Research Publishing

Open Journal of Safety Science and Technology, 2025, 15(1), 36-56
https://www.scirp.org/journal/ojsst
ISSN Online: 2162-6006
ISSN Print: 2162-5999

Analysis of Road Traffic Accident Using AI Techniques

Innocent Ekanem

Department HSE, PTAGINC, Seconded to "Environmental Health Safety Quality (SHEQ)" Unit, Rochester Gas & Electric, Rochester, NY, USA
Email: innocent_ekanem@rge.com

How to cite this paper: Ekanem, I. (2025) Analysis of Road Traffic Accident Using AI Techniques. *Open Journal of Safety Science and Technology*, **15**, 36-56.
https://doi.org/10.4236/ojsst.2025.151004

Received: December 30, 2024
Accepted: March 17, 2025
Published: March 20, 2025

Abstract

Road traffic accidents are one of the global safety and socioeconomic challenges. According to WHO (2024), it has caused over 1.19 million annual fatalities. It is also projected to cause economic losses, which are approximately $1.8 trillion between 2015 and 2030. In this research, machine learning (ML) approach was implemented to predict the severity of road traffic accidents and explore actionable insights for intervention. The dataset used in implementing machine learning models was collected from Victoria Road Crash incidence from the years 2012-2023. This dataset includes temporal, environmental, and infrastructure variables. The target variable is severity of the road accident which is in four classes: fatal, serious injury, minor injury, and property damage. The first part of the machine learning analysis involves feature analysis using feature importance by random forest and partial dependence plots. The feature analysis identified temporal factors like accident time and date as key influencing factors of severity. The significant peaks from feature analysis showed rush hours and late weekdays as major determinants of road accidents in Victoria. Similarly, speed zones also showed a significant influence on road accidents, and this emphasizes the correlation between higher speed limits and severe outcomes. Environmental and infrastructural factors, like lighting conditions and road geometry, showed comparatively lower impact. In the second part of the analysis, three machine learning models—Logistic Regression, Random Forest, and XGBoost—were implemented for predictive performance. Logistic Regression outperformed others with the classification of minor injuries (Class 3), with a recall of 100%. Random Forest showed slightly better balance across classes. However, all models struggled with minority classes, like fatal accidents (Class 1), due to class imbalance. Overall, the findings revealed the importance of targeted interventions during high-risk periods with stricter speed limit enforcement and improved lighting infrastructure.

Keywords

Safety, Machine Learning, Logistic Regression, Random Forest, XGBoost

1. Background

Road traffic accidents are one of the leading causes of death and injury across the world. This class of accidents constitutes a significant global safety issue, which results in almost 1.19 million deaths yearly and millions of injuries and disabilities [1]. Aside from the fatalities and injuries, road traffic accidents also impact the global economy negatively. As noted by [2], injuries and deaths caused by road accidents are projected to cost the world economy USD 1.8 trillion from 2015 to 2030. This is equivalent to an annual tax of 0.12% on global GDP. Undoubtedly, the socioeconomic burden of road accidents is huge as it affects the cost of healthcare, the productivity of people involved, and the emotional well-being. In this regard, predictive modelling of road traffic accident severity has become an important area of research for safety practitioners and other stakeholders seeking to mitigate the accidents and the consequences. The nature and severity of road traffic accidents are influenced by many factors. This includes environmental, vehicular, and human variables. Key determinants are road conditions, lighting, weather, vehicle speed, and driver behaviour and they all play impactful roles in shaping the outcomes of many road traffic accident outcomes [3] [4].

Accident severity ranges from minor injuries to severe outcomes and it is a key metric for understanding and mapping out road safety systems. Severity classification enables stakeholders to allocate resources effectively, prioritize high-risk areas, and implement targeted interventions. Given the complex interplay of factors that cause road traffic accidents, traditional statistical methods often fall short in capturing nonlinear relationships and high-dimensional data [5] [6]. Machine learning (ML) is a subset of AI and it offers a robust and efficient alternative for predicting accident severity by leveraging large datasets to identify hidden patterns [7]. Unlike traditional predictive approaches, ML algorithms possess the capability to model complex dependencies between environmental, vehicular, and human-related factors. Many studies have been conducted on the use of machine learning to predict road traffic accident. [8] explores machine learning techniques like KNN, AdaBoost, and Decision Tree for classifying accident severity but only focuses on model comparison without exploring the underlying relationships between predictors and crash severity. It is challenging to address the root causes of severity of road traffic accidents when influencing factors are not analysed. Similarly, [9] employed machine learning techniques like Logistic Regression, Decision Tree, and Random Forest to predict accident severity, but the research did not extensively analyse the predictors influencing the severity. While the models in the work of [9] achieved good accuracy, the neglect of model explainability limits their ability to provide actionable insights. Clearly, many of the existing

studies in the domain of road traffic accidents put more emphasis on predictive performance of machine learning models. They overlook influencing factors that can be used to interpret model predictions and translate them into practical interventions. Thus, it is imperative to bridge this gap by carrying out research that not only predicts road traffic accident severity but also extensively analyses the complex interdependencies of factors influencing the severity. With the adoption of machine learning techniques, this research aims to deliver accurate predictions and uncover hidden patterns within the road traffic accident data. More importantly, the research will provide actionable insights that can guide stakeholders based on the overall findings.

Research Questions

1) What are the significant factors influencing road crash severity?

2) How accurately can machine learning models predict the severity of road traffic accidents?

3) What actionable insights can be derived to reduce the occurrence of severe traffic accidents?

2. Literature Review

In the past, statistical approaches like regression, negative binomial models, and ordered probit models have been widely utilized to implement prediction of severity level of road traffic accidents and crashes [10] [11]. These models provided foundational insights, but they exhibit limitations when dealing with non-linear relationships and high-dimensional datasets [12]. Based on their computational design, regression model assumes linearity between predictors and the target variable [13]. This does not hold in complex road traffic accident data where variables like speed zones, light conditions, and road geometry interact dynamically. Similarly, ordered probit and logit models are constrained by their inability to model complex dependencies among features [14] [15]. This situation limits the strength of the conventional models in the domain of road traffic accidents. In recent years, machine learning models like Random Forest (RF), Support Vector Machines (SVM), Gradient Boosting, and Neural Networks have demonstrated superior performance with the prediction of crash severity. Unlike traditional models like regression models and probit and logit models, machine learning models are computationally effective at capturing non-linear relationships and complex interactions among influencing factors. Thus, machine learning techniques are well-suited for road traffic accident severity prediction. For instance, [16] in their work compared decision trees, Naïve Bayes, and Random Forest models. They concluded that Random Forest outperformed other techniques with the prediction of severity based on environmental factors like lighting and weather conditions. Similarly, [17] examined the performance of Random Forest, artificial neural networks (ANN), and decision trees with a focus on rainy conditions. In their research, Random Forest achieved the highest accuracy due to its robustness with homogeneity in road traffic accident data. [8] explored K-Nearest Neighbours

(KNN), AdaBoost, and decision trees for classification of road traffic accident severity but focused primarily on comparison of model performance. The investigation neglects analysis of underlying relationships between predictors and outcomes of road accident severity. Evidently, these studies highlight the predictive capabilities of machine learning models without considering the importance of understanding which factors trigger crash severity and how these insights will inform real-world interventions. Consequently, one of the critical challenges in existing studies lies in the detailed exploration of feature interpretability. Machine learning models, especially ensemble methods and deep learning, work as "black boxes". The models produce improved predictions without any explanation of how influencing features contribute to the severity like traditional models. As affirmed by [18], without interpretability and explainability, non-technical stakeholders using outcomes of machine learning models are left with a limited understanding of which factors contribute most to model outcomes. This limits the machine learning models' practical applicability.

The identification of influencing factors driving road traffic accident severity is essential for addressing the root causes of the accidents and developing targeted interventions. Existing research emphasizes the role of human, environmental, and vehicular factors in crash severity. Human behaviour like speeding, alcohol consumption, and fatigue, has been affirmed to increase the likelihood of severe crashes [19] [20]. Environmental conditions comprising poor lighting, adverse weather, and road geometry exacerbate the risk of traffic accidents as visibility and vehicle control is reduced [21]-[23]. Vehicular factors like vehicle type, speed zones, and road surface conditions also play a significant role in determining crash outcomes [17] [24].

