Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Predictive Limits of Roadway Infrastructure for Crash Injury Severity: Comparing Pre-Crash and Event-Informed Machine Learning Models

Mohammad Ridwan Tanvir

motan2@morgan.edu

Morgan State University

Fahim Jaman

Morgan State University

Research Article

Keywords:

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11435714/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Predictive Limits of Roadway Infrastructure for Crash Injury Severity: Comparing Pre-Crash and Event-Informed Machine Learning Models

Mohammad Ridwan Tanvir[1*] and Fahim Jaman[1]

[1*]Department of Civil, Environment and Transportation Engineering, Morgan State University, 1700 E Cold Spring Ln, Baltimore, Maryland, United States.

*Corresponding author(s). E-mail(s): motan2@morgan.edu;
Contributing authors: fajam1@morgan.edu;

Abstract

Crash-severity models often combine roadway characteristics with variables observed only during or after a collision, making it difficult to determine how much severity information is available for proactive safety applications. This study develops an information-availability framework for multiclass crash injury severity prediction using 2024 Ohio Highway Safety Information System (HSIS) data. Of 251,091 crash records, 219,491 (87.4%) were linked to roadway inventory segments. Severity was classified as fatal/serious injury (KA), minor/possible injury (BC), or property-damage-only (O). Four nested information sets were evaluated: roadway inventory attributes (M1), roadway and traffic exposure (M2), environmental, temporal, and crash-location context (M3), and crash-event and aggregated unit-record information (M4). LightGBM was applied across all models, with multinomial logistic regression and XGBoost additionally evaluated for M3. Cost-sensitive learning and route-grouped validation were used. On the test set, LightGBM KA ROC-AUC increased from 0.602 in M1 to 0.895 in M4, while KA PR-AUC increased from 0.044 to 0.282. M3 achieved KA ROC-AUC of 0.662 and PR-AUC of 0.066, with stable three-fold cross-validation performance. SHAP analysis identified hour of crash, posted speed, AADT, road contour, and road condition as leading M3 predictors. The results show that predictor availability has greater influence on performance than algorithm choice and should be aligned with the intended safety application. This distinction is particularly important when models are intended for proactive roadway screening.

Keywords: Traffic safety, Machine learning, Crash severity, Class Imbalance

1 Introduction

Crash injury severity remains central to transportation safety research as the consequences of a crash can vary substantially, ranging from property damage to fatal injury. These outcomes are influenced by interacting factors such as roadway design, traffic conditions, environmental conditions, vehicle characteristics, and human behavior (Liang et al. 2024). Crash-severity modeling is therefore widely used to identify factors associated with serious outcomes and to support safety-related decision making. Many crash-severity models combine roadway and traffic characteristics with variables that become known only during or after a collision. These may include the first or most harmful event, impact location, collision action, vehicle maneuver, and detailed contributing circumstances (Savolainen et al. 2011). Although such variables are useful for retrospective crash classification and injury assessment, they provide limited value for proactive roadway-safety planning because they are unknown when agencies evaluate roadway conditions before a crash occurs.

Roadway characteristics, in contrast, are properties that are generally known in advance of the actual crash event. Previous studies have reported the relationship between crash severity and roadway properties such as road geometry (Haghighi et al. 2018; Liang et al. 2024; Miaou and Lum 1993), surface characteristics (Lebaku et al. 2026), and terrain types (Awasthi et al. 2024; Choi et al. 2011). Variables such as posted speed limit, roadway width, number of lanes, functional classification, facility type, traffic volume, and surface characteristics can be evaluated independently of a particular crash incident. Environmental and temporal conditions, such as lighting, weather, road-surface condition, and time of day, may also be available before or at the onset of an event. Distinguishing these variables from crash-event characteristics is therefore important because a model that performs well using detailed collision information answers a different question from a model intended to identify roadway environments where severe outcomes may be more likely.

This distinction has become increasingly relevant as crash-severity analysis has shifted toward data-driven methods. Traditional studies have relied on discrete-outcome statistical models to estimate relationships between injury severity and roadway, driver, vehicle, traffic, and environmental characteristics (Savolainen et al. 2011). However, crash data often contain nonlinear relationships, interactions, and unobserved heterogeneity that are difficult to represent using a single predefined functional form (Liang et al. 2024; Mannering et al. 2016). These challenges have encouraged the use of machine-learning approaches, including tree-based models, gradient boosting, deep learning, ensemble methods, and explainable artificial intelligence, to improve predictive performance and model interpretation (Iranitalab and Khattak 2017; Madushani et al. 2023; Sattar et al. 2023; Somvanshi et al. 2026).

Crash-severity modeling is also complicated by class imbalance because fatal and serious-injury crashes are much less common than property-damage-only crashes, making overall accuracy an incomplete measure of model performance (Amiri et al. 2025; Fiorentini and Losa 2020). More importantly, high predictive performance may depend heavily on crash-event information that is unavailable before a collision occurs.

Although many studies have examined crash-severity prediction, relatively little attention has been given to separating information available before a crash from information generated during the crash event and quantifying the predictive contribution of each.

This study addresses that gap using the 2024 Ohio Highway Safety Information System (HSIS) crash, roadway, traffic, and vehicle-unit records. Crash records are linked to corresponding roadway inventory and vehicle-unit information, and injury severity is classified into three categories: fatal or serious injury, minor or possible injury, and property-damage-only crashes. The analysis uses four progressively expanded information sets. The first includes roadway infrastructure characteristics only; the second adds traffic exposure variables; the third incorporates environmental and temporal context; and the fourth adds crash-event and vehicle-unit information available during or after the collision. Comparing these nested information sets allows the study to quantify how predictive performance changes as increasingly event-specific information is introduced. The study proposes several key steps in its methodology. It separates roadway, traffic, contextual, and crash-event information; quantifies the added predictive value of each information layer; and applies explainable machine learning to identify factors associated with fatal and serious-injury crashes. Together, these contributions provide a more realistic assessment of the practical limits of crash-severity prediction for proactive transportation-safety planning.

