vehicles

MDPI

Article

Enhancing Traffic Accident Severity Prediction: Feature Identification Using Explainable AI

Jamal Alotaibi

Department of Computer Engineering, College of Computer, Qassim University, Buraydah 52571, Saudi Arabia; j.alotabi@qu.edu.sa

Abstract: The latest developments in Advanced Driver Assistance Systems (ADAS) have greatly enhanced the comfort and safety of drivers. These technologies can identify driver abnormalities like fatigue, inattention, and impairment, which are essential for averting collisions. One of the important aspects of this technology is automated traffic accident detection and prediction, which may help in saving precious human lives. This study aims to explore critical features related to traffic accident detection and prevention. A public US traffic accident dataset was used for the aforementioned task, where various machine learning (ML) models were applied to predict traffic accidents. These ML models included Random Forest, AdaBoost, KNN, and SVM. The models were compared for their accuracies, where Random Forest was found to be the best-performing model, providing the most accurate and reliable classification of accident-related data. Owing to the black box nature of ML models, this best-fit ML model was executed with explainable AI (XAI) methods such as LIME and permutation importance to understand its decision-making for the given classification task. The unique aspect of this study is the introduction of explainable artificial intelligence which enables us to have human-interpretable awareness of how ML models operate. It provides information about the inner workings of the model and directs the improvement of feature engineering for traffic accident detection, which is more accurate and dependable. The analysis identified critical features, including sources, descriptions of weather conditions, time of day (weather timestamp, start time, end time), distance, crossing, and traffic signals, as significant predictors of the probability of an accident occurring. Future ADAS technology development is anticipated to be greatly impacted by the study's conclusions. A model can be adjusted for different driving scenarios by identifying the most important features and comprehending their dynamics to make sure that ADAS systems are precise, reliable, and suitable for real-world circumstances.

Keywords: traffic accident severity prediction; explainable AI (XAI); machine learning; road safety; feature importance

Received: 3 March 2025
Revised: 22 April 2025
Accepted: 23 April 2025
Published: 28 April 2025

Citation: Alotaibi, J. Enhancing Traffic Accident Severity Prediction: Feature Identification Using Explainable AI. *Vehicles* **2025**, *7*, 38. https://doi.org/10.3390/vehicles7020038

1. Introduction

All around the world, more than a million people lose their lives in road accidents annually. The World Health Organization released research stating that road accidents cost the lives of 1.24 million people and 2.4 million injuries each year, the majority of victims being among the age group of 15–29 years, and more men are killed or injured (91%) compared to women worldwide. With the failure to handle and prevent accidents, the annual death toll from road accidents will likely reach approximately 1.9 million [1]. Since 2004, the World Health Organization has renamed World Health Day to Road Safety Day. Road accident incidents are higher than other fatal diseases, such as malaria, HIV/AIDS, and

tuberculosis. According to a United Nations report in 2009, road accident fatalities increased from 1.3 million between 2010 and 2020 to more than 1.9 million. Therefore, the Commission for Global Road Safety has suggested that necessary measures have to be taken to prevent increased road accidents by reducing their frequency. It can be estimated that over 5 million deaths and 50 million serious injuries globally, nationally, and regionally can be reduced by implementing proper safety measures in 2020 [2].

Modern cars are equipped with advanced driver-assistance systems, ADAS, which are sophisticated blends of various technologies aimed at improving vehicle comfort, safety, and efficiency. ADAS enhances driving comfort with fewer exposures to dangerous situations by taking over the actual driving tasks and issuing warnings when the situation is not safe [3]. It integrates a range of sensors with AI and advanced control systems for perception, decision-making, and control of movement. These technologies are broadly classified into these five categories: parking-assist systems, driving-control aid systems, collision-intervention systems, collision-warning systems, and other driver-assist systems [4]. Furthermore, besides machine learning-based techniques, spatial traffic analysis with underlying accident maps provides another perspective in existing road safety literature to effectively prevent road accidents. Tools such as road incident heatmaps and GIS-based crash visualization are useful for determining accident-prone areas and examining risk patterns unique to a certain site of accidents. In this regard, research work [5] examines how geospatial data can assist in evaluating traffic safety, especially for vulnerable road users. Future ADAS deployments may benefit from improved situational awareness and intervention tactics through the integration of such geographical knowledge with prediction algorithms.

Driver anomaly detection is one of the most crucial tasks in the ADAS, as it involves identifying deviations in the normal driving pattern, which may imply either impaired driving or a greater risk of having an accident. It aims to identify driving anomalies like drowsiness, distraction, and impairment to ensure the safe operation of the vehicle. According to a report, distracted driving is now one of the major causes of collisions [6]. This resulted in a dire need to automate the process of detecting and preventing driver anomalies so that loss of human life can be prevented. Enhanced use of ADAS has led to the availability of several public datasets related to traffic accident detection and prediction. The rise of such vast databases has resulted in an ever-increasing need for machine learning (ML) models to classify and analyze given datasets. With its ability to learn from enormous volumes of historical data, machine learning (ML) models offer a potent tool for accident prediction. These algorithms can find trends and risk variables that human observers might not notice right away by examining datasets that include comprehensive information on previous accidents. Predictive models can therefore help anticipate high-risk situations, allowing ADAS systems to take preventative measures. However, due to the black-box nature of ML algorithms, it is difficult to analyze how the model has reached a certain decision. This limitation may undermine the use of ML, as sometimes it becomes important to understand which features of the dataset are important for a model for a certain classification task. Limited work has been performed in the recent past to understand the explainability of ML models for traffic accident prediction.

This study uses a publicly available US accident dataset to investigate how well machine learning algorithms predict traffic accidents. To identify the model that offers the best accuracy in predicting traffic accidents, we specifically assess and compare several state-of-the-art classification algorithms: Random Forest, AdaBoost, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). The algorithm that performs the best among these models is Random Forest, which shows exceptional accuracy and dependability in forecasting incidents connected to accidents. It is worth mentioning that the dataset

does not include direct ADAS engagement indicators (e.g., ADAS ON/OFF), and, thus, the contribution is indirect: we provide evidence-backed feature importance analysis that could guide future sensor prioritization, alert systems, and driver state monitoring logic in ADAS. The objective is to employ data-driven research to uncover key factors determining accident severity, which can guide the design and prioritization of ADAS components, rather than to assess ADAS performance in and of itself.

The unique aspect of this research is to enhance the decision-making processes of the best-chosen ML model (Random Forest) regarding traffic anomaly detection using LIME and Explainable AI (XAI) technologies. These techniques provide insights into how and why traffic anomaly detection algorithms make decisions regarding their particular operation. This enables us to refine and enhance the performance of predictive models by increasing interpretability. The proposed work aims to ensure that the systems installed in cars are precise and understandable; these insights can also help advance ADAS. The proposed work can assist in the facilitation of ADAS feature design, such as prioritizing real-time risk factors (e.g., time of day, distance, location). The paper has been divided into multiple sections for better organization and understanding. Section 2 highlights the literature review, and Section 3 outlines the overall methodology adopted for the proposed work. Section 4 discusses the overall analysis of the results obtained after experimentation. Section 5 concludes the findings of the proposed work, and Section 6 discusses the future implications of the proposed work.

2. Literature Review

ADAS consists of automated emergency braking, adaptive cruise control, self-parking, lane-keeping assistance, and advanced variable emergency braking that will enhance safety and improve the efficiency of driving. The technology minimizes crashes and improves traffic efficiency due to the reduction in human error through real-time traffic information [7]. The advanced technology of ADAS is an automation that will eradicate human errors and thereby improve safety on the roads. The 2017 NHTSA report highlights how important automotive accessibility technologies can be for the reduction in collision numbers and the enhancement of traffic safety [8].

As automation takes over more driving jobs, it could greatly reduce accidents by eliminating of human error. The Center for Automotive Research (CAR) frequently releases roadmaps and white papers to educate the public on current and future automotive developments that include improved driver assistance, automation, and connectivity. These publications are the result of industry roundtables with important stakeholders. The study also gives a brief summary of the present situation and potential future course of the automotive industry by highlighting important difficulties in the development and implementation of advanced driver assistance systems (ADAS), autonomous vehicles (AV), and vehicle connectivity technologies [9]. When dealing with Automated Driving Systems to monitor road objects, it can be difficult to understand the AI models in Advanced Driver Assistance Systems. Road scene semantic segmentation faces several challenges like changing illumination, weather, shadows, false alarms, occlusions, detection errors, and the need for real-time data processing [10]. These challenges are made worse by the fact that the data need to be processed in real-time to make sure that the ADAS interventions are timely and accurate. With the wider adoption of ADAS and AV technologies, vast amounts of data have been generated from traffic-related activities. These data are being used to develop better models for predicting traffic accidents, making road safety a significant area of research. Initially, the prediction of accidents used traditional statistical methods; however, the advent of machine learning research has made way for newer predictive models.

Several studies in the recent past have examined the potential of machine learning (ML) models to predict and prevent traffic accidents, making this a crucial topic of research in the field of road safety. To predict traffic accidents [11], the researchers used decision tree methods and support vector machines (SVM). They discovered that both models were good at forecasting the chance of an accident, but decision trees were especially good at figuring out risk factors. Other researchers have explored the use of ensemble methods, such as Extreme Gradient Boosting (XGBoost), and other ensemble model learning methods are combined with deep learning models like DenseNet201, InceptionV3, and ResNet50 to detect driver anomalies in cars [12] with in-car cameras and Advanced Driver Assistance Systems. According to [6], the ensemble technique shows strong potential to achieve better road safety through its ability to prompt ADAS interventions to avert accidents and save lives [13,14]. While such methods may produce promising results, these machine learning models suffer from the drawback of not being interpretable. The increasing complexity of models like Random Forest, adaptive boosting (AdaBoost), and deep learning networks further complicates any understanding of how and why certain predictions are made [15–17]. Such a lack of transparency feeds into the challenge of trust concerning critical systems related to safety, especially ADAS, where it is imperative to know the justification for making certain decisions so as to facilitate corrections and build trust in technology.

Recent developments in deep learning and machine learning, especially with regard to ensemble methods and hybrid neural network designs, have greatly enhanced the ability to forecast the severity of accidents. The accuracy of the models was further improved by methods including features that may have a significant impact on traffic accidents. One noteworthy work used was a dataset of 13,546 motor vehicle crashes in Riyadh, Saudi Arabia, and the eXtreme Gradient Boosting (XGBoost) algorithm was applied to predict the severity of injuries. The model outperformed conventional techniques like logistic regression, random forest, and decision trees, emphasizing important characteristics like road conditions, lighting, and collision type as significant determinants of severe outcomes [18]. Similarly, three years of crash data from Mersin, Turkey, were used to create a hybrid deep learning model that combines Deep Neural Networks (DNN) and Random Forest (RF) [19]. Unconventional data sources like social media have also been used to investigate ensemble deep learning models. To improve forecast accuracy and enable real-time accident analysis, one such study combined several deep learning architectures to extract data from social media information about traffic [20]. Using actual data from Los Angeles County, a hybrid Convolutional Neural Network Long Short-Term Memory (CNN-LSTM) model was also put forth to simulate the effects of traffic accidents. This model is ideally suited for post-accident reaction prediction since it successfully incorporates spatial–temporal interdependence [21].