Similarly, road geometry provides insights about the structural design of the and most times, it includes aspects like intersections, slope, shoulders and curves. They contribute to crash risk due to vehicle instability [23] [25] [26]. Another variable that influences severity of road traffic accident is speed zone. This factor highlights speed limits at the location of road traffic accidents. It is a vital factor of severity as higher speed zone correlates with fatal outcomes based on impact forces caused by higher speed [27] [28]. However, while these studies identify the general importance of these factors, they fail to analyse the complexity around the interdependence of the accident factors. For instance, [29] implemented ANN, SVM, and decision trees in their work for traffic accident severity prediction but they did not explore how key features in their dataset like light conditions and vehicle speed interact to influence severity. Similarly, the work [17] focused on crash severity under rainy conditions, but the broader context of multiple contributing factors was not explored. Perhaps, while predictive performance metrics like accuracy, precision, recall, and F1-score have been utilized to evaluate ML models, they are not sufficient to address the practical needs of stakeholders. High accuracy alone does not translate into actionable insights unless the models provide clear explanations of their predictions. This main gap in the existing literature

forms the foundation for this research which is targeted at answering critical questions relating to road traffic accident severity prediction and comprehensive feature analysis.

3. Dataset, Pre-Processing and Transformation

Datasets are essential to building machine learning models for prediction, classification or segmentation [30]. They serve as the main source of information that models learn patterns to carry out subsequent analysis (prediction, classification, clustering). Dataset quality and richness of features in the dataset are critical for achieving high-performing machine learning models [31]. In this research, the used for implementing machine learning models was retrieved from Kaggle Victoria Road Crash Data (2012-2023). The dataset contains comprehensive information about road crashes in Victoria (Australia) from year 2012-2023. The dataset has 15 attributes detailing factors that influence road accidents in Victoria and 152,445 records. The dataset is rich in features as it incorporates temporal, environmental and infrastructure variables (Table 1).

Table 1. Details of the features in Victoria dataset.

Feature	Description
ACCIDENT_DATE	Date of accident
ACCIDENT_TIME	Time in which accident occurred
DAY_OF_WEEK	Day of the week in which the accident happened (in Figure)
DAY_WEEK_DESC	Day of the week in which the accident happened (in Words)
ACCIDENT_TYPE	Type of the accident that occurred
ACCIDENT_TYPE_DESC	Detail description of the accident that occurred
DCA_CODE	Code highlighting factors that caused the accident
DCA_DESC	Description of the factors that caused the accident
LIGHT_CONDITION	The lighting condition when the accident occurred
NODE_ID	Unique identifier for the spatial location of the accident
ROAD_GEOMETRY_DESC	Geometry of the road where the accident occurred
SEVERITY	Severity of the accident that occurred
SPEED_ZONE	Speed limit at the accident location
RMA	Road management area where the accident occurred

The dataset was loaded into the coding environment (Google Colab) using Pandas library in Python (Figure 1).

The Victoria dataset is pre-processed by checking for missing values using **df.isnull()** method in Pandas library.

The Victoria dataset contains features that are categorical in nature (Figure 1). The features are transformed to numerical values using LabelEncoder class from the sklearn.preprocessing module to encode each categorical feature by assigning unique integer labels to distinct category values (Figure 3).

```
import pandas as pd

# Load the dataset
file_path = '/content/Vic_Road_Crash_Data.csv'
df = pd.read_csv(file_path)
print("First 5 rows of the dataset:")
print(df.head())
```

```
First 5 rows of the dataset:
    ACCIDENT_NO ACCIDENT_DATE ACCIDENT_TIME  ACCIDENT_TYPE  \
0  T20120000060      1/1/2012      19:40:00              6
1  T20120000028      1/1/2012      04:00:00              4
2  T20120000021      1/1/2012      07:30:00              4
3  T20120000056      1/1/2012      16:15:00              4
4  T20120000018      1/1/2012      05:15:00              4

                 ACCIDENT_TYPE_DESC  DAY_OF_WEEK DAY_WEEK_DESC  DCA_CODE  \
0  Vehicle overturned (no collision)            1        Sunday       184
1      Collision with a fixed object            1        Sunday       183
2      Collision with a fixed object            1        Sunday       171
3      Collision with a fixed object            1        Sunday       183
4      Collision with a fixed object            1        Sunday       173
```

Figure 1. Data importation and inspection on Google Colab.

```
missing_values = df.isnull().sum()

# Display only columns with missing values
missing_values_summary = missing_values[missing_values > 0]

# Print the summary
if missing_values_summary.empty:
    print("\nNo missing values in the dataset.")
else:
    print("\nMissing values summary:")
    print(missing_values_summary)
```

```
No missing values in the dataset.
```

Figure 2. Checking for missing values in Victoria dataset.

```
import pandas as pd
from sklearn.preprocessing import LabelEncoder


# Select categorical features for encoding (assuming non-numeric columns are categorical)
categorical_features = df.select_dtypes(include=['object']).columns

# Initialize a dictionary to store encoders for each feature
label_encoders = {}

# Encode each categorical feature using LabelEncoder
for column in categorical_features:
    le = LabelEncoder()
    df[column] = le.fit_transform(df[column].astype(str))  # Apply label encoding
    label_encoders[column] = le  # Store the encoder for future use

# Display the first 5 rows of the encoded DataFrame
print("Encoded DataFrame (first 5 rows):")
print(df.head())
```

```
Encoded DataFrame (first 5 rows):
   ACCIDENT_NO  ACCIDENT_DATE  ACCIDENT_TIME  ACCIDENT_TYPE  \
0           15              0           1180              6
1            5              0            240              4
2            4              0            450              4
3           12              0            975              4
4            3              0            315              4
```

Figure 3. Data transformation with LabelEncoder.

From the features in the Victoria dataset, the target variable is severity. Severity categorizes accident that occurred based on fatality level. It is an important parameter to assess road safety. The dataset categorizes crash severity into four classes (Figure 4). Fatal class (Class 1) includes those accidents that result in loss of life and are the most severe. Serious injury (Class 2) is a class of accident that leads to major injuries. Minor injury (Class 3) accidents are those accidents that lead to injuries that are less severe. This class of injury requires minimal medical attention. Property Damage Only (PDO) (Class 4) is a class of accidents in which there are no injuries or fatalities but cause damage to vehicles or infrastructure.

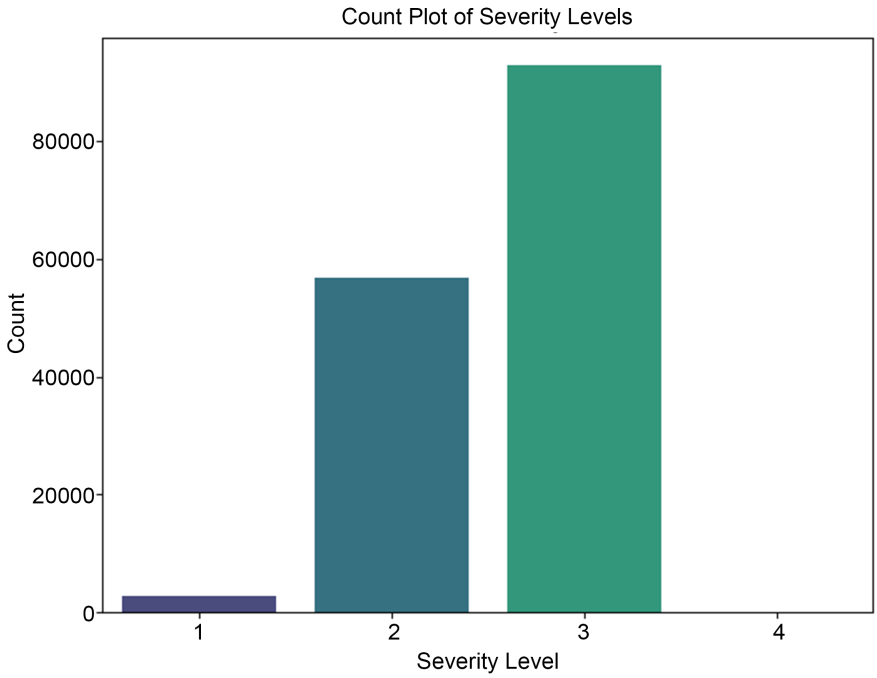


Figure 4. Classes of accident severity.