2 Literature Review

Crash severity modeling techniques are commonly employed in predicting discrete outcomes with ordered or unordered severity levels. Literature presents various statistical approaches such as ordered logit/probit, multinomial logit, nested structures, random-parameter specifications, latent-class models, and related extensions. Savolainen et al. (2011) presented that injury severity reflects complex interactions among human, vehicle, roadway, and crash mechanisms by reviewing the major methodological alternatives. Mannering et al. (2016) further argued that unobserved heterogeneity can materially affect crash-model estimates and transferability. These statistical approaches have remained relevant since their parameters are interpretable within a probabilistic framework, although parametric assumptions can be limited when relationships are highly nonlinear or interactive (Liang et al. 2024). Machine-learning methods have therefore become increasingly common in transportation-safety research because they can represent nonlinear relationships, complex interactions, and mixed data structures with relatively limited functional assumptions. Tree-based methods such as XGBoost and LightGBM have been widely adopted because of their predictive performance and scalability (Chen and Guestrin 2016; Ke et al. 2017). Deep-learning and hybrid approaches have also been applied to tabular crash data (Sattar et al. 2023; Somvanshi et al. 2026). Despite performance improvement of the crash modeling techniques, these approaches have shown limited explanatory power for all severity classes, especially for the critical but rare types of crashes. Class imbalance remains a major challenge because fatal and serious-injury crashes typically account for only a small proportion of reported crashes. Iranitalab and Khattak (2017) compared multinomial logit with several machine-learning classifiers and showed that model performance can

vary substantially across evaluation criteria, particularly when severe outcomes are much less frequent than lower-severity crashes. Fiorentini and Losa (2020) showed that ignoring this imbalance can lead to weak performance for minority classes, even when overall accuracy appears high. Subsequent studies have examined approaches such as resampling, class weighting, and threshold adjustment to improve severe-class identification (Amiri et al. 2025). These findings highlight the importance of evaluating crash-severity models with class-sensitive measures, including precision, recall, F1-score, and precision-recall metrics, rather than relying only on overall accuracy.

Explainable artificial intelligence has also become important as more complex models are used in transportation safety. SHapley Additive exPlanations (SHAP) provides feature-attribution values that indicate how individual variables contribute to model predictions and allows local explanations to be summarized into global importance measures (Lundberg and Lee 2017). Numerous transportation studies have applied SHAP and related methods to interpret roadway, environmental, vehicle, and crash-related factors (Madushani et al. 2023; Wei et al. 2023). These approaches improve model transparency, although feature importance should not be interpreted as evidence of causality.

Roadway characteristics have long been included in crash-severity models because operating speed, roadway cross-section, alignment, access, and facility type can influence both crash conditions and injury outcomes. Cheng et al. (2025) examined roadway determinants using a machine-learning framework and demonstrated the usefulness of nonlinear model interpretation. Gao et al. (2026) extended this direction by incorporating detailed roadway geometry with explainable ensemble learning and spatial and temporal validation. Previous studies have also investigated how the spatial configuration of roadway features (Awasthi et al. 2024; Choi et al. 2011) and deviations from geometric design standards (Malyshkina and Mannering 2010) influence crash occurrence and severity. These studies indicate that roadway attributes provide useful information for severity modeling and may exhibit complex nonlinear relationships with crash outcomes.

Model validation is another important issue in crash-severity analysis. Random train-test splits can place crashes from the same or similar roadway environments in both the training and testing samples, which may produce optimistic estimates of predictive performance. Spatial, temporal, or grouped validation provides a stricter assessment by testing whether a model can generalize to locations or periods that were not represented during training. This is particularly important when models are intended for roadway-level safety applications, where performance on previously unseen road segments or routes is more relevant than performance on randomly withheld crash records.

Despite these advances, an important distinction remains in the type of information available to crash-severity models. Roadway characteristics such as posted speed, functional class, roadway width, and traffic volume are known prior to crash incidents. Additional contextual information, including weather, lighting, road-surface condition, and time of day, may also be available without knowledge of the eventual collision mechanism. In contrast, variables such as harmful event, impact area, collision action, and detailed contributing circumstances become known only during or after

the crash. Pooling these variables can produce strong predictive models while obscuring how much severity information is actually available for proactive roadway-safety planning. Although previous research has extensively compared algorithms, imbalance treatments, and explainability methods, relatively limited attention has been given to quantifying predictive performance across these different information-availability levels. This study addresses that gap by systematically comparing roadway, traffic, contextual, and crash-event information within the proposed modeling and validation framework.

3 Data and Methodology

3.1 Data Preparation

The analysis used 2024 Ohio data obtained through the Federal Highway Administration Highway Safety Information System (FHWA 2026). The analyzed datasets contained 251,091 crash records, 405,667 roadway inventory segments, and 449,273 unit records. Crash records were linked to roadway inventory records using route identifier and milepost. A crash was assigned to a roadway segment when its *route_id* matched the segment *route_id* and its reported milepost fell between the segment beginning (*begmp*) and ending (*endmp*) mileposts. This procedure linked 219,491 crashes, representing 87.4% of the crash records. Vehicle-unit records were linked directly to crashes using the common *document_nbr* identifier. For crashes involving multiple units, unit-level characteristics were aggregated to the crash level rather than selecting a single vehicle.

