Predicting Injury Severity in Vehicle-Non-Motorist Crashes: A Comparative Machine Learning Framework with Interpretability Analysis

Parvez Anowar[1*]

[1]*Department of Civil, Environmental and Construction Engineering, University of Central Florida, Orlando, FL, USA*

Abstract

Pedestrians and bicyclists account for a disproportionate share of traffic fatalities, yet predicting crash severity remains challenging due to class imbalance and inconsistent benchmarking. This study analyzes 12,563 vehicle-non-motorist crashes from Florida's Signal4 database, comparing statistical models (Ordered Probit, Multinomial Logit), machine learning classifiers (Random Forest, XGBoost, LightGBM, SVM), and deep learning models (MLP, CNN) under identical conditions. Tree-based ensembles achieve the best performance (macro-F1: 0.48, ROC-AUC: 0.65 multiclass; 0.68 and 0.77 binary). Class-weighted training outperforms synthetic resampling, and tree ensembles match or exceed deep learning on tabular data. SHAP analysis identifies non-motorist age, violation history, lighting, and roadway type as the strongest severity predictors, with injury probability rising sharply beyond age 60. Calibration shows Gradient Boosting and SVM yield the most reliable probability estimates, while top-performing tree ensembles may need post-hoc calibration. The findings support prioritized infrastructure for elderly pedestrians and improved lighting on high-exposure corridors.

Keywords: Crash severity prediction; vulnerable road users; machine learning; SHAP interpretability; pedestrian safety; class imbalance; gradient boosting.

1. Introduction

Road traffic crashes remain a leading cause of death and disability globally, claiming approximately 1.19 million lives annually (WHO, 2019). Vulnerable road users (VRUs), including pedestrians and bicyclists, bear a disproportionate share of this burden, accounting for over half of all traffic fatalities. In the United States, Florida ranks among the highest pedestrian and bicyclist fatality rates, making the state a critical setting for studying vehicle-non-motorist crash severity.

A large body of research has examined factors influencing VRU injury severity. Studies report strong associations with lighting, roadway context, vehicle type, age of involved parties, and risky driving behaviors (Lee and Abdel-Aty, 2005; Sun et al., 2023; Zhao et al., 2024). At the area level, land use patterns, demographic composition, and built-environment features have also been linked to crash frequency and severity (Du et al., 2024; Yue et al., 2020; Ibrahim et al., 2026; Uddin et al., 2025). Research on pedestrian crashes at urban intersections further highlights the role of spatiotemporal correlations, crossing behavior, and vehicle type in determining injury outcomes (Zeng et al., 2023;

* Corresponding author: Parvez Anowar (pa545735@ucf.edu)

Hossain et al., 2024; Anowar et al., 2025). These findings establish that severity in vehicle-non-motorist crashes reflects the interplay of roadway design, environmental conditions, and human behavior.

Methodologically, two streams dominate the literature. Traditional discrete-outcome models, including logistic regression, multinomial logit (MNL), ordered probit (OP), and random-parameters extensions, have been widely applied to crash severity analysis (Abdel-Aty and Radwan, 2000; Iranitalab and Khattak, 2017), with random-parameters specifications capturing unobserved heterogeneity (Sun et al., 2023; Scarano et al., 2025; Hossain et al., 2024). These econometric approaches depend on distributional assumptions and can struggle with non-linearities and high-dimensional feature spaces (Li and Kockelman, 2022). Tree-based ensembles such as Random Forest, XGBoost, and LightGBM capture complex predictor interactions and have shown strong performance in crash-severity modeling (Zhang et al., 2018; Wen et al., 2021; Zhao et al., 2024; Santos et al., 2024; Kashifi, 2023; Hoque et al., 2025), and benchmark evidence indicates that tree-based models outperform deep networks on medium-sized tabular data (Grinsztajn et al., 2022; Ali et al., 2024). Comparative analyses also show ML classifiers exceeding OP and MNL on prediction tasks, while the statistical models retain value for directional interpretation (Iranitalab and Khattak, 2017; Ijaz et al., 2021; Komol et al., 2021).

Recent studies have increasingly adopted hybrid approaches combining the predictive power of ML with the interpretive strengths of statistical models. Sun et al. (2023) paired a Random Forest-SHAP framework with a random parameters logit model for VRU crash severity in Shenyang, China, showing that interaction effects between risk factors can reverse the direction of individual effects. Scarano et al. (2023, 2025) used complementary Random Forest and Random Parameters Logit models for cyclist crash severity in Great Britain. Sadeghi et al. (2024) combined XGBoost-SHAP with random parameters logit for mobility scooter crashes in the UK, also demonstrating temporal instability across COVID-era periods. Hossain et al. (2024) used XGBoost for variable selection before fitting random parameters logit models for pedestrian crashes on Louisiana interstates. These hybrid approaches underscore the complementary value of prediction- and explanation-oriented methods (Mannering et al., 2020; Wang et al., 2026; Tahmid et al., 2025).

