Transportation Research Part E 186 (2024) 103563

Contents lists available at ScienceDirect

Transportation Research Part E

journal homepage: www.elsevier.com/locate/tre

Explainable artificial intelligence in transport Logistics: Risk analysis for road accidents

Ismail Abdulrashid [a], Reza Zanjirani Farahani [b,*], Shamkhal Mammadov [c], Mohamed Khalafalla [d], Wen-Chyuan Chiang [a]

[a] *School of Finance and Operations Management, Collins College of Business, The University of Tulsa, 800 S Tucker Dr, Tulsa 74104 OK, United States*
[b] *Rennes School of Business, 2 rue Robert d'Arbrissel, Rennes 35065 France*
[c] *McDougall School of Petroleum Engineering, College of Engineering, The University of Tulsa, 74104 OK, United States*
[d] *Division of Engineering Technology, Florida A&M University, Tallahassee 32307 FL, United States*

A R T I C L E I N F O

Keywords:
Transport logistics
Explainable artificial intelligence
Road safety
Injury severity
Prediction and mitigation
Taxonomy

A B S T R A C T

Automobile traffic accidents represent a significant threat to global public safety, resulting in numerous injuries and fatalities annually. This paper introduces a comprehensive, explainable artificial intelligence (XAI) artifact design, integrating accident data for utilization by diverse stakeholders and decision-makers. It proposes responsible, explanatory, and interpretable models with a systems-level taxonomy categorizing aspects of driver-related behaviors associated with varying injury severity levels, thereby contributing theoretically to explainable analytics. In the initial phase, we employed various advanced techniques such as data missing at random (MAR) with Bayesian dynamic conditional imputation for addressing missing records, synthetic minority oversampling technique for data imbalance issues, and categorical boosting (CatBoost) combined with SHapley Additive exPlanations (SHAP) for determining and analyzing the importance and dependence of risk factors on injury severity. Additionally, exploratory feature analysis was conducted to uncover hidden spatiotemporal elements influencing traffic accidents and injury severity levels. We developed several predictive models in the second phase, including eXtreme Gradient Boosting (XGBoost), random forest (RF), deep neural networks (DNN), and fine-tuned parameters. Using the SHAP approach, we employed model-agnostic interpretation techniques to separate explanations from models. In the final phase, we provided an analysis and summary of the system-level taxonomy across feature categories. This involved classifying crash data into high-level causal factors using aggregate SHAP scores, illustrating how each risk factor contributes to different injury severity levels.

1. Introduction

Road transport crashes are a significant threat to public transport and logistics safety globally (Bhowmik et al., 2021), causing numerous injuries and fatalities annually. Understanding the relationship between injury severity outcomes and their contributing risk factors is crucial to mitigating the adverse impacts of road traffic accidents. While classic AI models can help with this through tasks

* Corresponding author at: Rennes School of Business, 2 rue Robert d'Arbrissel, Rennes, Brittany 35065 France.
E-mail addresses: ismail-abdulrashid@utulsa.edu (I. Abdulrashid), Reza.Zanjirani-Farahani@Rennes-sb.com (R. Zanjirani Farahani), shm3416@utulsa.edu (S. Mammadov), mohamed.ahmed@famu.edu (M. Khalafalla), wen-chyuan-chiang@utulsa.edu (W.-C. Chiang).

https://doi.org/10.1016/j.tre.2024.103563
Received 26 February 2024; Received in revised form 22 April 2024; Accepted 26 April 2024
Available online 30 April 2024
1366-5545/

such as learning, reasoning, and problem-solving based on human intelligence, this research utilizes the explainable AI (XAI) approach—as a subset of AI—that focuses on making decisions, providing more profound analysis, and ensuring transparency. This is particularly important in analyzing traffic accident data to discern causes and relationships between factors and develop improved predictive models for injury severity.

In the United States, according to the National Highway Traffic Safety Administration (NHTSA), over 6 million police-reported automobile crashes occurred in 2019, leading to approximately 36,000 fatalities and 4.4 million non-fatal injuries (Stewart, 2022). The severity of these injuries varies considerably, ranging from minor bruises to life-threatening conditions. As another relevant example, a recent study focusing on the metropolitan area level – as opposed to the national level, which may obscure risk variation across countries – found that pedestrians in San Francisco are 4.1 times more likely to be killed in a traffic accident than those in Stockholm. At the same time, bicyclists are 1.7 times more likely to be injured (McAndrews, 2011). Most of the existing analytical models are built on complex internal learning processes (i.e., they are 'black box' in nature) and calculations, which can be complex for humans to understand causes. Moving beyond the existing AI models, XAI attempts to explain such phenomena with more transparency easily. It offers a paradigm shift towards models that elucidate the reasoning behind their predictions. By leveraging XAI, this research aims to unveil the multifaceted causes of traffic accidents and the determinants of injury severity with unparalleled clarity. XAI's capability to demystify the models' decision-making processes enables stakeholders to identify high-risk scenarios and devise precisely targeted safety interventions (Lisboa et al., 2023). Unlike traditional AI methods, XAI uncovers the impact of each contributing factor in injury severity predictions, offering actionable insights for traffic safety management. This research utilizes XAI to obtain a deep understanding of crash causes and injury severity or to understand how the analytical models make predictions.

More studies have underscored the importance of incorporating XAI approaches (Tutun et al., 2023), which aim to provide responsible, interpretable, and transparent models (Dong et al., 2017). In the context of road traffic safety, promoting these approaches can improve predictions' accuracy and enhance decision-making transparency, thereby reducing the frequency and severity of future traffic accidents by identifying high-risk crash environments. The systematic literature review by Filom et al. (2022) on ML methods in port operations illustrates ML's growing impact on operational efficiency and decision-making. Similarly, Chen et al. (2024) and Yang et al. (2023) demonstrate AI's potential in harnessing spatial and temporal dependencies for various applications, ranging from loan default prediction to estimating heavy truck mobility flows. These advancements signal the broad applicability of ML and XAI in tackling complex challenges across transportation sectors, highlighting the value of such technologies in enhancing road traffic safety. Furthermore, the predictive models developed by Wei et al. (2022) demonstrate XAI's role in improving the interpretability and applicability of findings for traffic safety strategies.

Advanced technologies and data-driven analytical techniques for predicting injury severity in traffic accidents have shown promising results (Yang et al., 2022; Zheng & Sayed, 2020). In this context, Delen et al. (2017) identified numerous risk factors, such as behavioral and demographic characteristics of vehicle occupants (e.g., drug and/or alcohol levels, use of seat belts or other restraining systems, and the gender and age of the driver), as significant determinants influencing injury severity in automobile crashes. Ma et al. (2021) employed a deep learning model, specifically a stacked sparse autoencoder, to predict traffic accident injury severity, considering the spatial correlation of accident locations.

This paper contributes to the literature by formulating an XAI framework to assist in understanding and identifying causes of automobile crashes and predicting and mitigating injury severity risk factors in traffic accidents. The proposed methodology combines datasets from publicly available government sources, addresses data-related issues through feature selection and engineering, stratifies the dataset into subgroups based on regional emergencies, and then builds a predictive model using machine learning (ML) algorithms. The most effective model was used in an explanatory (model-agnostic) approach, revealing high-risk factors influencing injury severity. The interpretation stage provides a systems-level taxonomy for different components associated with varying levels of injury severity. The approach offers accurate predictions and a transparent and interpretable model, aiding decision-making in road safety management and reducing the occurrence and severity of future accidents.

Three major aspects of the current study are as follows. First, we highlight the variation in accident costs by injury severity level (see, for example, Narayanamoorthy et al., 2013). Second, the proposed approach adequately explains the relationship between contributing risk factors and injury severity. Finally, the XAI framework proposed in this paper accounts for the spatial correlation of these risk factors in predicting injury severity. We show that the interpretation of findings from ML explanation approaches, as well as the aggregation of resulting SHAP (SHapley Additive exPlanations) – an AI method for explaining ML models that are based on the cooperative game theory concept of Shapley values – scores, provides a breakdown of the systems-level taxonomy for various crashes and injuries. Our findings suggest that a comprehensive systems-level taxonomy may provide a high-level understanding of the root causes of traffic accidents, aiding policymakers.

The rest of this paper is organized as follows: Section 2 reviews the literature on XAI and ML studies in the context of automobile traffic crashes and varying levels of injury severity. Section 3 describes the structure and estimation procedure of the proposed framework. Section 4 discusses the data used to validate our approach, the study area description, and key sample characteristics. Section 5 presents and discusses the results of applying the proposed framework to our data on automobile accidents. Section 6 discusses the research implications and its practical contribution, while Section 7 concludes the paper and suggests future research directions.

2. Literature review

This section reviews state-of-the-art XAI approaches utilized in various fields and their potential to enhance the transparency and interpretability of ML models. We will highlight how the implementation of XAI should differ across different fields. Additionally, existing studies on applying data-driven ML models to analyze the characteristics of traffic crash risk factors are examined. Consequently, the contribution of this research from an AI perspective is also addressed.

2.1. Explainable AI: Tools and contexts

XAI is defined as the ability of an ML model to explain its decision-making process in a human-understandable manner (Masello et al., 2023). The need for XAI arises due to many ML models' black-box nature, making it difficult for humans to understand how decisions are made. XAI has been the focus of significant research efforts recently, particularly in mental health (Malhotra & Jindal, 2023) and in-demand forecasting (Jackson & Ivanov, 2023).

Several methodologies, including rule-based models (Arentze & Timmermans, 2007), tree-based models such as Decision Trees (DT) (Johnson et al., 2022; Yan et al., 2023), and Bayesian networks (Wu & Law, 2019) have been proposed to illustrate the use of XAI in the literature. Rule-based models are based on a set of 'if-then' statements, which humans can easily interpret (Huysmans et al., 2011). DT visualizes the decision-making process, making it easy for humans to understand (Yuan et al., 2022). Bayesian networks use probabilistic reasoning to explain the relationships between different variables (Dong, Su et al. 2017; Butz, Schulz et al. 2022). Moreover, recent research has focused on developing model-agnostic methods for XAI, which can be applied to any ML model. These methods include Local Interpretable Model-Agnostic Explanations (LIME) and SHAP. LIME generates local explanations by approximating a complex model with a simpler one, while SHAP uses cooperative game theory concepts to determine the contribution of each feature to the model's prediction (Lundberg & Lee, 2017; Masello et al., 2023).

Yuan et al. (2022) used random forest with SHAP to analyze real-time traffic-flow data in a connected vehicle for real-time safety evaluation, revealing the significant impact of lane difference regarding average speed on real-time safety and demonstrating the potential benefits of using CV data in active traffic management. Ebel et al. (2023) predicted the visual demand of in-vehicle touchscreen interactions. They provided insights into the effect of design factors such as driving automation and vehicle speed on driver distraction. The study demonstrates the potential of the method to inform designers about the implications of their design decisions.

XAI is increasingly important for developing trustworthy and effective ML systems. In this research, XAI is relevant for predicting injury severity in automobile crashes. The proposed model-agnostic approach seeks to provide a transparent and interpretable model that can be used to interpret the predictions of any complex ML model, regardless of its underlying structure, thereby aiding decision-making in road safety management. Furthermore, our XAI approach allows us to quantitatively evaluate overall model behavior and specific predictions. Unlike traditional XAI approaches such as DT and rule-based models, which may fail to capture complex interactions effectively, SHAP can be applied to black-box models, potentially providing insights into decision-making processes. Furthermore, while DT and rule-based models may provide global interpretability, they lack the flexibility to provide local explanations. By understanding specific risk factors contributing to injury severity, decision-makers can develop targeted interventions to reduce the severity of future accidents (Lin et al., 2021). Overall, our XAI has the potential to improve the transparency and interpretability of the proposed ML models, enabling stakeholders and policymakers to understand the decision-making process and build trust in the models' outcomes.