3.2 Crash Severity Classification and Analytical Sample

The HSIS crash severity information is coded in five ordered categories: fatal, suspected serious injury, suspected minor injury, possible injury, or property-damage-only. For modeling purpose, these five levels were consolidated into three classes consistent with a KABCO-style severity hierarchy: KA (fatal or suspected serious injury), BC (suspected minor or possible injury), and O (property-damage-only). This consolidation preserves a policy-relevant severe-injury class while providing sufficient observations for multiclass machine learning. As shown in Table 1, the linked sample contained 6,508 KA crashes (2.97%), 50,473 BC crashes (23.00%), and 162,510 O crashes (74.04%). Table 2 presents the descriptive statistics of selected roadway and traffic characteristics across the three crash-severity classes. KA crashes were associated with higher posted speed limits (mean 41.42 mph), narrower roadways (mean 41.93 ft), and lower traffic volumes (mean AADT 12,966) compared with BC and O crashes, suggesting potential differences in roadway environments across severity classes.

3.3 Predictor Sets and Data-quality Screening

The methodological design uses nested predictor sets based on information availability relative to the crash, named as M1, M2, M3 and M4 (Table 3). M1 contains roadway infrastructure characteristics, M2 adds traffic exposure, and M3 adds contextual conditions that exist before or at the onset of a crash and can, in principle, be observed

Table 1 Severity distribution in the roadway-linked analytical sample

Severity class	Definition	N	Percent (%)
KA	Fatal and suspected serious injury	6,508	2.97
BC	Suspected minor and possible injury	50,473	23.00
O	Property damage only	162,510	74.04
Total		219,491	100.00

Table 2 Descriptive statistics of selected roadway and traffic variables by crash-severity class

	KA		BC		O	
Variable	Mean	Median	Mean	Median	Mean	Median
Posted speed (mph)	41.42	45	39.07	35	38.05	35
Number of lanes	2.77	2	2.98	2	2.91	2
Roadway width (ft)	41.93	32	46.67	40	46.98	38
AADT	12,966	6,434	16,619	9,452	16,262	9,105
Truck share (%)	6.76	4.44	5.98	3.70	6.27	3.77

without knowing the collision mechanism. M4 adds event/unit variables that describe the crash mechanism or involved unit and therefore become available during or after the event. The sets are nested, therefore, M1 is a subset of M2, M2 a subset of M3, and M3 a subset of M4. Incremental changes in performance therefore quantify the predictive information added by each layer.

Candidate variables were screened for missing values and redundancy before modeling. Posted speed (1.5% missing), lane count (0.2%), functional class (0.2%), facility type (0.2%), and surface type (1.2%) were nearly complete; roadway width and AADT had 17.2% and 18.1% missingness, respectively, while truck AADT was missing for approximately 31%. By contrast, median type/width were missing for approximately 88% of linked crashes, and outside shoulder width and access control were missing for approximately 59%. Variables with significant number of missing values were excluded from the final model rather than allowing structural missingness to become a dominant signal. Truck percentage was derived as $100 \times$ truck AADT / total AADT where both quantities were valid. Traffic control was retained in M4 because it was recorded as event information rather than as a complete roadway-inventory field; in a deployment with a comprehensive control inventory, that variable could be classified as pre-event information.

3.4 Modeling Framework

LightGBM was used as the primary modeling framework for the nested information sets M1–M4. Using the same algorithm across all four specifications allowed changes in predictive performance to be attributed primarily to the additional information

Table 3 Nested information sets used for model comparison

Feature set	Variables	Information type
M1 (Infrastructure)	Posted speed; number of lanes; roadway width; divided/undivided; functional class; facility type; surface type	Roadway inventory attributes
M2 (Exposure)	M1 + AADT; truck percentage	Traffic exposure and vehicle mix
M3 (Context)	M2 + weather; lighting; road condition; road contour; hour; day of week	Environmental, temporal, and crash-location context
M4 (Event)	M3 + number of units; pedestrians; unit type; pre-crash action; contributing circumstance; collision action; impact area; most harmful event; occupants; vehicle year; traffic control	Crash-event and unit-record information

introduced at each stage rather than to differences in modeling technique. LightGBM was selected because of its suitability for large tabular datasets and its ability to accommodate nonlinear relationships, categorical predictors, and missing values (Ke et al. 2017).

To assess whether the findings were sensitive to algorithm choice, M3, was additionally estimated using multinomial logistic regression and XGBoost. M3 was selected for this comparison as it incorporates roadway inventory, traffic exposure, and environmental and temporal context while excluding the crash-event and unit-record information introduced in M4. Multinomial logistic regression provided a conventional statistical baseline, while XGBoost provided an alternative gradient-boosting approach (Chen and Guestrin 2016). The M3 comparison therefore examined whether improvements in the proactive prediction model were primarily attributable to the available information or to the choice of algorithm. M4 was retained as an event-informed benchmark and was estimated using LightGBM only.

Extensive hyperparameter optimization was not within the scope of the study. Instead, moderate model complexity was maintained to support consistent comparison across the nested information sets. The LightGBM specification used 60 boosting iterations, a learning rate of 0.06, 31 leaves, row and feature subsampling rates of 0.85, L1 regularization of 0.05, and L2 regularization of 1.0. XGBoost used 100 boosting iterations, a maximum depth of 5, a learning rate of 0.06, row and feature subsampling rates of 0.85, and a multiclass probabilistic objective.