4. Feature Analysis

Feature selection techniques help to identify the most predictive features [32]. While analysis of feature enables improvement in model performance, it is also important to show interaction of influencing variables to non-technical stakeholders. There are various techniques to analyse data features and assess interactions and level of influence. Commonly used techniques feature importance derived from tree-based models like Random Forest or XGBoost. These models rank features based on the level of their contribution to the improvement of predictive accuracy. Tree-based feature importance methods are computationally efficient and valuable when handling high-dimensional datasets [33] [34]. They provide a straightforward measure for level of relevance among data features. Thus, feature importance technique is a practical choice for large datasets like the Victoria dataset. Perhaps, one of the weaknesses of feature importance technique is its bias towards features with many unique values and it must be carefully considered

when interpreting outcomes [35]. Partial Dependence Plots (PDPs) is also an effective technique for feature interpretability. The technique works by creating visuals of the marginal effect of data features on the predicted outcome. PDPs are particularly valuable for assessing non-linear relationships and interactions between features [36]. This makes it important to use the Victoria Road dataset. PDP is a step further compared to tree-based models with capability to capture how changes in a feature influence the target variable.

4.1. Feature Importance by Random Forest

The setup for feature importance using a Random Forest technique involves instantiation of random forest model. Random Forest classifier is instantiated with 100 estimators and trained on the training dataset (**X_train and y_train**). After training the model, the **feature_importances_ attribute** of the Random Forest model is prompted to provide the importance score of each feature based on the decrease in node impurity.

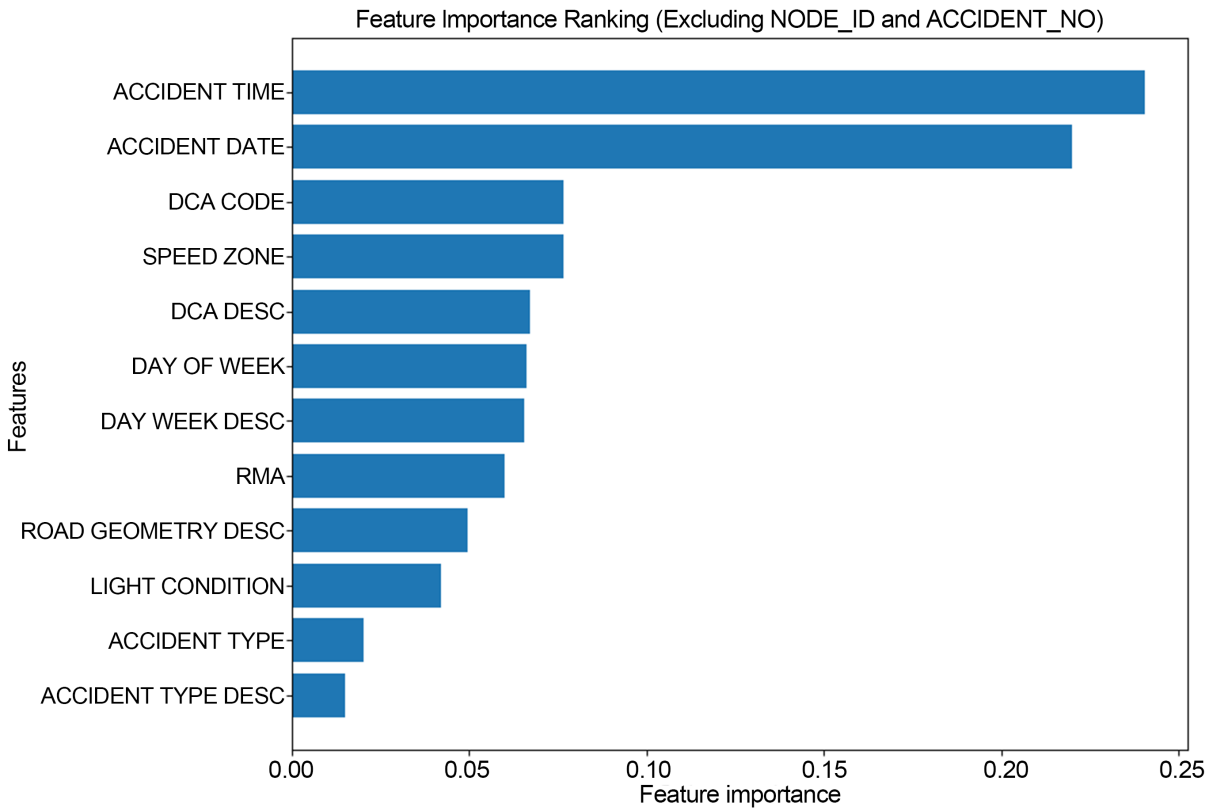


Figure 5. Feature importance plot.

From the feature importance plot, the most influential factor is ACCIDENT_TIME. This suggests that the time of the accident plays a key role in determining the level of severity of the accident outcomes. The feature ACCIDENT_DATE is the second most important, as seen in **Figure 5**. The first two features in the ranking were explored to validate their level impact on the severity of accident in Victoria.

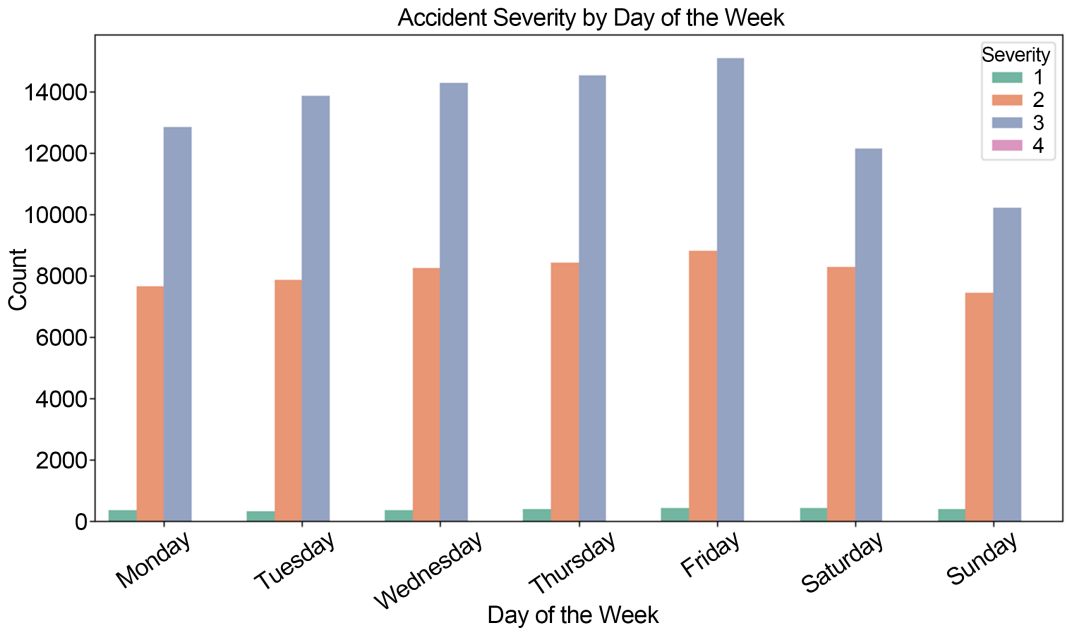


Figure 6. Distribution of accident by day of the week.