2.2. Traffic accident injury severity

Automobile accidents pose a significant threat to public health, and understanding the factors that contribute to injury severity is essential for developing targeted interventions to reduce the severity of future accidents (Topuz & Delen, 2021). Predicting traffic crash frequency (Castro et al., 2012) and injury severity is crucial for emergency responders, medical personnel, and traffic engineers. Researchers have recently explored using ML and data analytics techniques to develop predictive models for injury severity in traffic crashes.

To address this challenge, researchers have utilized various ML methods, including feature selection and XAI approaches, to improve the accuracy and interpretability of injury severity prediction models. For example, Li et al. (2021) used a DT model to predict injury severity in automobile accidents based on factors such as vehicle type, road type, weather conditions, and driver behavior. Similarly, Azhar et al. (2022) developed a random forest model to predict injury severity in traffic crashes based on driver-related factors such as age, gender, and driving experience. However, these models lack transparency in how the predictions are made, limiting their practical applications.

Rule-based and DT methods: Explainable analytics approaches can help to overcome this challenge by providing insights into the decision-making process of these models (Irarrázaval et al., 2021). Song et al. (2021) developed a DT-based model to predict injury severity in traffic crashes, which can identify the most significant factors contributing to the predictions. In their comparison of different techniques, Taamneh et al. (2017) developed a rule-based model to predict injury severity in automobile accidents based on driver-related factors such as age, gender, and driving experience. The model uses a set of rules that specify the conditions under which the predictions are made, making it easy to understand and interpret. However, DT and rule-based models are specific algorithms that may not be easily applicable to a wide range of models. In contrast, SHAP is model-agnostic and thus can be applied to any ML model, including complex ones like neural networks and ensemble methods.

Gaussian process: We also explored using multi-task learning and Gaussian processes for injury severity prediction in automobile crashes (Yuan et al., 2022) and ML algorithms for injury severity prediction in motorcycle crashes. Yang et al. (2022) used a multi-task learning approach to predict injury severity based on factors such as weather, road conditions, and driver behavior. Nasernejad et al. (2021) and Quintero Mínguez et al. (2019) proposed an approach to injury severity prediction based on Gaussian processes. Wahab and Jiang (2020) used an ML approach to predict injury severity in motorcycle crashes. These studies highlight the importance of injury prediction and provide valuable insights into the factors contributing to injury severity. However, using the Gaussian process has been shown to have limitations in providing interpretability for highly complex models. Furthermore, it may not explicitly capture feature interactions like SHAP does, so their interpretability may be more concerned with understanding uncertainty than individual feature contributions.

Feature selection: Feature selection has also been a recent research focus in injury severity prediction. Li et al. (2021) used sequential feature selection methods to identify the most important factors for predicting injury severity in multi-vehicle crashes. Xie & Zhang (2022) proposed a feature selection approach based on a multi-task learning model for predicting injury severity in pedestrian accidents. Furthermore, in recent studies, XAI has been applied to enhance the identification of risk features associated with injury severity prediction. Angarita-Zapata et al. (2021) proposed an XAI approach based on DT to predict injury severity in bicycle accidents. Amini et al. (2022) used SHAP values to explain the contribution of different risk factors to injury severity prediction in pedestrian accidents.

Several other recent studies have focused on injury severity prediction in various contexts. For example, Asscheman et al. (2023) argue for reconsidering injury severity beyond the maximum abbreviated injury score, considering other risk factors such as hospitalization time and quality of life. Rahimi et al. (2020) investigated the injury severity of single-vehicle truck crashes in a developing country. Ma et al. (2021) proposed a deep-learning analytic framework to predict traffic accident injury severity based on contributing risk factors.

We combined traditional XAI models with well-known statistical methods to determine the best risk features for our prediction models. Furthermore, we utilized some of the ML methodological approaches to injury severity prediction and addressed the nature of the count of automobile accidents by injury severity level. As noted by Narayanamoorthy et al. (2013), identifying a measure of exposure to traffic crash risk and injury severity is an important aspect of modeling crash frequency and the associated level of injury. In the current context, feature selection and engineering, predictive modeling, and XAI approaches have been utilized to provide accuracy and interpretability of traffic accident injury severity prediction. Our study aims to provide valuable insights into risk factors contributing to crash-related injuries and offer a framework for developing targeted interventions to reduce the occurrences and severity of future accidents. The current paper, therefore, has the potential to be of assistance in the evaluation of public policies concerning the management of road safety. This is accomplished through the identification and quantification of measures of exposure to these crash-related risk factors. It may eventually affect the transportation economy by lowering the costs associated with automobile insurance and individual liabilities incurred due to injuries sustained.

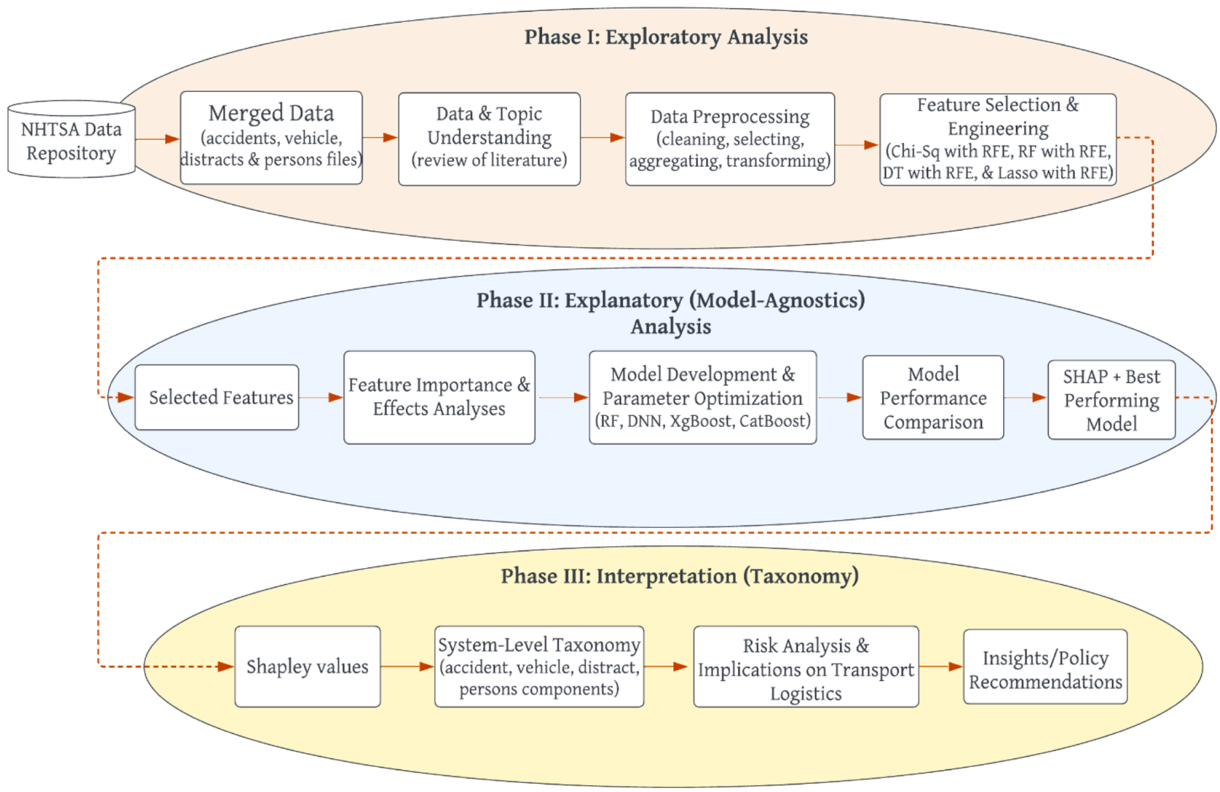


Fig. 1. Display of the proposed XAI methodology.

2.3. Summary of contributions

This research proposes a framework for investigating, analyzing, and explaining individual-level contributing risk factors for injury severity to develop policies to reduce the occurrence and severity of automobile traffic crashes. Insights from this framework can be used to create tailored regional policy plans. We add to the body of knowledge on automobile traffic crashes and injury severity by identifying contributing factors to car accidents. Additiognally, we use local explanation approaches to provide insights into the various causes of automobile crashes as well as the relevant risk factors for each class of injury (i.e., no injury, minor injury, and severe injury), taking into account spatio-temporal and demographic features associated with five constructs: vehicle, crash, human, road, and environmental elements.

3. Methodology

3.1. Overview of the methodology

This paper introduces a three-step, data-driven XAI methodology aimed at pinpointing the principal reasons behind traffic accidents and uncovering the factors that affect the severity of injuries incurred in car accidents. The objective is to enhance road safety and management by clearly understanding these contributing factors. We constructed our methodological artifact based on three analytics building blocks: (i) exploratory, (ii) explanatory (model-agnostic), and (iii) interpretation, which provided a systems-level taxonomy for different components associated with various levels of injury.

The initial exploratory phase involves amalgamating datasets from varied sources, including vehicle, personal, environmental, and accident records, to diagnose and rectify data-related issues, perform data engineering, and identify the most effective features for model construction. This phase also includes mapping out key characteristics of crash causes along with other important attributes. The second phase, explanation, focuses on developing a predictive model utilizing the significant features selected earlier to uncover the complex stochastic relationships among all factors influencing crash causes and injury severities. The final phase, interpretation, uses model-agnostic methods and taxonomy strategies to elucidate each feature's relative importance in the model's overall predictions and the interpretation of the results. Fig. 1 illustrates the analytical methodology employed in this study, detailing the three distinct phases.

3.2. Feature analysis and selection

Feature selection is an important part of model development because it reduces the computational cost of model training by removing unnecessary features and improving model generalization. In the collected and preprocessed dataset, features in our data correspond to variables that appear to influence our target outcome (injury severity). High-quality features are characterized by their informativeness, relevance, interpretability, and lack of redundancy, laying a strong foundation for effective modeling and problem-solving, thereby yielding robust and convincing outcomes (Ma et al., 2021). Hence, conducting feature analysis and selection is essential for refining model performance. In this study, we employed techniques such as data exploration and visualization, filtering, and wrapper methods to delve into the causes of traffic accidents and pinpoint crucial features that affect injury severity in car accidents.

First, we utilized visualization tools to examine spatiotemporal attributes. We conducted a basic descriptive statistical analysis of the features to gain insights into crash causes and potentially influential features on injury severity. Second, we applied an ML strategy through filter methods—CATBoost combined with Shapley values—to prioritize features by their significance and relationship with factors contributing to injury severity. CATBoost is an ML technique based on gradient-boosted decision trees (GBDTs), suitable for feature extraction (Chu et al., 2024), regression (Arora et al., 2021), and classification (Kriebel & Stitz, 2022). Additionally, CatBoost efficiently processes categorical data with a reduced risk of overfitting and adeptly handles datasets with mixed feature types, noisy data, and complex dependencies (Nweke et al., 2019).

In the final step, we implemented a sequential feature selection strategy, training and evaluating ML models on subsets of features and choosing those that enhance accuracy. Specifically, we utilized recursive feature elimination (RFE), a process that iteratively builds models with different feature subsets. This method specifies the number of features to retain as a parameter and uses the entire dataset for training a classifier, which then assigns weights to each feature. Features with the lowest weights are progressively removed. In linear models, weights might correspond to model coefficients or feature importance in tree-based models like DT and RF. This procedure is repeated until the predefined number of features is selected with fitting algorithms such as Chi-square, Lasso, RF, and DT employed in the RFE algorithm execution.