Feature engineering has become one of the tools to enhance the interpretability of ML models. It focuses on interpreting and analyzing raw data by creating and refining several meaningful variables. It works by highlighting the prime features that affect the outcome significantly. In this way, model complexity is lowered, and it becomes easier to analyze how the input affects the model's output. Predicting traffic accidents depends on a number of variables, such as behavioral, environmental, and infrastructure data. To find the most important factors in accident prediction, several feature selection approaches have been applied extensively. Refs. [22,23] assert that variables like weather, time of day, and kind of road are repeatedly found to have a major impact on accident rates. To find out how each feature contributes to the model's performance, feature engineering [24–26] and selection techniques like correlation analysis and permutation importance are frequently employed. For instance, ref. [27] identified key accident predictors in a publicly available dataset of

traffic events in the US using feature selection and permutation importance techniques. However, one of the major drawbacks of manual feature engineering is that it can be biased by the designer's assumptions, potentially missing out on important patterns in the data. It also relies heavily on domain expertise, which might lead to overlooking more complex relationships that automated methods can uncover.

To address the above challenges, Explainable Artificial Intelligence (XAI) has risen to the occasion. XAI's particular methods offer a more systematic and data-driven approach to feature extraction compared to manual feature engineering. Unlike manual methods, which rely on domain knowledge and heuristic assumptions, XAI provides automated, transparent insights into how each feature contributes to model predictions. This allows for more accurate, consistent, and objective feature importance evaluation, helping uncover hidden relationships in the data that may be overlooked in traditional feature engineering. Moreover, it addresses the issues of interpretability in complex machine learning models. XAI aims to render machine learning models more transparent and understandable, most especially in high-stakes applications such as autonomous driving, healthcare, and finance. SHAP (Shapley Additive exPlanations), one of the most popular XAI approaches, offers a consistent framework for enhancing the interpretability of the results of any machine learning model. It is based on cooperative game theory. SHAP values give each feature a "Shapley value" that indicates how much it contributes to a model's prediction. This enables both global interpretability, which provides insights into the behavior of the model as a whole, and local interpretability, which illustrates how certain properties affect individual predictions. Because of its reliability, consistency, and capacity to operate with intricate, black-box models, SHAP has emerged as a cutting-edge technique.

The creation of straightforward, interpretable models that roughly mimic the behavior of intricate, black-box models locally around a particular prediction is the goal of LIME (Local Interpretable Model-agnostic Explanations), another noteworthy XAI technique. By choosing data points close to the instance being predicted and training a local, explainable model, LIME creates interpretable surrogate models. The method is quite versatile and usable in a variety of fields because it is independent of the underlying machine learning model. LIME and SHAP offer important insights into how various factors, such as weather or road type, affect accident risk forecasts in the context of traffic accident prediction [28,29]. These cutting-edge XAI methods are becoming more widely recognized as crucial resources for guaranteeing the moral and secure application of machine learning models, in addition to enhancing model transparency. Additional explainable AI techniques include Integrated Gradients, Counterfactual Explanations, Partial Dependence Plots (PDPs), and Permutation Feature Importance. Methods like Integrated Gradients and PDPs allow for to visualizing what features contributed to the prediction, and Counterfactual Explanations give information on what would have to change in features for a different predicted outcome. Permutation importance determines feature importance by evaluating the effects of randomizing features on model performance. The kind and quantity of the dataset, model complexity, computing cost, interpretability requirements, and the real-time nature of the application are some of the variables that influence the choice of an acceptable XAI approach [9,30,31]. These factors aid in choosing the most practical and successful explanation strategy.

The reviewed literature highlights the increasing role of machine learning and explainable AI in traffic accident prediction. While traditional statistical methods have been used historically, the recent advances leverage ensemble methods and hybrid deep learning architectures for improved accuracy. However, many of these models lack interpretability, especially when applied to critical safety systems like ADAS. Few studies integrate XAI techniques to explain predictions and guide sensor prioritization. This gap forms the

foundation of the proposed study, which combines robust ML modeling with transparent interpretation to contribute toward safer, smarter vehicle systems.

3. Methodology

In this section, we describe the approach taken in this current work to identify the likelihood of traffic accident severity using machine learning methods and explainable AI.

3.1. Dataset Description

The publicly accessible US dataset that is being used in this study [27] consists of over 1.5 million traffic accident reports from 49 US states, which have been steadily gathered since February 2016. The data come from a variety of sources, such as APIs that compile reports from law enforcement, transportation authorities, traffic cameras, and road sensors. Every record has comprehensive data, including timestamps, geolocation, weather, accident severity (ranked from 1 to 4), and surrounding points of interest. Category 1 corresponds to minor accident severity, 2 to moderate accident severity, 3 to major accident severity, and 4 to severe accident severity. These categories are annotated by experts who were involved in the collection of the dataset.

With over 1.5 million accident records, the dataset is a sizable and extensive analytical resource. It has 49 features that cover a lot of information, including the location, time, date, weather, accident severity, and surrounding areas of interest. These characteristics offer a thorough understanding of the circumstances behind each accident. Table 1 depicts the distribution of various important features across the given dataset.

Table 1. Distribution and Variability of Key Features within the Dataset.

Attribute	Count	Mean	Std Dev	Min	25%	50% (Median)	75%	Max
Severity	2,845,342	2.14	0.48	1	2	2	2	4
Distance (mi)	2,845,342	0.70	1.56	0.00	0.05	0.24	0.76	155.19
Temperature (°F)	2,776,068	61.79	18.62	−89.00	50.00	64.00	76.00	196.00
Wind Chill (°F)	2,375,699	59.66	21.16	−89.00	46.00	63.00	76.00	196.00
Humidity (%)	2,772,250	64.37	22.87	1.00	48.00	67.00	83.00	100.00
Pressure (in)	2,786,142	29.47	1.05	0.00	29.31	29.82	30.01	58.90
Visibility (mi)	2,774,796	9.10	2.72	0.00	7.00	10.00	10.00	140.00
Wind Speed (mph)	2,687,398	7.40	5.53	0.00	3.50	7.00	10.00	1087.00
Precipitation (in)	2,295,884	0.007	0.093	0.00	0.00	0.00	0.00	24.00

These statistics provide insight into the distribution and variability of key features within the dataset, which are crucial for modeling and analysis purposes.

3.2. Data Pre-Processing

The following pre-processing steps were performed for the given dataset:

- Data Cleaning: The dataset initially contained missing values. Rows with more than 40 missing values were removed, while the remaining missing values were handled using forward and backward filling methods. Specifically, the features Wind_Chill (F), Wind_Speed (mph), and Precipitation (in) were retained without imputation, as they contained non-trivial missing values.
- Feature Selection: Two columns (End_Lat and End_Lng) were dropped from the dataset as they were irrelevant to the severity prediction task. Additionally, features

such as Start_Lat, Start_Lng, and other spatial or categorical features were retained as they may provide important context in predicting accident severity.

- Categorical Encoding: Several categorical columns were encoded using Label Encoding. This is necessary for machine learning models like Random Forest, which require numerical inputs. All categorical columns, including Weather Condition, Road Condition, and Visibility (mi), were encoded into numeric values using Label Encoder from sklearn.preprocessing.
- Data Split: The dataset was split into training and test sets using a 70–30 split (train_test_split), where 30% of the data was used for model evaluation, and the remaining 70% was used for training. Moreover, K-fold validation with 5 folds was applied before reporting the results in the proposed work to ensure reliability and accuracy.

3.3. Application of Various Machine Learning Models

In this study, we applied various machine learning models—Random Forest, Adaboost, SVM, and KNN—to our dataset, comparing their performance in terms of Precision, Recall, and F1-Score across four severity levels. Random Forest emerged as the best-performing model, achieving high precision and recall values, especially in classifying Severity Levels 2 and 3, with F1-scores of 0.97 and 0.85, respectively. Adaboost performed similarly but with slightly lower values across all classes. SVM and KNN showed relatively lower metrics, especially for Severity Levels 1 and 4, with KNN showing the lowest performance overall. This assures us that Random Forest is the best algorithm to use when dealing with our big dataset because of its solid ensemble learning method. Table 2 depicts the evaluation results that were obtained after experimentation.

Table 2. Evaluation Results After Applying Various Machine Learning Algorithms.

Model	Severity	Precision	Recall	F1-Score
Random Forest	1	0.86	0.76	0.81
	2	0.96	0.98	0.97
	3	0.87	0.84	0.85
	4	0.65	0.43	0.52
Adaboost	1	0.83	0.73	0.78
	2	0.94	0.96	0.95
	3	0.84	0.82	0.83
	4	0.61	0.40	0.48
SVM	1	0.81	0.70	0.75
	2	0.93	0.94	0.93
	3	0.80	0.79	0.82
	4	0.37	0.45	0.48
KNN	1	0.78	0.68	0.73
	2	0.90	0.92	0.91
	3	0.79	0.76	0.77
	4	0.55	0.33	0.41

3.4. Model Selection

Random Forest Classifier was selected for predicting traffic accident severity due to its capability to deal with big data and identify correlations between features. Random Forest is a kind of bagging technique in which many decision trees are created and all their results are combined to make decisions less sensitive to individual trees and thus more accurate. First, each tree in the forest is learned on a different random sample of the data, and each node in the tree uses a random feature subset, which improves the model applied to large and high-dimensional data.

We also tried to use a Decision Tree Classifier as another method of data representation to judge the influence of particular characteristics on the forecasts. However, due to random selection and data aggregation across various trees, Random Forest was selected as the best option.

3.5. Hyperparameter Optimization

A Grid Search method was used to methodically investigate a variety of hyperparameter combinations to maximize the Random Forest classifier's performance. After extensive experimentation, the following hyperparameters were identified as optimal parameters for the selected Random Forest model:

min_samples_leaf = 0.1, min_samples_split = 2, boostrap = True, min_leaf_nodes = 10, n_estimators = 100, max_depth = 16. The excellent classification performance and generalization capabilities led to the selection of this fine-tuned set of parameters, which enhanced the accuracy and resilience of the model.

The Random Forest classifier's learning curve in Figure 1 shows how well the model performs with different training set sizes. To ensure the robust evaluation of the model, we used 5-Fold cross-validation (cv = 5). This method divides the dataset into five equal-sized subsets at random. The model is trained on four subsets and validated on the fifth, repeating the process five times so that each subset is used for validation exactly once. When compared to a single train–test split, this method provides a more reliable and objective evaluation of the model's performance. Training datasets of varying sizes of 10%, 50%, and 100% as depicted in Table 3 were used to construct the learning curve. For each section of data, 5-fold cross-validation was applied, and the training and validation accuracies were averaged across the five folds to assess how the model's performance evolves with increasing data.

