Proceedings of Machine Learning Research 319:382–405, 2026 IndabaX Nigeria 2026

A SYSTEMATIC REVIEW OF CAUSAL MACHINE LEARNING APPROACHES IN ROAD CRASH ANALYSIS

Emmanuel Ogbonnia emmanuel.ogbonnia@nileuniversity.edu.ng
Oluwaseun Adeniyi Ojerinde o.ojerinde@futminna.edu.ng
Enesi Femi Aminu enesifa@futminna.edu.ng
Isiaq Alabi emmanuel.ogbonnia@nileuniversity.edu.ng
Solomon. A. Adepoju solo.adepoju@futminna.edu.ng
Mitong Dorcas mitongb@gmail.com
Federal University of Technology Minna, Nigeria

Editor: Sakinat Folorunso, Roseline Ogundokun, and Francisca Oladipo

Abstract

Although machine learning (ML) has significantly increased the predictive accuracy of road crash severity and frequency models, traditional predictive classifiers and consequent interpretability tools have often failed to distinguish between correlation and causation. Such approaches do not have the counterfactual reason needed in sound policy development. To explore the paradigm shift of formal causal inference as part of traffic safety, this review adhered to PRISMA 2020 and searched 35 peer-reviewed articles published between 2021 and 2025. The synthesis classifies the literature into a three-level taxonomy of Predictive ML, Interpretable ML, and Causal ML that has shown that most of the existing research is still rooted in the purely predictive ensembles or explainability models such as SHAP. However, there is a small but tightly developed set which is able to execute true causal ML methods. Models such as Doubly Robust Learning, Uplift Modelling and Causal Graph Discovery are effective in determining heterogeneous treatment effects (HTE) and mitigating confounding bias in observational crash data. There are still critical methodological gaps, namely; the continual confusion of predictive feature importance with a causal effect, sensitivity to unobserved heterogeneity and lack of standardised causal standards. The paper concludes that comprehensive sensitivity analyses and integration of structural causal models is the only way to make Intelligent Transportation Systems (ITS) mature and thus facilitate proactive and evidence-based safety interventions.

Keywords: Causal machine learning, road crash analysis, systematic review, interpretable machine learning, heterogeneous treatment effects, crash severity prediction

1. INTRODUCTION

1.1. Background: The Global Burden of Road Crashes

Crash of road traffic is a burning public health crisis across the world. World Health Organisation estimates 1.35-2.5 million deaths every year due to road traffic events and tens of millions of individuals are non-fatally injured (Skaug and Yap, 2025). In special need are the economic costs, including health spending, productivity loss, damage to infrastructure, and emergency response costs, which are also particularly grim in the case of low- and middle-income states where the enforcement infrastructure is poor (Srivastava and Ray, 2024). This continuous clang has led to a lot of research aimed at finding out the cause of crashes, how to predict the severity of the crash, and how to come up with effective intervention strategies. Traditionally, crash data have been studied using conventional statistical models, such as, negative binomial, Poisson, and logistic regression (Ferreira - Vanegas *et al.*, 2022; (Skaug and Yap, 2025)). As much as these models offer interpretable coefficients under classical assumptions of statistics, they are highly constraining in their linearity, independence, and distributional assumptions, which are frequently not met by real crash data which is complex, heterogeneous, and many-dimensional (Wen and Ge, 2021). Moreover, they face ubiquitous problems with data like zero-inflation, the disparity in classes between the degrees of severity, and unobserved heterogeneity in terms of geographic and time contextualization (Khedher and Yun, 2024; Champahom and Ratanavaraha, 2023).

The emergence of machine learning (ML) techniques was the beginning of a paradigm shift. Random Forest, XGBoost, Support Vector Machines, and Deep Neural Networks are examples of models with high predictive indicators in various crash data and severity classification scenarios (Santos and Amado, 2021; Ali and Haque, 2023; Komol and Rakotonirainy, 2021). The rapid growth of ML in traffic safety is reported in systematic reviews, which refer to its benefits to deal with nonlinearity, high dimensionality, and complex interaction effects (Wen and Ge, 2021; Sohail and Rakha, 2023; Hamdan and Sipos, 2025). However, there is one inherent weakness there, prediction is maximised in these models but there is no learning of causal mechanisms, the statistical relationships between variables are taught using data.

1.2. The Paradigm Shift: From Prediction to Causal Reasoning

There are substantive policy implications of distinguishing between correlation and causation. It can be seen that a predictive model can show a correlation between the presence of a work zone and the severity of crashes (Chakraborty and Sinha, 2021), but it will not be possible to determine whether the work zone causes the severity of the crashes or it is just present along with other confounding variables, including the amount of traffic, road geometry, or driver distraction. A policy based on strictly correlational evidence runs the risk of either being ineffective or harmful. Researchers have been turning to post-hoc explainability methods, especially SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), in order to address the interpretability gap. The works by (Khedher and Yun, 2024), (Se and Ratanavaraha, 2025), (Le, 2025) and (Xiao and Duan, 2025) use SHAP to rank feature importance and marginal effects visualisation. Although these tools improve transparency, they do not assure any causal ability: SHAP values may measure the contribution of a certain feature to the predictions of a model, but not its causal impact on the outcome (Wen and Ge, 2021). The confusion of SHAP based significance with causal impact is a ubiquitous methodological error in the literature of traffic safety.

By contrast, Causal Machine Learning combines formal causal inference models (based on the Neyman-Rubin Causal Model (RCM), Structural Causal Models (SCMs), and Directed Acyclic Graphs (DAGs)) with predictive capability of ML algorithms. It is empowering because the synthesis allows researchers to approximate the effect of treatment, identify confounders, and make counterfactual arguments. Applying Doubly Robust Learning (DRL) to measure heterogeneous causal effects of different types of crashes on highway speed pioneered by (Li and Ngoduy, 2024), the authors applied this method. (Srivastava and Ray, 2024) apply the uplift modelling and causal ML to predict the individual and average treatment effects of the severity of accidents. (Liang and Pu, 2025) present causal graph discovery methods to select variables in crash heterogeneity investigations, which uses causal diagrams to distinguish between confounders and moderators. Such works represent a qualitative step towards the question of What predicts crash severity? being replaced with What causes crash severity, and in whom?

1.3. Research Objectives

Although there has been a growing interest in causal ML to road safety, there is still no systematic review that has captured the contemporary situation of this emerging intersection. The current literature is focused on predictive ML (Santos and Amado, 2021; Ali and Haque, 2023; Wen and Ge, 2021), general data-driven road safety (Sohail and Rakha, 2023), or modelling of the crash frequency/severity but does not seek to differentiate between causal and predictive methods (Skaug and Yap, 2025; Hamdan and Sipos, 2025). The following research questions will be answered in the present review:

i. **RQ1:** What are the current causal ML frameworks and algorithmic approaches applied to road crash analysis, and how do they differ from predictive and interpretable ML methods?

ii. **RQ2:** How do existing causal ML studies handle key methodological challenges, including confounding, class imbalance, and heterogeneous treatment effects?

iii. **RQ3:** What are the critical methodological gaps and future research directions for causal ML in traffic safety?