3.5 Class Imbalance and Validation

As fatal and serious-injury (KA) crashes represented only 2.97% of the linked sample, overall accuracy was considered insufficient for evaluating model performance. Final model performance was therefore assessed using Macro F1, balanced accuracy, KA

precision and recall, KA ROC-AUC, and KA precision-recall area under the curve (PR-AUC). For class-weight selection, mean Macro F1 across the route-grouped cross-validation folds was used as the primary criterion.

Class imbalance was addressed using cost-sensitive learning, in which larger training weights were assigned to the less frequent severity classes. In general, the weighted training objective can be expressed as

$$L_{\text{weighted}} = \sum_{i=1}^{N} w_{y_i} L\left(y_i, \hat{p}_i\right), \tag{1}$$

where N is the number of training observations, y_i is the observed severity class for crash i, $\hat{p}_i$ represents the predicted class probabilities, $L(\cdot)$ is the underlying classification loss, and w_{y_i} is the weight assigned to the observed severity class.

Five O:BC:KA weighting schemes were evaluated: 1:1:1, 1:2:5, 1:2:8, 1:2:12, and 1:2:16. The schemes were compared using three-fold route-grouped cross-validation (CV) within the training sample. Mean Macro F1 was used as the selection criterion because it gives equal importance to the performance of each severity class regardless of class frequency. For the three severity classes, Macro F1 was calculated as

$$\text{Macro F1} = \frac{1}{3}\left(F1_O + F1_{BC} + F1_{KA}\right). \tag{2}$$

As shown in Table 4, the 1:2:8 weighting scheme achieved the highest mean Macro F1 of 0.334 and was therefore selected for the final models. Increasing the KA weight beyond 8 improved KA recall but reduced KA precision and overall multiclass performance. For example, increasing the KA weight from 8 to 16 raised KA recall from 0.100 to 0.336, while KA precision decreased from 0.114 to 0.072 and Macro F1 decreased from 0.334 to 0.311. This pattern indicates that increasingly large KA weights improved severe-crash detection at the cost of a greater false-positive burden.

Table 4 Class-weight sensitivity analysis using route-grouped cross-validation

O:BC:KA	Macro F1	KA Precision	KA Recall	KA F1
1:1:1	0.283	0.000	0.000	0.000
1:2:5	0.316	0.136	0.016	0.028
1:2:8	**0.334**	**0.114**	**0.100**	**0.106**
1:2:12	0.323	0.085	0.220	0.122
1:2:16	0.311	0.072	0.336	0.118

Model generalization was evaluated using a route-grouped train–test design. Approximately 80% of the linked crashes were assigned to the training set and 20% to the test set, with each route identifier assigned entirely to one partition. This approach prevented crashes from the same route from appearing in both training and test sets, reducing potential spatial information leakage. The resulting samples contained 175,020 training crashes and 44,471 test crashes. KA crashes represented 2.98% of the

training set and 2.91% of the test set. Within the training sample, three route-grouped CV folds were used for class-weight selection and internal validation. All crashes associated with a given route remained within the same fold. The test routes were not used during class-weight selection or model development. After the modeling choices were established, the final models were fitted using the complete training sample and evaluated once on the test set.

3.6 Performance Measures and Explainability

Model performance was evaluated using multiclass macro ROC-AUC, KA-versus-rest ROC-AUC, KA precision-recall area under the curve (PR-AUC), balanced accuracy, Macro F1, and KA-specific precision, recall, and F1. These measures were selected to account for the substantial class imbalance in the severity outcome and to evaluate both overall multiclass performance and the ability of the models to identify the relatively rare KA crashes. PR-AUC is particularly informative for the KA class because its no-skill baseline corresponds to the prevalence of the positive class, which was approximately 0.029 in the test set. Predicted class labels were assigned according to the class with the highest predicted probability. No post-hoc decision-threshold optimization was performed using the test set.

Model explainability was examined using SHAP values for the LightGBM M3 model (Lundberg and Lee 2017). M3 was selected for interpretation because it represents the most comprehensive specification based on roadway, traffic, environmental, and temporal information without introducing detailed crash-event and unit-record variables. KA-class feature contributions were obtained using LightGBM's native contribution calculation for a sample of 1,000 observations from the test set.

Global feature importance was summarized using the mean absolute SHAP value. This measure represents the average magnitude of a feature's contribution to the KA model score, regardless of direction. For selected roadway variables, directional patterns were also examined by averaging KA-class SHAP values across observed categories or posted-speed values. Positive SHAP values indicate that a feature value increases the model score for KA relative to the model baseline, whereas negative values decrease that score. SHAP values describe the behavior of the fitted predictive model and should therefore be interpreted as model-based associations rather than causal effects.

4 Results

4.1 Nested information value

Table 5 presents test-set performance across the four nested LightGBM information sets. Incremental KA-class ROC-AUC and precision-recall AUC across the four nested information sets are presented in Figure 1 and 2. The roadway-inventory model (M1) provided limited discrimination of KA crashes, with a KA ROC-AUC of 0.602 and KA PR-AUC of 0.044. Adding traffic exposure in M2 produced essentially no improvement:

KA ROC-AUC was 0.600 and KA PR-AUC remained 0.044. The addition of environmental, temporal, and crash-location context in M3 produced a clearer improvement, increasing KA ROC-AUC to 0.662 and KA PR-AUC to 0.066.

The largest performance increase occurred when crash-event and aggregated unit-record information was introduced in M4. KA ROC-AUC increased to 0.895 and KA PR-AUC to 0.282. Relative to the KA prevalence of approximately 0.029 in the test set, the PR-AUC of M3 was about 2.3 times the no-skill baseline, whereas M4 was about 9.7 times the baseline. Macro ROC-AUC similarly increased from 0.562 in M1 to 0.818 in M4. This pattern indicates that the largest gain in discrimination occurred after detailed event and unit-record information became available.