SHAP-based interpretation has become standard in crash-severity research, supporting global feature rankings and local prediction explanations (Zahid et al., 2024; Kashifi, 2023; Sun et al., 2023). Despite these advances, systematic reviews identify several persistent gaps (Ali et al., 2024; Wen et al., 2021). First, class imbalance is pervasive in severity datasets, yet over 70% of published studies do not explicitly address it, and no consensus exists on the best handling strategy (Santos et al., 2024; Zahid et al., 2024). Second, evaluation practice is inconsistent: at least 23 different training-testing splits are documented, accuracy is privileged over per-class metrics, and probability calibration is almost never assessed despite its importance for operational deployment (Komol et al., 2021; Sadeghi et al., 2024). Third, few studies benchmark statistical baselines, tree-based ensembles, and deep learning under identical preprocessing and metric conditions, making cross-family comparisons unreliable.

This study addresses these gaps using Florida's Signal4 crash database. The contributions are fourfold:

1) A unified pipeline trains, tunes, and evaluates statistical models (OP, MNL), machine learning classifiers (Random Forest, XGBoost, LightGBM, Logistic Regression, SVM, kNN, ANN-MLP), and deep learning models (MLP, CNN) under identical

preprocessing, splitting, and metric conditions, enabling cross-family comparison that is rare in the published literature.

2) A systematic comparison of class imbalance strategies, including class weighting, SMOTE-NC at varying intensity levels, random under-sampling, and combined schemes, is conducted to determine their impact on minority-class detection, directly addressing a gap identified by recent reviews (Ali et al., 2024; Wen et al., 2021).
3) The predictive value of alternative target formulations is assessed by comparing multiclass (three-level) severity prediction against binary (Severe vs. Non-Severe) reformulation, with emphasis on the binary task most relevant to high-risk operational screening.
4) Model interpretation through SHAP, partial-dependence, and permutation importance is used to identify actionable predictors of severe injury, and the findings are contextualized against comparable studies across multiple jurisdictions.

2. Data description

The crash data used in this study originate from the Signal4 Analytics system for the state of Florida, extracted in 2024 and filtered to isolate crashes involving exactly one motor vehicle and a single vulnerable road user (pedestrian or bicyclist). Each record represents a single crash event containing linked information from police reports on the roadway environment, people involved, and crash circumstances. After applying filters to remove multi-vehicle events, non-motor-vehicle crashes, and records with irreconcilable severity coding, the final working dataset comprises 12,563 vehicle-non-motorist crashes. The observations are randomly partitioned into training (70%, n = 8,794), validation (15%, n = 1,884), and test (15%, n = 1,885) sets, with stratification to preserve the overall severity distribution across splits.

Table 1: Overview of predictor variables used in the crash-severity models.

Category	*Variables*
Temporal	YEAR, DAY, SEASON, WEEKEND, SIN_HOUR, COS_HOUR
Crash context	LOCATION, INTERCHANGE, INTERSECTION_RELATED, IMPACT_ZONE, HIT_&_RUN, LANE_DEPARTURE_RELATED, SPEEDING_RELATED, AGGRESSIVE_DRIVING
Roadway	ROAD_TYPE, INTERSECTION_TYPE, SHOULDER_TYPE, ROAD_SURFACE_CONDITION, RURAL_OR_URBAN
Environment	LIGHT_CONDITION, WEATHER_CONDITION, VISION_OBSTRUCTED
Traffic control	TRAFFIC_CONTROL
Driver	DRIVER_SEX, DRIVER_AGE, DRIVER_CONDITION, DRIVING_LICENSE_TYPE, DRIVER_RE_EXAM_RECOMMENDED, DRIVER_DISTRACTION_TYPE, DRIVER_VIOLATION, COUNT_DRIVER_VIOLATION, ALCOHOL_RELATED, DRUG_RELATED, RESTRAINT_SYSTEM, UNRESTRAINED, AIR_BAG_DEPLOYED, DRIVER_EJECTED, DRIVER_AGE_MISSING
Vehicle	VEHICLE_TYPE, VEHICLE_MOVEMENT_TYPE, TRAILER_COUNT, MOTORCYCLE_RELATED, CMV_INVOLVED, SCHOOL_BUS_RELATED, PASSENGER_PRESENT
Non-motorist	NM_TYPE, NM_SEX, NM_AGE, NM_LOCATION, NM_VIOLATION, COUNT_NM_VIOLATION, NM_SAFETY_EQUIPMENT

The target variable is crash injury severity for the non-motorist. Original Signal4 categories based on KABCO-type police coding are mapped to a three-level ordered outcome: Severe (fatal and incapacitating injury), Moderate (non-incapacitating injury), and Minor/None (possible injury and no injury). In the final dataset, approximately 47.7% of crashes fall in Minor/None, 35.8% in Moderate, and 16.5% in Severe. The Severe class forms a substantial minority, motivating careful treatment of class imbalance and evaluation beyond overall accuracy.