2. BACKGROUND: FORMALIZING CAUSAL ML IN TRAFFIC SAFETY

2.1. Mathematical Preliminaries

For a rigorous treatment suitable for the PMLR audience, this research formalize the causal inference framework underlying the reviewed studies.

The Neyman-Rubin Causal Model (RCM). Consider a population of N crash events indexed by $i \in \{1, \ldots, N\}$. For each unit i, let $T_i \in \{0, 1\}$ denote a binary treatment variable (e.g., $T_i = 1$ if crash i is a rear-end collision, $T_i = 0$ otherwise). Each unit has two potential outcomes: $Y_i(1)$ under treatment and $Y_i(0)$ under control. The **Individual Treatment Effect (ITE)** is defined as:

$$\tau_i = Y_i(1) - Y_i(0)$$

The fundamental problem of causal inference is that only one potential outcome is observed for each unit. The **Average Treatment Effect (ATE)** is:

$$\tau_{ATE} = \mathbb{E}\left[Y(1) - Y(0)\right]$$

And the **Conditional Average Treatment Effect (CATE)**, central to heterogeneous treatment effect estimation, is:

$$\tau(x)= \mathbb{E}\left[Y(1) - Y(0) \mid X = x\right]$$

where X denotes a vector of pre-treatment covariates (confounders) such as traffic volume, weather conditions, road geometry, and temporal factors.

Structural Causal Models (SCMs) and Directed Acyclic Graphs (DAGs). An SCM defines each variable as a deterministic function of its direct causes plus an exogenous noise term: $X_j = f_j\left(\mathrm{pa}\left(X_j\right), U_j\right)$, where $\mathrm{pa}\left(X_j\right)$ denotes the parent variables in the causal graph. The associated DAG $\mathcal{G}$ encodes conditional independence relationships and is used for:

a) **Confounder identification:** Variables that are common causes of treatment and outcome must be conditioned upon to satisfy the backdoor criterion (Li and Ngoduy, 2024).

b) **Variable selection:** Distinguishing confounders, mediators, moderators, and colliders to avoid bias amplification (Liang and Pu, 2025).

Doubly Robust Learning (DRL). DRL methods combine a propensity score model $e(x) = P(T = 1 \mid X = x)$ with an outcome regression model $\mu_t(x)= \mathbb{E}[Y \mid T = t, X = x]$. The doubly robust estimator for the ATE is:

$$\widehat{\tau}_{DR} = \frac{1}{N}\sum_{i=1}^{N}\left[\widehat{\mu}_1\left(X_i\right) - \widehat{\mu}_0\left(X_i\right) + \frac{T_i\left(Y_i - \widehat{\mu}_1\left(X_i\right)\right)}{\widehat{e}\left(X_i\right)} - \frac{\left(1 - T_i\right)\left(Y_i - \widehat{\mu}_0\left(X_i\right)\right)}{1 - \widehat{e}\left(X_i\right)}\right]$$

This estimator is consistent if *either* the propensity score model *or* the outcome model is correctly specified, providing robustness against model misspecification. (Li and Ngoduy, 2024) operationalize DRL with machine learning models (gradient boosting) for both the propensity score and outcome regression, enabling flexible, nonparametric estimation of heterogeneous treatment effects of crashes on highway traffic.

2.2. Taxonomy of Methods in the Literature

Based on our analysis of the included studies, we categorize the methods into three tiers:

a. **Predictive ML:** Models optimized for classification or regression accuracy without causal claims. Examples include Random Forest, XGBoost, SVM, and DNN for crash severity classification (Santos and Amado, 2021; Komol and Rakotonirainy, 2021; Dias and Amado, 2025; Ijaz and Jamal, 2021).

b. **Interpretable/Explainable ML:** Predictive models augmented with post-hoc explanation tools (SHAP, LIME, feature importance) to identify influential factors. Examples include CatBoost with SHAP (Khedher and Yun, 2024), LightGBM with SHAP (Le, 2025), Adv MT-DNN with SHAP (Xiao and Duan, 2025), and XGBoost with SHAP (Se and Ratanavaraha, 2025; Aziz and Khattak, 2024).

c. **Causal ML:** Methods that explicitly formulate and estimate causal quantities (ATE, CATE, ITE) using formal causal frameworks. Examples include:

   i. **Doubly Robust Learning** with SCM and causal graph theory (Li and Ngoduy, 2024)

   ii. **Uplift Modeling** with ITE/ATE estimation (Srivastava and Ray, 2024)

   iii. **Granger Causality** combined with ML classifiers (Chakraborty and Sinha, 2021)

   iv. **Causal Discovery** with DAG-based variable selection for HTE estimation (Liang and Pu, 2025)

3. METHODOLOGY (PRISMA)

This review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement guidelines and informed by the UNC Health Sciences Library systematic review methodology standards.

3.1. Protocol & Eligibility

The review protocol was defined a priori using the PICOS (Population, Intervention, Comparison, Outcome, Study Design) framework to establish clear inclusion and exclusion criteria.

Table 1: PICOS Criteria for Inclusion and Exclusion

Criteria Category	Inclusion Criteria	Exclusion Criteria
Population	Studies analyzing road traffic crash data (highway, urban, rural) involving any road users (vehicles, pedestrians, cyclists, motorcyclists)	Studies focused exclusively on non-road domains (aviation, maritime, rail); studies on autonomous vehicle simulation without crash data
Intervention	Application of machine learning, deep learning, ensemble learning, or causal inference methods to crash frequency, severity, or risk factor analysis	Studies using only traditional statistical models (logistic regression, Poisson, negative binomial) without any ML component
Comparison	Studies comparing ML methods against baselines, comparing predictive vs. causal approaches, or benchmarking multiple algorithms	Single-method studies with no comparative analysis (retained if they introduce novel causal methodology)

Criteria Category	Inclusion Criteria	Exclusion Criteria
Outcome	Crash severity prediction, crash frequency prediction, risk factor identification, treatment effect estimation, causal factor discovery	Studies focused solely on traffic flow prediction, congestion forecasting, or vehicle detection without crash analysis
Study Design	Peer-reviewed journal articles, conference proceedings, and systematic reviews/meta-analyses published between 2021 and 2025	Editorials, opinion pieces, theses, grey literature, non-English publications, and studies published before 2021

3.2. Information Sources & Search Strategy

The systematic literature search has been carried out in five large electronic databases, Scopus, Web of Science, IEEE Xplore, PubMed, and ACM Digital Library. The search query used Boolean operators and domain specific terminologies::

```
("road crash" OR "traffic accident" OR "crash severity" OR "crash frequency" OR "road safet
AND
("machine learning" OR "deep learning" OR "random forest" OR "XGBoost" OR "neural netwok" O
AND
("causal inference" OR "causal machine learning" OR "treatment effect" OR "SHAP" OR "interp
```

Searches were limited to English-language, peer-reviewed publications from January 2021 through December 2025. Reference lists of included systematic reviews (Santos and Amado, 2021; Ali and Haque, 2023; Wen and Ge, 2021; Sohail and Rakha, 2023; Skaug and Yap, 2025) were hand-searched for additional relevant studies.