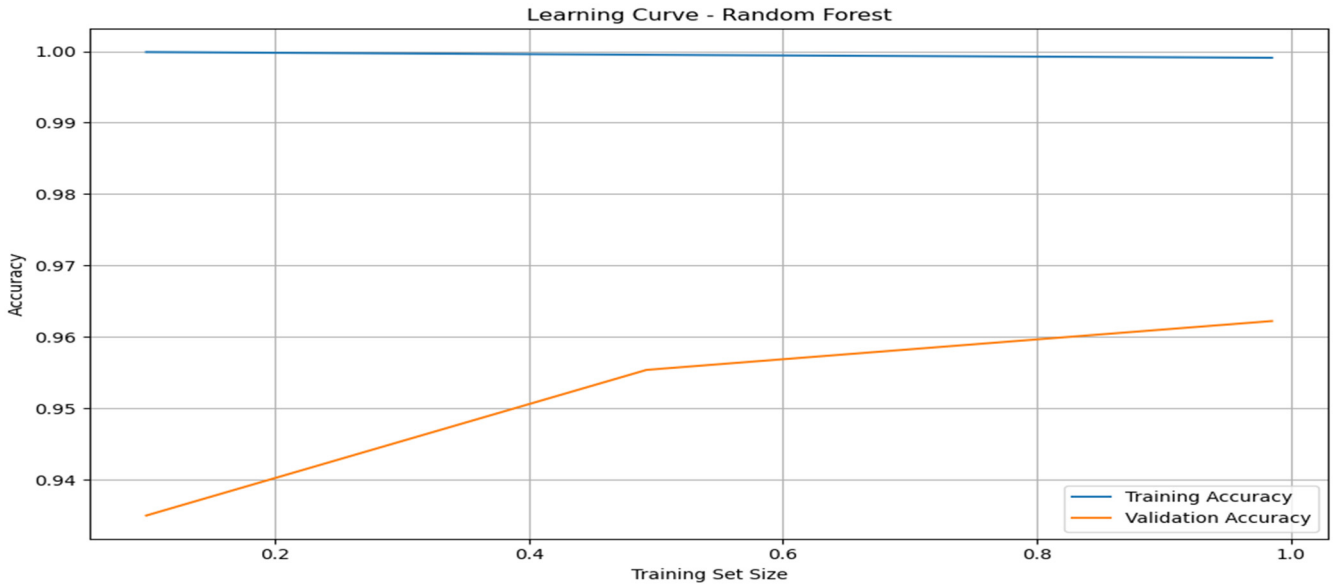

Figure 1. Learning Curve for the Selected Random Forest Model.

Table 3. Evaluation Results for K-Fold Validation Across Training Sets of Varying Sizes.

Training Set Size	Mean Training Accuracy	Mean Validation Accuracy
10%	1.00	0.931
50%	1.00	0.944
100%	1.00	0.962

After applying K-fold validation, the obtained pattern suggests mild overfitting, evident from the small gap between the training and the validation accuracy, yet the model maintains strong generalization performance. The use of 5-fold cross-validation ensures that the reported results are statistically robust and less sensitive to data splits, enhancing the reliability of the evaluation.

3.6. Application of SMOTE

The dataset used in the experimentation depicts a clear imbalance in the distribution of accident severity. Severity Level 2 was the most common class of mild accidents, accounting for about 81.7% of the dataset. The most serious accidents, Severity Level 4, accounted for only 2.7% of the dataset, but Severity Level 3 accounted for roughly 17.2%. Severity Level 1 was the least represented, accounting for only 0.9% of all cases. Machine learning models, which favor the dominant class during training, are challenged by this high skew in class distribution. To overcome this challenge, we have used SMOTE to balance the classes distribution within the dataset and analyze the performance of the Random Forest before and after the application of SMOTE.

3.7. Explainable AI Techniques

To increase the level of interpretability, the Random Forest was further supplemented with the Explainable AI (XAI) methods [2,32]. Two primary methods were utilized: the permutation importance and the Local Interpretable Model-agnostic Explanations (LIME) [33]. To determine how each feature affected the model prediction, permutation importance was used. The importance of each of the features was obtained by systematically permuting the data in each feature and measuring the deterioration in the performance of the model as a result of such permuting. High-influence and critical feature sets with a marked decline in accuracy during the model's decision-making process were pinpointed, proving that these features shape the model's ability to accurately predict specific accident severity levels. LIME was then applied to explain predictions made by the model and localize these explanations. This method constructs substitute linear models for the given cases and shows the impact of features on the forecasted value. The use of LIME allowed us to better understand the role of features such as time of day, weather, and distance in the identified samples and see exactly how the model arrived at its conclusion.

These techniques were performed using Python (version 3.10) with corresponding libraries of permutation importance from scikit-learn and the LIME. Feature ranking and its visualization [34] as well as representations of variations in impact showed the broad tendencies and specifics of concrete instances, which allowed for a deeper penetration into the model actions and, thus, added more transparency to the predictive system. Figure 2 depicts the overall methodology that was adapted for the proposed research work.

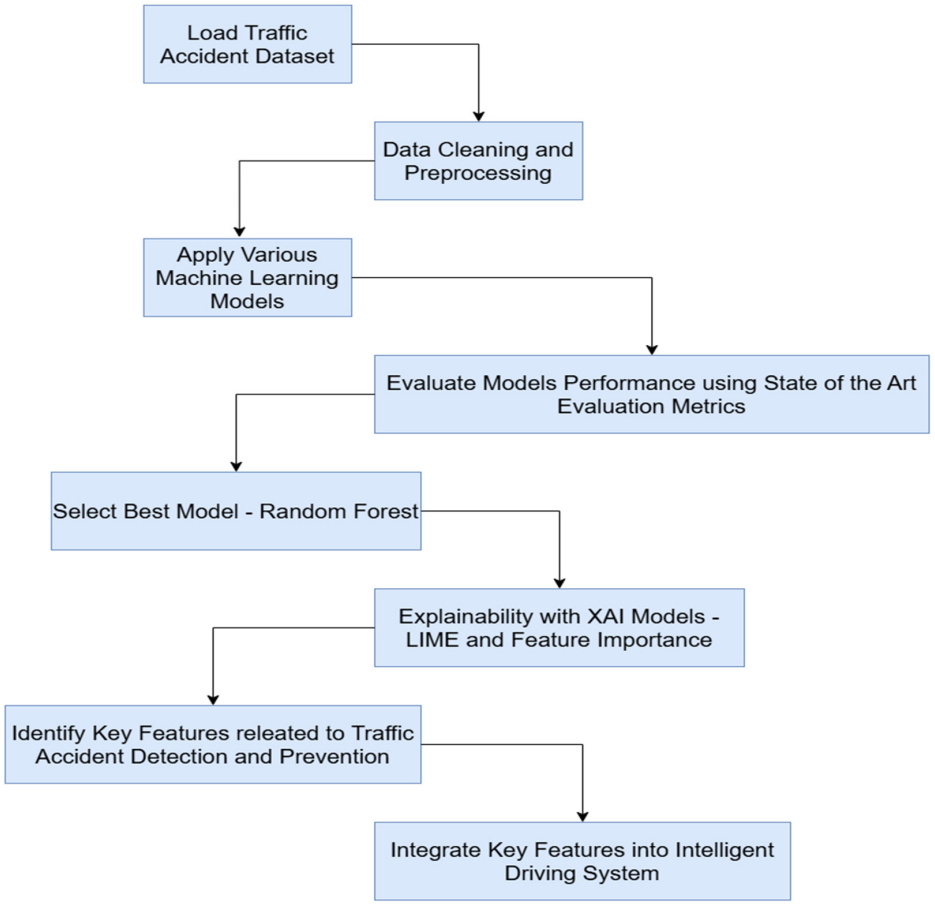

Figure 2. Proposed Methodology.

4. Results and Discussion

After training the Random Forest classifier on the pre-processed dataset, we evaluated its performance using several standard classification metrics: accuracy, confusion matrix, classification report, ROC curve, and log loss.

4.1. Accuracy

The accuracy of the whole model was found to be 0.96 (96%) which means that the model had accurately predicted the traffic accident severity in 96% of the model testing cases. From this high accuracy, it is very clear that the model is able to differentiate the different severity levels very well.

4.2. Confusion Matrix

The confusion matrix provides a detailed breakdown of how the model performed across the four severity classes.

4.3. ROC Curve

The ROC (Receiver Operating Characteristic) curve shows how well a classification model distinguishes between classes by plotting the true positive rate against the false positive rate. A higher area under the curve (AUC) indicates better performance.

4.4. Log Loss

Log loss measures the accuracy of a classification model based on predicted probabilities. It penalizes confident but wrong predictions more heavily. Lower log loss means better model performance.

After thorough experimentation, the evaluation of the proposed work was performed across all four evaluation measures. Figure 3 shows the confusion matrix obtained for all four classes.

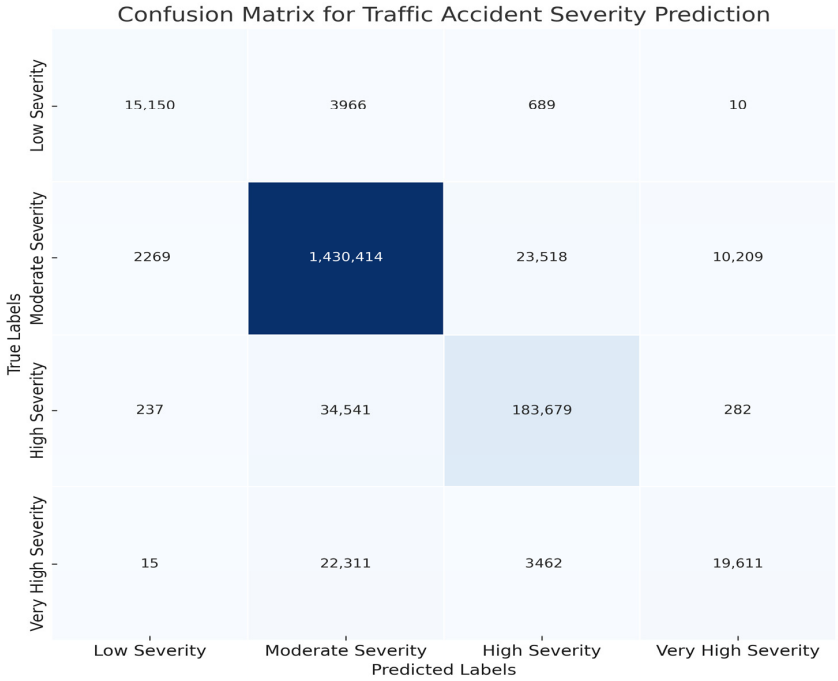

Figure 3. Confusion Matrix Obtained After Applying Random Forest on the Given Dataset.

- Class 1 (Low Severity): The model correctly predicted 15,150 instances of low severity but made 3966 false positives, where accidents were incorrectly classified as having low severity.
- Class 2 (Moderate Severity): The model performed exceptionally well with the moderate severity class, correctly predicting 1,430,414 instances. The false positives (2269) and misclassifications into other classes were relatively low.
- Class 3 (High Severity): While the model performed well with high severity accidents (183,679 correctly predicted), it had higher misclassifications (34,541 false positives) compared to moderate severity accidents.
- Class 4 (Very High Severity): The model struggled the most with predicting very high severity accidents. Only 43% of such accidents were correctly classified, as indicated by the recall of 0.43, with a notable number of false positives (22,311 misclassified as moderate severity) and false negatives (3462 instances of very high severity misclassified into other classes).

4.5. Classification Report

The classification report includes key performance metrics such as precision, recall, and F1-score for each severity class, as shown in Table 4.

Table 4. Classification Report Generated After Applying Random Forest.