The class-prediction results further distinguish ranking ability from final classification performance. M1 did not identify KA crashes as the most probable class, yielding zero KA recall and F1. M2 detected only a small proportion of KA crashes, with recall of 0.004 and F1 of 0.007. M3 improved KA recall to 0.110 and KA F1 to 0.112 under the selected class weighting, while M4 increased KA recall to 0.464 and KA F1 to 0.333. Thus, the pre-crash information sets contained useful ranking signal, but substantially stronger severe-crash classification became possible only after event-specific information was introduced.

Table 5 Test-set performance across nested information sets (LightGBM)

Set	Macro ROC-AUC	KA ROC-AUC	KA PR-AUC	Bal. acc.	Macro F1	KA Recall	KA F1
M1	0.562	0.602	0.044	0.333	0.285	0.000	0.000
M2	0.563	0.600	0.044	0.334	0.287	0.004	0.007
M3	0.604	0.662	0.066	0.367	0.338	0.110	0.112
M4	0.818	0.895	0.282	0.580	0.542	0.464	0.333

4.2 Algorithm comparison for the pre-crash/context model

Since M3 is the richest specification that does not rely on detailed crash-event or unit-record information, it was used for the algorithm comparison. LightGBM provided the strongest overall performance, with KA ROC-AUC of 0.662, KA PR-AUC of 0.066, balanced accuracy of 0.367, and Macro F1 of 0.338. XGBoost was close, with KA ROC-AUC of 0.660, KA PR-AUC of 0.065, and Macro F1 of 0.324. In contrast, multinomial logistic regression was weaker, with KA ROC-AUC of 0.643, KA PR-AUC of 0.055, and Macro F1 of 0.312 (Table 6).

The relatively small difference between LightGBM and XGBoost suggests that the substantive conclusion for M3 may not depend on a single boosting implementation. The multinomial logistic model also retained discrimination above chance, indicating that part of the pre-crash severity signal can be captured without nonlinear boosting, although the tree-based models provided a modest improvement in overall and KA-specific performance.

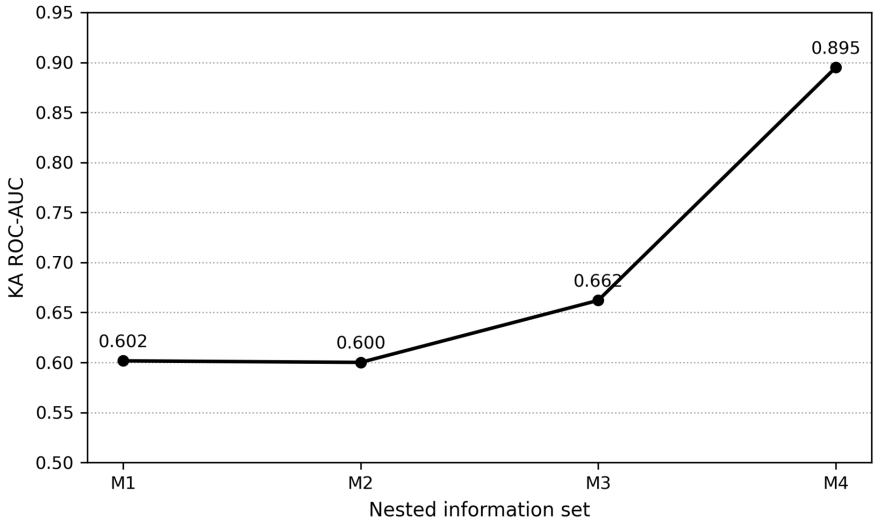

Fig. 1 Incremental KA-class ROC-AUC across the four nested information sets

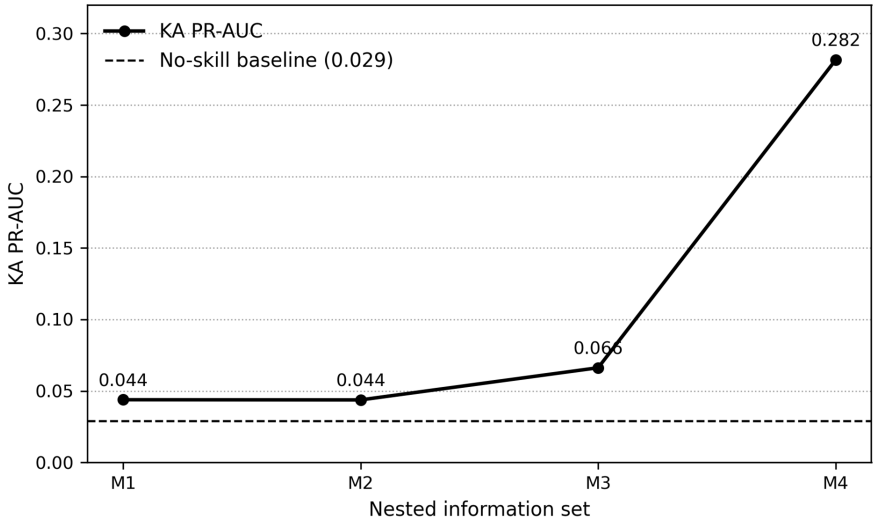

Fig. 2 Incremental KA-class precision-recall AUC across the four nested information sets

Table 6 Test-set algorithm comparison for M3

Model	Macro ROC-AUC	KA ROC-AUC	KA PR-AUC	Bal. acc.	Macro F1	KA Recall	KA F1
Multinomial logistic	0.589	0.643	0.055	0.350	0.312	0.067	0.078
XGBoost	0.601	0.660	0.065	0.359	0.324	0.090	0.098
LightGBM	0.604	0.662	0.066	0.367	0.338	0.110	0.112