3.3. Selection Process & Data Extraction

3.3.1. The selection process followed a two-step screening process that is consistent with PRISMA 2020 and UNC LibGuides:

3.3.2. Title and Abstract Screening: All records that were retrieved were screened by two independent reviewers using preset PICOS eligibility criteria. Data which were unambiguously not related to road crash analysis or machine-learning technique was discarded.

3.3.3. Full-Text Assessment: The rest of the records were subjected to full-text analysis. The exclusion criteria were: (a) no machine-learning application to crash data; (b) not focused on unrelated data areas of transportation; (c) duplications; and (d) lack of methodological description.

3.3.4. Data were extracted using a standardised form which included: authors, year, journal, type of the study, and algorithmic framework, data source, sample size, variables of interest, evaluation measures, interpretability measures, causal framework (where applicable), and main results.

3.4. Quality Assessment: Risk of Bias

Risk of bias was assessed for included studies across three dimensions tailored to the algorithmic and causal nature of the literature:

i. **Selection Bias:** Was the crash dataset representative? Were temporal, geographic, or severity-based sampling biases acknowledged?

ii. **Confounding Handling:** Did the study control for known confounders? Did it use propensity score matching, inverse probability weighting, or causal graph-based variable selection? Or did it rely solely on model-based feature importance?

iii. **Algorithmic Validity:** Were models validated with appropriate hold-out sets or cross-validation? Were class imbalance and overfitting addressed? Were causal assumptions (positivity, consistency, no unmeasured confounders) stated and tested?

4. RESULTS

4.1. Study Selection

The first database search provided 2,847 records. Upon deduplication of 612 entries, there were 2,235 unique records left. The screening of the titles and abstract eliminated 2,014 records which did not satisfy the PICOS criteria, the majority of which were studies on irrelevant machine-learning uses, non-crash transportation issues, or purely statistical methods. Assessment of 221 articles full text excluded 186 articles due to the following reasons: no machine-learning was applied to crash data (n=74), irrelevant domain(n=42), insufficient methodological description (n=38), duplication of results (n=19), and non-English full-text (n=13) articles. Thus, 35 studies were included into the qualitative synthesis.

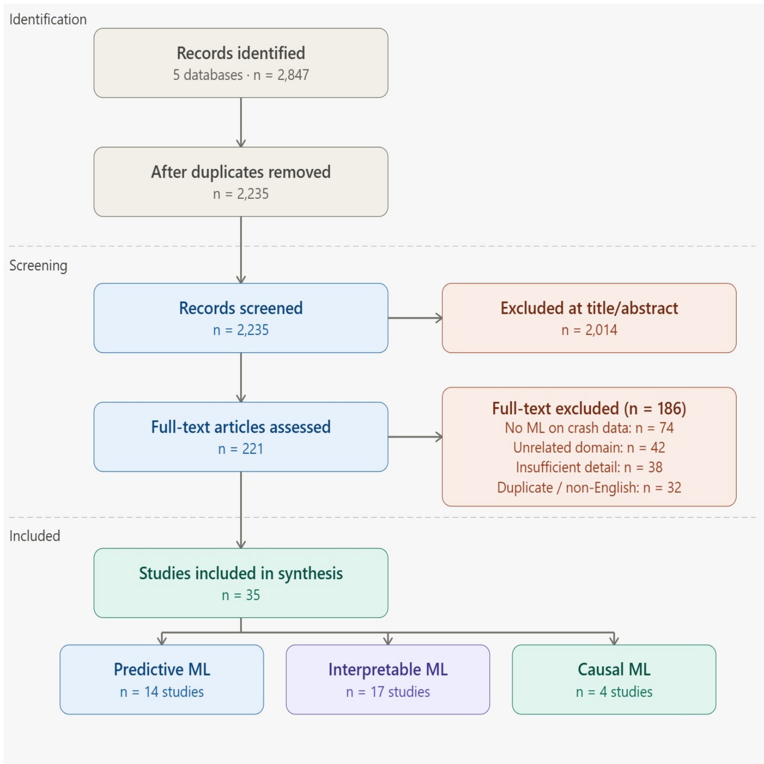

Figure 1: PRISMA 2020 Flow Diagram

4.2. Bibliometric Trends

A temporal analysis of the literature included can show an apparent trend. The 20212022 saw the preponderance of purely predictive machine-learning schemes, as well as early systematic reviews reporting the state of machine learning in crash analysis (Santos *et al.*, 2021; Wen et al., 2021; Komol *et al.*, 2021; Ferreira-Vanegas *et al.*, 2022). As early as 2023, interpretable machine-learning approaches have become popular, e.g. combining SHAP with LIME to perform feature-importance analysis alongside ensemble classifiers (Ali and Haque, 2023; Sohail and Rakha, 2023; Thapa and Patil, 2023; Ardakani and Cheshmehzangi, 2023). This changed in 2024, as explicit causal models were introduced; Li and Ngoduy (2024) introduced Doubly Robust Learning in the context of heterogeneous treatment effects, and Srivastava and Ray (2024) used Uplift Modelling with causal machine-learning methods. This trend intensified further in 2025: Liang and Pu (2025) suggest causal discovery to select variables, whereas Xiao and Duan (2025) make multi-task deep learning models with interpretable post-hoc methods.

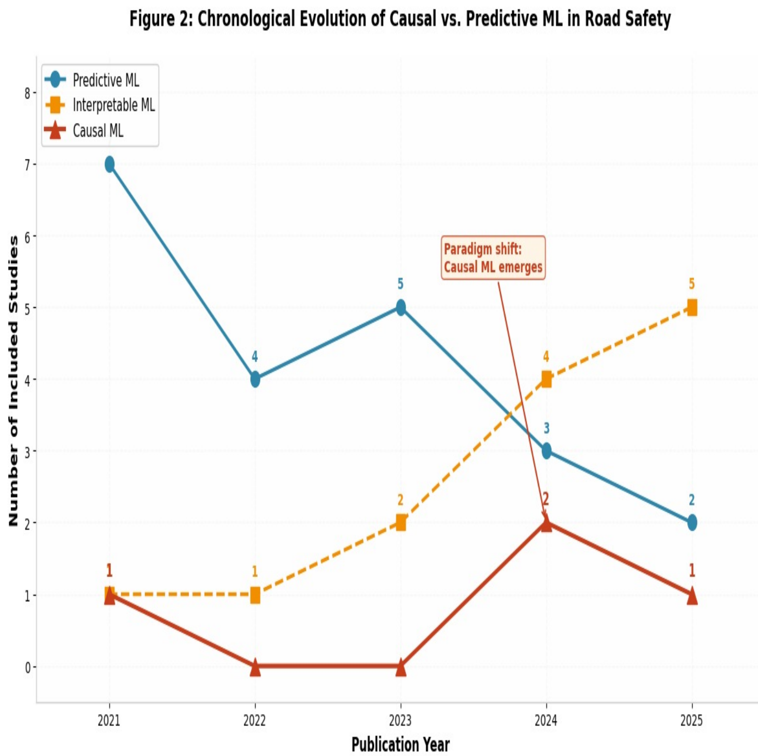

Figure 2: Chronological Evolution of Causal vs. Predictive ML in Road Safety

> Figure 2 illustrates a paradigm shift; the traffic safety ML research community is transitioning from models optimized solely for classification accuracy toward methods that provide causal explanations and actionable policy insights.