Severity	Precision	Recall	F1-Score
1	0.86	0.76	0.81
2	0.96	0.98	0.97
3	0.87	0.84	0.85
4	0.65	0.43	0.52

Table 4. *Cont.*

Severity	Precision	Recall	F1-Score
Accuracy		0.96	
Macro Accuracy	0.83	0.75	0.79
Weighted Accuracy	0.94	0.94	0.94
ROC Curve		0.97	
Log Loss		0.19	

- Class 1 (Low Severity): The model achieved a precision of 0.86, meaning that 86% of instances predicted as low severity were correct. The recall was 0.76, meaning 76% of all true low severity accidents were correctly classified. The F1-score for this class was 0.81, indicating a good balance between precision and recall.
- Class 2 (Moderate Severity): The moderate severity of the accident was well predicted by the model with a precision of 0.96, recall of 0.98, and F1-score of 0.97. This indicates that the model is quite well suited for estimating this kind of class and does not have a tendency to misclassifications.
- Class 3 (High Severity): When it comes to high severity accidents, the model's precision was 0.87, and the recall was 0.84; thus, the F1-score was 0.85. Despite the decrease in terms of precision and recall, calculated for documents with moderate severity, the result remains rather high.
- Class 4 (Very High Severity): The worst performance was observed for very high severity accidents, namely, a precision of 0.65 and a recall of 0.43. The low recall therefore means that many very high severity accidents have been misplaced in the low severity category. The F1-score of 0.52 also shows the difficulty in rightly identifying this particular class.
- With a high ROC AUC score of 0.97, the model demonstrated exceptional capacity to differentiate between normal and abnormal cases. The log loss value of 0.19 further indicates how accurate and well-calibrated the model's projected probabilities are. These metrics imply that the model produces reliable predictions in addition to performing classification tasks efficiently. In a nutshell, the model shows excellent predictive performance appropriate for real-world use. The ROC curve for all four classes is shown in Figure 4 below.

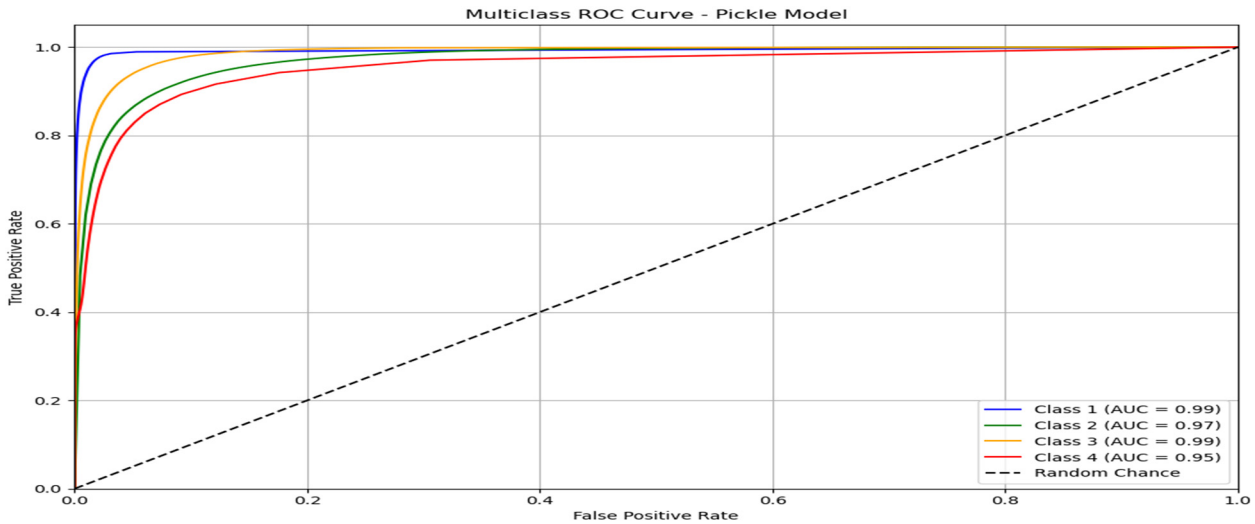

Figure 4. ROC Curve Obtained After Applying Random Forest on the Given Dataset.

4.6. Result Discussion After the Application of SMOTE Technique

Before applying SMOTE, the model exhibited strong performance on the majority class (Severity 2), with very high precision and recall. However, it struggled to effectively predict minority classes, especially Severity Level 4, with a recall of only 0.43 and an F1-score of 0.52. This imbalance reflects the skewed nature of the dataset, which could result in biased outcomes in real-world scenarios where predicting severe accidents is crucial.

After applying SMOTE to balance all classes as highlighted in Table 5, the model became significantly more sensitive to underrepresented severity levels. The recall for Severity Level 4 improved from 0.43 to 0.58, and the F1-score rose from 0.52 to 0.62. Similarly, Severity Level 1 saw an improvement in recall (from 0.76 to 0.81) and a stable F1-score. While the overall accuracy slightly decreased due to the synthetic nature of balanced data, the macro average F1-score increased from 0.79 to 0.80, indicating better performance across all classes. These enhancements underline the effectiveness of SMOTE in improving the fairness and reliability of the model in real-world applications. The model maintained excellent class discrimination with an ROC AUC score of 0.97 after using SMOTE to address the class imbalance. The log loss, however, rose to 0.24, suggesting that although the model continues to distinguish between classes effectively, its confidence in probability estimates has somewhat declined. This is to be expected since SMOTE adds artificial examples, which can complicate the calibration of probability. For classification tasks, the model is still very successful. Figures 5 and 6 show the confusion matrix and ROC curve values that are obtained for the SMOTE-based balanced dataset.

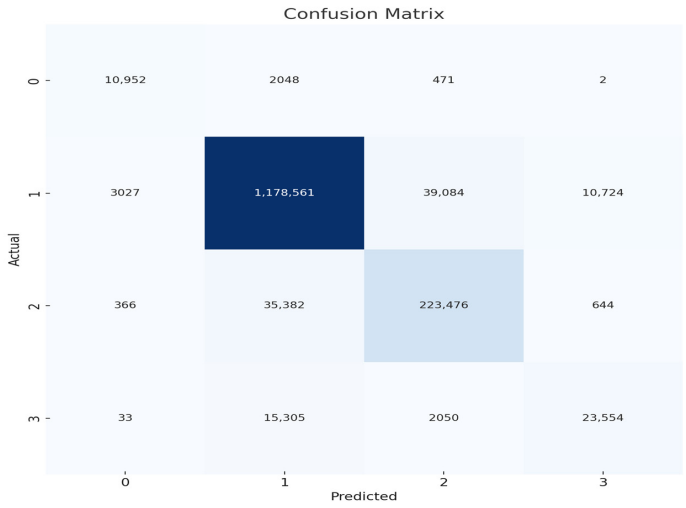

Figure 5. Confusion Matrix Obtained After Applying Random Forest on the SMOTE-Based Balanced Dataset.

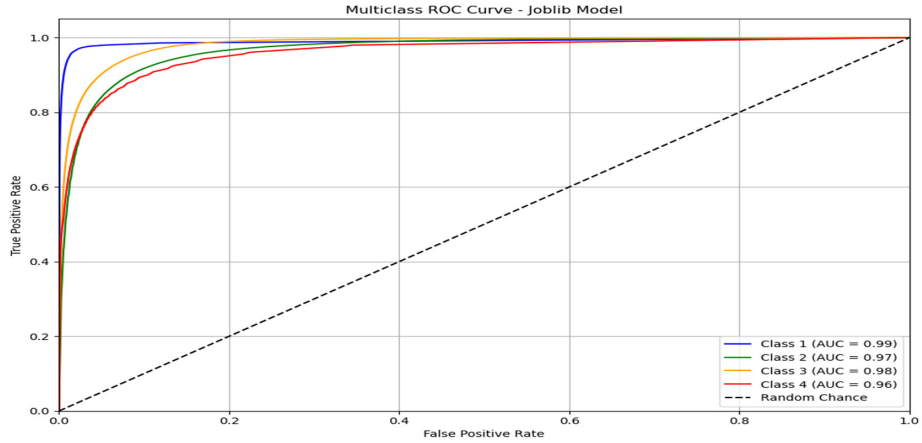

Figure 6. ROC Curve Obtained After Applying Random Forest on the SMOTE-Based Balanced Dataset.

Table 5. Classification Report Generated After Applying Random Forest on the SMOTE-Based Balanced Dataset.

Severity	Precision	Recall	F1-Score
1	0.76	0.81	0.79
2	0.96	0.96	0.96
3	0.84	0.86	0.85
4	0.67	0.58	0.62
Accuracy		0.95	
Macro Average	0.81	0.80	0.80
Weighted Average	0.93	0.93	0.93
ROC Curve		0.97	
LOG Loss		0.24	

4.7. Comparison with Existing Systems

We evaluated the performance of our suggested model by comparing it with six previous studies that made use of the US accident dataset, as indicated in Table 6. With an overall accuracy of 0.96 (96%), our Random Forest model performed better in the majority of cases. With an excellent precision, recall, and F1-score for Severity Level 2, which represents the majority class in the sample, the model demonstrated very good class-wise results.

Table 6. Comparison with Existing Systems.

Existing Work	Model(s) Used	Accuracy	Precision	Recall	F1-Score
Al Dahash et al. [34]	Random Forest	0.80	0.81	0.792	0.80
Gutierrez et al. [19]	AutoML	0.81	Not mentioned	Not mentioned	Not mentioned
Sajadi. et al. [20]	OLM, OPM, SVM	0.71	Not mentioned	Not mentioned	Not mentioned
Vinta al. [35]	Random Forest, SVM	0.85	Not mentioned	Not mentioned	Not mentioned
Proposed	Random Forest	0.96	0.83	0.75	0.79

Most of the existing studies sometimes lacked comprehensive class-wise evaluation and possessed lower accuracies, ranging from 0.71 to 0.85 (71.3% to 85%). These results demonstrate how well our approach predicts the severity of accidents on large, unbalanced datasets. It is worth mentioning that the main aim of the study was to perform automated feature extraction for the given dataset using XAI techniques. A model with high accuracy and good interpretability is the ultimate outcome of this study. The unique aspect of the proposed work is the application of XAI that was missing in the majority of previous studies on the existing dataset.

4.8. Application of Explainable AI

A major limitation associated with most machine learning models is the fact that they are so-called "black-boxes"; one cannot understand how the decision was arrived at. This weakness can negatively affect trust if not well explained in areas such as medicine or business. This is something that Explainable AI (XAI) responds to by allowing for interpretable ways of how those models make decisions, allowing for the feature importance and logical understanding of the predictions made to be given. Feature importance and interpretability methods are used in this case to explain and build confidence in the model's

decision-making process. Due to this, we have used different XAI models for the trained Random Forest to explain how decision-making was performed by the trained model.

4.8.1. Model Explainability Using LIME

To achieve interpretability, we used LIME (Local Interpretable Model-agnostic Explanations) on the trained Random Forest model. LIME lets us explain the individual predictions by mimicking the complex model with simple, explainable models for each of the predictions.

- Global Feature Importance: To identify the global feature importance, we applied LIME to generate explanations for multiple instances from the test set. The weights of each feature were added and a horizontal bar plot was produced to demonstrate the most important features with regards to the entire dataset. This helps in identifying which features had the biggest effect on the prediction of the accident severity.
- Local Explanation for Individual Predictions: Moreover, LIME was used to create local explanations of individual predictions. For instance, we randomly selected an instance from the test set and applied LIME to explain which features contributed to the predicted severity of that given accident. These local explanations are more useful for domain experts and decision-makers because they give details of why certain predictions are being made at the local level.