4.3 Route-grouped cross-validation stability

As illustrated in Table 7, three-fold route-grouped cross-validation within the training sample produced stable M3 performance across alternative route partitions. KA ROC-AUC values were 0.671, 0.678, and 0.675, yielding a mean of 0.675 ± 0.004. KA PR-AUC averaged 0.066 ± 0.001, Macro ROC-AUC averaged 0.607 ± 0.002, balanced accuracy averaged 0.364 ± 0.003, and Macro F1 averaged 0.334 ± 0.003. The final test-set KA ROC-AUC for M3 (0.662) was somewhat lower than the mean internal cross-validation value (0.675), suggesting that the independent test routes were modestly more challenging. Nevertheless, the narrow variation across the three route-grouped folds indicates that the M3 results were not driven by a single favorable training partition.

Table 7 Three-fold route-grouped cross-validation results for M3 LightGBM

Fold	Macro ROC-AUC	KA ROC-AUC	KA PR-AUC	Bal. acc.	Macro F1
1	0.605	0.671	0.064	0.364	0.333
2	0.609	0.678	0.066	0.361	0.331
3	0.607	0.675	0.066	0.367	0.337
Mean ± SD	0.607 ± 0.002	0.675 ± 0.004	0.066 ± 0.001	0.364 ± 0.003	0.334 ± 0.003

4.4 Explainable machine learning results

The explainability analysis focused on the M3 LightGBM model because M3 is the most comprehensive specification based on roadway, traffic, environmental, temporal, and crash-location information without detailed event or unit-record variables. Figure 3 summarizes global KA-class feature importance. Hour of crash ranked first (mean absolute contribution 0.143), followed by posted speed (0.129) and AADT (0.084). Road contour (0.066), road condition (0.064), day of week (0.054), lighting (0.046), and roadway width (0.041) formed the next tier of predictors.

Among the roadway-inventory variables, posted speed was the dominant contributor. The posted-speed attribution pattern was nonlinear (Fig. 4). Within the 1,000-observation SHAP sample, and retaining posted-speed groups represented by at least 30 observations, mean KA-class contributions were negative at 25, 35, and 40 mph, changed sign at 45 mph, and were positive at 55 and 65 mph. The largest mean contribution among the retained speed groups occurred at 55 mph (0.223). This pattern should not be interpreted as a causal speed-limit threshold because posted speed is correlated with facility type, access environment, roadway geometry, and operating conditions. Rather, it shows how the fitted M3 model used the posted-speed environment when assigning KA severity scores after accounting for the other included predictors.

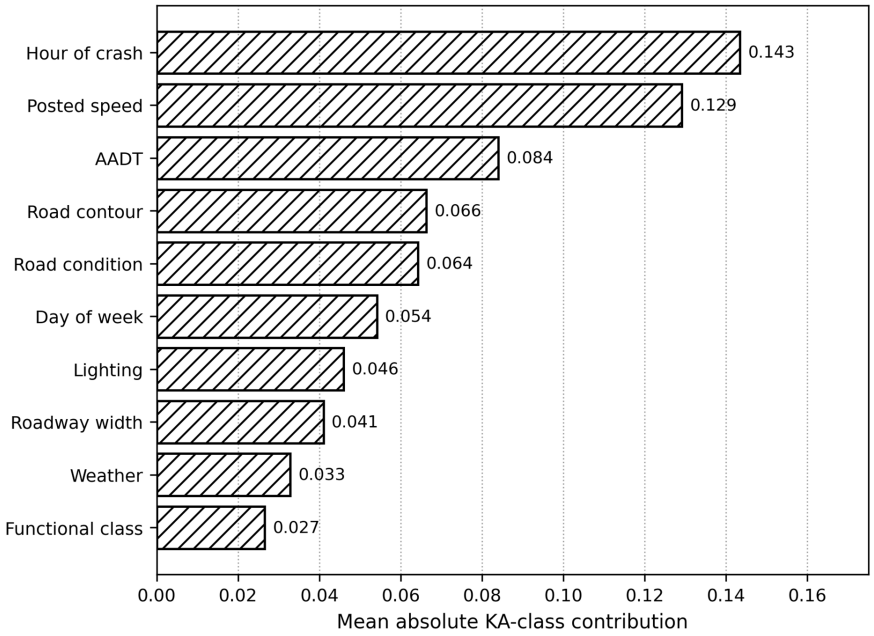

Fig. 3 Global mean absolute SHAP values for the KA class in the M3 LightGBM model

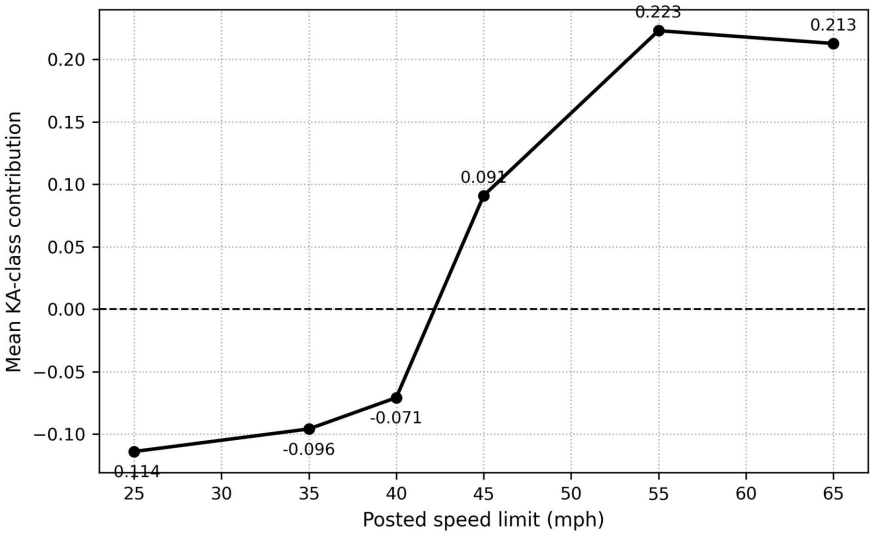

Fig. 4 Directional relationship between posted speed and mean KA-class contribution in the M3 model