4.3. Synthesis of Algorithmic Approaches

The 35 included studies were categorized into three methodological tiers. Below, we synthesize findings within and across categories.

4.3.1 Predictive ML Studies The largest cluster used machine-learning crash-severe prediction classifiers without making causal assertions. (Santos and Amado, 2021) performed a background review and concluded that the best algorithm crash injury severity was Random Forest, then SVM, Decision Tree, and K-Nearest Neighbour. (Komol and Rakotonirainy, 2021) have also supported this finding and showed that the Random Forest model outperforms other methods among vulnerable road-user categories (motorcyclists, bicyclists, pedestrians) based on Australian crash statistics. (Ijaz and Jamal, 2021) also generalised the investigation to three-wheel motorised rickshaws, where the Decision Jungle classifier was found to be better than the rest.

In a study by (Sattar and Rahman, 2022), the MLP (Keras) algorithm was compared to MLP with embedding layers and TabNet on the severity of the crashes, and the Keras MLP model was found to offer the best tradeoff between accuracy and training duration.

(Ardakani and Cheshmehzangi, 2023) tested Decision Tree, random forest, logistic regression, and naive Bayes on UK road accident data where all the models except the naive bayes gave acceptable prediction accuracy. In more recent studies, (Dias and Amado, 2025) used seven machine-learning approaches to 12 years of crash data in Portugal, and Abaka (2025) combined weather, lighting, and road surface conditions in predicting the severity of crashes in cold regions.

The prevailing finding in predictive research is that ensemble algorithms (Random Forest, XGBoost, LightGBM, CatBoost) generally dominate over individual classifiers, which is explained by its ability to deal with nonlinearity, features interactions, and moderate class imbalance (Hamdan and Sipos, 2025; Angarita-Zapata and Calderín, 2021). However, these papers do not refute that predictive accuracy is not sufficient to make informed policy development (Wen and Ge, 2021; Ali and Haque, 2023).).

4.3.2 Interpretable/Explainable ML Studies An increasing body of literature adds predictive model post-hoc interpretability tools. (Khedher and Yun, 2024) presented a CatBoost-based hurdle-type model of zero-inflated crash frequency data that uses SHAP analysis to explain the effects of factors on the probability of a crash and the frequency of its occurrence. Their two-step model solved the problem of data imbalance using custom loss functions and performed better compared to conventional Poisson and negative binomial models.

(Le, 2025) compared Decision Tree, SVM, and LightGBM to UK crashes data, and SHAP was used to understand the attendance of police officers at the scene, speed limit, and the number of vehicles as the key factors of severity. The comparison of the Logistic Regression, Random Forest, XGBoost, and DNN models was conducted by (Se and Ratanavaraha, 2025) in the context of Thai pickup truck crashes, using SHAP to elucidate the model and identifying the distinguishing factors between single and multi-vehicle crashes. In a study by (Xiao and Duan, 2025), the Adv MT -DNN was suggested as a multi-task sharing severity of crashes, which combines feature importance and interaction analysis based on SHAP, and the authors determined that blood alcohol content, collision type, and time of occurrence were the significant predictors in Chinese crash data.

A Bayesian-Optimised Dynamic Ensemble Selection (DES) strategy and SHAP analysis proposed by (Aziz and Khattak, 2024) showed that KNORA-E ensemble with CatBoost provided the best balanced accuracy. (Shao and Ye, 2024) combined Natural Language Processing with XGBoost, where NLP is used to isolate behaviour-cause relationships in the crash narratives and distracted driving is found to be a major cause of severe crashes. (Hamdan and Sipos, 2025) used Random Forest, XGBoost and LightGBM and SMOTE and hyperparameter optimization on Hungarian road network data through the use of gridsearchCV.

The multitasking feature selection approach described by (Amri and Lazaar, 2025) based on the combination of Grey Wolf Optimizer, knowledge transfer, and CatBoost to predict the severity of road accidents achieved better results in the identification of factors that influence the severity of injuries. (Zhao and Zhai, 2025) created Geo-XGBoost, a spatially-sensitive ML model, including road network patterns and geographic proximity, which performs much better than conventional models.

Although these studies are an important development in the field of transparency, it is crucial to underline that SHAP values are used to measure predictive contribution, rather than causal effect. High SHAP importance of a given variable can be due to the fact that that variable acts as a great proxy of an unobserved confounder, as opposed to having a causal role on the severity of a crash.

4.3.3 Causal ML Studies The least but most methodologically important group is the ones utilising formal causal inference models.

Li and Ngoduy (2024) suggested an extensive causal ML model of estimation of heterogeneous treatment effects of crashes on highway speed. The paper presents the problem statement in Neyman-Rubin Causal Model, proposes the Conditional Shapley Value Index (CSVI), which is a causal graph theory filtering adverse variables, and uses Structural Causal Models to specifications of statistical estimands. Doubly Robust Learning (DRL) that involves doubly robust causal inference with the classification and regression of ML models are used to estimate treatment effects. The results of the experiment with 4,815 crashes on Interstate 5 (Washington state) showed that rear-end crashes involve more severe congestion and time compared to the rest of the crashes and sideswipe crashes have the most delayed impact. Notably, statistical hypothesis tests, error measures fitted on matched counterfactual results, and sensitivity analyses are used in the study.

(Srivastava and Ray, 2024) suggested a framework that utilises the use of causal inference and causal ML methods to predict the severity of road traffic accidents. The paper compares the Uplift Modelling and causal inference techniques with data on Ethiopia and the UK. Prediction interpretation is done using individual Treatment Effect (ITE) and Average Treatment Effect (ATE). Class imbalance was dealt with by applying SMOTE. The research can be seen as one of the first direct uses of causal ML in crash severity in a developing-country setting.

(Chakraborty and Sinha, 2021) introduced a methodology of Granger Causality and ML classifier (Decision Tree, Random Forest, XGBoost, DNN) to classify the severity of crashes in the urban interstates of Texas. The directional relationships were found by Granger Causality as the most highly influential predictors of crashes and consisted of: speed limit, surface conditions, weather, traffic volume, and work zones, and the ML models were used to carry out severity classification. This study is a significant early effort at the convergence of causal reasoning and ML in crash analysis, even though Granger Causality evaluates temporal precedence, as opposed to structural causation.