3.3. Traffic crash pattern prediction

ML techniques have gained popularity due to their impressive prediction capabilities. However, they have been criticized for their incomprehensible "black box" techniques and lack of transparency (Dong et al., 2018). Recent research has proposed many approaches, such as LIME and Layer-Wise Relevance Propagation (Bach et al., 2015), to increase the interpretable nature of ML models. These techniques aim8 to apply the standard Shapley regression values in complicated ML models to the cooperative game theory equations to determine each feature's relative importance. Lundberg & Lee (2017) presented SHAP, a novel method that combines Shapley values with agnostic techniques such as LIME and Deep Learning Important FeaTures (DeepLIFT). Using an additive significance score, SHAP determines how much the model's projected outcome would shift if a given feature were considered. The SHAP technique considers all possible feature combinations (without Q). It calculates the average deviation from the prediction when Q(k) is added to the combination in order to determine the importance of a specific instance k of feature Q (i.e., Q(k)). This method trains 2^P models using the entire data set and consistent hyperparameters to ensure that all possible feature combinations are accounted for. To illustrate, suppose models D1 and D2 are trained using the same features, with the addition of feature *g*, for model D2. The SHAP score of feature *g* is determined by subtracting the expected values of D1 and D2 for each data instance. The SHAP score of features is the weighted average of the marginal contributions of *g* to the prediction for every possible combination of features. This is mathematically represented by the following equation (Lundberg & Lee, 2017)

$$SHAP_g(y) = \sum_{\Gamma:g\in\Gamma} \frac{M_\Gamma(y) - M_{\Gamma-g}(y)}{|\Gamma| \times \binom{G}{\Gamma}} \tag{1}$$

Using the predicted outcome of an instance (*y*), the equation calculates the importance score of a feature utilizing a model trained with a subset of features, Γ, that includes the feature in question, g, denoted by $M_\Gamma(y)$. The magnitude of the feature set is denoted by $|\Gamma|$, and the total number of features is denoted by G. The importance score, also known as the change in the expected model prediction, quantifies how many accounts must be taken of the feature before reliable prediction can be made. SHAP extends traditional Shapley values by computing a unique marginal contribution for each feature instance within the dataset rather than assigning a single importance score to the feature as a whole (Lundberg & Lee, 2017). This approach allows for an additive compilation of marginal contributions for each instance, which, when summed together, provides the prediction for that specific instance. This distinctive feature of SHAP enables independent interpretation of each instance, providing a deeper comprehension of the significance of each feature at both the high-level and instance levels. In the present investigation, the predicted risk of sustaining severe injury can be broken down for each driver into the marginal contribution of each predictor. Thus, SHAP values offer a more comprehensive analysis of the feature importance scores.

However, the SHAP analysis can only be trusted if founded on a solid predictive model. As a result, we trained four conventional ML models (XgBoost, CatBoost, Deep Neural Network, and Random Forest) and assessed their performance before calculating the SHAP scores. In order to determine if a given injury resulting from a crash is just a property damage, a minor injury, or severe based on their demographic and other contributing factors, we employ several tried-and-true ML methods. Several methods have been presented in the literature for evaluating the efficacy of prediction models with a multi-class response variable. To evaluate the efficacy of the models, we adopted three criteria from the existing literature, as suggested by Calvino et al. (2022): Measures of within-class accuracy (sensitivity and specificity) and combined performance metrics (F-measure and G-Mean) are discussed, along with the percentage of properly categorized cases (for each class and on average). In order to determine these values, we built a confusion matrix (versus the other two groups) for each category, where *k* represents the number of injury severity classes. We then generated the following performance measures for each class, knowing that true positive (TP) for a class is the total number of instances in which all labels are correctly classified, false positive (FP) for a class is the sum of all instances in the target class's column of a confusion matrix that does not include the TP, true negative (TN) for a given class is the sum of all instances in which a class is classified using labels other than the target label, and false negative (FN) for a class is the sum of all instances in the target class's row in a confusion matrix that do not include the TP, computed as follows:

$$Accuracy_k = \frac{TP_k + TN_k}{TP_k + TN_k + FP_k + FN_k}$$

$$Sensitivity_k = \frac{TP_k}{TP_k + FN_k}$$

$$Specificity_k = \frac{TN_k}{TN_k + FP_k}$$

$$F - measure_k = \frac{2 \times TP_k}{2 \times TP_k + FP_k + FN_k}$$

$$G - mean_k = \frac{TP_k}{\sqrt{(TP_k + FP_k)(TP_k + FN_k)}}$$

Since our study deals with three different classes of injury severity, the measures were generalized into multi-class performance metrics.

3.4. Hyper-Parameter optimization

ML algorithms heavily rely on hyperparameters because they exert direct control over the behaviors of training algorithms and substantially affect the performance of the resulting models (Wu Jia et al., 2019). This issue becomes critical in ML development due to the extensive range of hyperparameters related to network architecture—like the number of layers, neurons per layer, and activation functions—as well as training hyperparameters, including the choice of optimizer algorithm, learning rate, and its decay, batch size, number of training epochs, loss function, and regularization weights. These parameters require constant tuning and specification during the training of ML algorithms. To address this issue, both manual and automated search strategies have been implemented. Several automatic search approaches, such as Bayes optimization (Daziano & Achtnicht, 2014) and grid search (Ameli et al., 2022), have been proposed in the literature to overcome the disadvantages of manual search, which is dependent on the basic knowledge and experience of expert users and is difficult to reproduce. This study used a grid search approach to optimize the training and network architecture hyperparameters. Grid search trains an ML model on the training set with every possible combination of hyperparameter values, evaluates performance on a cross-validation set using a predefined metric, and outputs hyperparameters that achieve the best performance. Since grid search is known to suffer from the curse of dimensionality, we decided not to consider varying each hyperparameter to maintain the algorithm's efficiency by keeping the number of permutations within a reasonable range. In addition, this

Table 1
Distribution of the percentage of injured motorists by injury severity level.

Feature	Category	South	Midwest	West	Northeast
Injury Severity	Class 0	54.17 %	17.40 %	15.84 %	12.58 %
	Class 1	53.79 %	15.80 %	19.46 %	10.96 %
	Class 2	52.71 %	17.04 %	19.63 %	10.62 %
AIR_BAG_DEPL	Not deployed	53.66 %	16.93 %	16.31 %	13.10 %
	Deployed	55.53 %	17.43 %	19.33 %	7.71 %
REST_TYP	Shoulder lap	54.97 %	15.70 %	16.86 %	12.46 %
	Not applicable	46.75 %	23.93 %	18.73 %	10.59 %
	Child rest	49.99 %	26.40 %	15.25 %	8.36 %
	Other rest	35.92 %	42.78 %	14.07 %	7.24 %
ACC_CONF	D	57.71 %	15.39 %	15.60 %	11.31 %
	M	53.10 %	15.38 %	19.00 %	12.52 %
	J	53.02 %	16.97 %	17.75 %	12.26 %
	K	55.36 %	17.01 %	15.25 %	12.38 %
	L	54.87 %	17.55 %	15.24 %	12.33 %
	F	51.49 %	17.70 %	17.79 %	13.02 %
	C	39.73 %	21.31 %	23.05 %	15.91 %
	A	52.05 %	20.74 %	16.36 %	10.85 %
	B	53.12 %	20.24 %	16.84 %	9.81 %
	I	55.40 %	19.39 %	15.17 %	10.04 %
	G	56.37 %	18.77 %	13.31 %	11.55 %
	H	56.33 %	13.29 %	16.46 %	13.92 %
	E	74.31 %	10.09 %	3.67 %	11.93 %
IMPACT1_IM	Front	53.85 %	17.80 %	16.11 %	12.24 %
	Rear	56.79 %	15.43 %	15.98 %	11.79 %
	Left	54.59 %	18.03 %	17.47 %	9.90 %
	Right	54.82 %	18.36 %	17.11 %	9.70 %
	Right Front	44.37 %	15.40 %	23.27 %	16.96 %
	Left Front	46.22 %	14.38 %	23.32 %	16.08 %
	No Collision	42.18 %	17.52 %	25.27 %	15.03 %
	Left Back	47.36 %	13.29 %	21.43 %	17.92 %
	Right Back	50.73 %	14.51 %	23.00 %	11.76 %
	Undercarriage	60.68 %	18.89 %	10.88 %	9.55 %
	Parts in Motion	52.26 %	20.14 %	15.45 %	12.15 %
	Top	56.10 %	22.06 %	18.20 %	3.64 %
BDYTYP_IM_UPD	Passenger cars	53.76 %	15.85 %	16.58 %	13.81 %
	Light trucks vans	54.90 %	18.30 %	16.68 %	10.11 %
	Large trucks	53.31 %	19.20 %	16.80 %	10.69 %
	Motorcycles	47.62 %	16.48 %	25.12 %	10.78 %
	Buses	49.08 %	17.65 %	14.62 %	18.66 %
	Other	43.88 %	27.94 %	23.09 %	5.08 %
MODELYEAR	2015 & newer	54.56 %	14.99 %	16.39 %	14.06 %
	b/w 2010 & 2014	54.82 %	17.67 %	15.04 %	12.48 %
	b/w 2005 & 2009	53.42 %	18.98 %	16.26 %	11.34 %
	b/w 2000 & 2004	52.45 %	18.82 %	19.39 %	9.35 %
	1999 & older	52.60 %	16.71 %	23.97 %	6.72 %
AGE_IM	44 or less	54.09 %	16.62 %	17.05 %	12.24 %
	25 or less	54.90 %	16.84 %	17.73 %	10.53 %
	65 or less	53.23 %	17.33 %	16.08 %	13.35 %
	84 or less	52.50 %	18.50 %	15.10 %	13.90 %
	85 or more	53.50 %	14.22 %	16.03 %	16.25 %
REST_MIS	No	54.86 %	16.44 %	16.64 %	12.06 %
	Not vehicle	45.22 %	23.28 %	19.11 %	12.39 %
	Yes	54.41 %	14.08 %	17.16 %	14.35 %
SEX_IM	Male	53.02 %	17.01 %	17.47 %	12.50 %
	Female	55.20 %	17.04 %	16.13 %	11.64 %

enabled the grid search to be accomplished in an acceptable time span, given the computing system's technical limitations.

4. Study area description and data

This study sourced its accident data from NHTSA database, a program under the United States Department of Transportation. This database, which spans from 1979 to the present, compiles collision data from diverse sources, such as police reports, covering a range of incidents from minor property damages to fatal crashes. The Comprehensive State-by-State Report (CRSS), representing a nationwide sample, includes data from 60 selected locations across the U.S. drawn from nearly seven million police-reported car accidents annually. Data collectors meticulously process thousands of crash reports yearly to guarantee accuracy and reliability. Once encoded and validated, this data becomes publicly available in an electronic file format that supports up to 120 data points.

Our research utilized data from four distinct repositories over three years (2018–2020), covering accidents, distractions, individuals, and vehicles. This data encompasses the extent of property damage, driver specifics, crash scenarios, environmental conditions, and the severity of injuries. Information on driver distractions includes activities like texting, eating, or adjusting the stereo. Demographic and situational details of individuals, as well as vehicle specifics, are also included. A comprehensive dataset for each crash event was created by merging data using unique identifiers. Initially, the dataset contained 372,375 records and 167 variables before cleanup and preprocessing.

These variables cover various accident aspects, including vehicle and driver demographics, road and traffic conditions, environmental factors, and accident circumstances. Given the categorical nature of most variables, dummy variables and ML encoding techniques were employed. Despite being time-intensive, data preparation was crucial for addressing missing data, considered missing at random (MAR), allowing for systematic data imputation, as demonstrated by Ahmed et al. (2023).

Utilizing MAR to address missing data systematically has led to significant advancements in data imputation. For instance, if "DISTRACTED DRIVING" data are missing at random, it could be due to other factors such as "EVENT_IM_MANCOL" (first harmful event, i.e., describing the first injury or damage-producing event of the crash), "INT_HWY" (whether the crash occurred on an interstate highway), and "VSURCOND" (road surface condition). By assuming MAR, it is possible to estimate the missing values in "DISTRACTED_DRIVING" with some accuracy using the known values of these variables.