5 Discussion

The results show that crash-severity prediction is strongly influenced by the information available to the model. Roadway inventory alone provided limited KA discrimination (ROC-AUC = 0.602), and adding traffic exposure in M2 produced essentially no improvement (0.600). Contextual information in M3 increased KA ROC-AUC to 0.662 and PR-AUC to 0.066, whereas adding crash-event and unit-record information in M4 produced a much larger increase to 0.895 and 0.282, respectively. Thus, the largest gain occurred only after information describing the crash mechanism became available. This does not imply that roadway characteristics are unimportant; rather, their ability to classify the severity of an individual crash is inherently more limited than that of event-specific information.

This distinction is important for model deployment. Models intended for roadway screening or infrastructure prioritization should be evaluated using predictors available before, or independently of, the crash event. Variables such as pre-crash action, impact area, contributing circumstance, and most harmful event can substantially improve prediction but are generally unavailable for proactive screening. The algorithm comparison supports this interpretation: LightGBM and XGBoost produced nearly identical M3 performance (KA ROC-AUC = 0.662 and 0.660), while the gain from M3 to M4 was much larger. Route-grouped cross-validation was also stable, with mean KA ROC-AUC of 0.675 ± 0.004, suggesting that the main conclusion was not dependent on a particular route partition or boosting implementation.

Class imbalance remained an important challenge. Equal class weighting failed to identify KA crashes during cross-validation, whereas the selected 1:2:8 weighting scheme improved Macro F1 and enabled severe-crash detection. Higher KA weights increased recall but reduced precision and overall balanced performance. This illustrates why ROC-AUC alone is insufficient: M1 and M2 retained some ranking ability even though KA classification under the maximum-probability rule was essentially zero.

SHAP analysis of M3 showed that hour of crash, posted speed, AADT, road contour, and road condition were the most influential predictors. Posted speed was the strongest persistent roadway attribute, with higher-speed environments contributing more positively to KA predictions. These patterns should be interpreted as predictive associations rather than causal effects because roadway speed, geometry, facility type, and operating conditions are interrelated.

Several limitations should be considered. The analysis was based on a single year of Ohio data, so temporal and geographic generalizability remain uncertain. Roadway linkage was achieved for 87.4% of crashes, and several roadway variables with substantial missingness were excluded. Extensive hyperparameter optimization was also intentionally avoided because the study focused on comparing information availability across nested feature sets rather than maximizing algorithm-specific performance. A consistent model specification helped maintain comparability across M1–M4 and reduce the risk of overfitting. Future work could examine temporal and external-state validation and assess whether more extensive tuning changes the observed predictive patterns.

6 Conclusion

This study evaluated how crash-severity prediction changes as progressively richer information becomes available, using 2024 Ohio HSIS crash records. Results showed that roadway inventory and traffic-exposure variables alone provided limited discrimination of fatal and serious-injury (KA) crashes, while adding environmental, temporal, and crash-location context improved performance moderately. The largest gain occurred when crash-event and aggregated unit-record information were introduced. The findings emphasize that crash-severity models should be evaluated using predictors consistent with their intended application. Models for proactive roadway screening should rely on information available before or independently of the crash event, even if this results in lower predictive performance than event-informed models. Explicitly distinguishing these information sets provides a more realistic assessment of what roadway and contextual data can predict before the crash mechanism is known.

Declarations

- **Funding:** No funding was received for conducting this study.
- **Competing interests:** The authors have no relevant financial or non-financial interests to disclose.
- **Ethics approval:** Not applicable. This study used secondary administrative crash and roadway records and involved no recruitment, intervention, or direct interaction with human participants.
- **Consent to participate:** Not applicable.
- **Consent for publication:** Not applicable.
- **Data availability:** The Highway Safety Information System (HSIS) data used in this study are available from the Federal Highway Administration through the HSIS data-request process and are subject to FHWA data-use conditions. The processed analytical dataset is therefore not redistributed with this manuscript.
- **Materials availability:** Not applicable.
- **Code availability:** The analysis code is available from the corresponding author upon reasonable request.
- **Authors' contributions:** Mohammad Ridwan Tanvir: Conceptualization, methodology, data preparation, formal analysis, model development, visualization, and writing–original draft. Fahim Jaman: Conceptualization, literature review, and writing–review and editing. All authors reviewed and approved the final manuscript.