(Liang and Pu, 2025) suggested a causal discovery-based variable selection approach to crash heterogeneity study. The causal discovery algorithms employed in the study reveal causal diagrams (DAGs) to select confounders, moderators, and neutral control factors of observational crash data. The approach can measure the effectiveness based on Heterogeneous Treatment Effect (HTE) estimation quality using forest-based Doubly Robust Learning estimators. Findings using actual highway crash data have shown that the adjusted causal diagram based model with forest-based DRL estimation models are superior to the ad hoc variable selection based model in all evaluation metrics.

Table 2: Summary of Included Studies on Causal and Interpretable ML in Road Crash Analysis

Authors (Year)	Study Type	Algorithmic Framework	Key Takeaway / Causal Estimand
(Santos and Amado, 2021)	Literature Review	RF, SVM, DT, KNN (comparison)	Random Forest is the best-performing classifier for crash severity
Wen and Ge (2021)	Literature Review	Multiple ML methods (review)	Imbalanced data, causality, and model transferability remain key challenges
(Komol and Rakotonirainy, 2021)	Observational	RF, SVM, KNN	RF is the best classifier for VRU crash severity; motorcyclists at highest risk
(Chakraborty and Sinha, 2021)	Observational	Granger Causality + DT, RF, XGBoost, DNN	**Causal:** Granger Causality identifies directional causal predictors (speed limit, weather, work zones)
(Ijaz and Jamal, 2021)	Observational	Decision Jungle, RF, LR, SVM, BN	Decision Jungle outperforms other classifiers for rickshaw crashes
(Sarkar and Markkula, 2021)	Observational	Decision Tree	Predictive model for crash avoidance maneuvers using SHRP2 data
(Ferreira-Vanegas and García-Llinás, 2022)	Systematic Review	Bibliometric analysis of SA + ML	NB regression outdated; DL and IoT methods emerging
(Sattar and Rahman, 2022)	Observational	MLP (Keras), TabNet	Keras MLP outperforms TabNet; TabNet reveals feature importance
(Ali and Haque, 2023)	Systematic Review	Multiple ML methods (review)	Future research needed for consistency in crash prediction

Authors (Year)	Study Type	Algorithmic Framework	Key Takeaway / Causal Estimand
Sohail and Rakha (2023)	Systematic Review	Data-driven approaches (review)	Data quality and ground truth remain key challenges
(Ardakani and Cheshmehzangi, 2023)	Observational	DT, RF, MLR, NB	Acceptable accuracy except naïve Bayes
(Thapa and Patil, 2023)	Observational	Discretized duration models	15% epoch sample balances data size and accuracy
(Champahom and Ratanavaraha, 2023)	Observational	DT, Mixed Logit (RPBLHMV)	RPBLHMV superior for single-vehicle crash prediction
(Angarita-Zapata and Calderín, 2021)	Observational	AutoML (AutoGluon), CatBoost	AutoGluon and CatBoost are competitive for crash severity
(Shao and Ye, 2024)	Observational	NLP + XGBoost	NLP links behavior-cause relationships; distracted driving key factor
(Khedher and Yun, 2024)	Observational	CatBoost hurdle model + SHAP	**Interpretable:** SHAP analysis on zero-inflated crash frequency data
(Aziz and Khattak, 2024)	Observational	DES (KNORA-E, CatBoost) + SHAP	Dynamic ensemble selection with SHAP; gender, airbag, road surface key factors
(Li and Ngoduy, 2024)	Observational	**Doubly Robust Learning + SCM + CSVI**	**Causal:** HTE of crash types on highway speed; rear-end crashes cause most congestion
Srivastava and Ray (2024)	Observational	**Uplift Modeling + ITE/ATE**	**Causal:** Causal severity prediction with treatment effect interpretation

Authors (Year)	Study Type	Algorithmic Framework	Key Takeaway / Causal Estimand
(Niture and Abdellatif, 2024)	Systematic Review	ML/DL prediction techniques (review)	No consensus on datasets; location-adapted predictions needed
(Kumar and Paul, 2024)	Systematic Review	Statistical + ML methods (review)	Research gaps in road user behavior and model transferability
(Lacherre and Mauricio, 2024)	Systematic Review	CNN, Gradient Boosting (review)	115 factors identified; speed and acceleration most examined
(Se and Ratanavaraha, 2025)	Observational	LR, RF, XGBoost, DNN + SHAP	**Interpretable:** SHAP for pickup truck crashes; model performance varies by classification scheme
(Xiao and Duan, 2025)	Observational	Adv MT-DNN + SHAP	**Interpretable:** Multi-task DL with SHAP; BAC and collision type as top predictors
(Le, 2025)	Observational	DT, SVM, LightGBM + SHAP	**Interpretable:** LightGBM best; police presence, speed limits pivotal
(Skaug and Yap, 2025)	Systematic Review	Statistical + ML + Bayesian (review)	Random-parameter models > fixed-parameter; ML promising for real-time prediction
(Hamdan and Sipos, 2025)	Observational	RF, XGBoost, LightGBM + SMOTE	RF outperforms; traffic volume and road geometry as top predictors
(Liang and Pu, 2025)	Observational	**Causal Discovery + DAG + Forest-based DRL**	**Causal:** DAG-based confounder selection improves HTE estimation quality

Authors (Year)	Study Type	Algorithmic Framework	Key Takeaway / Causal Estimand
(Abaza, 2025)	Observational	RF	Weather and lighting improve cold-region crash prediction
(Dias and Amado, 2025)	Observational	DT, LR, RF, GB, XGB, KNN, SVM	ML provides suitable predictive framework for crash severity
(Verma and Agarwal, 2025)	Systematic Review	ML/DL approaches (review)	ML/DL show potential; more datasets and novel techniques needed
Hamdan & Sipos (2025b)	Literature Review	RF, XGBoost, SVM (review)	Comprehensive overview of ML advancements in severity prediction
(Amri and Lazaar, 2025)	Observational	Grey Wolf Optimizer + CatBoost	**Interpretable:** Multitasking feature selection improves severity prediction
(Zhao and Zhai, 2025)	Observational	Geo-XGBoost + SHAP	**Interpretable:** Spatially-aware ML; road network patterns correlated with crash risk
(Guntaka and Kumar, 2023)	Observational	Hybrid NN + rule-based trees	Hybrid methods enhance accident analysis