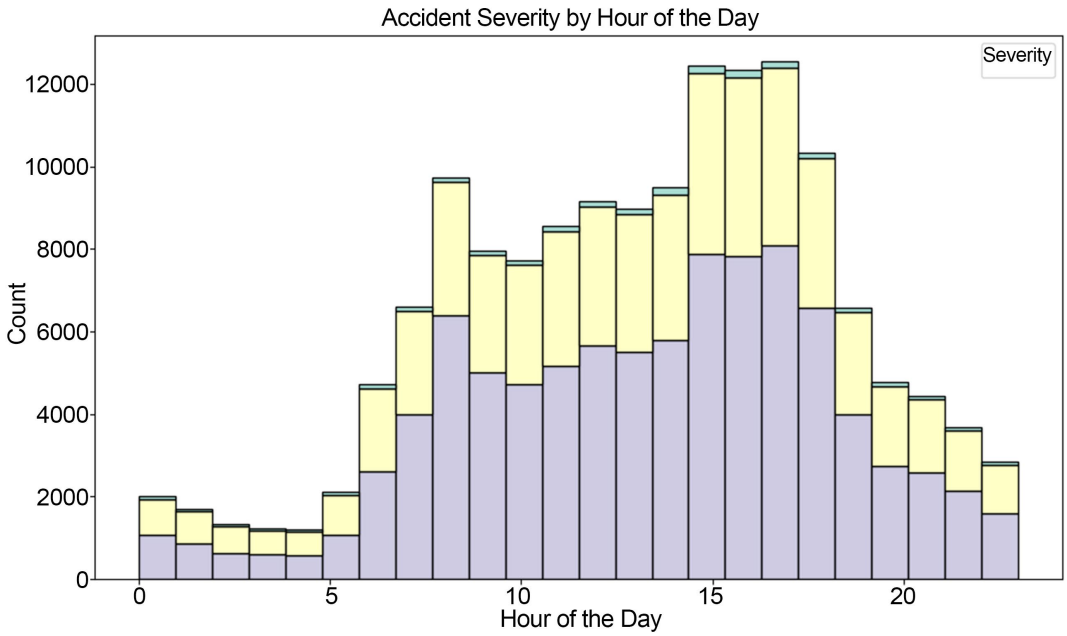


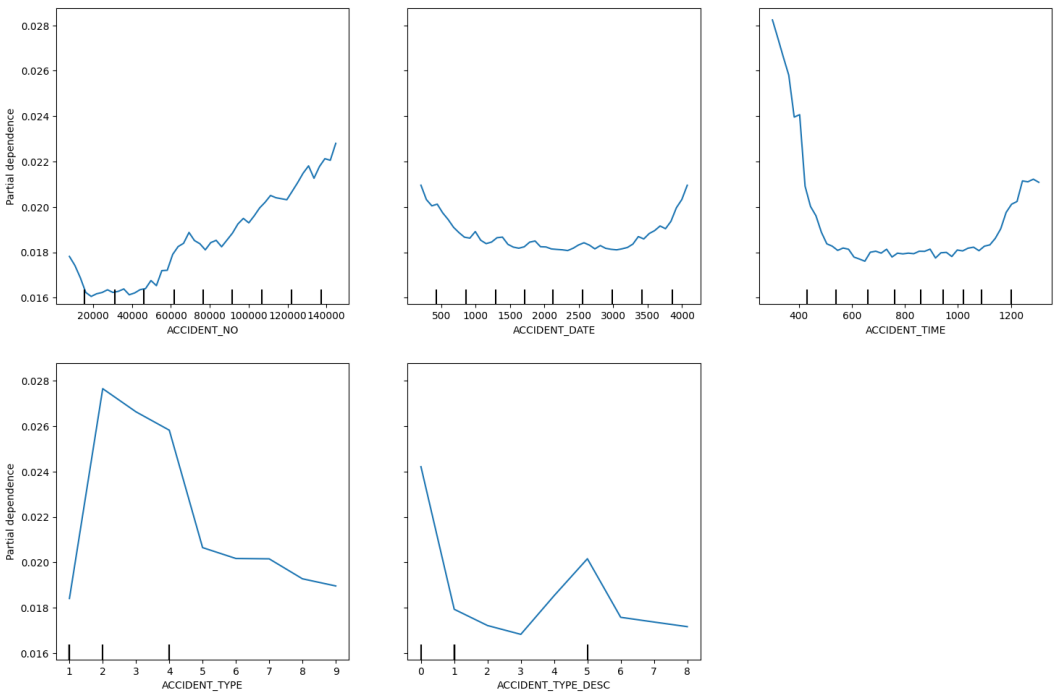
Figure 7. Distribution of accident severity by hour of the day.

The feature exploration in Figure 3 shows that severity Class 3 dominates as the most frequent type of accident. It is followed by Severity 2, regardless of the day. Also, weekdays (Monday to Friday) show a slightly higher number of accidents compared to weekends (Figure 6). This is likely due to increased road activity during workdays. Also, at a more granular level of interpretation, more accidents occur on Fridays. For hourly distribution (Figure 7), the frequency of accidents shows noticeable peaks during the late afternoon (around 3 PM to 6 PM). These

peaks coincide with regular rush hours. This implies higher road activity and congestion are influencing factors for accident. While severity Class 3 remains the most frequent type during all hours, severity Class 2 increases during peak traffic times. Additionally, early morning hours (midnight to 5 AM) show the least accident counts. This reflects reduced road activity during this particular hour. From Figure 2, DCA_CODE and SPEED_ZONE also rank high. The features show that specific accident types and speed limits are key determinants of severity. These features capture specific road rules and driver behaviours that directly affect accident outcomes. Features like ROAD_GEOMETRY_DESC and LIGHT_CONDITION are at the bottom of the ranking scale. Road geometry and lighting are static factors compared to the dynamic nature of time and traffic conditions, which change frequently and have a more immediate impact on accident likelihood. While road geometry and lighting conditions can exacerbate the severity of crashes under specific situations, they generally play a major role in accident causation compared to driver actions and environmental conditions during peak traffic periods.

4.2. Feature Importance: Partial Dependence Plots (PDP)

The **PartialDependenceDisplay.from_estimator** method from **sklearn.inspection** is used to generate PDPs for features in the Victoria dataset. The trained model, test data and target class (severity) were set as inputs to compute the partial dependence for each feature. The PDP results (Figure 8) show the influence of features like time, speed zone, accident type, and lighting conditions on accident severity.



(a)

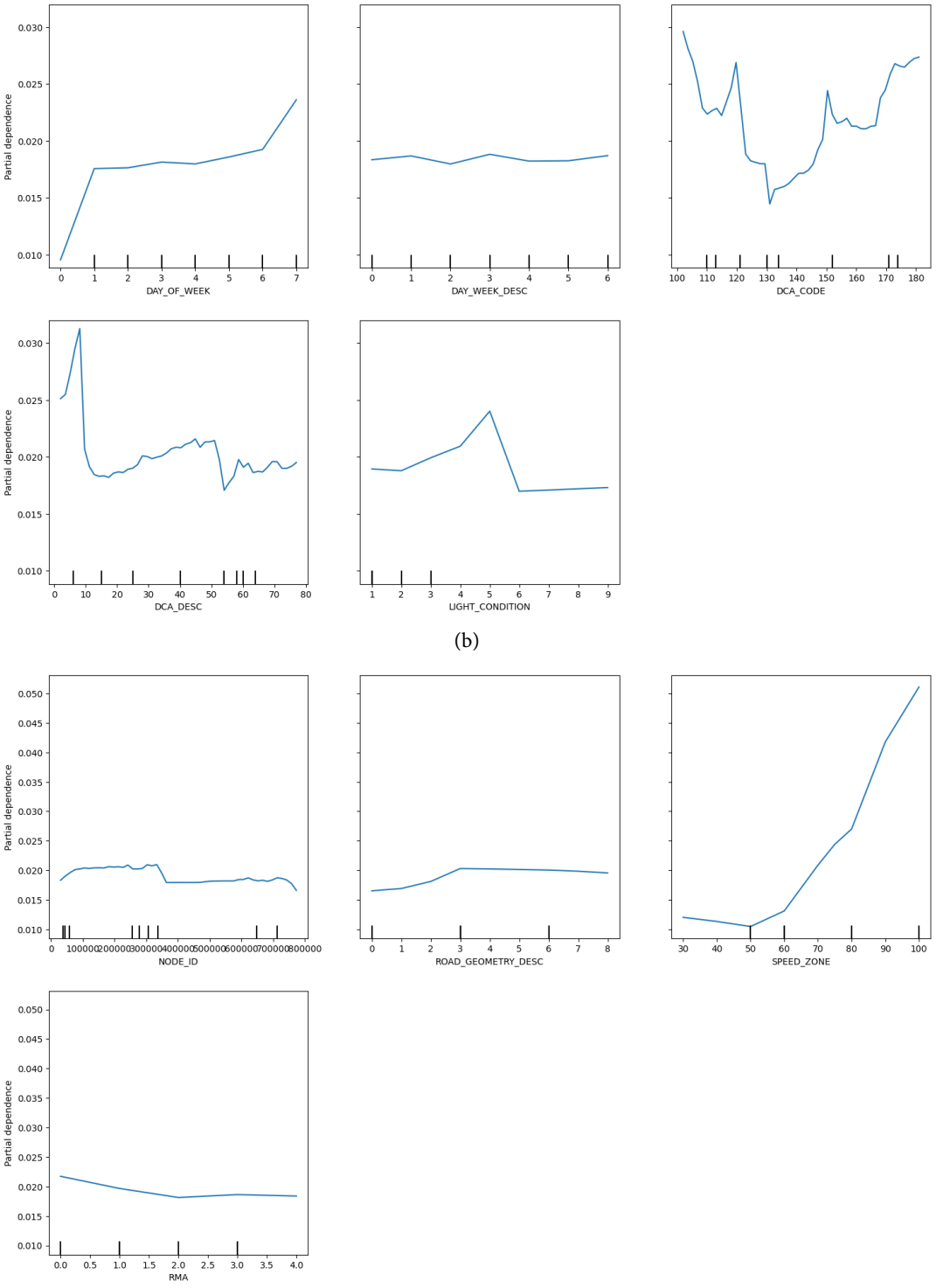


Figure 8. Partial dependence plots for features in Victoria dataset.