References

Amiri MA, Afshari S, Soltani A (2025) Machine learning approaches to traffic accident severity prediction: Addressing class imbalance. Machine Learning with Applications 22:100792. https://doi.org/10.1016/j.mlwa.2025.100792, URL https://linkinghub.elsevier.com/retrieve/pii/S2666827025001756

Awasthi D, Parti R, Mahajan K (2024) Effect of spatial relationship between curves on crash severity at horizontal curves in a mountainous terrain. Accident Analysis

& Prevention 206:107714. https://doi.org/10.1016/j.aap.2024.107714, URL https://www.sciencedirect.com/science/article/pii/S0001457524002598

Chen T, Guestrin C (2016) XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, San Francisco California USA, pp 785–794, https://doi.org/10.1145/2939672.2939785, URL https://dl.acm.org/doi/10.1145/2939672.2939785

Cheng C, Chen S, Ma Y, et al (2025) Crash severity prediction and interpretation for road determinants based on a hybrid method. Journal of Transportation Safety & Security 17(1):30–56. https://doi.org/10.1080/19439962.2024.2364661, URL https://www.tandfonline.com/doi/full/10.1080/19439962.2024.2364661

Choi J, Kim S, Heo TY, et al (2011) Safety effects of highway terrain types in vehicle crash model of major rural roads. KSCE Journal of Civil Engineering 15(2):405–412. https://doi.org/10.1007/s12205-011-1124-x, URL https://doi.org/10.1007/s12205-011-1124-x

FHWA (2026) Highway Safety Information System (HSIS) Overview | FHWA. URL https://highways.fhwa.dot.gov/turner-fairbank-highway-research-center/safety/hsis

Fiorentini N, Losa M (2020) Handling Imbalanced Data in Road Crash Severity Prediction by Machine Learning Algorithms. Infrastructures 5(7):61. https://doi.org/10.3390/infrastructures5070061, URL https://www.mdpi.com/2412-3811/5/7/61

Gao Z, Easa SM, Liu Y, et al (2026) Road traffic accident severity prediction based on large-scale data and multi-dimensional factors: an explainable ensemble learning approach. Accident Analysis & Prevention 233:108581. https://doi.org/10.1016/j.aap.2026.108581, URL https://linkinghub.elsevier.com/retrieve/pii/S0001457526001909

Haghighi N, Liu XC, Zhang G, et al (2018) Impact of roadway geometric features on crash severity on rural two-lane highways. Accident Analysis & Prevention 111:34–42. https://doi.org/10.1016/j.aap.2017.11.014, URL https://www.sciencedirect.com/science/article/pii/S0001457517304013

Iranitalab A, Khattak A (2017) Comparison of four statistical and machine learning methods for crash severity prediction. Accident Analysis & Prevention 108:27–36. https://doi.org/10.1016/j.aap.2017.08.008, URL https://linkinghub.elsevier.com/retrieve/pii/S0001457517302865

Ke G, Meng Q, Finley T, et al (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: Advances in Neural Information Processing Systems, vol 30. Curran Associates, Inc., URL https://papers.nips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

Lebaku PKR, Gao L, Sun J, et al (2026) Assessing the Influence of Pavement Performance on Road Safety Through Crash Frequency and Severity Analysis. International Journal of Pavement Research and Technology 19(8):2190–2211. https://doi.org/10.1007/s42947-025-00568-7, URL https://doi.org/10.1007/s42947-025-00568-7

Liang Y, Yuan H, Wang Z, et al (2024) Nonlinear effects of traffic statuses and road geometries on highway traffic accident severity: A machine learning approach. PLOS ONE 19(11):e0314133. https://doi.org/10.1371/journal.pone.0314133, URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0314133

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, NIPS'17, pp 4768–4777, URL https://dl.acm.org/doi/10.5555/3295222.3295230

Madushani JS, Sandamal RK, Meddage D, et al (2023) Evaluating expressway traffic crash severity by using logistic regression and explainable & supervised machine learning classifiers. Transportation Engineering 13:100190. https://doi.org/10.1016/j.treng.2023.100190, URL https://linkinghub.elsevier.com/retrieve/pii/S2666691X23000301

Malyshkina NV, Mannering FL (2010) Empirical assessment of the impact of highway design exceptions on the frequency and severity of vehicle accidents. Accident Analysis & Prevention 42(1):131–139. https://doi.org/10.1016/j.aap.2009.07.013, URL https://www.sciencedirect.com/science/article/pii/S000145750900178X

Mannering FL, Shankar V, Bhat CR (2016) Unobserved heterogeneity and the statistical analysis of highway accident data. Analytic Methods in Accident Research 11:1–16. https://doi.org/10.1016/j.amar.2016.04.001, URL https://linkinghub.elsevier.com/retrieve/pii/S2213665716300100

Miaou SP, Lum H (1993) Modeling vehicle accidents and highway geometric design relationships. Accident Analysis & Prevention 25(6):689–709. https://doi.org/10.1016/0001-4575(93)90034-T, URL https://www.sciencedirect.com/science/article/pii/000145759390034T

Sattar K, Chikh Oughali F, Assi K, et al (2023) 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, URL https://link.springer.com/10.1007/s00521-022-07769-2

Savolainen PT, Mannering FL, Lord D, et al (2011) The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. Accident Analysis & Prevention 43(5):1666–1676. https://doi.org/10.1016/j.aap.2011.03.025, URL https://linkinghub.elsevier.com/retrieve/pii/S0001457511000765

Somvanshi S, Baitullah A, Rafat SA, et al (2026) Predicting Glare-Related Traffic Outcomes with Transformer-Based Explainable Tabular Deep Learning. Data Science for Transportation 8(2):25. https://doi.org/10.1007/s42421-026-00163-7, URL https://link.springer.com/10.1007/s42421-026-00163-7

Wei Z, Zhang Y, Das S (2023) Applying Explainable Machine Learning Techniques in Daily Crash Occurrence and Severity Modeling for Rural Interstates. Transportation Research Record: Journal of the Transportation Research Board 2677(5):611–628. https://doi.org/10.1177/03611981221134629, URL https://journals.sagepub.com/doi/10.1177/03611981221134629