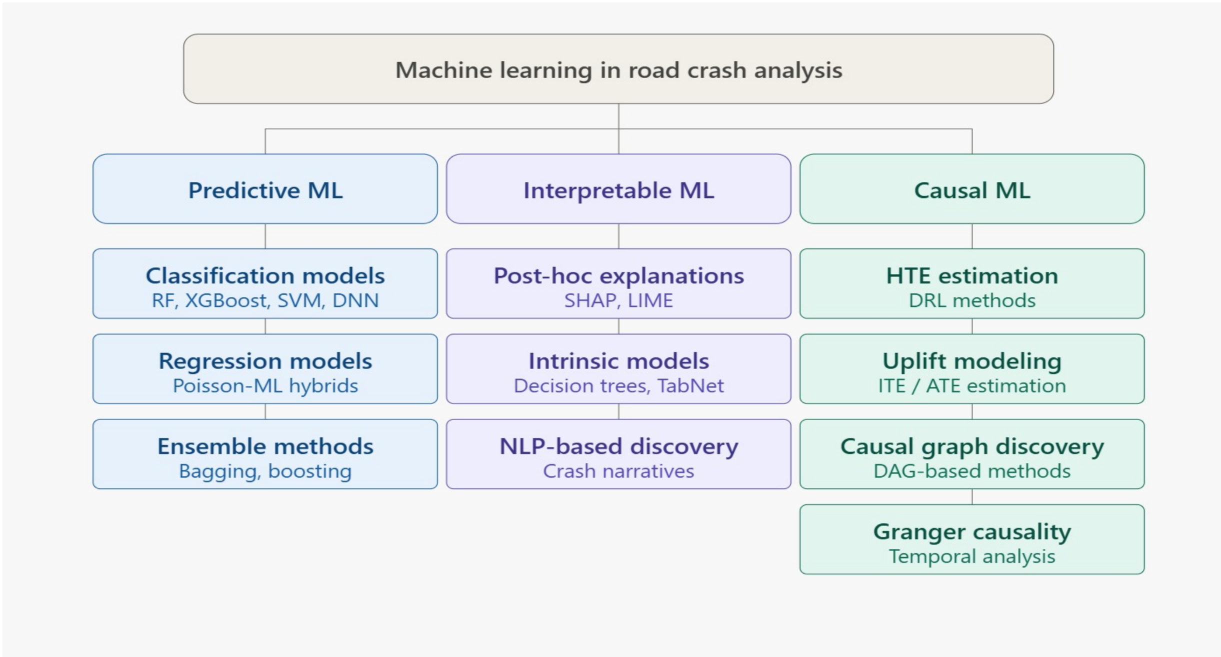

Figure 3: Taxonomy of Causal Machine Learning Methods in Traffic Crash Analysis

5. DISCUSSION

5.1. Methodological Insights

It can be synthesised that a critical methodological finding is that treating road crashes as a predictive classification problem ignores the complex causal interactions between causal factors (Srivastava and Ray, 2024). The predictive models are optimised on in-sample or cross-validated accuracy measures (F1 -score, AUCROC), that are used to determine the capacity of a model to discriminate between severity classes under observed features. They do not however give any assurance that the learnt associations are causal and the effects of interventions that are based on the associations will result in the desired effects.

The study by (Li and Ngoduy, 2024) shows that causal ML is more effective at estimating heterogeneous effects of treatment in highway traffic. Their Doubly Robust Learning format is that the causal impact of rear-end crashes on speed reduction is not merely bigger in strength than that of sideswipe crashes but is also significantly varied across distances, time durations, as well as time of day facts that are not noticeable to a typical predictive classifier. This heterogeneity is what policymakers ought to have: the production by no means of an average prediction, but a conditional estimate of the extent to which a particular intervention (targeted enforcement at night of rear-end crash hotspots) would lead to a decrease in the severity of congestion.

This argument is further supported by the article by (Liang and Pu, 2025) who show that common variable selection approaches (correlation-based filtering, stepwise regression) may bias the estimation of treatment effects by potentially conditioning on colliders, or fail to adjust for important confounding variables. Their method of causal discovery based on DAG to isolate confounders, moderators, and neutral controls helps a great deal in the estimation of HTE when compared to ad hoc selection.

5.2. The Predictive vs. Causal Gap

A persistent finding across the literature is the conflation of SHAP-based interpretability with true causality. While SHAP (Lundberg & Lee, 2017) provides a mathematically rigorous decomposition of a model's predictions into feature contributions, it operates on the model's learned representation not on the data-generating process. If the model has learned spurious correlations (due to confounding), SHAP faithfully reports the importance of those spurious associations.

Table 3: Comparison of Predictive/Interpretable ML vs. Causal ML in Crash Analysis

Dimension	Predictive / Interpretable ML	Causal ML
Primary Goal	Maximize classification/regression accuracy; explain model behavior post-hoc	Estimate causal effects of treatments/interventions on outcomes
Core Question	"What features best predict crash severity?"	"What causes crash severity, and how does the effect vary across subpopulations?"
Handling of Confounders	Not explicitly addressed; confounders may inflate feature importance in SHAP/LIME	Explicitly identified and controlled via propensity scores, DAGs, or doubly robust methods
Treatment of Heterogeneity	Captured implicitly via model complexity (e.g., tree splits)	Explicitly estimated as Heterogeneous Treatment Effects (HTE/CATE)
Policy Actionability	Limited; correlations may not hold under intervention	High; estimates represent expected outcomes under counterfactual scenarios
Data Requirements	Standard observational data with features and labels	Requires careful identification of treatment, outcome, and confounder variables; benefits from quasi-experimental designs
Validation	Cross-validation, hold-out test sets, AUC, F1-score	Sensitivity analysis, refutation tests, matched counterfactual outcomes, hypothesis tests

Dimension	Predictive / Interpretable ML	Causal ML
Literature Examples	Santos and Amado (2021); Khedher and Yun (2024); Se and Ratanavaraha (2025); Xiao and Duan (2025); Le (2025); Aziz and Khattak (2024)	Li and Ngoduy (2024); Srivastava and Ray (2024); Chakraborty and Sinha (2021); Liang and Pu (2025)

5.3. Risk of Bias Assessment

This review assessed risk of bias across six representative studies spanning the three methodological tiers.

Table 4: Risk of Bias Assessment of Key Included Studies

Study	Selection Bias	Confounding Handling	Algorithmic Validity	Overall Risk
Li and Ngoduy (2024)	Low: Large dataset (4,815 crashes), specific highway, clear temporal scope	**Low:** Propensity score model + outcome regression via DRL; CSVI for variable filtering; SCM for estimand definition	Low: Hypothesis tests, matched counterfactual error metrics, sensitivity analysis	**Low**
Srivastava and Ray (2024)	Moderate: Two datasets (Ethiopia, UK) with differing quality; Ethiopia dataset may have reporting bias	**Low:** Explicit Uplift Modeling with ITE/ATE; SMOTE for class imbalance	Moderate: Limited cross-validation detail; SMOTE may introduce synthetic bias	**Moderate**