Figure 7 shows the global feature importance for the given traffic accident severity prediction model, as interpreted by LIME on a Random Forest classifier. After careful analysis, the following observations have been observed. Among all the features, "Source" has the highest importance, suggesting that the origin or type of data source (e.g., traffic cameras, sensors, law enforcement reports) significantly predicts accident severity. This might be because certain sources report accidents differently or with varying levels of detail, impacting the model's classification accuracy. Moreover, the weather timestamp and end time, both of these being time-related features, are also highly important. This could indicate that the timing of an accident (particularly when it ends) correlates with the severity level, perhaps due to factors like traffic congestion or response time. Furthermore, among other remaining features, start time and distance (mi) have been observed as important ones. The start time of the accident also holds notable importance, as it aligns with typical traffic patterns where certain times of day (like rush hours) might see more severe accidents. Furthermore, the distance traveled before the accident occurred might correlate with the severity, as longer travel distances could mean higher speeds or a greater potential for severe collisions. Moreover, the analysis highlights that Start Longitude (Lng) and Civil Twilight moderately contribute to predicting accident severity, suggesting the influence of geographic location and visibility under low-light conditions. Similarly, Timezone and Zipcode reflect regional and localized factors like traffic patterns, infrastructure, and weather conditions that can impact accident outcomes. In contrast, features such as Amenity, Precipitation (in), Station, Railway, Nautical Twilight, Sunrise Sunset, and Temperature (F) show minimal importance, indicating that while they may influence traffic conditions, they are not significant predictors of accident severity in this dataset.

Certain weather-related features, such as Precipitation and Temperature, surprisingly show low importance in predicting accident severity, which suggests the potential for further exploration, such as creating engineered features (e.g., combining temperature and humidity) to enhance their predictive power. Additionally, some features like Astronomical Twilight and Nautical Twilight appear redundant or minimally informative, making them candidates for removal to streamline the model and reduce noise.

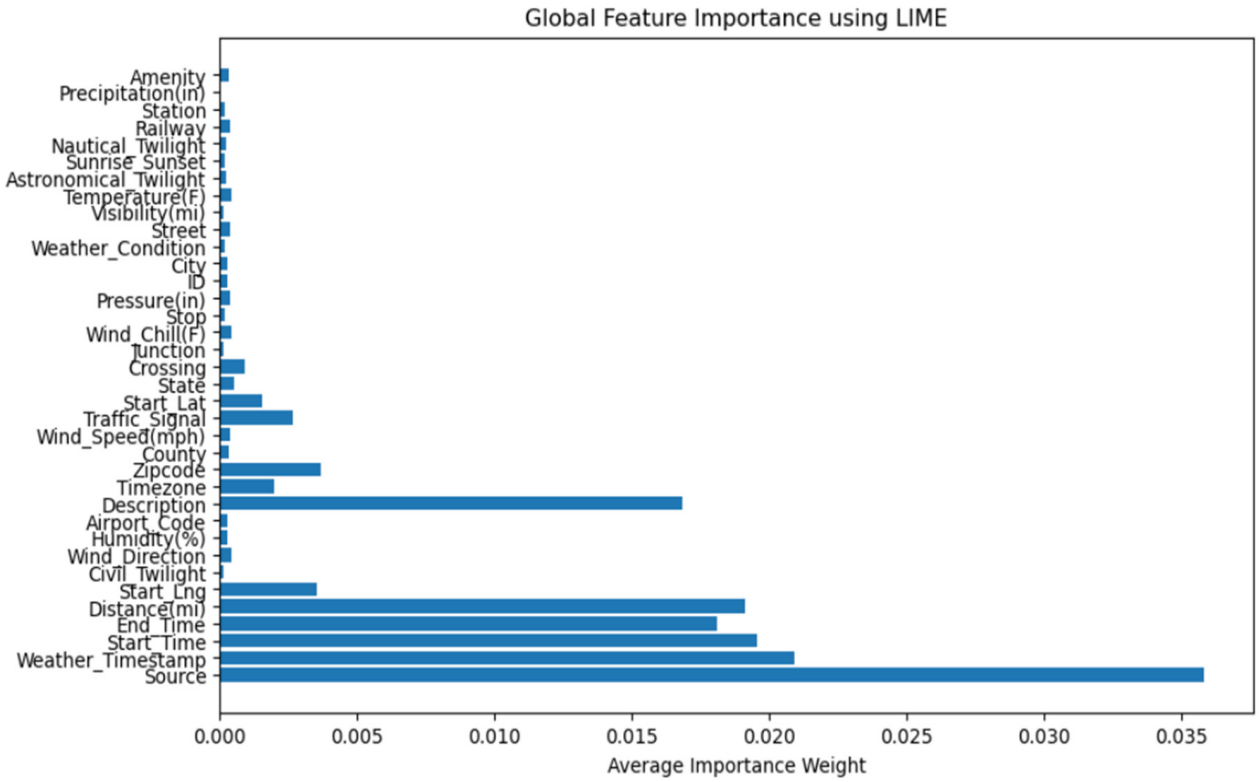

Figure 7. Global Feature Importance Using LIME.

Figure 8 depicts the LIME explanation of a specific data instance and its predicted class. It gives useful information about a prediction made by the Random Forest model with 99% certainty that the instance belongs to Class 2. Other features that are input for the accident severity are Source, which scored 0.04. The time keys that are Weather Timestamp, Start Time, and End Time all scored 0.02, which shows the importance of timing and data source on the accident occurrence. Distance (mi) also has a small coefficient of 0.02, which indicates its importance; it might be associated with travel speed and risk. Small effects, such as the Traffic Signal and Timezone (0.01 each), may indicate regional or traffic control influences.

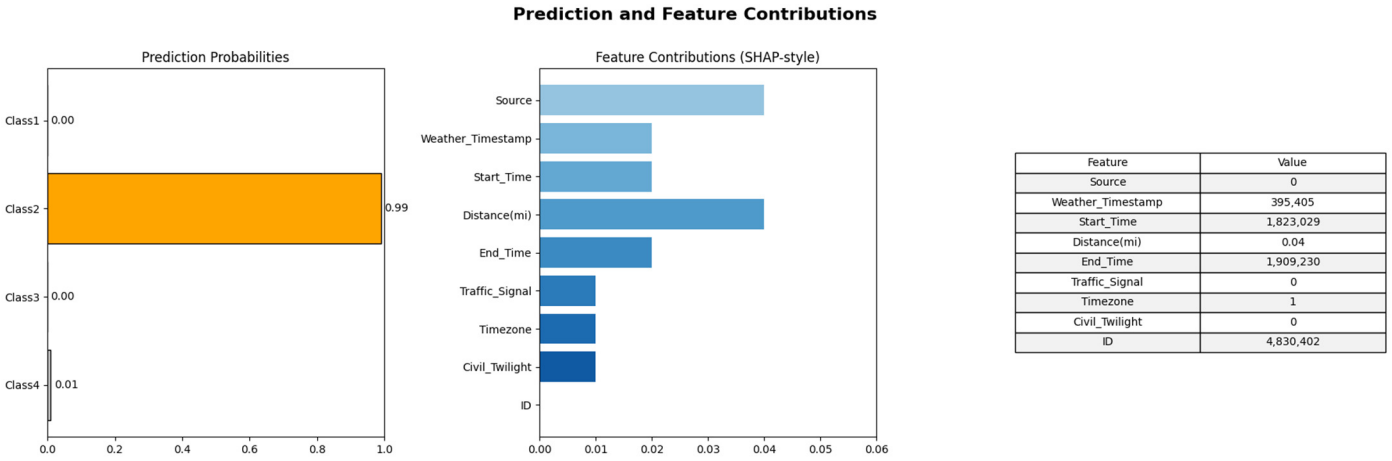

Figure 8. Local Explanation of an Instance Using LIME.

Far more negligible features include Civil Twilight (0.01) and even near-zero features like ID and Street for this particular prediction. The categories of Source and Traffic Signal (0) and Timezone (1) distinguish some accident patterns and correspond to the traffic conditions during the instance.

The highlighted time-related aspects stress their relevance to the characterization of traffic flows and the severity of accidents; furthermore, the moderate but stable contribution of Distance (mi) and Source proves the constancy of their role in the predictions.

In conclusion, it can be seen that the sample was assigned to Class 2 mainly because of Timing features (Weather Timestamp, Start Time, End Time, and Data Source) with Distance (mi) and traffic features (Traffic Signal, Timezone) as the marginal contributors. This conforms to traditional traffic severity prediction trends that show that aspects of time and place are of high importance.

Using feature contributions, the LIME feature weight graph in Figure 9 helps us to understand the model's prediction for Class 2. It is clear that the positive weights, such as End Time, are shifting the classification far into Class 2 in the direction of the class, which may be caused by certain time patterns inherent to the class. Similarly, Start Time and Weather Timestamp influence the prediction positively, but less strongly, which shows that they play a moderate part in the classification.

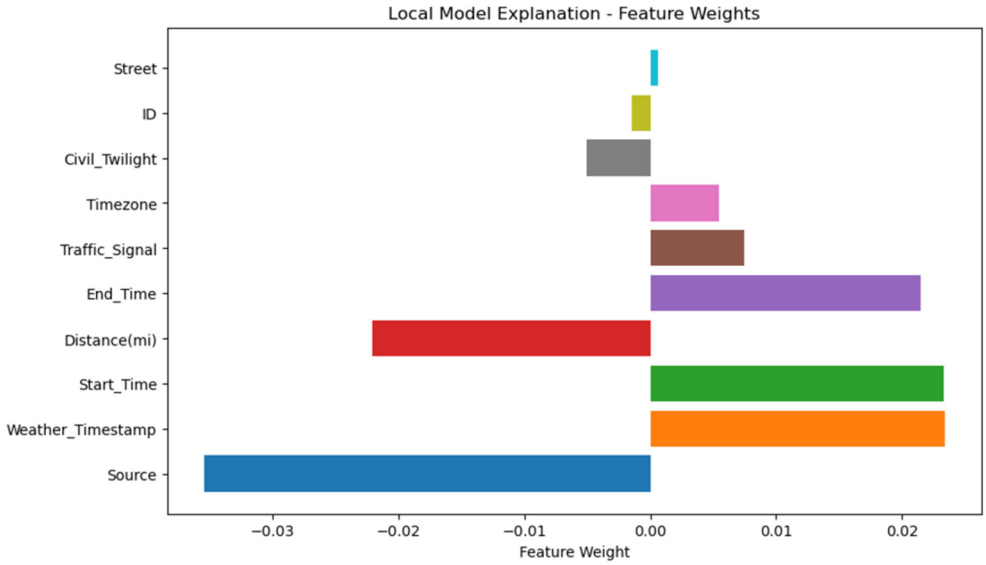

Figure 9. Feature Weight Graph Using LIME.

On the other hand, we have features with negative weights, such as Source and Distance (mi) that oppose the classification to Class 2 and map these features to other classes. First, it is seen that Source has the most significant negative influence, pointing to its close relation to the other classifications. Distance (mi) is also somewhat important in voting against Class 2, probably because of the distance ranges associated with other classes. From the analysis of the features, it is evident that almost all the features have a very small negative impact, which is negligible in the context of the final classification; and for features like Civil Twilight and ID, they almost do not affect the classification at all.