From **Figure 5**, ACCIDENT_DATE plot displays a U-shaped trend. This partial dependence is lower for mid-range dates and increases at both earlier and later periods. This shows temporal variations in accident severity influenced by sea-

sons. The plot for ACCIDENT_TIME shows a sharp decline in partial dependence which is observed during early hours. This decline is followed by a gradual rise later in the day. This trend aligns with peak traffic periods (rush hour) as deduced under feature importance analysis in Section 4.2. The DAY_OF_WEEK plot reveals partial dependence that increases steadily from Monday to Sunday. This implies that accidents later in the week have a higher likelihood of being severe. This may be due to increased fatigue as the week runs out. The DCA_CODE plot shows significant variation with certain DCA codes contributing more to accident severity. This might be linked to specific road conditions or accident scenarios. In the LIGHT_CONDITION visual, spike is observed for specific lighting conditions. This means that poor or changing lighting significantly impacts accident severity. There is a steep positive trend in the SPEED_ZONE plot. This implies that higher speed zones correlate with increased severity. This underscores the role of speed in influencing the severity of outcomes. There is poor variation in the plot of ROAD_GEOMETRY_DESC. This means that road geometry has limited influence on severity. RMA: The partial dependence of RMA slightly decreases. The plot shows a negative relationship with severity. This means improved road management procedures reduce accident and accident severity.

5. Machine Learning Models

By design and configuration, machine learning (ML) involves training algorithms on datasets to identify patterns, relationships, and trends, which are subsequently applied to new, unseen data [37] [38]. Unlike traditional statistical method, machine learning can handle large volume of datasets and features to capture simple and complex interactions [39]. As road accidents are caused by many factors which often do not exist in linear relationship, machine learning becomes applicable in understanding these complex interactions to make predictions. To explore the utility of machine learning in road accident domain, specific ML models such as the Logistic Regression Model, Random Forest (RF) Model, and XGBoost Model offer distinct advantages.

5.1. Logistic Regression

Logistic Regression is common statistically rooted technique in machine learning domain for binary and multi-class classification tasks [40]. Unlike linear regression, which predicts a continuous target variable, logistic regression estimates the probability of the class to which data instances belong by modelling the log-odds of the target variable as a linear function of the influencing features [41] [42]. The logistic function (sigmoid function) transforms the log-odds into probabilities that are between 0 and 1. This makes logistic regression suitable for classification tasks. One key strength of logistic regression is its simplicity and interpretability [43]. The coefficient in logistic regression is directly interpreted as the change in the log-odds of the target variable per unit increase in the influencing variable(s). For Victoria dataset, logistic regression is implemented in multinomial form as

target variable (severity) as more than two classes. The multinomial logistic utilizes SoftMax function to predict the probabilities of each class under severity.

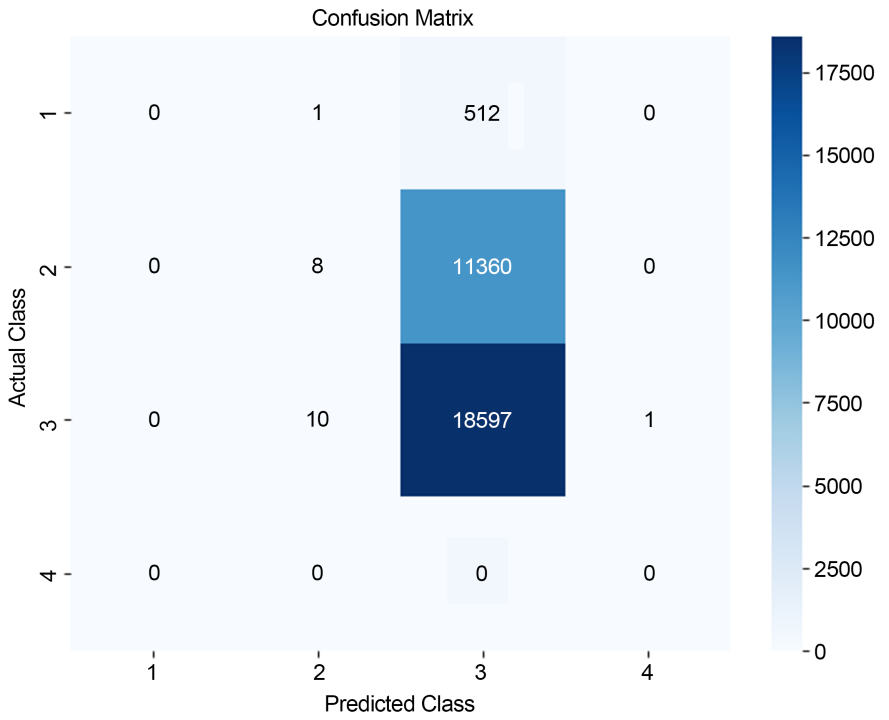


Figure 9. Confusion matrix for logistic regression.

The confusion matrix in Figure 9 shows the performance of the logistic regression model for predicting accident severity across four classes. Class 1 (severe) has no true positives, and all its actual cases are misclassified into other classes. This shows that logistic regression struggles significantly with the identification of Class 1. For Class 2, 11,360 out of its 11,368 actual occurrences were correctly classified. The result shows a high recall but limited precision. Class 3 dominates the confusion matrix, with 18,597 out of 18,608 instances. Logistic regression produces a strong performance for this class. Class 4 has zero predictions and zero correct classifications. Logistic regression model fails to properly recognize this class due to the class imbalance and its very small instances in the dataset. Overall, the logistic regression model performs best for Class 3, moderately for Class 2, and poorly for Classes 1 and 4.

5.2. Random Forest

Random Forest is a robust ensemble learning technique that is constructed from multiple decision trees [44] [45]. Each tree in the Random Forest is trained on a bootstrap sample. During the splitting process, a random subset of features is considered for the best split to minimize overfitting. The aggregation of outcomes from multiple trees strengthens the robustness of random forest models and improves the accuracy of the outcome [46]. The strength of Random Forest lies in

its ability to handle high-dimensional datasets and complex interactions between the influencing variables. For the Victoria Dataset, **RandomForestClassifier** was initialized with **n_estimators = 100** and **random_state = 42**. With this, the model was built with 100 decision trees, and the randomness was fixed to ensure reproducibility. The Random Forest model was trained on the training set (**X_train** and **y_train**) and predictions were made on the test set.

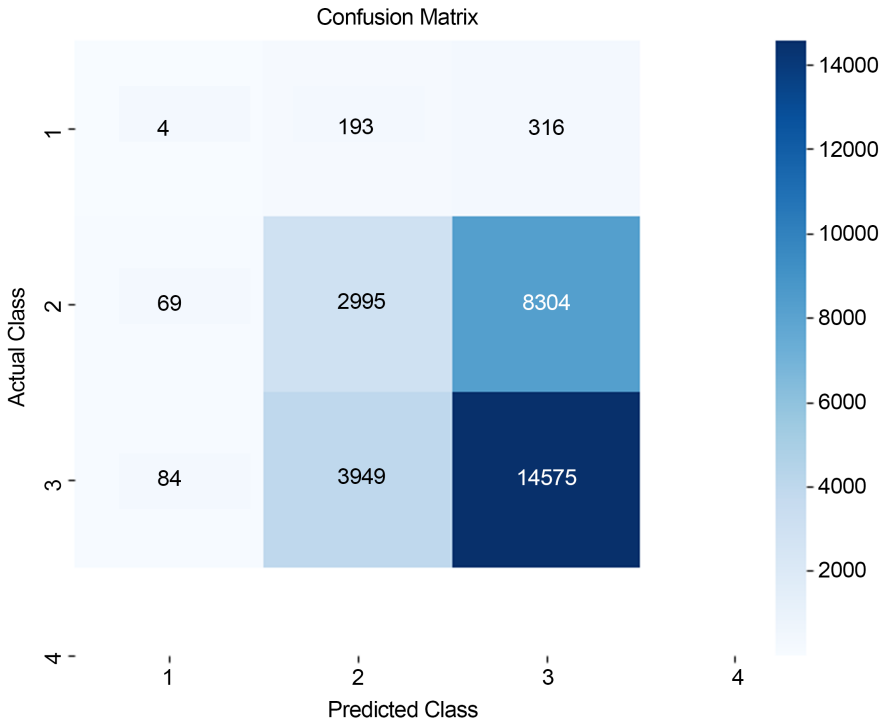


Figure 10. Confusion matrix for random forest.