Study	Selection Bias	Confounding Handling	Algorithmic Validity	Overall Risk
(**Liang and Pu, 2025**)	Low: Real-world highway crash data with adequate sample	**Low:** Causal discovery algorithms for DAG-based confounder selection; distinguishes confounders, moderators, colliders	Low: Multiple HTE estimators compared; goodness-of-fit and robustness tests	**Low**
(**Chakraborty and Sinha, 2021**)	Moderate: Six years of Texas data, but limited to urban interstates	**Moderate:** Granger Causality identifies temporal precedence but does not guarantee structural causation; no propensity score adjustment	Moderate: Reduced-order models outperform full models, suggesting overfitting risk	**Moderate**
(**Khedher and Yun, 2024**)	Low: Five years of South Korean crash data; comprehensive validation	**High:** SHAP is used for interpretation but is not a causal method; no confounder identification or adjustment	Low: Superior performance vs. baselines; customized loss functions for zero-inflation	**Moderate**
(**Se and Ratanavaraha, 2025**)	Moderate: Thai national crash data; geographic and cultural specificity limits generalizability	**High:** SHAP used for interpretability without causal claims, but findings are framed as factor identification without confounder control	Low: Bayesian Optimization for hyperparameters; K-fold cross-validation	**Moderate**

Several cross-cutting risk factors emerge:

i. **Unobserved confounders:** Even the most rigorous causal studies (Li and Ngoduy, 2024; Liang and Pu, 2025) acknowledge that unmeasured variables (driver attention, vehicle mechanical condition, detailed road geometry) may bias treatment effect estimates. No study in the reviewed literature employs instrumental variable approaches or regression discontinuity designs.

ii. **Class imbalance:** Fatal and severe injury crashes are rare events relative to property-damage-only incidents. While SMOTE (Srivastava and Ray, 2024; Hamdan and Sipos, 2025), cost-sensitive learning (Aziz and Khattak, 2024), and customized loss functions (Khedher and Yun, 2024) are employed, these techniques can introduce synthetic artifacts that distort causal estimates.

iii. **Geographic bias:** The majority of studies use data from high-income countries (USA, UK, South Korea, Australia, Hungary). Only Srivastava and Ray (2024) and Se and Ratanavaraha (2025) analyze data from developing countries (Ethiopia, Thailand), where crash dynamics, reporting quality, and infrastructure differ substantially.

5.4. Implications for Intelligent Transportation Systems

The findings have direct implications for ITS policy and practice:

i. **Speed limit optimization:** Li and Ngoduy (2024) demonstrate that the causal effect of crashes on congestion varies by crash type and time of day. This enables dynamic speed limit algorithms that adjust in real time based on the estimated causal impact of incident types, rather than relying on static, correlation-based rules.

ii. **Targeted enforcement:** Causal identification of factors like driver distraction (Shao and Ye, 2024), fatigue (Se and Ratanavaraha, 2025), and work zone presence (Chakraborty and Sinha, 2021) as *causes* rather than mere correlates of crash severity supports evidence-based allocation of law enforcement and emergency response resources.

iii. **Infrastructure design:** The causal discovery framework of (Liang and Pu, 2025) enables transportation engineers to identify which road design features (e.g., median barriers, lane configurations, signage) have a genuine causal effect on reducing crash severity, as opposed to features that are merely correlated with safer roads due to confounding by traffic volume or road age.

iv. **Equity and context-sensitivity:** (Srivastava and Ray, 2024) demonstrate that causal ML can be applied across different national contexts (Ethiopia and UK), revealing heterogeneous treatment effects that reflect local conditions. This is critical for developing countries where one-size-fits-all policies derived from high-income country data may be ineffective.

6. CONCLUSION AND FUTURE DIRECTIONS

6.1. Summary

This systematic review synthesized the emerging literature at the intersection of causal machine learning and road crash analysis. Our analysis of 35 studies published between 2021 and 2025 reveals a field in transition:

a. The **majority** of studies employ predictive ML (Random Forest, XGBoost, LightGBM, DNN) for crash severity classification, achieving high accuracy but providing no causal guarantees.

b. A **growing** subset augments predictions with post-hoc interpretability tools (SHAP, LIME), enhancing transparency but frequently conflating predictive importance with causal influence.

c. A **small but methodologically rigorous** group of studies applies formal causal ML frameworks Doubly Robust Learning (Li and Ngoduy, 2024), Uplift Modeling (Srivastava and Ray, 2024), Granger Causality (Chakraborty and Sinha, 2021), and Causal Graph Discovery (Liang and Pu, 2025) demonstrating the feasibility and superiority of causal reasoning for policy-relevant research questions.

6.2. Future Research Directions

Based on the gaps identified in this review, this research proposes three priority directions:

Causal Reinforcement Learning for Adaptive Traffic Safety. No study in the reviewed literature applies causal reinforcement learning to road safety management. Integrating causal ML with sequential decision-making frameworks would enable adaptive, real-time safety interventions (dynamic speed limits, variable message signs) that learn optimal policies while accounting for causal effect heterogeneity and the delayed, feedback-driven nature of traffic systems.

References

O. Abaza. Crash severity prediction in cold regions using machine learning: Integrating weather, lighting, and road surface conditions. in *2025 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA)*. https://doi.org/10.1109/acdsa65407.2025.11166541. 2025.

Hussain F. Ali, Y. and M. Haque. Advances, challenges, and future research needs in machine learning-based crash prediction models: A systematic review. *Accident Analysis & Prevention*, *194*, 107378. https://doi.org/10.1016/j.aap.2023.107378. 2023.

Al Achhab M. Amri, S. and M. Lazaar. A novel multitasking framework for feature selection in road accident severity analysis. *International Journal of Advanced Computer Science and Applications*, *16*(4). https://doi.org/10.14569/ijacsa.2025.0160429. 2025.

Maestre-Góngora G. Angarita-Zapata, J. S. and J. F. Calderín. A bibliometric analysis and benchmark of machine learning and automl in crash severity prediction: The case study of three colombian cities. *Sensors*, *21*(24), 8401. https://doi.org/10.3390/s21248401. 2021.

Liang X.-Mengistu K. T. So R. S. Wei X. He B. Ardakani, S. P. and A. Cheshmehzangi. Road car accident prediction using a machine-learning-enabled data analysis. *Sustainability*, *15*(7), 5939. https://doi.org/10.3390/su15075939. 2023.

Chen F. Khan-I. Khahro S. H. Muhammad A. M. Memon Z. A. Aziz, K. and A. Khattak. Road traffic crash severity analysis: A bayesian-optimized dynamic ensemble selection guided by instance hardness and region of competence strategy. *IEEE Access*, *12*. https://doi.org/10.1109/access.2024.3465489. 2024.

Gates T. Chakraborty, M. and S. Sinha. Causal analysis and classification of traffic crash injury severity using machine learning algorithms. *Data Science for Transportation*, *5*(2). https://doi.org/10.1007/s42421-023-00076-9. 2021.

Wisutwattanasak P. Se-C. Banyong C. Jomnonkwao S. Champahom, T. and V. Ratanavaraha. Analysis of factors associated with highway personal car and truck run-off-road crashes: Decision tree and mixed logit model with heterogeneity in means and variances approaches. *Informatics*, *10*(3), 66. https://doi.org/10.3390/informatics10030066. 2023.

Araújo G. Dias, J. and C. Amado. Crash injury severity prediction for passenger cars with machine learning algorithms. *The European Journal of Public Health*, *35*(suppl. 1). https://doi.org/10.1093/eurpub/ckaf165.111. 2025.