All in all, the Source and Distance (mi) of an incident are significant to classify it into various classes, whereby the temporal features, End Time, Start Time, and Weather Timestamp, are most relevant for Class 2. This present analysis shows how features are ranked; timing aspects contribute to prediction, while Source and Distance (mi) are significant predictors of other classes. The LIME model is often used more frequently to evaluate the effectiveness of features locally. However, understanding the features' importance globally across the whole dataset can certainly help us in understanding their relevance to traffic accident prediction. For this reason, we have also used permutation feature importance to understand the global impact of features on traffic accident prediction.

4.8.2. Permutation Feature Importance

In order to confirm the importance of each feature, permutation importance was used for the evaluation of the model's performance after shuffling the feature values. This method determines critical features by quantifying the reduction in model performance due to shuffling, since larger reductions point toward important features.

Some of the test data were randomly chosen and then the permutation importance was calculated in multiple trials so that there would be some measure of accuracy in determining the contribution of each feature. These were followed by the ranking of the features and a bar plot to show the rankings. This analysis offered a better perspective of the model's stability together with the individual contribution of the features to the prediction task and further enabled the identification of the important features that caused high variance in the model.

Figure 10 shows how many features influence the Random Forest model performance. The first two features, which are the most important, are Source and Description because they possess the highest importance coefficients, suggesting their centrality in the classification of target outcomes. One can only assume that these features contain some patterns or information that this model interprets correctly for the predictions. Moreover, Distance (mi) resurfaces as a factor, which means that the size or extent of an event is useful in the decision-making process.

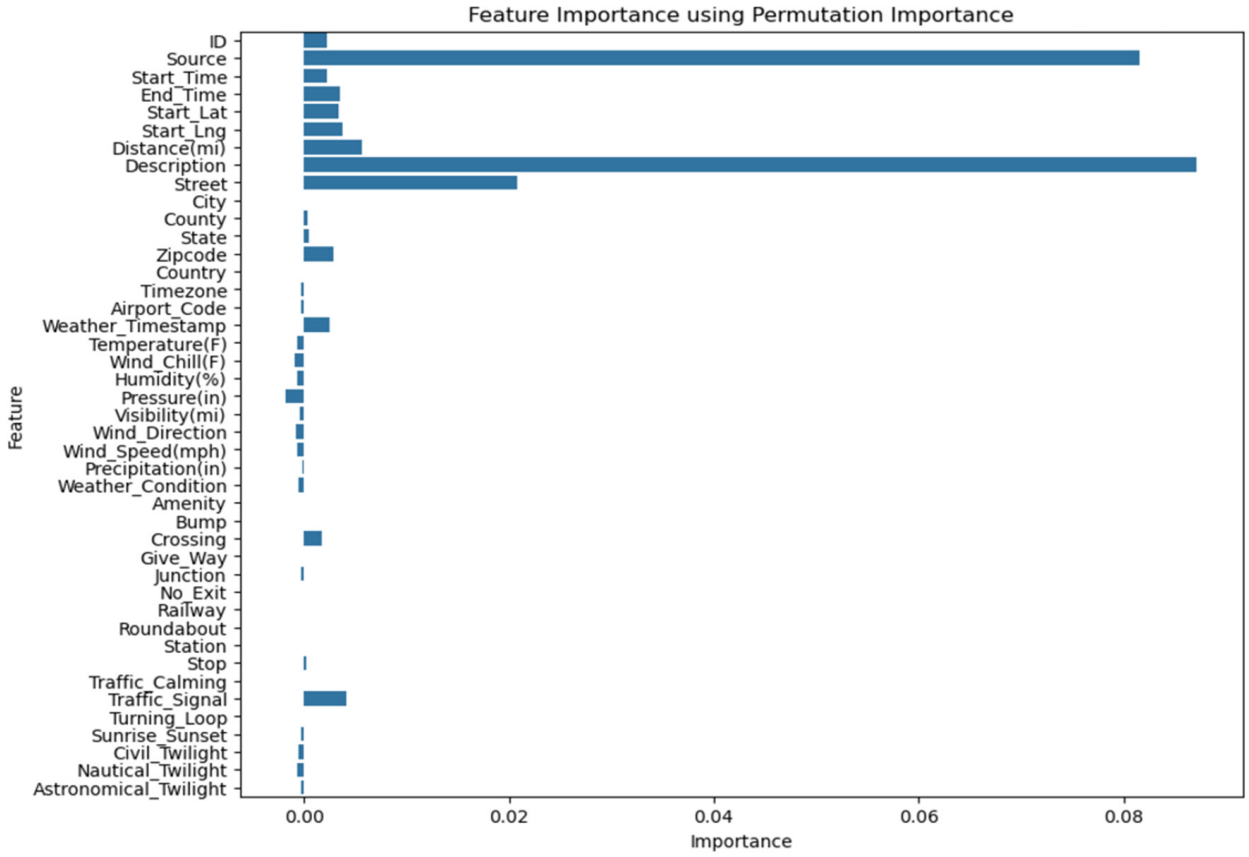

Figure 10. Global Feature Importance Using Permutation Importance Analysis.

Features with a moderate level of importance are Street, which has a small predictive significance, while City and especially State have a very small influence on the target variable. However, attributes like Humidity (%), Visibility (mi), Weather Condition, and Wind Chill (F) turned out to be unimportant, thus suggesting that they have very little use in making predictions. This implies that weather-related variables could either be of less

significance in this regard or might warrant further processing or feature engineering to make them more effective.

In general, the feature importance is quite uneven, and the most important features are Source, Description, and Distance. The other features' contribution is less than 10 per cent, which makes it possible to remove low-usefulness variables from the model to enhance the model's efficiency, while not necessarily decreasing its effectiveness.

4.8.3. Decision Tree

Further to an explanation of global and local feature importance, we also presented feature importance from a single decision tree from the Random Forest model. Although Random Forest includes a number of trees, visualization of one tree allows for contemplating the model as it makes decisions at each node. From the feature splits, we can obtain an understanding of how certain features, such as weather conditions or road types, can affect the predicted severity.

The decision tree was plotted with the help of the plot_tree function in the sklearn.tree where each node depicts a decision that is made on the basis of a feature, and the colors shown at the level represent the class or severity level. This offers a perfect example on how the model processes data and comes up with a decision that will be implemented.

Figure 11 represents a single tree with a Random Forest model having an upper limit on the depth equal to 3 which makes it easily understandable as to how the tree is formed and how it makes its decisions on a certain path. The tree splits on the root node in the feature Street $\leq$ 913284, which shows that this feature provides the greatest information gain at that level and plays an important decision-making role in the tree. This shows that the Street feature has a critical role in determining the outcome of the model.

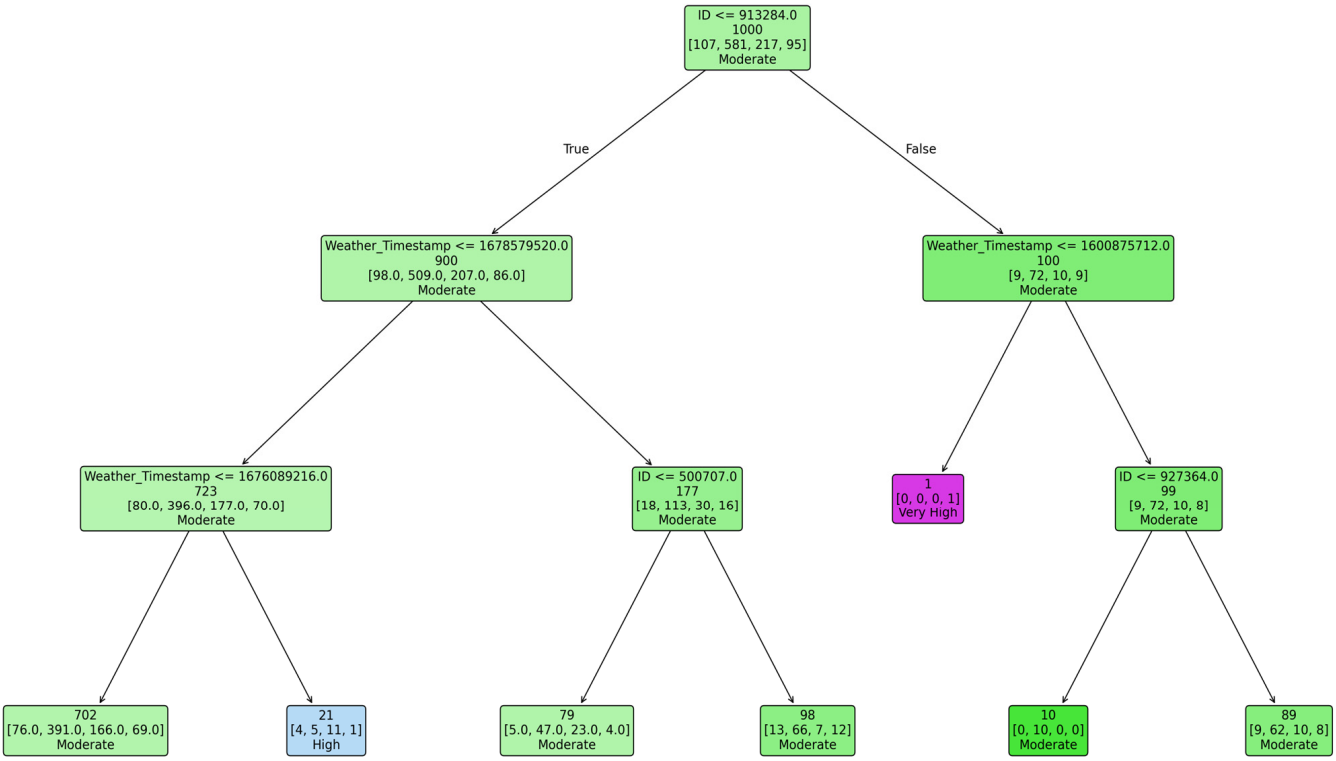

Figure 11. Single Tree Representation with Random Forest Model.

Other nodes that are in between are additional splits based on ID and Weather Timestamp. Every node gives the split criterion (e.g., Street ID $\leq$ 913284), Gini, the number of

data points that reached this node, and the target distribution for the classification task. They explain how the model refines its decision-making as it moves deeper into the tree while using thresholds on features to obtain better separation of the data.

Leaf nodes take the terminals of the tree, and no split takes place at a particular depth within such trees. These nodes hold the target values or the probability of the class membership of the samples reaching them. In this tree, the items at the extremes of the branches are the decisions made after the previous splits of the branches. They show how the model determines, in this case, classifies or predicts the outcome once the decision path is over.

The attributes closer to the top of the tree, such as Street, Weather Timestamp, and ID, are generally more important because they affect more of the data. The decision path along the branches from the root node to the end nodes, representing the structure of a tree, is known as the decision thresholds that lead to the determination of the particular output.

This tree is just a single tree representing just one portion of the Random Forest model, which combines several trees to provide more accurate results via averaging or voting. Reducing the "depth" to 3 makes it easier to understand and reduces overfitting, while still giving us a view of how the model uses features at a high level regarding how the model makes its decisions. Table 7 highlights a summary of the XAI techniques used in the proposed work.