Injury severity was initially recorded on an eight-point scale, from no injury to fatal. This study, recognizing the importance of differentiating injury severity for road safety analysis, categorized injury severity into three levels: no injury, minor injury, and severe injury, excluding data points for categories with unknown or pre-crash fatalities to enhance dataset consistency (Ma et al., 2021).

Prior studies have focused on binary injury severity (Amini et al., 2022), which may limit the amount of useful information derived from the data. The refined dataset for this analysis contained 358,962 records and 66 variables, facilitating an in-depth examination of

Table 2
Selected features from at least three methods.

Feature	Chi-Sq RFE	RF RFE	DT RFE	Lasso RFE
REGION	√	√	√	
URBANICITY	√	√	√	
PERMVIT	√	√	√	
MONTH	√	√	√	
TYP_INT	√	√	√	
LGTCON_IM	√	√	√	
NUMOCCS	√	√	√	
TRAV_SP	√	√	√	
ROLLOVER	√		√	√
VTRAFWAY	√	√	√	
VNUM_LAN	√	√	√	
VSPD_LIM	√	√	√	√
VPROFILE	√	√	√	
VTRAFCON	√	√	√	√
IMPACT1_IM	√	√	√	√
PCRASH1_IM	√	√	√	√
SEAT_IM	√	√	√	√
AGE_IM	√	√	√	√
ACC_TYPE_CAT	√	√	√	√
ACC_CONF	√	√	√	√
VEH_AGE	√	√	√	√
MODELYEAR	√	√	√	√
AIR_BAG_DEPL	√	√	√	√
BDYTYP_IM_UPD	√	√	√	√
EVENT1_IM_MANCOL	√	√		√
HOUR_DAYTIME	√	√	√	√
REST_TYP	√	√	√	√
WEATHER_IM	√	√	√	√
WEEKDAY_UPD	√	√	√	√
DISTRACTED_DRIVING	√	√	√	√
VE_TOTAL		√	√	√

the factors influencing injury severity. This approach aligns with other studies that utilize ordinal scales for injury severity, providing a detailed framework for understanding and mitigating the impact of traffic accidents on public safety. Other existing studies in the literature that have utilized three ordinal scales include (Ahmed et al., 2023; Topuz & Delen, 2021).

Table 1 shows the distribution of the number of injured motorists by severity level (across all regional tracts, which is a representative sample of all crashes in the United States). The dominant injury types in traffic accidents in the United States are “just property damage” and “minor” injuries, which are mostly in the Southern and Western regions, with a lower share of “minor” and “severe” injuries occurring in the Northeastern region. The southern region witnessed 54 % of all road traffic accidents, while the northeast region was responsible for 12.11 %.

We compare the sample distribution of motorized traffic crashes in the United States’ four major regions. Airbags were not deployed in most traffic accidents across the four regions. AIR_BAG_DEPL (a feature that indicates whether or not an airbag is deployed during a specific crash) is known to indicate the possibility of a severe injury. Other factors suspected of contributing to the injury severity in car accidents include the area of impact on the vehicle, IMPACT1_IM, and the type of restraint used by the driver. The southern region has nearly 60.68 % of traffic accidents involving undercarriage. Moreover, Table 1 illustrates that the number and distribution of injuries across census tracts differ markedly based on region, feature type, and severity level. This highlights the necessity of analyzing injury counts according to the contributing risk feature and severity level rather than aggregating all injuries into a single category.

5. Results

5.1. Exploratory analysis

The results of the feature analysis and selection algorithms utilized in the present study are provided in Table 2. A total of thirty-one (31) features were chosen by at least three out of the four feature techniques that were implemented.

First, we examine the distribution of temporal traffic accidents over three years from 2018 to 2020, revealing distinct levels of

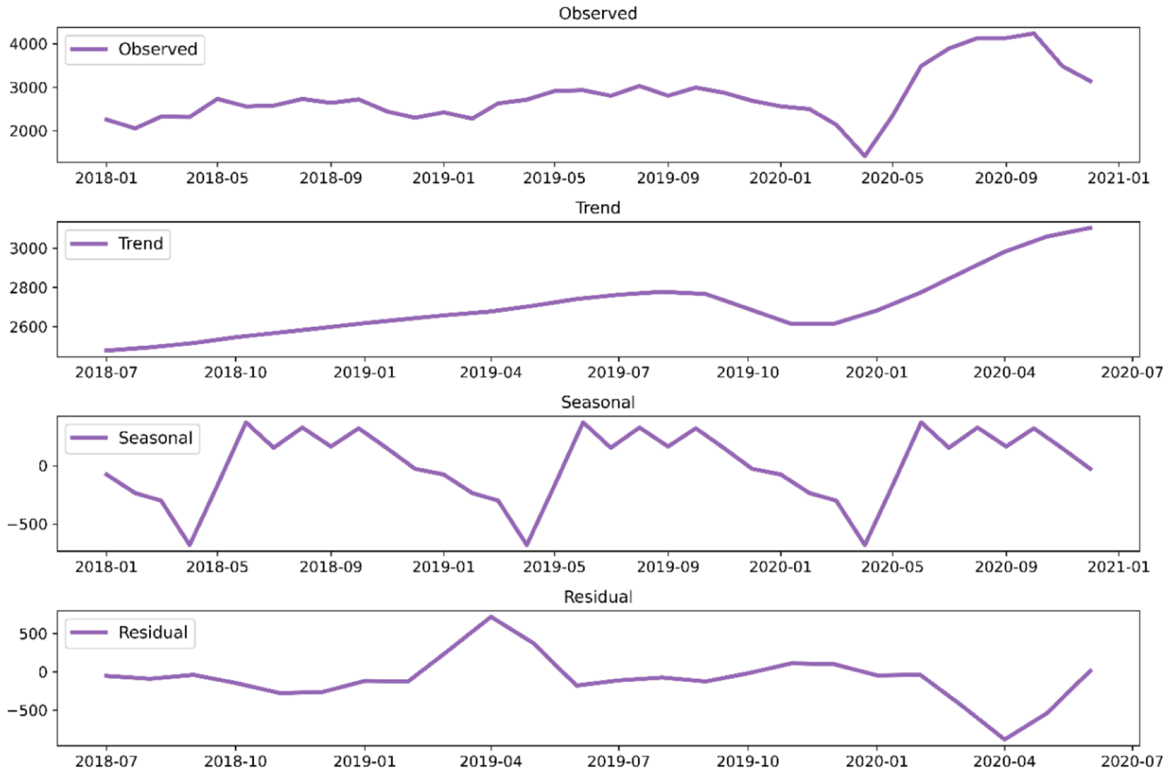
a) Annual crash frequency by severity level

b) Monthly crash frequency by severity level

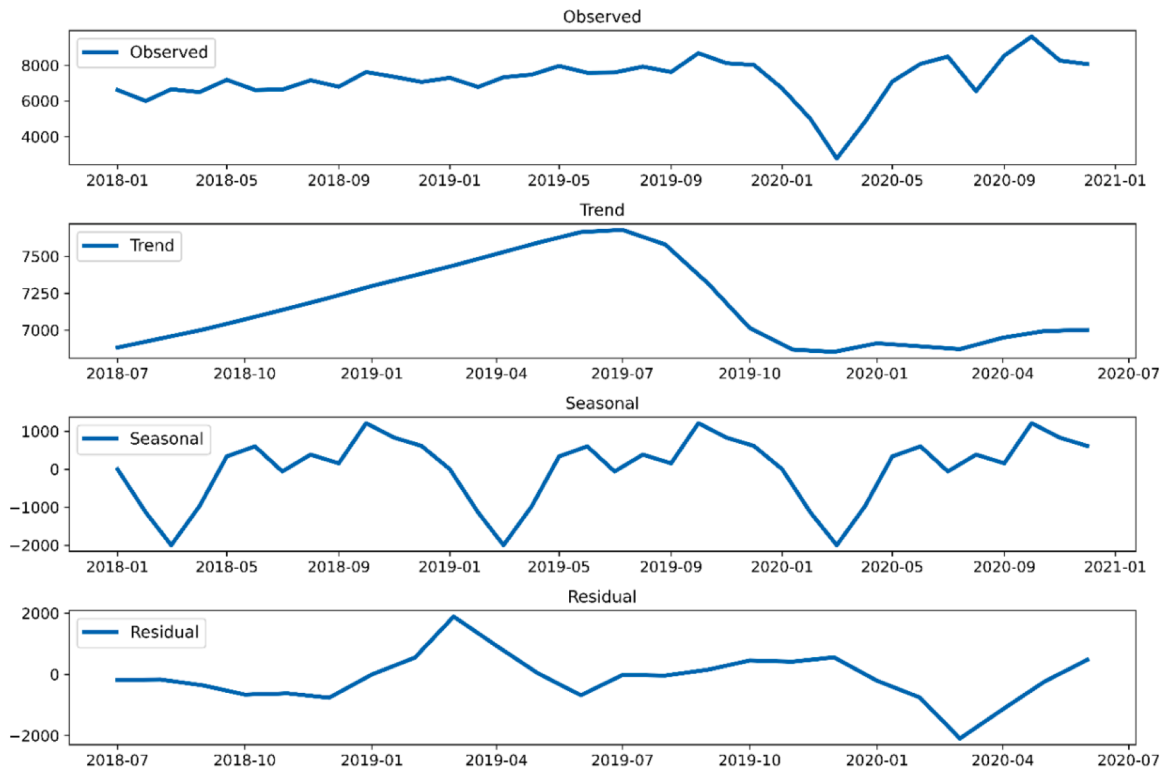
c) Weekly crash frequency by severity level

d) Hour-of-day crash frequency by severity level

Fig. 2. Temporal distribution of traffic injury severity by hour, day, week, month, and year.



a) Accidents that result in minor and severe injuries.



b) Accidents with no injuries (i.e., just property damage).

Fig. 3. Monthly decomposition of traffic accident.

injury severity. Fig. 2(a) depicts the annual distribution of traffic crashes with varying levels of injury severity. It shows that the number of crashes with minor to severe injuries increased yearly, while the category of just property damage (i.e., class 0) remained consistent over the three years. Fig. 2(b) shows a continuous increase in the number of crashes from May to October, with October having the highest crash occurrence and April having the lowest. In addition, traffic crashes resulting in minor to severe injuries occur between May and October, with the highest number occurring in August and October. Most traffic crashes during these months appear to have resulted in no injuries (just property damage). Fig. 2(c) shows that weekdays have more traffic accidents than weekends.

Furthermore, traffic accidents leading to injuries ranging from minor to severe were most frequent on Fridays, significantly surpassing those occurring on Sundays, which saw the lowest number of traffic accidents and injuries of any severity. According to Fig. 2 (d), the highest occurrence of traffic accidents was observed between 2:00p.m. and 4:00p.m., with subsequent peaks from 5:00p.m. to 7:00p.m. and from 11:00 a.m. to 1:00p.m. These times align with periods of commuting, lunch breaks, and the evening return from work. The afternoon and evening accidents had a higher likelihood of fatalities than those in the early morning. This result suggests that fatal crashes are more common in the early afternoon. During this time of day, speeds are typically higher, and the impact of distracted driving is even more significant.

Moving forward, we aim to expand our understanding of the patterns within our automobile traffic crash time-series data by breaking down the time series. This step will allow us to better grasp the trends involved. Fig. 3 depicts the temporal pattern of various levels of traffic accident injury severity. The trend for traffic crashes that resulted in minor- and severe injury over the three years is shown in Fig. 3a), while the trend for no-injury traffic crashes (i.e., traffic crashes that resulted in only property damage) is depicted in Fig. 3b). Both show significant increasing trends and seasonal trends until around March 2020, when the United States government declared a nationwide lockdown due to the outbreak of COVID-19 disease. During this time, there was a significant decrease in the number of traffic crashes and injuries; however, traffic crashes showed a significantly increased trend beginning in May 2020.