Vélez-J. I. Ferreira-Vanegas, C. M. and G. A. García-Llinás. Analytical methods and determinants of frequency and severity of road accidents: A 20-year systematic literature review. *Journal of Advanced Transportation*, *2022*, 7239464. https://doi.org/10.1155/2022/7239464. 2022.

Prakash R. Karunakar-V. J. Gade R. Sagar P. V. Guntaka, J. and A. D. Kumar. Leveraging machine learning for road accident analysis. *International Journal of Membrane Science and Technology*, *10*(4). https://doi.org/10.15379/ijmst.v10i4.2223. 2023.

N. Hamdan and T. Sipos. Predicting crash severity on the hungarian road network: An ensemble machine learning approach with resampling and hyperparameter tuning. *IEEE Access*, *13*. https://doi.org/10.1109/access.2025.3630827. 2025.

Lan L. Zahid-M. Ijaz, M. and A. Jamal. A comparative study of machine learning classifiers for injury severity prediction of crashes involving three-wheeled motorized rickshaw. *Accident Analysis & Prevention*, *154*, 106094. https://doi.org/10.1016/j.aap.2021.106094. 2021.

M. B. B. Khedher and D. Yun. An interpretable machine learning-based hurdle model for zero-inflated road crash frequency data analysis: Real-world assessment and validation. *Applied Sciences*, *14*(23), 10790. https://doi.org/10.3390/app142310790. 2024.

Hasan-M. M. Elhenawy M. Yasmin S. Masoud M. Komol, M. M. R. and A. Rakotonirainy. Crash severity analysis of vulnerable road users using machine learning. *PLoS ONE*, *16*(8), e0255828. https://doi.org/10.1371/journal.pone.0255828. 2021.

Tiwari G. Kumar, P. and S. B. Paul. Road safety studies at micro, meso, and macroscopic levels: A systematic review. *IATSS Research*, *49*(1). https://doi.org/10.1016/j.iatssr.2024.12.001. 2024.

Castillo Sequera J. L. Lacherre, J. and D. Mauricio. Factors, prediction, and explainability of vehicle accident risk due to driving behavior through machine learning: A systematic literature review, 2013–2023. *Computation*, *12*(7), 131. https://doi.org/10.3390/computation12070131. 2024.

K. Le. Improving road safety: Supervised machine learning analysis of factors influencing crash severity. *Scientific Journal of Silesian University of Technology. Series Transport*, *127*, 97–112. https://doi.org/10.20858/sjsutst.2025.127.8. 2025.

Pu Z. Cui Z. Lee S. Guo X. Li, S. and D. Ngoduy. Inferring heterogeneous treatment effects of crashes on highway traffic: A doubly robust causal machine learning approach. *ArXiv*. https://doi.org/10.48550/arxiv.2401.00781. 2024.

Li S. Xu N. Guo X. Liang, X. and Z. Pu. A heterogeneous effects analysis method of highway crash factors based on causal framework. *SAE Technical Paper Series*. https://doi.org/10.4271/2025-01-7185. 2025.

N. Niture and I. Abdellatif. A systematic review of factors, data sources, and prediction techniques for earlier prediction of traffic collision using ai and machine learning. *Multimedia Tools and Applications*, *83*(30). https://doi.org/10.1007/s11042-024-19599-6. 2024.

Dias J. Santos, K. and C. Amado. A literature review of machine learning algorithms for crash injury severity prediction. *Journal of Safety Research*, *80*, 254–269. https://doi.org/10.1016/j.jsr.2021.12.007. 2021.

Hickman J. McDonald A. D. Huang W. Vogelpohl T. Sarkar, A. and G. Markkula. Steering or braking avoidance response in shrp2 rear-end crashes and near-crashes: A decision tree approach. *Accident Analysis & Prevention*, *154*, 106055. https://doi.org/10.1016/j.aap.2021.106055. 2021.

Oughali F. C. Assi K. Ratrout N. Jamal A. Sattar, K. and S. M. Rahman. Transparent deep machine learning framework for predicting traffic crash severity. *Neural Computing and Applications*, *35*(2), 1535–1547. https://doi.org/10.1007/s00521-022-07769-2. 2022.

Champahom T. Jomnonkwao S. Boonyoo T. Karoonsoontawong A. Se, C. and V. Ratanavaraha. Pickup truck crash severity analysis via machine learning: Policy insights for developing countries. *International Journal of Injury Control and Safety Promotion*, *32*(1). https://doi.org/10.1080/17457300.2025.2504975. 2025.

Shi X. Zhang Y. Shiwakoti N. Xu Y. Shao, Y. and Z. Ye. Injury severity prediction and exploration of behavior-cause relationships in automotive crashes using natural language processing and extreme gradient boosting. *Engineering Applications of Artificial Intelligence*, *133*, 108542. https://doi.org/10.1016/j.engappai.2024.108542. 2024.

Nojoumian M. Dang N. Skaug, L. and A. Yap. Road crash analysis and modeling: A systematic review of methods, data, and emerging technologies. *Applied Sciences*, *15*(13), 7115. https://doi.org/10.3390/app15137115. 2025.

Cheema M. A. Ali M. E. Toosi A. Sohail, A. and H. A. Rakha. Data-driven approaches for road safety: A comprehensive systematic literature review. *Safety Science*, *158*, 105949. https://doi.org/10.1016/j.ssci.2022.105949. 2023.

Gohil B. N. Srivastava, N. and S. Ray. Road traffic accident severity prediction using causal inference and machine learning. in *Proceedings of the 8th International Conference on Data Science and Management of Data (12th ACM IKDD CODS and 30th COMAD).* https://doi.org/10.1145/3703323.3703749. 2024.

Mishra S. Velaga N. Thapa, D. and G. R. Patil. Advancing proactive crash prediction: A discretized duration approach for predicting crashes and severity. *Accident Analysis & Prevention*, *190*, 107407. https://doi.org/10.1016/j.aap.2023.107407. 2023.

R. Verma and M. M. Agarwal. A systematic review on road accident prediction: A special attention toward machine learning and deep learning approaches. *Journal of Circuits, Systems and Computers*, *34*(5). https://doi.org/10.1142/s0218126625300090. 2025.

Xie Y. Jiang L. Pu Z. Wen, X. and T. Ge. Applications of machine learning methods in traffic crash severity modelling: Current status and future directions. *Transport Reviews*, *41*(6), 855–879. https://doi.org/10.1080/01441647.2021.1954108. 2021.

Y. Xiao and Z. Duan. An explainable multi-task deep learning framework for crash severity prediction using multi-source data. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-09226-1. 2025.

Li Z. Liu P. Chen K. Zhao, J. and X. Zhai. Impact of road network patterns on traffic crash prediction: An interpretable geography machine learning approach. *Journal of Transportation Safety & Security*. https://doi.org/10.1080/19439962.2025.2529832. 2025.