Table 7. Summary of the XAI Techniques Used in the Proposed Work.

XAI Method	Strengths	Limitations	Application in Proposed Work
LIME (Local Interpretable Model-agnostic Explanations)	Provides detailed local explanations for individual predictions using simple surrogate models	May produce inconsistent results across similar samples; sensitive to perturbations	Used to interpret how specific features influence individual accident severity predictions
Permutation Feature Importance	Model-agnostic, easy to implement, and quantifies the impact of each feature on overall model performance	Can be affected by correlated features and requires multiple evaluations	Applied to identify globally important features influencing the Random Forest model
Decision Tree-based Feature Importance	Fast and natively supported in tree-based models; intuitive and easy to visualize	Not ideal for non-tree models; can undervalue correlated or low-variance features	Applied to visualize how an individual tree decides certain criteria

4.9. Conclusive Insights

After analyzing the results from LIME, permutation importance, and decision tree feature analysis, several key features emerge as significant in predicting accident severity. Temporal features such as Start Time, End Time, Timezone, and Weather Timestamp consistently show strong positive contributions, particularly in the LIME analysis. This highlights the impact of time-based factors on accident severity, possibly due to traffic patterns, rush hours, or weather conditions at specific times. Similarly, Source and Distance (mi) play dual roles—while Source is shown to have a strong negative weight against Class 2 in LIME, it emerges as one of the most important features in permutation importance, indicating its strong role in classification. Distance (mi), though moderately important, influences accident severity predictions by potentially linking longer travel distances to increased risk factors.

Furthermore, the permutation importance and decision tree analysis reinforce Source and Description as the most critical features, suggesting that these attributes contain valuable patterns that the Random Forest model leverages effectively. Street also appears as a significant decision-making feature in the decision tree model, acting as the primary split criterion in a simplified tree, which indicates that location-based attributes hold predictive value in accident severity classification. Crossing and traffic signals also play a role in predicting accident severity highlighting the importance of spatial features. On the other hand, features such as humidity, visibility, weather conditions, and wind chill are found to be less important in determining accident severity, possibly due to their weaker correlations with accident outcomes. The combined explainability results indicate that time-related, source-related, and geographical attributes are the most influential in predicting accident severity, while weather-related factors contribute minimally. The proposed study aims to look for contextual and environmental characteristics linked to serious accident outcomes by analyzing data from actual traffic accidents. The analysis's conclusions can be applied to improve ADAS systems' efficacy and reactivity, allowing them to more accurately predict and respond to high-risk driving situations. Moreover, the key predictors can inform ADAS risk assessment modules. Integrating machine learning models into ADAS allows for timely alerts or automated responses. These insights help prioritize which environmental and situational sensors to develop or enhance. Ultimately, they support more accurate and context-aware safety interventions.

It is crucial to point out that the features that our XAI approaches have extracted are the features that may be unique to the dataset utilized in this investigation and should not be considered as causal correlations. The dataset that was used for the given research is publicly available. The main effort from our side was to explore its hidden patterns and understand the impact of various features on accident prediction. This will give future insights to researchers who want to work on this dataset. It is worth mentioning that the feature importance rankings are influenced by the structure, reporting standards, and contextual variables inherent to the dataset, and, thus, interpretations must be made with caution. However, despite these drawbacks, the current study shows how interpretable machine learning can assist risk assessments in real-life scenarios by offering important and insightful information on the features linked to the severity of traffic accidents. A more thorough assessment of feature stability, generalizability, and multidimensional influences on accident severity can be achieved by expanding the methodology to include multiple datasets from various time periods, geographies, or domains in subsequent research.

4.10. Example Scnerio to Demonstrate XAI's Extracted Features Can Contribute Toward Predicting Traffic Accident Severity

The top features that are obtained after applying various XAI models are the following:

- Source
- Description
- Start_Time
- End_Time
- Time Zone
- Weather Timestamp
- Crossing
- Traffic Signals

In order to demonstrate how these features contribute to predicting accident severity, consider a scenario about a traffic accident that occurred on a big city road in the evening hours. A local traffic monitoring firm (Source), renowned for its prompt and thorough data

reporting, has recorded the occurrence. The accident occurs close to a pedestrian crossing (Crossing), and it is noteworthy that there is no traffic signal in the region (Traffic_Signal), which raises the possibility of crashes because of uncontrolled traffic flow. A higher impact event, maybe involving numerous vehicles or significant vehicle movement after point of contact, is suggested by the reported distance affected by the accident (Distance (mi)), which is comparatively long.

The accident may have occurred during a period of high traffic intensity, which may have coincided with the daily rush hour, according to the temporal attributes Start_Time, End_Time, and Timezone. In contrast to the structural and time-based parameters, the XAI results indicate that weather-related features such as Wind_Chill (F), Visibility (mi), and Humidity (%) had little effect on severity prediction. Furthermore, the model probably detects the patterns or keywords in the Description column, which summarizes incident information, that are linked to more serious outcomes. While features such as Source and Description emerged as highly important in the XAI analysis, it is important to note that they may not be directly linked to accident severity in a causal sense. Instead, their influence on the model's accuracy may stem from dataset-specific patterns or the quality and structure of how incidents are recorded. This can also give us insight about features like Wind_Chill, Visibility, and Humidity that may appear to be very relevant to accident prediction, but, due to their low underlying correlation with accident severity for the given dataset, they have not contributed significantly to predicting accident severity. However, the correlation among these variables is explored in Section 4.7 to understand how these features may contribute toward a driver's driving pattern. This interpretation suggests that the features extracted via XAI cannot be generalize for all cases. In fact, in order to improve the reliability of extracted features, various diverse datasets should be explored, which may help improve model generalizability, which is a future direction of the proposed work.

4.11. Driver Behavior Pattern

The dataset being used in the proposed work lacks attributes that may directly map to driver behavior patterns. This resulted in a need to analyze attributes that may indirectly help in assessing driver behavior patterns. To address the absence of explicit driver fault indicators in the dataset, we explored indirect behavioral and environmental variables that may serve as surrogates for driver decision-making. The pairwise correlations between important characteristics, including Weather Condition, Visibility, Traffic Signal, Junction, and Distance, are displayed in Figure 12. The fact that the majority of factors have minimal linear correlation (values near 0) indicates that they most likely influence accident severity either independently or through non-linear interactions. For example, there is a little negative association (-0.09) between visibility and weather conditions, suggesting that bad weather may somewhat impair visibility. Furthermore, there are weak negative associations between Distance and Traffic_Signal and Junction, which could indicate more intricate spatial patterns in crashes in urban than rural areas. Despite the dataset's absence of explicit behavioral annotations, our analysis validates our method of leveraging these contextual variables to infer patterns of driver behavior, such as reaction to environmental cues or intersection control.

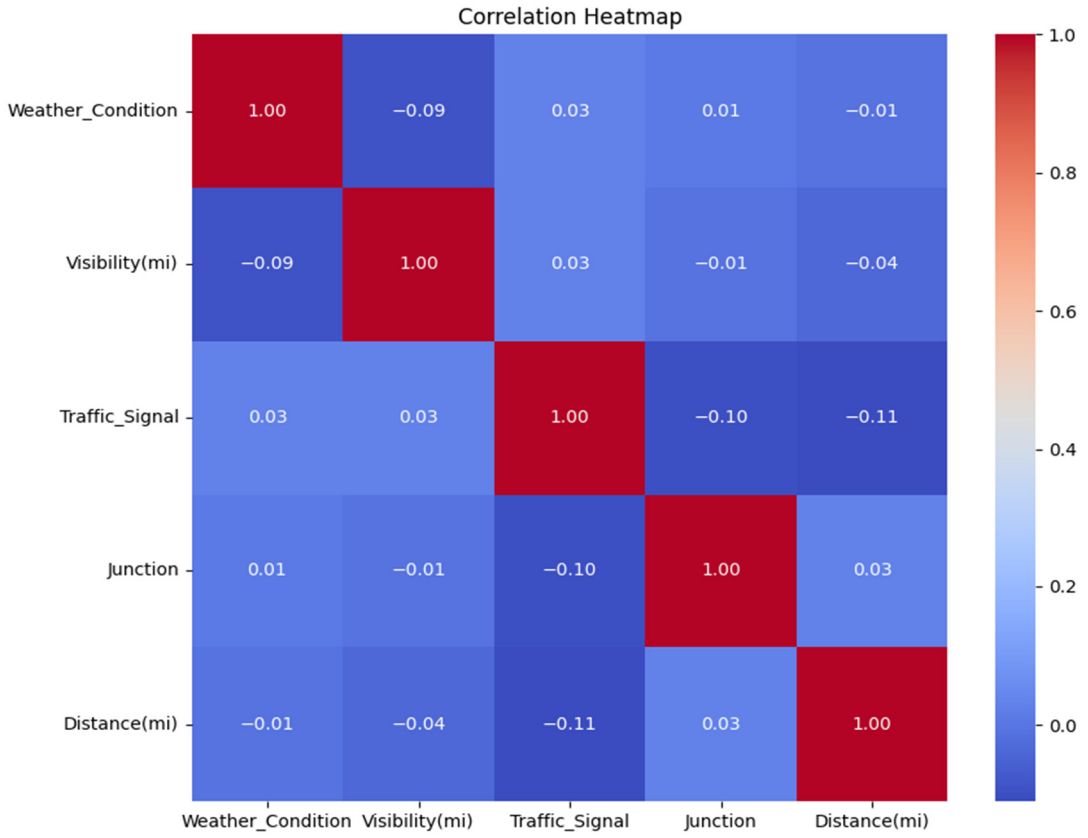

Figure 12. Correlation Heatmap of Features for the Indirect Assessment of Driver Behavior.

5. Limitations

Despite showing excellent prediction ability in classifying the severity of accidents, the proposed work has a number of limitations. The dataset's class imbalance is a major drawback, especially the underrepresentation of Severity Level 4 (extremely high severity), which can have a substantial impact on recall and make it more difficult for the model to accurately identify the most important cases. Furthermore, the analysis's feature set was devoid of crucial human-related characteristics like driver conduct or fault signs, which could have biased feature selection and limited the model's ability to identify causal elements. Another drawback is that the model may not be as transferable or reliable when used in different geographic or regulatory contexts due to variations in accident reporting criteria, data quality, and classification between states. Moreover, the dataset does not contain any metadata related to ADAS; therefore, the results cannot infer direct ADAS assessment; however, they can be used to inform ADAS algorithms, especially in real-time decision-making and risk classification logic.

6. Conclusions

This study shows how machine learning (ML) models can be used to improve the safety of Advanced Driver Assistance Systems (ADAS) and anticipate traffic accidents in advance. Numerous classification models were tested on a publicly available US accident dataset, including AdaBoost, K-Nearest Neighbors (KNN), Random Forest, and Support Vector Machines (SVM). The main aim was to find the best model that can efficiently classify the given dataset with high accuracy. The model's transparency was further enhanced by the application of explainable AI (XAI) approaches like SHAP and permutation importance. These XAI techniques enable us to extract features which are highly relevant with regard to

the traffic accidents' prediction. This enhances model transparency, building user trust and enabling us to understand the model's decision-making criteria.