From results presented in Figure 10, Class 1 shows poor performance with only 4 correct predictions out of 513 actual instances. This shows random forest struggles with minority classes. Class 2 shows moderate performance and random forest made 2995 correct predictions out of 11,368 actual instances. Class 3 dominates the performance results with the highest accuracy, as 14,575 out of 18,608 instances were correctly classified. Random forest model did not make any predictions for Class 4. This observation shows that random forest model could not be learned from the class due to its small number of counts.

5.3. Extreme Gradient Boosting (XGBoost)

XGBoost (Extreme Gradient Boosting) is a gradient boosting framework optimized for speed and performance. Like random forest, XGBoost builds an ensemble of decision trees sequentially. With the sequential boosting concept, each new tree corrects the errors of the previous ones [47] [48]. Thus, loss function is minimized using gradient descent while employing regularization techniques (L1 and L2 penalties) to prevent overfitting. XGBoost offers several benefits, and this in-

cludes scalability, efficiency, and the ability to handle imbalanced datasets through weighting [49] [50]. For the accident severity dataset, XGBoost's ability to model non-linear relationships and capture interactions between features makes it an ideal machine learning model. The XGBoost model was instantiated using the **XGBClassifier** class from the **xgboost** library. During initialization, **objective = 'multi:softmax'** was used to configure the model for multi-class classification and outputs the class with the highest probability. Also, the class in the target variable was specified dynamically with the use of syntax **num_class = len(np.unique(y)).** The XGBoost model was trained on the training set (**X_train and y_train**) and final predictions were made on the test set (**X_test**).

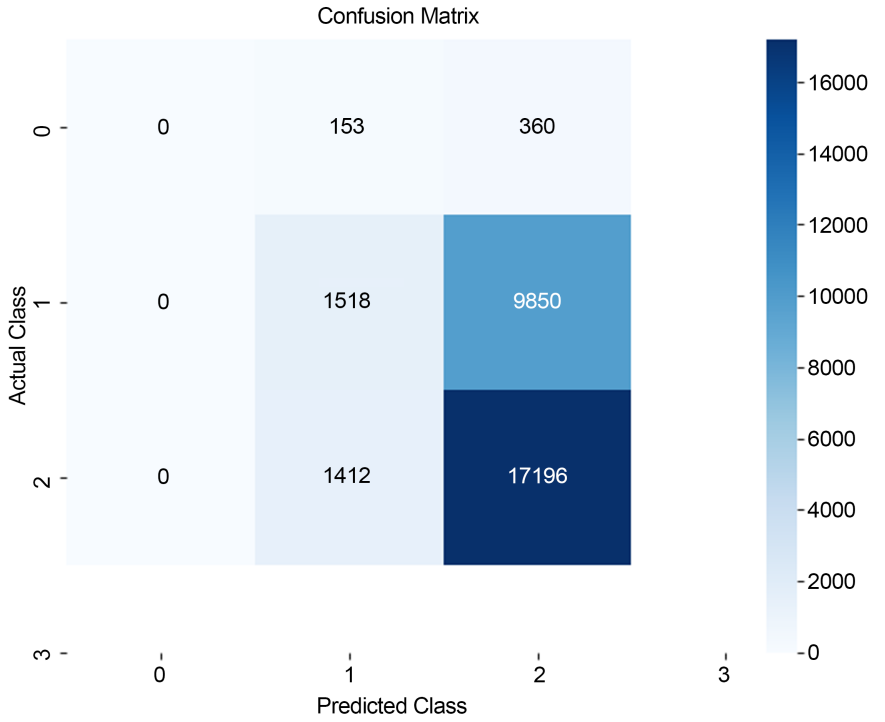


Figure 11. Confusion Matrix for XGBoost.

From the result of XGBoost model, Class 0, which is Class 1 in the Victoria dataset, shows no correct predictions. The complete failure of XGBoost to identify Class 1 indicates that the model finds it difficult to classify the class. Severity Class 2 shows moderate performance, with 1518 correct predictions out of 11,368 actual instances. Class 3 shows dominant performance with XGBoost with the highest accuracy making 17,196 correct predictions out of 18,608 actual instances (**Figure 11**). No predictions were made for Class 4. Overall, the outcome reveals that the XGBoost model heavily favours class 3 predictions while underperforming for Classes 1, 2 and 4.

6. Machine Learning Results: Comparison

All three machine learning models performed poorly with Class 1 (**Table 2**). The

data instance of Class 1 is significantly small, and all the models struggle to effectively predict instances in the class. Random Forest slightly outperformed Logistic Regression and XGBoost in terms of precision (3% against 0%). This difference is minimal and inconsequential. For Class 2, the models show better performance. XGBoost achieved the highest precision (49%) compared to Logistic Regression and Random Forest (both at 42%). However, the recall for XGBoost model is considerably lower (13%) than Random Forest (26%). This observation reflects the trade-off between precision and recall. Logistic Regression completely failed with handling recall for Class 2 (0%). For Class 3, Logistic regression achieved the highest recall (100%). With this result, Logistic regression outperforms both XGBoost (92%) and Random Forest (78%). All models achieved very close precision values for Class 3 (around 63%). The recall strength of Logistic regression indicates that the model captured more true positives for Class 3. Random Forest achieved a slightly lower F1-Score (70%) compared to XGBoost (75%). Both XGBoost and Random Forest are outperformed by Logistic Regression (76%). Looking at the accuracy, Logistic Regression and XGBoost both achieved 61%. This is slightly higher than Random Forest at 58%. This means that while Logistic Regression and XGBoost made more accurate predictions. Perhaps Random Forest's performance across individual classes was more balanced.

Table 2. Machine learning models results.

Metric	Logistic Regression (LR)	Random Forest (RF)	XGBoost
Precision—Class 1	0	0.03	0
Precision—Class 2	0.42	0.42	0.49
Precision—Class 3	0.61	0.63	0.63
Recall—Class 1	0	0.01	0
Recall—Class 2	0	0.26	0.13
Recall—Class 3	1	0.78	0.92
F1-Score—Class 1	0	0.01	0
F1-Score—Class 2	0	0.32	0.21
F1-Score—Class 3	0.76	0.7	0.75
Accuracy	0.61	0.58	0.61
Macro Avg	0.25	0.35	0.32
Weighted Avg	0.46	0.55	0.53

7. Results and Safety Implications

[51] highlights that road geometry plays a crucial role in influencing crash frequency. This is common when road segments are poorly designed, or they exist in hazardous conditions. Similarly, [52] demonstrates that poor lighting conditions, especially at night without proper illumination, are associated with severe injuries. Despite being static factors, road geometry and lighting are critical factors influencing accident outcomes under specific circumstances. Unlike [51] and

[52], feature analysis carried out in this research ranks ROAD_GEOMETRY_DESC and LIGHT_CONDITION lower. The discrepancy is due to differences in dataset locations as variations in road design standards, lighting infrastructure, and environmental factors significantly influence the importance of these features in different regions. Also, Ahmed *et al.* (2021) emphasize the predictive importance of temporal factors like ACCIDENT_TIME. In their research, they confirm that accidents occurring during peak hours are more likely to have severe outcomes. The ranking of this feature in this research aligns with [53] findings. The explicit implementation of feature analysis in the current research provides more detailed understanding of the relative importance of different features when predicting accident severity. By ranking features based on their predictive power, feature analysis helps prioritize features that are more relevant to accident likelihood and severity. Notably, it improves model interpretability, as shown in **Figures 5-8**.

The findings of this research reveal significant safety implications for stakeholders. The dominance of temporal factors like accident time and accident date highlights the need for targeted interventions during high-risk periods like late afternoons and Fridays. These periods coincide with rush hours, and it emphasizes the need for road safety strategies like increased traffic policing, better signal coordination, and public awareness campaigns to reduce congestion. Also, the importance of speed zones as a determinant of accident severity pinpoints the vital role of speed management in the reduction of fatalities and severe injuries. Higher speed zones are strongly linked to fatal outcomes due to increased impact forces. Based on the findings in this work, stricter enforcement of speed limits is suggested in areas prone to high-speed crashes. The significant influence of lighting conditions on accident severity implies that improving road lighting infrastructure will substantially reduce severe accidents. Similarly, the findings on road management areas (RMA) indicate that well-maintained roads with effective management systems contribute to lower severity. This reinforces the need for sustained infrastructure investment. Machine learning models (Logistic Regression, Random Forest and XGBoost) performed well in predicting accident severity for the majority of classes, and this shows their individual strengths in analysing complex road traffic datasets. However, their poor performance with minority classes, like fatal accidents (Class 1) and property damage only (Class 4) signals demand resolution with class balancing methods. Oversampling or synthetic data generation techniques will improve prediction for these key classes and ensure better resource allocation for high-risk areas. From a policy perspective, integration of predictive modelling into road safety strategies enables proactive measures rather than reactive measures. Authorities can use machine learning outputs to identify high-risk zones and implement localized interventions like stricter speed enforcement and improved traffic flow management.