Additionally, there was a notable seasonal decline in the number of injuries and crashes at the beginning of each spring, which was followed by an exponential rise in traffic accidents from June through October. This could be due to more travel during the summer and fall seasons. The increasing trend is stronger for crashes resulting in minor and severe injuries; additionally, traffic accident occurrences appear to peak around the fall and then decrease in the winter. Between 2018 and 2020, traffic accidents that resulted in only property damage or incidents without injuries experienced three peaks during the fall and winter seasons.

5.2. Explanatory analysis

To investigate and understand risk features likely influencing crash causes and contributing to the increase in the level of injury sustained by automobile occupants, we utilized the Shapley value in conjunction with our best-performing ML model obtained in this study. However, first, we performed data preprocessing, which included data splitting, normalization, and resampling. We used label encoding to convert our data's categorical features into numerical features by assigning each category a unique integer value. This technique preserves the ordinal relationship between categories and can improve overall prediction accuracy. Furthermore, one-hot encoding was used to convert categorical features into binary features usable in ML models. In one-hot encoding, a binary feature is created for each category, with a value of one indicating that the sample belongs to that category and zero indicating that it does not. This technique prevents ML models from assuming an ordinal relationship between categories, which can improve prediction accuracy overall.

Second, we normalized the features to ensure they were comparable and had the same scale. This technique prevents ML models from being biased toward features with higher values, which can improve prediction accuracy overall. In this study, we specifically used the StandardScaler technique, which scales the features to have a mean equal to zero and a standard deviation equal to one.

Third, we address the target class's imbalance by oversampling the data to increase the number of other classes, making it more

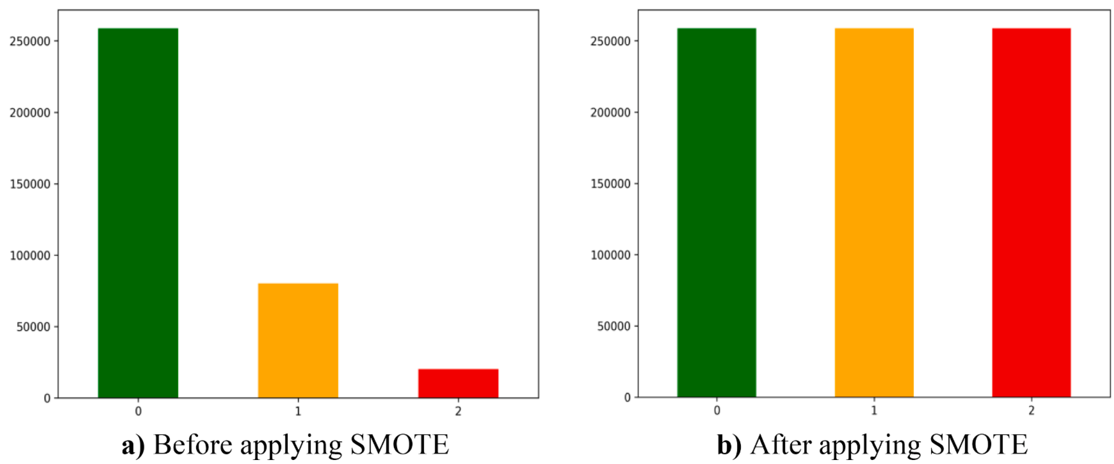


Fig. 4. Traffic accident injury severity instances across three target classes.

representative of the real-world scenario. The oversampling technique entails duplicating samples from the minority class at random until the class distribution is balanced. This technique prevents ML models from being biased towards the majority class and can improve prediction accuracy overall. Because the dataset contains more traffic crashes that resulted in no injury, we used the Synthetic Minority Oversampling Technique (SMOTE) (Zhang et al., 2022). Prior to the technique, the proportion of reported cases with severe injuries was much lower than the proportion of cases with minor injuries or no injuries at all, which stood at 72.08 % for no injury, 22.33 % for minor injury, and 5.59 % for severe injury cases (i.e., 258,750 no injury cases, 80,146 minor injury cases, and 20,066 cases of severe injuries). After applying the SMOTE technique to the data, the total number of injuries across all three classes is 258,750 (33.33 %, 33.33 %, and 33.33 %).

Fig. 4 shows the number of reported cases in the three target classes before and after using the SMOTE technique.

The data was then split into 80 % training, 20 % testing, and multiple classification models were built using a variety of conventional ML algorithms before selecting the most accurate one as the premise for calculating SHAP scores. Table 3 presents the ML models' parameters with the possible range of values.

Next, we trained and validated RF, DNN, XGBoost, and CATBoost models to represent search- and optimization-based algorithms. We then utilized a 10-fold cross-validation strategy to ensure the models' performances were comparable. The performance of the predictive models on the test dataset at each distinct level of the response variable is displayed in Table 4 below.

In Table 4, we presented the overall performance of the models on the test dataset for four different methods, revealing that the RF model has significantly higher accuracy than XGBoost, DNN, and CATBoost, especially when considering the three combined performance metrics for each level of the response variable: accuracy, sensitivity, and specificity. Consequently, we utilized the RF algorithm to conduct SHAP analysis and ascertain the significance of predictive factors. To quantify the impact of our model, we compare the current study's predictive performance to previous studies on traffic crash injury severity prediction in Table 5 below.

Despite a significant reduction in risk features in the current study, our findings were consistent with and improved upon recent related studies. Furthermore, some of these studies used binary response variables, whereas in our case, we included three different traffic crash injury severity levels.

Since the SHAP algorithm requires a binary classification problem to compute feature importance scores, we converted our multi-class classification problem into three binary classifications, each classifying instances in one class against the other two classes. Using K-Means clustering to divide the sample into a relatively small number of clusters and then selecting a random instance from each cluster as the representative is a common method for SHAP analysis with large samples. To represent the actual distribution of the sample more accurately, the number of instances in each cluster is used as the representative's weight in the SHAP score calculation procedure. Here, we grouped the sample (40,000 instances) into 50 clusters and selected random representatives for the SHAP analysis. Fig. 5 displays the SHAP summary plot, illustrating the importance and impact of input features on the severity of injuries resulting from traffic accidents.

According to the feature importance ranking detailed above, the four factors most influencing the severity of injuries in automobile accidents are Air Bag Deployment (indicating if an airbag was deployed during the crash), Restraint Types (such as seat belts), Accident Configuration, and BDYTYP IM UPD (referring to the body type of vehicles involved in the accident). Environmental factors, including weather conditions and the state of the road surface, play a much smaller role in determining the severity of crash injuries. The vehicle's make is the feature having the least impact.

In addition, Fig. 6(a) shows a SHAP summary plot for each class for the top ten features. It emphasizes the significance of characteristics that determine different levels of injury severity for each class. Airbag Deployment, Restriction Type, and Accident

Table 3
Range of possible ML hyper-tuned parameters.

Model Name	Hyper-tuned Parameters	Values range
XGBoost	Learning rate	0.01—0.1
	Max depth	16—20
	Gamma	0.1—0.4
	Colsample by tree	0.4—0.9
	Reg alpha	0.01—10
	Reg lambda	0.01 – 10
CatBoost	Iterations	600—1000
	Tree depth	10—16
	Learning rate	0.001 – 1
	L2 leaf regularization	0.1 – 2
RF	Max depth	10 – 30
	Max features	'auto', 'sqrt', 'log2', 'None'
	Min sample leaf	0 – 0.5
	Min sample split	0 – 1
	Number of estimators	100 – 1500
DNN	Number of hidden layers	2 – 10
	Hidden layer size	6 – 512
	Learning rate	0.0001 – 1
	Dropout rate	– 0.5
	epoch	100—10000
	Batch size	16—1028

Table 4
Performance comparison of different ML models.

Algorithm	Overall accuracy	Severity level	Accuracy	Sensitivity	Specificity	F-Measure	G-Mean
XGBoost	0.9443	No injury	0.946	0.886	0.976	0.916	0.930
		Minor	0.947	0.946	0.948	0.923	0.947
		Severe	0.995	1.0	0.993	0.993	0.996
RF	0.9532	No injury	0.951	0.908	0.972	0.924	0.939
		Minor	0.951	0.937	0.958	0.928	0.948
		Severe	0.996	1.0	0.993	0.993	0.997
DNN	0.7200	No injury	0.858	0.666	0.955	0.758	0.797
		Minor	0.860	0.902	0.839	0.811	0.870
		Severe	0.992	0.999	0.989	0.990	0.994
CATBoost	0.9484	No injury	0.95	0.888	0.981	0.922	0.933
		Minor	0.951	0.957	0.948	0.929	0.952
		Severe	0.996	1.0	0.994	0.994	0.997

Table 5
Comparison of the predictive performance of present and prior studies.

Study	Accuracy	Sensitivity	Specificity	AUC
Present study*	**0.9532**	**0.948**	**0.974**	**0.961**
Zhang & Abdel-Aty (2022)	0.773	0.772	0.942	0.857
Delen et al. (2017)	0.904	0.8855	0.9207	0.928
Candefjord et al. (2021)	NA	NA	NA	0.86
Formosa et al. (2020)	0.895	0.85	NA	0.935
P. Li et al. (2020)	NA	0.868	NA	0.932
Zeng & Huang (2014)	0.5491	NA	NA	NA
Talmor et al. (2006)	NA	NA	NA	0.7940
Wang et al. (2021)	0.854	0.853	0.854	NA
C. Chen et al. (2016)	0.6273	NA	NA	0.627
Jianfeng et al. (2019)	0.8831	NA	NA	NA
Mujalli & de Oña (2011)	0.60	0.70	0.48	0.63

* The table reports the average of each measure across the three response variable levels.

Configuration are the most affected features for determining whether a crash resulted in only property damage (i.e., no injury) or severe injury classes. The main contributors to minor injury are found to be airbag deployment and the gender of the driver and occupants in the vehicles. Fig. 6(b) illustrates the effect range and distribution of various features on traffic accident injury severity across different levels. The scatter plot represents the Shapley values for each feature, with points colored from blue (low value) to red (high value) according to their impact. The density of these points indicates how frequently they occur in the dataset. For example, a higher number of vehicle occupants during a crash correlates with higher Shapley values, suggesting a significant influence on the severity of injuries from accidents. A similar pattern is observed with the number of vehicles being towed, supporting the notion that an increase in the number of vehicles involved elevates the risk of fatal injuries. Moreover, airbag deployment is linked to higher Shapley values, indicating a greater potential for causing severe injuries in accidents.

5.3. Interpretation: Taxonomy

In this section, we use a systems-level taxonomy to classify contributing risk factors in order to better understand the crash causes that resulted in various levels of injury during automobile traffic accidents. These elements are divided into five categories: vehicle, road, environment, crash, and human. This all-encompassing approach allows us to understand the mechanisms underlying crash causes and identify the high-level causal factors related to different injury severity classes. We use the SHAP method to learn how each construct contributes to injury severity. This approach allows us to examine the influence of each systematic construct in aggregate (across the three severity classes) and individually for each severity class, allowing us to study the specific risk factors contributing to each injury severity class. Using the system-level taxonomy, dissecting the severity of injuries sustained in automobile crashes into several distinct elements is possible. When these elements are subdivided into more manageable categories, researchers and policy-makers can concentrate on the most important aspects of vehicle safety and contribute to reducing accident rates and severity.

Fig. 7 provides a comprehensive system-level taxonomy for different components associated with various levels of severity of injuries. Aggregate SHAP scores for each component variable were utilized to analyze their impact on each injury severity class.