The study concludes that Random Forest provides a very accurate and interpretable framework for predicting the severity of traffic accidents when combined with explainable AI techniques. Temporal variables (Start Time, End Time), spatial markers (Street, Crossing), and incident metadata (Description, Source) were the most significant predictors among the features that were extracted. The investigation also revealed that, despite their obvious relevance, weather-related parameters had no effect on this dataset. This study closes the gap between actionable insights for safety systems and black-box machine learning models by addressing feature interpretability. This study does not directly evaluate ADAS performance. Instead, it focuses on analyzing real-world accident data to identify conditions associated with severe outcomes. These insights can inform enhancements in ADAS responsiveness and decision-making. These findings can guide the development of context-aware ADAS systems that prioritize sensor inputs, evaluate risks in real time, and respond adaptively in a variety of scenarios. The results imply that future ADAS technology can be tailored for practical uses by knowing which factors affect the frequency of accidents.

7. Future Work

In order to increase the predictive potential of the model, future studies should look into enlarging the dataset to include more variables, such as traffic volume and driver behavior. The capacity of the model to adjust to changing driving conditions may also be improved by real-time data from sensors and traffic infrastructure. Geospatial analysis offers valuable insights by identifying high-risk accident zones based on location patterns.

Although this is not addressed in the current study, it is suggested as a direction of research for the future. Model accuracy and contextual awareness may be improved by methods like clustering and spatial machine learning. For preventative safety measures, this strategy can aid in the creation of location-aware ADAS features. Moreover, more datasets will be explored that may contain behavioral features related to the driver driving patterns in order to asses how much the driver's fault contributes to the occurrence of an accident.

By identifying intricate patterns in the data, investigating cutting-edge machine learning approaches like deep learning may increase prediction accuracy even more. A more complete safety system would also be produced by adding human factors like driver weariness and distraction to predictive algorithms. Maintaining the efficacy and reliability of ADAS technology in practical applications will require ongoing improvements in model interpretability and openness.

Funding: This research was supported by the Deanship of Graduate Studies and Scientific Research at Qassim University for financial support (QU-APC-2025).

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: The Researcher would like to thank the Deanship of Graduate Studies and Scientific Research at Qassim University for financial support (QU-APC-2025).

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ADAS	Advanced Driver Assistance Systems
ML	Machine Learning

LIME	Local Interpretable Model-agnostic Explanations
XAI	Explainable Artificial Intelligence
SVM	Support Vector Machines
K-NN	K-Nearest Neighbor
DNN	Deep Neural Networks

References

1. Briggs, A.M.; Cross, M.J.; Hoy, D.G.; Sànchez-Riera, L.; Blyth, F.M.; Woolf, A.D.; March, L. Musculoskeletal Health Conditions Represent a Global Threat to Healthy Aging: A Report for the 2015 World Health Organization World Report on Ageing and Health. *Gerontologist* **2016**, *56*, S243–S255. [CrossRef] [PubMed]
2. Čabarkapa, M. *Road Safety: From Global to Local and Vice Versa*; Cambridge Scholars Publishing: Newcastle Upon Tyne, UK, 2019.
3. Masello, L.; Castignani, G.; Sheehan, B.; Murphy, F.; McDonnell, K. On the road safety benefits of advanced driver assistance systems in different driving contexts. *Transp. Res. Interdiscip. Perspect.* **2022**, *15*, 100670. [CrossRef]
4. Jain, A.; Goyal, V.; Sharma, K. A Comprehensive Analysis of AI/MLenabled Predictive Maintenance Modelling for Advanced Driver-Assistance Systems. *J. Electr. Syst.* **2024**, *20*, 486–507.
5. Macioszek, E.; Wyderka, A.; Jurdana, I. The bicyclist safety analysis based on road incidents maps. *Sci. J. Silesian Univ. Technol. Ser. Transp.* **2025**, *126*, 129–147. [CrossRef]
6. Chengula, T.J.; Mwakalonge, J.; Comert, G.; Siuhi, S. Improving road safety with ensemble learning: Detecting driver anomalies using vehicle inbuilt cameras. *Mach. Learn. Appl.* **2023**, *14*, 100510. [CrossRef]
7. Souweidane, N.; Smith, B. *State of ADAS, Automation, and Connectivity*; Center for Automotive Research: Ann Arbor, MI, USA, 2023; pp. 1–40.
8. Yang, C.D.; Ozbay, K.; Ban, X. *Developments in Connected and Automated Vehicles*; Taylor & Francis: Abingdon-on-Thames, UK, 2017; Volume 21, pp. 251–254.
9. Chengula, T.J.; Mwakalonge, J.; Comert, G.; Sulle, M.; Siuhi, S.; Osei, E. Enhancing advanced driver assistance systems through explainable artificial intelligence for driver anomaly detection. *Mach. Learn. Appl.* **2024**, *17*, 100580. [CrossRef]
10. Sakib, N.; Bashar, S.; Rahman, A. Road Accident Analysis of Dhaka City Using Counter Propagation Network. In Proceedings of the International Symposium on Ubiquitous Networking, Montreal, QC, Canada, 25–27 October 2022; pp. 164–179.
11. Hansen, J.H.; Busso, C.; Zheng, Y.; Sathyanarayana, A. Driver modeling for detection and assessment of driver distraction: Examples from the UTDrive test bed. *IEEE Signal Process. Mag.* **2017**, *34*, 130–142. [CrossRef]
12. Saravanarajan, V.S.; Chen, R.-C.; Dewi, C.; Chen, L.-S.; Ganesan, L. Car crash detection using ensemble deep learning. *Multimed. Tools Appl.* **2024**, *83*, 36719–36737. [CrossRef]
13. Santos, K.; Firme, B.; Dias, J.P.; Amado, C. Analysis of motorcycle accident injury severity and performance comparison of machine learning algorithms. *Transp. Res. Rec.* **2024**, *2678*, 736–748. [CrossRef]
14. Fan, S.; Yang, Z. Accident data-driven human fatigue analysis in maritime transport using machine learning. *Reliab. Eng. Syst. Saf.* **2024**, *241*, 109675. [CrossRef]
15. Ali, Y.; Hussain, F.; Haque, M.M. Advances, challenges, and future research needs in machine learning-based crash prediction models: A systematic review. *Accid. Anal. Prev.* **2024**, *194*, 107378. [CrossRef] [PubMed]
16. Owais, M.; Alshehri, A.; Gyani, J.; Aljarbou, M.H.; Alsulamy, S. Prioritizing rear-end crash explanatory factors for injury severity level using deep learning and global sensitivity analysis. *Expert Syst. Appl.* **2024**, *245*, 123114. [CrossRef]
17. Balasubramani, S.; Aravindhar, J.; Renjith, P.; Ramesh, K. DDSS: Driver decision support system based on the driver behaviour prediction to avoid accidents in intelligent transport system. *Int. J. Cogn. Comput. Eng.* **2024**, *5*, 1–13.
18. Jamal, A.; Zahid, M.; Tauhidur Rahman, M.; Al-Ahmadi, H.M.; Almoshaogeh, M.; Farooq, D.; Ahmad, M. Injury severity prediction of traffic crashes with ensemble machine learning techniques: A comparative study. *Int. J. Inj. Control Saf. Promot.* **2021**, *28*, 408–427. [CrossRef]
19. Acı, Ç.İ.; Mutlu, G.; Ozen, M.; Acı, M. Enhanced Multi-Class Driver Injury Severity Prediction Using a Hybrid Deep Learning and Random Forest Approach. *Appl. Sci.* **2025**, *15*, 1586. [CrossRef]
20. Gutierrez-Osorio, C.; González, F.A.; Pedraza, C.A. Deep Learning Ensemble Model for the Prediction of Traffic Accidents Using Social Media Data. *Computers* **2022**, *11*, 126. [CrossRef]
21. Sajadi, P.; Qorbani, M.; Moosavi, S.; Hassannayebi, E. Accident Impact Prediction based on a deep convolutional and recurrent neural network model. *arXiv* **2024**, arXiv:2411.07537.
22. Elvik, R. Problems in determining the optimal use of road safety measures. *Res. Transp. Econ.* **2014**, *47*, 27–36. [CrossRef]
23. Adewopo, V.; Elsayed, N.; Elsayed, Z.; Ozer, M.; Zekios, C.L.; Abdelgawad, A.; Bayoumi, M. Big Data and Deep Learning in Smart Cities: A Comprehensive Dataset for AI-Driven Traffic Accident Detection and Computer Vision Systems. In *SoutheastCon*; IEEE: New York, NY, USA, 2024.

24. Grigorev, A.; Saleh, K.; Ou, Y.; Mihaita, A.-S. Integrating Large Language Models for Severity Classification in Traffic Incident Management: A Machine Learning Approach. *arXiv* **2024**, arXiv:2403.13547.
25. Adewopo, V.A.; Elsayed, N. Smart city transportation: Deep learning ensemble approach for traffic accident detection. *IEEE Access* **2024**, *12*, 59134–59147. [CrossRef]
26. Zahid, A.; Qasim, T.; Bhatti, N.; Zia, M. A data-driven approach for road accident detection in surveillance videos. *Multimed. Tools Appl.* **2024**, *83*, 17217–17231. [CrossRef]
27. Tang, J.; Zheng, L.; Han, C.; Yin, W.; Zhang, Y.; Zou, Y.; Huang, H. Statistical and machine-learning methods for clearance time prediction of road incidents: A methodology review. *Anal. Methods Accid. Res.* **2020**, *27*, 100123. [CrossRef]
28. Abdulrashid, I.; Farahani, R.Z.; Mammadov, S.; Khalafalla, M.; Chiang, W.-C. Explainable artificial intelligence in transport Logistics: Risk analysis for road accidents. *Transp. Res. Part E Logist. Transp. Rev.* **2024**, *186*, 103563. [CrossRef]
29. Aboulola, O.I.; Alabdulqader, E.A.; Alarfaj, A.A.; Alsubai, S.; Kim, T.-H. An Automated Approach for Predicting Road Traffic Accident Severity Using Transformer Learning and Explainable AI Technique. *IEEE Access* **2024**, *12*, 61062–61072. [CrossRef]
30. Abdollahi, A.; Li, D.; Deng, J.; Amini, A. An explainable artificial-intelligence-aided safety factor prediction of road embankments. *Eng. Appl. Artif. Intell.* **2024**, *136*, 108854. [CrossRef]
31. Khan, M.A.; Khan, M.; Dawood, H.; Dawood, H.; Daud, A. Secure Explainable-AI Approach for Brake Faults Prediction in Heavy Transport. *IEEE Access* **2024**, *12*, 114940–114950. [CrossRef]
32. Das, A.; Rad, P. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv* **2020**, arXiv:2006.11371.
33. Saarela, M.; Jauhiainen, S. Comparison of feature importance measures as explanations for classification models. *SN Appl. Sci.* **2021**, *3*, 272. [CrossRef]
34. Salimiparasa, M.; Sedig, K.; Lizotte, D. Unlocking the Power of Explainability in Ranking Systems: A Visual Analytics Approach with XAI Techniques. In *International Workshop on Explainable Artificial Intelligence in Healthcare*; Springer: Berlin/Heidelberg, Germany, 2023.
35. Vinta, S.R.; Rajarajeswari, P.; Kumar, M.V.; Kumar, G.S.C. BConvLSTM: A deep learning-based technique for severity prediction of a traffic crash. *Int. J. Crashworthiness* **2024**, *29*, 1051–1061. [CrossRef]