8. Future Works

As the models implemented in this research struggle with minority classes, future

works should focus on addressing the challenges of class imbalance to improve the prediction accuracy for the minority classes. In this regard, advanced resampling techniques, like SMOTE-ENN or adaptive synthetic sampling, should be experimented with to handle data balancing. Also, external datasets like weather reports, traffic density, and driver behaviour should be incorporated to provide more comprehensive insights into the factors influencing road crash severity. The adoption of deep learning models like Convolutional Neural Networks (CNNs) or Long Short-Term Memory (LSTM) networks will further improve prediction accuracy by capturing complex temporal and spatial relationships in the road traffic dataset. Lastly, future research should aim to test the generalizability of the models by applying outcomes to road accident data from other regions. This will ensure scalability and broader applicability.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] WHO (2024) Road Traffic Injuries. Road Safety.

[2] Chen, S., Kuhn, M., Prettner, K. and Bloom, D.E. (2019) The Global Macroeconomic Burden of Road Injuries: Estimates and Projections for 166 Countries. *The Lancet Planetary Health*, **3**, e390-e398. https://doi.org/10.1016/s2542-5196(19)30170-6

[3] Sharma, B.R. (2008) Road Traffic Injuries: A Major Global Public Health Crisis. *Public Health*, **122**, 1399-1406. https://doi.org/10.1016/j.puhe.2008.06.009

[4] Akallouch, M., Fardousse, K., Bouhoute, A. and Berrada, I. (2023) Exploring the Risk Factors Influencing the Road Accident Severity: Prediction with Explanation. 2023 *International Wireless Communications and Mobile Computing* (*IWCMC*), Marrakesh, 19-23 June 2023, 763-768. https://doi.org/10.1109/iwcmc58020.2023.10182749

[5] Li, X., Lord, D., Zhang, Y. and Xie, Y. (2008) Predicting Motor Vehicle Crashes Using Support Vector Machine Models. *Accident Analysis & Prevention*, **40**, 1611-1618. https://doi.org/10.1016/j.aap.2008.04.010

[6] Wang, Y., Zhai, H., Cao, X. and Geng, X. (2023) Cause Analysis and Accident Classification of Road Traffic Accidents Based on Complex Networks. *Applied Sciences*, **13**, Article 12963. https://doi.org/10.3390/app132312963

[7] Tamascelli, N., Campari, A., Parhizkar, T. and Paltrinieri, N. (2024) Artificial Intelligence for Safety and Reliability: A Descriptive, Bibliometric and Interpretative Review on Machine Learning. *Journal of Loss Prevention in the Process Industries*, **90**, Article ID: 105343. https://doi.org/10.1016/j.jlp.2024.105343

[8] Ballamudi, V.K.R. (2019) Road Accident Analysis and Prediction Using Machine Learning Algorithmic Approaches. *Asian Journal of Humanity, Art and Literature*, **6**, 185-192. https://doi.org/10.18034/ajhal.v6i2.529

[9] Vanitha, R. and Swedha, M. (2023) Prediction of Road Accidents Using Machine Learning Algorithms. *Middle East Journal of Applied Science & Technology*, **6**, 64-75. https://doi.org/10.46431/mejast.2023.6208

[10] Ogungbire, A., Kalambay, P., Gajera, H. and Pulugurtha, S. S. (2023) Deep Learning, Machine Learning, or Statistical Models for Weather-Related Crash Severity Prediction. Mineta Transportation Institute. https://doi.org/10.31979/mti.2023.2320

[11] Kamali, R., Mazaheri, A. and Rahimi, A. (2024) Analysis of Crash Severity at Intersections and Roundabouts Using Ordered and Generalized Ordered Probit Models. *Proceedings of the 8th International Conference on Road and Rail Infrastructure* (*CETRA* 2024), Cavtat, 15-17 May 2024. https://doi.org/10.5592/co/cetra.2024.1664

[12] Mannering, F., Bhat, C.R., Shankar, V. and Abdel-Aty, M. (2020) Big Data, Traditional Data and the Tradeoffs between Prediction and Causality in Highway-Safety Analysis. *Analytic Methods in Accident Research*, **25**, Article ID: 100113. https://doi.org/10.1016/j.amar.2020.100113

[13] Roustaei, N. (2024) Application and Interpretation of Linear-Regression Analysis. *Medical Hypothesis Discovery and Innovation in Ophthalmology*, **13**, 151-159. https://doi.org/10.51329/mehdiophthal1506

[14] Johnston, C., McDonald, J. and Quist, K. (2019) A Generalized Ordered Probit Model. *Communications in Statistics—Theory and Methods*, **49**, 1712-1729. https://doi.org/10.1080/03610926.2019.1565780

[15] Ding, P., Imbens, G., Qu, Z. and Ye, Y. (2024) Computationally Efficient Estimation of Large Probit Models. arXiv: 2407.09371. https://doi.org/10.48550/arXiv.2407.09371

[16] Pourroostaei Ardakani, S., Liang, X., Mengistu, K.T., So, R.S., Wei, X., He, B., *et al.* (2023) Road Car Accident Prediction Using a Machine-Learning-Enabled Data Analysis. *Sustainability*, **15**, Article 5939. https://doi.org/10.3390/su15075939

[17] Lee, J., Yoon, T., Kwon, S. and Lee, J. (2019) Model Evaluation for Forecasting Traffic Accident Severity in Rainy Seasons Using Machine Learning Algorithms: Seoul City Study. *Applied Sciences*, **10**, Article 129. https://doi.org/10.3390/app10010129

[18] Salih, A.M., Raisi-Estabragh, Z., Galazzo, I.B., Radeva, P., Petersen, S.E., Lekadir, K., *et al.* (2024) A Perspective on Explainable Artificial Intelligence Methods: SHAP and Lime. *Advanced Intelligent Systems*, **7**, Article ID: 2400304. https://doi.org/10.1002/aisy.202400304

[19] Bhavyasree, N. and Dhanusree, N. (2024) Prediction of Road Accidents Using Machine Learning Algorithm. *International Journal of Advance Research and Innovative Ideas in Education*, **10**, 154-159.

[20] Bener, A., Yildirim, E., Özkan, T. and Lajunen, T. (2017) Driver Sleepiness, Fatigue, Careless Behaviour and Risk of Motor Vehicle Crash and Injury: Population Based Case and Control Study. *Journal of Traffic and Transportation Engineering* (*English Edition*), **4**, 496-502.

[21] Rais, W., Oulha, R., Rahal, D.D. and Lallam, M. (2024) Analysis of the Environmental Factors Contributing to Road Traffic Accidents Involving School-Aged Pedestrian Children in Urban Algerian Settings. *Studies in Engineering and Exact Sciences*, **5**, e9521. https://doi.org/10.54021/seesv5n2-378

[22] Bakutin, Y. (2024) Visibility and Safety of Vehicle Traffic in the Dark (Current Reality, Future Projections). *Visegrad Journal on Human Rights*, **3**, 13-19. https://doi.org/10.61345/1339-7915.2024.3.2

[23] Bozorg, S., Tetri, E., Kosonen, I. and Luttinen, T. (2018) The Effect of Dimmed Road Lighting and Car Headlights on Visibility in Varying Road Surface Conditions. *Leukos*, **14**, 259-273. https://doi.org/10.1080/15502724.2018.1452152

[24] Assi, K., Rahman, S.M., Mansoor, U. and Ratrout, N. (2020) Predicting Crash Injury Severity with Machine Learning Algorithm Synergized with Clustering Technique: A Promising Protocol. *International Journal of Environmental Research and Public Health*, **17**, Article 5497. https://doi.org/10.3390/ijerph17155497

[25] Wang, L. (2023) Safety Evaluation for Highway Geometric Design Based on Spatial