In Fig. 7a), we present a summary and dissection of the interpretation of the system-level taxonomy across all feature categories using aggregated SHAP scores. The figure provides insight into how individual factors influence the severity of traffic accidents across different levels, emphasizing the value of applying normalized aggregated SHAP values for each construct. This method is crucial for grasping the comprehensive impact of each category within the systems-level taxonomy on various injury severity levels, showcasing the significance of each element in determining the severity outcomes of traffic incidents. Thus, based on the heights of the aggregate bars, we observed that human factors contribute the most to injuries, while road factors contribute the least to injuries sustained in an

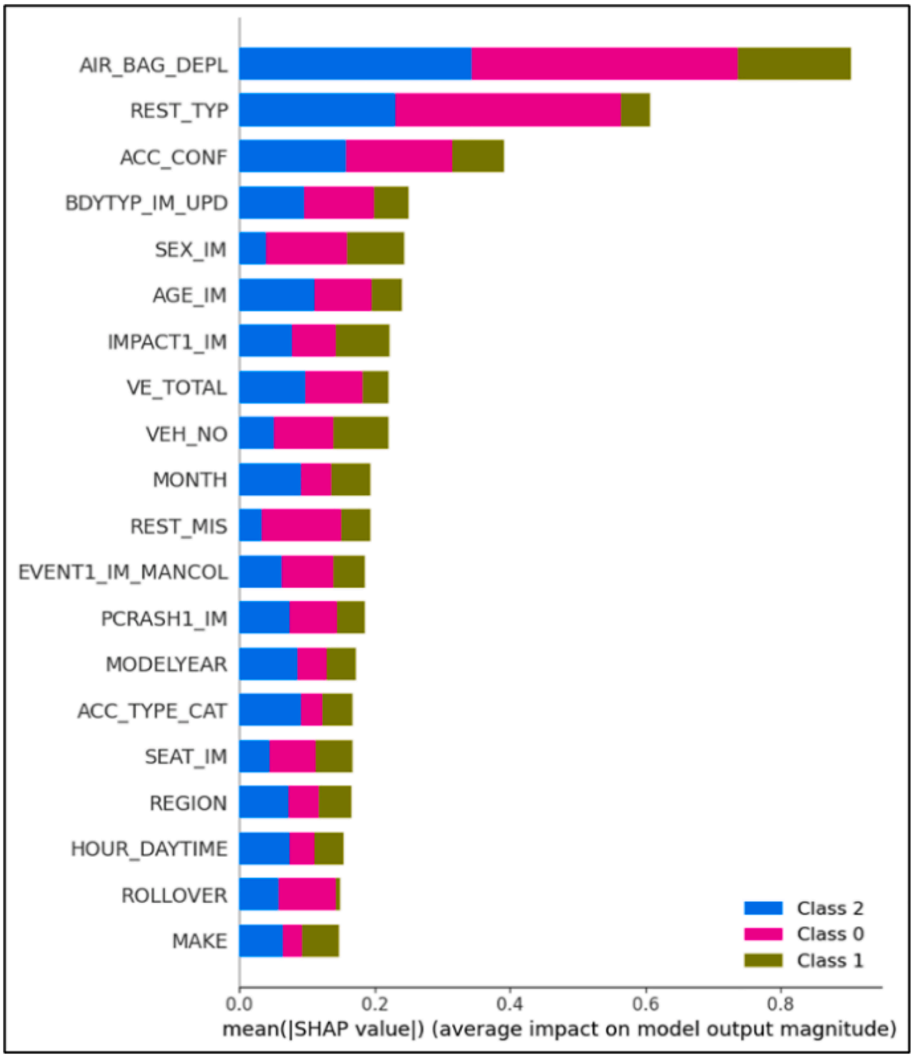


Fig. 5. Aggregate SHAP scores for top 20 influential features.

automobile traffic accident. The SHAP score determines the severity level, which subdivides each bar into classes. Traffic accidents that result in only property damage (class 0) and those leading to severe injuries (class 2) tend to have higher scores for the majority of factors compared to the other category (minor injuries). Meanwhile, minor injuries (class 1) receive the lowest scores across the elements.

Furthermore, we divide the system-level taxonomy into different elements for each type of injury, as shown in Fig. 7b), 7c), and 7d). This allows us to gain a comprehensive understanding of the risk factors that contribute to each injury in each class. Based on our findings, traffic accidents with just property damages are mostly caused by human error because it is the bar with the highest aggregate SHAP score. This is largely due to human factors such as distracted driving, the influence of drug or alcohol use while driving, and the driver's use of a different seating position. In addition, crash elements and vehicle elements have been found to have contributed to minor (class 1) and severe (class 2) injuries. The road and human elements contribute the least to class 1 and 2 injuries sustained by drivers and passengers.

6. Discussions

This paper presents a three-stage decision support framework that employs the use of XAI to create an approach for identifying contributing risk variables for automobile traffic accidents and injury severity in the United States. The framework presented here was informed and guided by CRSS, which served as a representative sample of the entire country and was evaluated by traditional ML models to achieve better prediction accuracy to determine the severity of injuries in automobile collisions.

The framework presented in this research aligns with the growing imperative for transparency and interpretability in AI models, a theme echoed by Dong et al. (2018). These studies call for AI that not only predicts but also explains its reasoning, making it invaluable

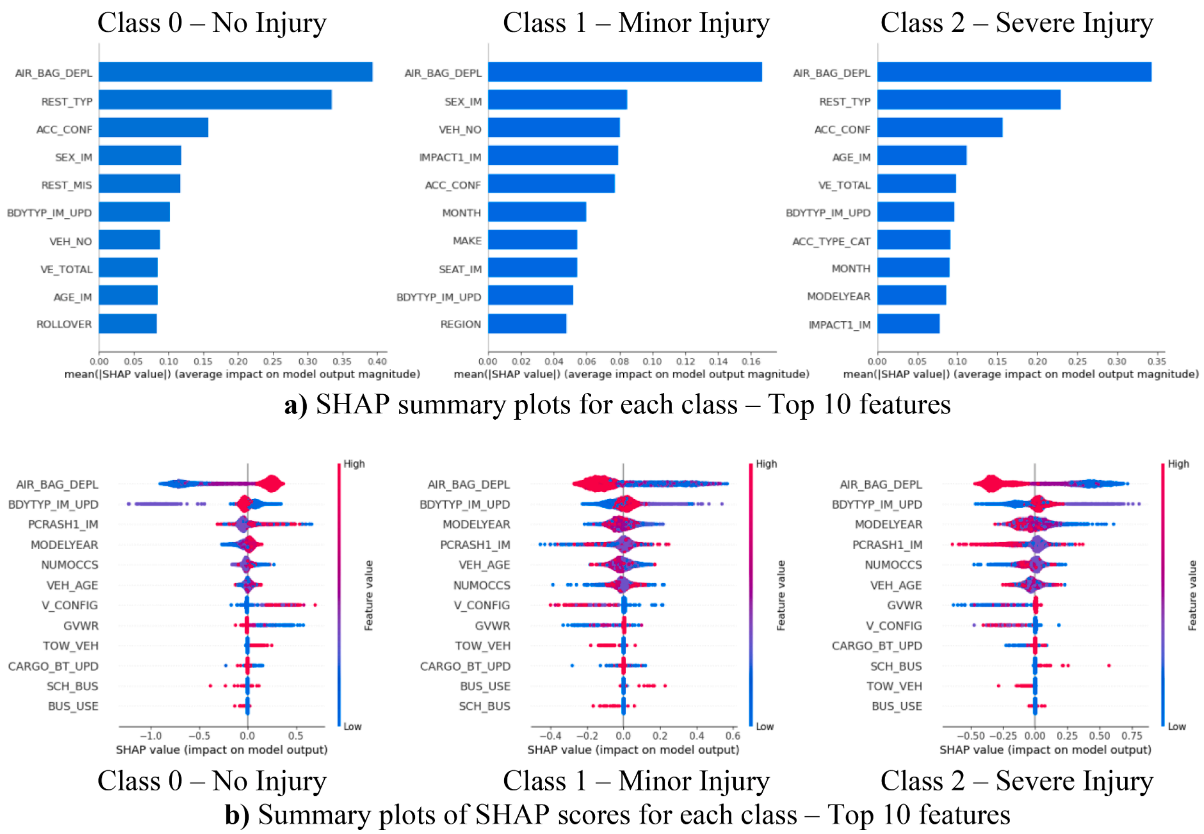


Fig. 6. SHAP summary plots and scores for each class.

for applications in which understanding the 'why' behind predictions is as important as the predictions themselves. This study's use of XAI approaches addresses this need by dissecting complex, non-linear relationships between a plethora of risk factors and injury outcomes. The use of SHAP in this work represents an application of model-agnostic interpretability in traffic safety, enabling stakeholders to make sense of predictions across a variety of ML algorithms. This is particularly relevant in the context of the findings by Lisboa et al. (2023), which advocate for the integration of XAI in domains where decision-making has profound implications, such as public safety and healthcare. By providing interpretability, this study offers an analytical lens through which policymakers can view and assess traffic safety interventions.

Our study contributes uniquely to the field by combining traditional statistical methods with XAI to unravel the significance of risk features for injury severity prediction. While previous studies, such as those by Filom et al. (2022), Chen et al. (2024), and Yang et al. (2023), have applied AI in various contexts, our research extends the interpretability to feature-specific contributions to injury severity. This specificity allows for targeted safety measures, potentially reducing the frequency and severity of traffic accidents—a contribution that stands out in the landscape of traffic safety literature.

Moreover, this research highlights the importance of human factors in crash risks and injury severity, affirming the findings of Wei et al. (2022), while providing novel insights into the causative factors leading to different injury severities. The study's findings also resonate with recent discussions in the field, such as the works of Ma et al. (2021), which emphasize the need for granular and interpretable models in traffic safety.

6.1. Insights for researchers

This paper demonstrates the design of computational algorithms using a "white box" XAI approach to address injury severity issues caused by automobile traffic accidents. We built a unique XAI pipeline that combines and configures various components of ML algorithms, such as XGBoost, CatBoost, DNN, RF, and SHAP. White box approaches are a significant advancement over the black box approaches commonly used in most AI/ML applications, and they are critical for AI transparency, interpretability, and acceptance in mission-critical applications, as well as for ensuring confidence, trust, and acceptance in practical road safety settings.

Our research shows that XAI-enabled tools integrate a single analysis of injury severity from a traffic accident into a comprehensive analytic framework and performing the corresponding processing and analysis of accident data using various ML approaches is the core concept of the method. In the proposed method, (1) different factors to injury severity are analyzed by combining an ML approach (i.e., RF) with the Shapley value to understand the important features and contribution; (2) data-related issues, data engineering and

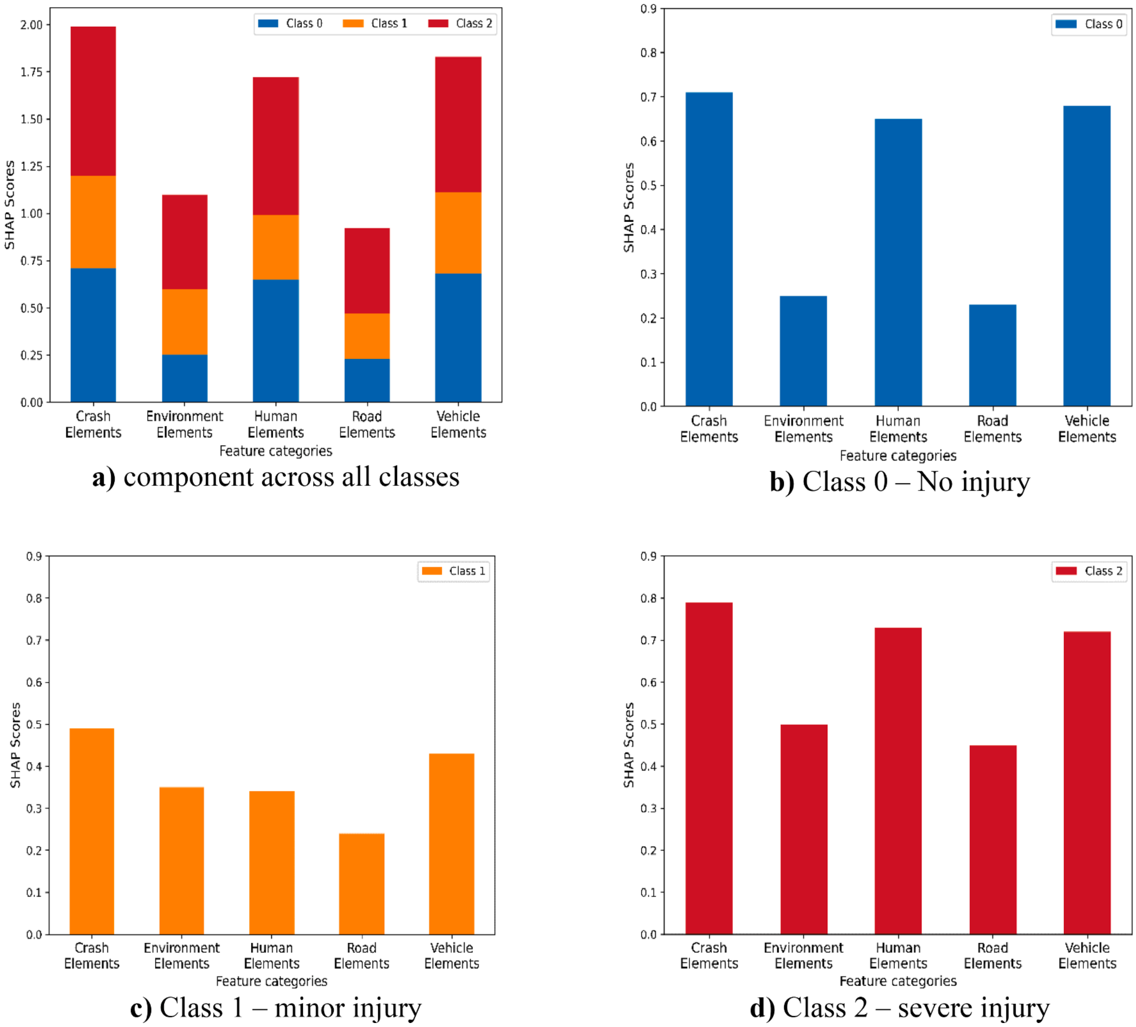


Fig. 7. System-level classification of injury severity-related components.

the selection of the optimal set of features for model building are addressed. In addition to plotting influential characteristics of crash causes alongside significant factors, (3) multiple traditional ML models were employed to predict the absence of injury, minor injuries, and severe injuries for each data class.

The results of analyzing a three-year traffic accident dataset from the United States indicate that (1) the importance and dependence of contributing risk factors can be obtained using the RF model and the Shapley value, (2) the RF model outperformed baseline deep learning models, and (3) SHAP was employed and provided an understanding of the hidden stochastic relationships between all factors that are associated with crash causes. The proposed analytic framework is applicable to other datasets that contain more information about automobile collisions, thereby mitigating the negative effects of traffic accidents on the safety of road users.

6.2. Practical contributions

We suggest policy options for five key feature components, namely, crash, environmental, human, road, and vehicle components that decision makers in public transportation agencies could implement to reduce the risk of on-road crashes. This study discovered that human component features directly influence an increase in crash risk and the severity of injury for all three categories. Several human factors influencing the likelihood of a road traffic crash, or a serious injury have been identified. Nevertheless, driver distraction is one of the leading causes of road traffic fatalities and injuries.

Other studies have also shown that driver distraction is a result of manual distractions (e.g., hands are off the wheel for other non-driving activities), visual distractions (e.g., not watching the road), cognitive distractions (e.g., the driver is not focused on the driving task), or the use of technology or inattention has led to more crashes and injuries (Michelaraki et al., 2023). However, none of these studies specifically identified the severity of the injury to which the human components contribute. We have shown that this category's features are more likely to cause property damage than minor or severe injury. Furthermore, when developing any public safety or technology-based system interventions, crash and vehicle components must be considered, as the feature sets from these components significantly contribute to determining minor injuries.

Additionally, measures such as enforcing regulations to ensure automobile companies follow certain standards during the manufacturing process, prohibiting drivers from engaging in certain activities while driving, organizing regular training for truck drivers to improve safe driving practices and fatigue management, and so on, could be implemented to reduce potential road crashes and thus reduce the rate of severe injuries.

7. Conclusions

This paper has advanced traffic safety research by developing and implementing a novel XAI framework that improves our understanding of the factors contributing to injury severity in car accidents. During the preprocessing phase, we used MAR with Bayesian dynamic conditional imputation to address missing records and SMOTE to address data imbalances. Our approach has highlighted the complex interplay of various risk factors, including driver behavior, vehicle characteristics, road conditions, and environmental influences. To strike a balance between explainability and prediction accuracy, we used advanced machine learning algorithms (XGBoost, DNN, RF, and CatBoost) in conjunction with the powerful interpretative tool SHAP. We provided an impressively accurate predictive model and achieved a high level of interpretability, bridging the gap between data-driven insights and practical knowledge.

Our findings provide a nuanced view into the dynamics of traffic accidents, revealing how certain environmental, vehicle, and driver characteristics variables, such as airbag deployment (AIR_BAG_DEPL), rest type (REST_TYP), the body type of vehicles involved in the accident (BDYTYP_IM_UPD), gender (SEX_IM), age (AGE_IM), and the driver's seating position at the time of the crash (SEAT_IM), play critical roles in the severity of injuries sustained during crashes. Using the SHAP scores, we used a model-agnostic interpretation approach to separate explanations and models by developing a systems-level taxonomy, which has enabled us to group these risk factors into coherent categories, resulting in a better understanding of their individual and collective effects. Such insights are invaluable to policymakers and public transportation agencies, as they provide a solid foundation for developing targeted safety interventions and regulatory measures to reduce the risk of road traffic accidents.

Furthermore, this study demonstrates that using black-box ML models in conjunction with SHAP can aid in extracting useful explanations from complex models, thus enhancing transparency and trust in AI-driven decision support systems. By providing a clear exposition of the decision-making process, XAI paves the way for more informed and confident applications of AI in critical domains in which understanding the 'why' behind predictions is as important as the predictions themselves.

While this study presents a comprehensive analysis leveraging XAI to understand traffic accident risk factors, it has limitations. The reliance on historical data may not fully capture the rapidly evolving nature of driving behaviors, especially under unique circumstances like the COVID-19 pandemic. Future research should consider real-time data analysis to capture more dynamic changes in driving patterns. Additionally, incorporating more granular data regarding road conditions, driver-specific variables (such as psychological factors), and the impact of emerging technologies (e.g., autonomous vehicles) could further refine our understanding of injury severity predictors.

Future research could explore the long-term impact of the pandemic on traffic crash changes by severity level, exploring how changes in travel behavior influenced by the pandemic may persist or evolve (see, for example, Fig. 3(a) and Fig. 3(b)). Investigating the effectiveness of policy interventions implemented during the pandemic and their lasting effects on traffic safety could provide valuable lessons for managing future public health crises. Moreover, the role of technology, including the use of AI to enhance road safety during and after the pandemic, presents a fruitful area for exploration. The pandemic has accelerated the adoption of technology in monitoring and managing traffic flows; thus, assessing its effectiveness and potential for integration into standard traffic safety practices could be another direction for research.

Furthermore, as highlighted in our findings, the variation in accident costs by injury severity level suggests an avenue for economic analysis of traffic accidents in the post-pandemic era. Evaluating the cost-effectiveness of different safety measures and interventions could support more informed decision-making by transportation authorities and policymakers.

CRediT authorship contribution statement

Ismail Abdulrashid: Conceptualization. **Reza Zanjirani Farahani:** Supervision. **Shamkhal Mammadov:** Formal analysis. **Mohamed Khalafalla:** Validation. **Wen-Chyuan Chiang:** Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

REFERENCES

Ahmed, A., Topuz, K., Moqbel, M., Abdulrashid, I., 2023. What makes accidents severe! Explainable analytics framework with parameter optimization. Eur. J. Oper. Res. https://doi.org/10.1016/j.ejor.2023.11.013.

Ameli, M., Faradonbeh, M.S.S., Lebacque, J.-P., Abouee-Mehrizi, H., Leclercq, L., 2022. Departure time choice models in urban transportation systems based on mean field games. Transp. Sci. 56 (6), 1483–1504. https://doi.org/10.1287/trsc.2022.1147.

Amini, M., Bagheri, A., Delen, D., 2022. Discovering injury severity risk factors in automobile crashes: a hybrid explainable AI framework for decision support. Reliab. Eng. Syst. Saf. 226, 108720 https://doi.org/10.1016/j.ress.2022.108720.

Angarita-Zapata, J.S., Maestre-Gongora, G., Calderín, J.F., 2021. A bibliometric analysis and benchmark of machine learning and AutoML in crash severity prediction: the case study of three colombian cities. Sensors 21 (24), Article 24. https://doi.org/10.3390/s21248401.

Arentze, T., Timmermans, H., 2007. Parametric action decision trees: Incorporating continuous attribute variables into rule-based models of discrete choice. Transp. Res. B Methodol. 41 (7), 772–783. https://doi.org/10.1016/j.trb.2007.01.001.

Arora, R., Agrawal, A., Arora, R., Poonia, R.C., Madaan, V., 2021. Prediction and forecasting of COVID-19 outbreak using regression and ARIMA models. J. Interdiscip. Math. 24 (1), 227–243. https://doi.org/10.1080/09720502.2020.1840075.

Asscheman, S., Versteeg, M., Panneman, M., Kemler, E., 2023. Reconsidering injury severity: Looking beyond the maximum abbreviated injury score. Accid. Anal. Prev. 186, 107045 https://doi.org/10.1016/j.aap.2023.107045.

Azhar, A., Ariff, N.M., Bakar, M.A.A., Roslan, A., 2022. Classification of driver injury severity for accidents involving heavy vehicles with decision tree and random forest. Sustainability 14 (7), Article 7. https://doi.org/10.3390/su14074101.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., Samek, W., 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One 10 (7), e0130140.

Calvino, J.J., López-Haro, M., Muñoz-Ocaña, J.M., Puerto, J., Rodríguez-Chía, A.M., 2022. Segmentation of scanning-transmission electron microscopy images using the ordered median problem. Eur. J. Oper. Res. 302 (2), 671–687. https://doi.org/10.1016/j.ejor.2022.01.022.

Candefjord, S., Sheikh Muhammad, A., Bangalore, P., Buendia, R., 2021. On scene injury severity prediction (OSISP) machine learning algorithms for motor vehicle crash occupants in US. J. Transp. Health 22, 101124. https://doi.org/10.1016/j.jth.2021.101124.

Castro, M., Paleti, R., Bhat, C.R., 2012. A latent variable representation of count data models to accommodate spatial and temporal dependence: application to predicting crash frequency at intersections. Transp. Res. B Methodol. 46 (1), 253–272. https://doi.org/10.1016/j.trb.2011.09.007.

Chen, C., Zhang, G., Yang, J., Milton, J.C., Alcántara, A., "Dely", 2016. An explanatory analysis of driver injury severity in rear-end crashes using a decision table/ Naïve Bayes (DTNB) hybrid classifier. Accid. Anal. Prev. 90, 95–107. https://doi.org/10.1016/j.aap.2016.02.002.