Path Properties. *Journal of Advanced Transportation*, **2023**, Article ID: 6685010. https://doi.org/10.1155/2023/6685010

[26] Burlacu, A. and Mihai, A. (2023) Applications of Differential Geometry of Curves in Roads Design. *Romanian Journal of Transport Infrastructure*, **12**, 1-13. https://doi.org/10.2478/rjti-2023-0010

[27] Islam, M. (2023) An Exploratory Analysis of the Effects of Speed Limits on Pedestrian Injury Severities in Vehicle-Pedestrian Crashes. *Journal of Transport & Health*, **28**, Article ID: 101561. https://doi.org/10.1016/j.jth.2022.101561

[28] Jung, S. and Qin, X. (2023) Identifying the Local Impacts of Speed-Related Factors on Tunnel Entrance Crash Severity. *Transportation Research Record*: *Journal of the Transportation Research Board*, **2677**, 730-742. https://doi.org/10.1177/03611981231167156

[29] Mohanta, B.K., Jena, D., Mohapatra, N., Ramasubbareddy, S. and Rawal, B.S. (2022) Machine Learning Based Accident Prediction in Secure IoT Enable Transportation System. *Journal of Intelligent & Fuzzy Systems*, **42**, 713-725. https://doi.org/10.3233/jifs-189743

[30] Ursu, E., Minnegalieva, A., Rawat, P., Chernigovskaya, M., Tacutu, R., Sandve, G.K., Robert, P.A. and Greiff, V. (2024) Training Data Composition Determines Machine Learning Generalization and Biological Rule Discovery. bioRxiv. https://doi.org/10.1101/2024.06.17.599333

[31] Mazurek, S. and Wielgosz, M. (2023) Assessing Dataset Quality through Decision Tree Characteristics in Autoencoder-Processed Spaces. arXiv: 2306.15392. https://doi.org/10.48550/arXiv.2306.15392

[32] Rout, S., Mallick, R. and Kumar Sahu, S. (2023) Exploring the Significance of Feature Analysis in AI/ML Modeling. 2023 *OITS International Conference on Information Technology* (*OCIT*), Raipur, 13-15 December 2023, 580-585. https://doi.org/10.1109/ocit59427.2023.10431396

[33] Little, C.O., Lina, D.H. and Allen, G.I. (2023) Fair Feature Importance Scores for Interpreting Tree-Based Methods and Surrogates. arXiv: 2310.04352. https://doi.org/10.48550/arXiv.2310.04352

[34] Doyen, S., Taylor, H., Nicholas, P., Crawford, L., Young, I. and Sughrue, M.E. (2021) Hollow-tree Super: A Directional and Scalable Approach for Feature Importance in Boosted Tree Models. *PLOS ONE*, **16**, e0258658. https://doi.org/10.1371/journal.pone.0258658

[35] Ewald, F.K., Bothmann, L., Wright, M.N., Bischl, B., Casalicchio, G. and König, G. (2024) A Guide to Feature Importance Methods for Scientific Inference. In: Longo, L., Lapuschkin, S. and Seifert, C., Eds., *Explainable Artificial Intelligence*, Springer, 440-464. https://doi.org/10.1007/978-3-031-63797-1_22

[36] Jain, N., Ghosh, S., Murthy, C.A. and Ghosh, A. (2023) A Relative Density-Based Bclustering Method for Identifying Non-Linear Feature Relations. *SSRN Journal*. https://doi.org/10.2139/ssrn.4607100

[37] Adeyemi, T.S. (2024) Defect Detection in Manufacturing: An Integrated Deep Learning Approach. *Journal of Computer and Communications*, **12**, 153-176. https://doi.org/10.4236/jcc.2024.1210011

[38] Janiesch, C., Zschech, P. and Heinrich, K. (2021) Machine Learning and Deep Learning. *Electronic Markets*, **31**, 685-695. https://doi.org/10.1007/s12525-021-00475-2

[39] Tufail, S., Riggs, H., Tariq, M. and Sarwat, A.I. (2023) Advancements and Challenges in Machine Learning: A Comprehensive Review of Models, Libraries, Applications, and Algorithms. *Electronics*, **12**, Article 1789.

https://doi.org/10.3390/electronics12081789

[40] Kravets, P., Pasichnyk, V. and Prodaniuk, M. (2024) Mathematical Model of Logistic Regression for Binary Classification. Part 1. Regression Models of Data Generalization. *Visnik Nacional'nogo universitetu "L'vivs'ka politehnika". Serià Ìnformacijni sistemi ta mereži*, **15**, 290-321. https://doi.org/10.23939/sisn2024.15.290

[41] Zaidi, A. and Al Luhayb, A.S.M. (2023) Two Statistical Approaches to Justify the Use of the Logistic Function in Binary Logistic Regression. *Mathematical Problems in Engineering*, **2023**, Article ID: 5525675. https://doi.org/10.1155/2023/5525675

[42] Sun, Y., Zhang, Z., Yang, Z. and Li, D. (2019) Application of Logistic Regression with Fixed Memory Step Gradient Descent Method in Multi-Class Classification Problem. 2019 6*th International Conference on Systems and Informatics* (*ICSAI*), Shanghai, 2-4 November 2019, 516-521. https://doi.org/10.1109/icsai48974.2019.9010220

[43] Kumar, N.A., Jangale, L., Sathe, V., Shelke, A. and Redij, T. (2024) Study of Supervised Logistic Regression Algorithm. *Alochana Journal*, **13**, 227-230.

[44] Ganiyu, A., Darvishi, I., Addo-Quaye, R., Yeboah-Ofori, A., Asare, B.T. and Oguntoyinbo, O. (2024) Classification Algorithms Using Ensemble Methods. 2024 11*th International Conference on Future Internet of Things and Cloud* (*FiCloud*), Vienna, 19-21 August 2024, 168-175. https://doi.org/10.1109/ficloud62933.2024.00033

[45] Salman, H.A., Kalakech, A. and Steiti, A. (2024) Random Forest Algorithm Overview. *Babylonian Journal of Machine Learning*, **2024**, 69-79. https://doi.org/10.58496/bjml/2024/007

[46] Thomas, N.S. and Kaliraj, S. (2024) An Improved and Optimized Random Forest Based Approach to Predict the Software Faults. *SN Computer Science*, **5**, Article No. 530. https://doi.org/10.1007/s42979-024-02764-x

[47] Natras, R., Soja, B. and Schmidt, M. (2022) Ensemble Machine Learning of Random Forest, AdaBoost and XGBoost for Vertical Total Electron Content Forecasting. *Remote Sensing*, **14**, Article 3547. https://doi.org/10.3390/rs14153547

[48] Sahin, E.K. (2020) Assessing the Predictive Capability of Ensemble Tree Methods for Landslide Susceptibility Mapping Using XGBoost, Gradient Boosting Machine, and Random Forest. *SN Applied Sciences*, **2**, Article No. 1308. https://doi.org/10.1007/s42452-020-3060-1

[49] Velarde, G., Sudhir, A., Deshmane, S., Deshmukh, A., Sharma, K. and Joshi, V. (2023) Evaluating XGBoost for Balanced and Imbalanced Data: Application to Fraud Detection. arXiv: 2303.15218. https://doi.org/10.48550/arXiv.2303.15218

[50] Khan, A.A., Chaudhari, O. and Chandra, R. (2024) A Review of Ensemble Learning and Data Augmentation Models for Class Imbalanced Problems: Combination, Implementation and Evaluation. *Expert Systems with Applications*, **244**, Article ID: 122778. https://doi.org/10.1016/j.eswa.2023.122778

[51] Dong, C., Xie, K., Sun, X., Lyu, M. and Yue, H. (2019) Roadway Traffic Crash Prediction Using a State-Space Model Based Support Vector Regression Approach. *PLOS ONE*, **14**, e0214866. https://doi.org/10.1371/journal.pone.0214866

[52] Pourroostaei Ardakani, S., Liang, X., Mengistu, K.T., So, R.S., Wei, X., He, B., *et al.* (2023) Road Car Accident Prediction Using a Machine-Learning-Enabled Data Analysis. *Sustainability*, **15**, Article 5939. https://doi.org/10.3390/su15075939

[53] Ahmed, S., Hossain, M.A., Bhuiyan, M.I.I. and Ray, S.K. (2021) A Comparative Study of Machine Learning Algorithms to Predict Road Accident Severity. 2021 20*th International Conference on Ubiquitous Computing and Communications* (*IUCC/CIT/DSCI/SmartCNS*), London, 20-22 December 2021, 390-397.