Chu, N., Ng, K.K.H., Liu, Y., Hon, K.K., Chan, P.W., Li, J., Zhang, X., 2024. Assessment of approach separation with probabilistic aircraft wake vortex recognition via deep learning. Transport. Res. Part e: Logist. Transport. Rev. 181, 103387 https://doi.org/10.1016/j.tre.2023.103387.

Daziano, R.A., Achtnicht, M., 2014. Forecasting adoption of ultra-low-emission vehicles using bayes estimates of a multinomial probit model and the GHK simulator. Transp. Sci. 48 (4), 671–683. https://doi.org/10.1287/trsc.2013.0464.

Delen, D., Tomak, L., Topuz, K., Eryarsoy, E., 2017. Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods. J. Transp. Health 4, 118–131. https://doi.org/10.1016/j.jth.2017.01.009.

Dong, C., Shao, C., Clarke, D.B., Nambisan, S.S., 2018. An innovative approach for traffic crash estimation and prediction on accommodating unobserved heterogeneities. Transp. Res. B Methodol. 118, 407–428. https://doi.org/10.1016/j.trb.2018.10.020.

Ebel, P., Lingenfelder, C., Vogelsang, A., 2023. On the forces of driver distraction: Explainable predictions for the visual demand of in-vehicle touchscreen interactions. Accid. Anal. Prev. 183, 106956 https://doi.org/10.1016/j.aap.2023.106956.

Filom, S., Amiri, A.M., Razavi, S., 2022. Applications of machine learning methods in port operations – A systematic literature review. Transport. Res. Part e: Logist. Transport. Rev. 161, 102722 https://doi.org/10.1016/j.tre.2022.102722.

Formosa, N., Quddus, M., Ison, S., Abdel-Aty, M., Yuan, J., 2020. Predicting real-time traffic conflicts using deep learning. Accid. Anal. Prev. 136, 105429 https://doi.org/10.1016/j.aap.2019.105429.

Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., Baesens, B., 2011. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. Decis. Support Syst. 51 (1), 141–154. https://doi.org/10.1016/j.dss.2010.12.003.

Irarrázaval, M.E., Maldonado, S., Pérez, J., Vairetti, C., 2021. Telecom traffic pumping analytics via explainable data science. Decis. Support Syst. 150, 113559 https://doi.org/10.1016/j.dss.2021.113559.

Jackson, I., Ivanov, D., 2023. A beautiful shock? Exploring the impact of pandemic shocks on the accuracy of AI forecasting in the beauty care industry. Transport. Res. Part e: Logist. Transport. Rev. 180, 103360 https://doi.org/10.1016/j.tre.2023.103360.

Jianfeng, X., Hongyu, G., Jian, T., Liu, L., Haizhu, L., 2019. A classification and recognition model for the severity of road traffic accident. Adv. Mech. Eng. 11 (5) https://doi.org/10.1177/1687814019851893, 1687814019851893.

Johnson, M., Albizri, A., Simsek, S., 2022. Artificial intelligence in healthcare operations to enhance treatment outcomes: A framework to predict lung cancer prognosis. Ann. Oper. Res. 308 (1), 275–305. https://doi.org/10.1007/s10479-020-03872-6.

Kriebel, J., Stitz, L., 2022. Credit default prediction from user-generated text in peer-to-peer lending using deep learning. Eur. J. Oper. Res. 302 (1), 309–323. https://doi.org/10.1016/j.ejor.2021.12.024.

Li, P., Abdel-Aty, M., Yuan, J., 2020. Real-time crash risk prediction on arterials based on LSTM-CNN. Accid. Anal. Prev. 135, 105371 https://doi.org/10.1016/j.aap.2019.105371.

Li, J., Fang, S., Guo, J., Fu, T., Qiu, M., 2021. A motorcyclist-injury severity analysis: a comparison of single-, two-, and multi-vehicle crashes using latent class ordered probit model. Accid. Anal. Prev. 151, 105953 https://doi.org/10.1016/j.aap.2020.105953.

Lin, K.-Y., Liu, Y., Li, L., Dou, R., 2021. A Review of Explainable Artificial Intelligence. In: Dolgui, A., Bernard, A., Lemoine, D., von Cieminski, G., Romero, D. (Eds.), Advances in Production Management Systems. Artificial Intelligence for Sustainable and Resilient Production Systems. Springer International Publishing, pp. 574–584. https://doi.org/10.1007/978-3-030-85910-7_61.

Lisboa, P.J.G., Saralajew, S., Vellido, A., Fernández-Domenech, R., Villmann, T., 2023. The coming of age of interpretable and explainable machine learning models. Neurocomputing 535, 25–39. https://doi.org/10.1016/j.neucom.2023.02.040.

Lundberg, S.M., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions. Adv. Neural Inf. Proces. Syst. 30. https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Ma, Z., Mei, G., Cuomo, S., 2021. An analytic framework using deep learning for prediction of traffic accident injury severity based on contributing factors. Accid. Anal. Prev. 160, 106322 https://doi.org/10.1016/j.aap.2021.106322.

Malhotra, A., Jindal, R., 2023. XAI transformer based approach for interpreting depressed and suicidal user behavior on online social networks. Cogn. Syst. Res. 101186 https://doi.org/10.1016/j.cogsys.2023.101186.

Masello, L., Castignani, G., Sheehan, B., Guillen, M., Murphy, F., 2023. Using contextual data to predict risky driving events: A novel methodology from explainable artificial intelligence. Accid. Anal. Prev. 184, 106997 https://doi.org/10.1016/j.aap.2023.106997.

Mujalli, R.O., de Oña, J., 2011. A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks. J. Saf. Res. 42 (5), 317–326. https://doi.org/10.1016/j.jsr.2011.06.010.

Narayanamoorthy, S., Paleti, R., Bhat, C.R., 2013. On accommodating spatial dependence in bicycle and pedestrian injury counts by severity level. Transp. Res. B Methodol. 55, 245–264. https://doi.org/10.1016/j.trb.2013.07.004.

Nasernejad, P., Sayed, T., Alsaleh, R., 2021. Modeling pedestrian behavior in pedestrian-vehicle near misses: A continuous Gaussian Process Inverse Reinforcement Learning (GP-IRL) approach. Accid. Anal. Prev. 161, 106355 https://doi.org/10.1016/j.aap.2021.106355.

Nweke, H.F., Teh, Y.W., Mujtaba, G., Al-garadi, M.A., 2019. Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions. Inform. Fus. 46, 147–170. https://doi.org/10.1016/j.inffus.2018.06.002.

Quintero Mínguez, R., Parra Alonso, I., Fernández-Llorca, D., Sotelo, M.Á., 2019. Pedestrian path, pose, and intention prediction through gaussian process dynamical models and pedestrian activity recognition. IEEE Trans. Intell. Transp. Syst. 20 (5), 1803–1814. https://doi.org/10.1109/TITS.2018.2836305.

Rahimi, E., Shamshiripour, A., Samimi, A., Mohammadian, A., (Kouros), 2020. Investigating the injury severity of single-vehicle truck crashes in a developing country. Accid. Anal. Prev. 137, 105444 https://doi.org/10.1016/j.aap.2020.105444.

Song, L., Fan, W. (David), & Li, Y. (2021). Time-of-day variations and the temporal instability of multi-vehicle crash injury severities under the influence of alcohol or drugs after the Great Recession. Anal. Methods Accid. Res., *32*, 100183. https://doi.org/10.1016/j.amar.2021.100183.

Taamneh, M., Alkheder, S., Taamneh, S., 2017. Data-mining techniques for traffic accident modeling and prediction in the United Arab Emirates. J. Transport. Saf. Secur. 9 (2), 146–166. https://doi.org/10.1080/19439962.2016.1152338.

Talmor, D., Thompson, K.M., Legedza, A.T.R., Nirula, R., 2006. Predicting severe head injury after light motor vehicle crashes: Implications for automatic crash notification systems. Accid. Anal. Prev. 38 (4), 767–771. https://doi.org/10.1016/j.aap.2006.01.008.

Topuz, K., Delen, D., 2021. A probabilistic Bayesian inference model to investigate injury severity in automobile crashes. Decis. Support Syst. 150, 113557 https://doi.org/10.1016/j.dss.2021.113557.

Tutun, S., Tosyali, A., Sangrody, H., Khasawneh, M., Johnson, M., Albizri, A., Harfouche, A., 2023. Artificial intelligence in energy industry: Forecasting electricity consumption through cohort intelligence & adaptive neural fuzzy inference system. J. Business Anal. 6 (1), 59–76. https://doi.org/10.1080/2573234X.2022.2046514.

Wahab, L., Jiang, H., 2020. Severity prediction of motorcycle crashes with machine learning methods. Int. J. Crashworthiness 25 (5), 485–492. https://doi.org/10.1080/13588265.2019.1616885.

Wang, Q., Gan, S., Chen, W., Li, Q., Nie, B., 2021. A data-driven, kinematic feature-based, near real-time algorithm for injury severity prediction of vehicle occupants. Accid. Anal. Prev. 156, 106149 https://doi.org/10.1016/j.aap.2021.106149.

Wei, Z., Das, S., Zhang, Y., 2022. Short duration crash prediction for rural two-lane roadways: applying explainable artificial intelligence. Transp. Res. Rec. 2676 (12), 535–549. https://doi.org/10.1177/03611981221096113.

Wu, C.-L., Law, K., 2019. Modelling the delay propagation effects of multiple resource connections in an airline network using a Bayesian network model. Transport. Res. Part e: Logist. Transport. Rev. 122, 62–77. https://doi.org/10.1016/j.tre.2018.11.004.

Xie, S., & Zhang, J. (2022). A Novel Variable Selection Approach Based on Multi-criteria Decision Analysis. In D. Ciucci, I. Couso, J. Medina, D. Ślęzak, D. Petturiti, B. Bouchon-Meunier, & R. R. Yager (Eds.), Information Processing and Management of Uncertainty in Knowledge-Based Systems (pp. 115–127). Springer International Publishing. https://doi.org/10.1007/978-3-031-08974-9_9.

Yan, R., Wang, S., Zhen, L., 2023. An extended smart “predict, and optimize” (SPO) framework based on similar sets for ship inspection planning. Transport. Res. Part e: Logist. Transport. Rev. 173, 103109 https://doi.org/10.1016/j.tre.2023.103109.

Yang, Y., Jia, B., Yan, X.-Y., Chen, Y., Song, D., Zhi, D., Wang, Y., Gao, Z., 2023. Estimating intercity heavy truck mobility flows using the deep gravity framework. Transport. Res. Part e Logist. Transport. Rev. 179, 103320 https://doi.org/10.1016/j.tre.2023.103320.

Yang, Z., Zhang, W., Feng, J., 2022. Predicting multiple types of traffic accident severity with explanations: A multi-task deep learning framework. Saf. Sci. 146, 105522 https://doi.org/10.1016/j.ssci.2021.105522.

Yuan, C., Li, Y., Huang, H., Wang, S., Sun, Z., Wang, H., 2022. Application of explainable machine learning for real-time safety analysis toward a connected vehicle environment. Accid. Anal. Prev. 171, 106681 https://doi.org/10.1016/j.aap.2022.106681.

Zeng, Q., Huang, H., 2014. A stable and optimized neural network model for crash injury severity prediction. Accid. Anal. Prev. 73, 351–358. https://doi.org/10.1016/j.aap.2014.09.006.

Zhang, S., Abdel-Aty, M., 2022. Real-time crash potential prediction on freeways using connected vehicle data. Anal. Methods Accid. Res. 36, 100239 https://doi.org/10.1016/j.amar.2022.100239.

Zhang, W., Yan, S., Li, J., Tian, X., Yoshida, T., 2022. Credit risk prediction of SMEs in supply chain finance by fusing demographic and behavioral data. Transport. Res. Part e Logist. Transport. Rev. 158, 102611 https://doi.org/10.1016/j.tre.2022.102611.