The predictor set comprises 52 variables covering seven thematic categories, summarized in Table 1. Most predictors are categorical with multiple levels, while a smaller subset (e.g., ages, violation counts, derived temporal measures) are numeric. In the raw Signal4 data, missing or unknown values appear in many fields; during preprocessing, these are recoded as explicit "Unknown" categories for categorical variables or imputed using simple, reproducible rules for numeric variables, ensuring that all models can learn from the full sample.

3. Methodology

Figure 1 presents the end-to-end workflow adopted in this study, encompassing data filtering, severity recoding, feature engineering, model development, evaluation, and interpretability analysis.

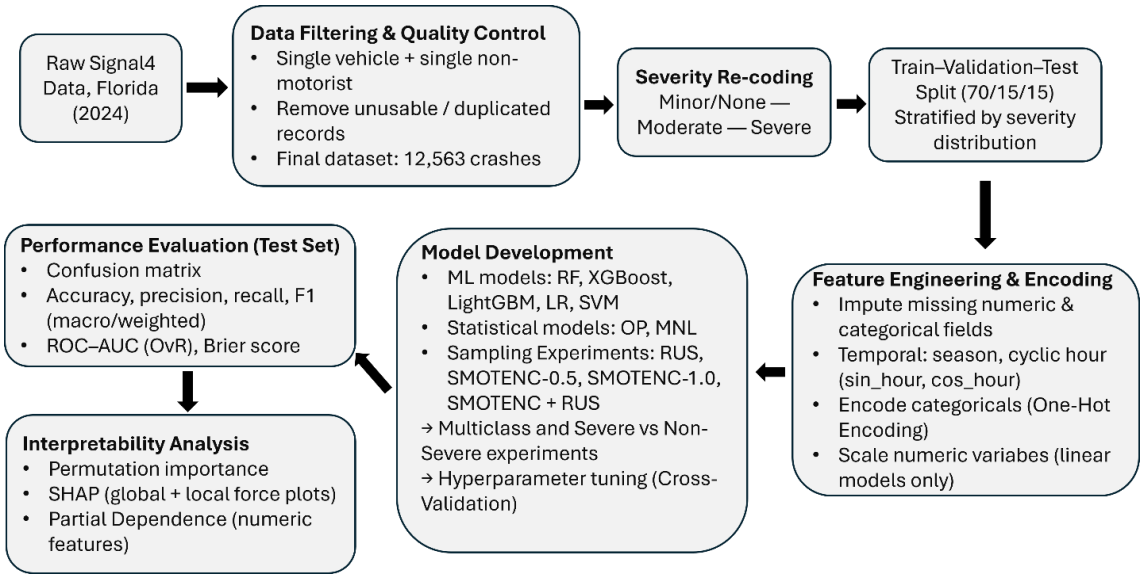

Figure 1: Overview of the crash-severity modeling workflow.

3.1 Data preprocessing and feature engineering.

The study follows a supervised learning framework in which crash injury severity for the non-motorist is predicted from crash, roadway, environmental, vehicle, and person-level variables. The full dataset is randomly split into training, validation, and test sets (70-15-15) with stratification by severity level to ensure that class proportions are preserved across splits. All model development, including hyperparameter tuning and interpretability analyses, is conducted on the training and validation data, while a final unbiased assessment is obtained on the held-out test set.

Temporal variables (year, month, date, time of day, day of week) are converted into derived features including hour of day, weekend and season indicators, and cyclic encodings, specifically sin(hour) and cos(hour), that preserve the circular 24-hour structure while permitting flexible non-linear effects. The remaining categorical predictors (e.g., roadway type, intersection type, lighting condition, vehicle type, non-motorist location) are handled using one-hot encoding, with explicit "Unknown" levels created where police reports contain missing or ambiguous entries. Numeric variables (e.g., ages, violation counts) are left on their natural scale for tree-based models and standardized to zero mean and unit variance for linear and kernel methods. Simple and transparent imputation rules are applied: medians for numeric variables and the most frequent level for categorical variables.

3.2 Statistical baseline models.

Two classical discrete-outcome models serve as interpretable baselines: an Ordered Probit (OP) model and a Multinomial Logit (MNL) model. The OP model treats injury severity as an ordered outcome arising from an underlying continuous propensity $Y^* = X\beta + \varepsilon$, where X is the predictor matrix, β is the coefficient vector, and ε follows a standard normal distribution. Observed severity levels correspond to Y^* crossing ordered threshold parameters, with category probabilities given by differences of the cumulative normal distribution evaluated at those thresholds (Abdel-Aty and Radwan, 2000). The MNL model, by contrast, treats severity levels as nominal, estimating a separate utility expression for each category relative to a reference level. This structure permits different predictors to influence each injury level in distinct ways, providing greater flexibility when the ordinal assumption does not hold strictly. Both models are estimated via maximum likelihood.

3.3 Machine-learning classifiers.

The primary predictive analysis employs eight classifiers: Random Forest (RF), XGBoost, LightGBM, Gradient Boosting, multinomial Logistic Regression (LR), linear-kernel Support Vector Machine (SVM), k-nearest neighbors (kNN), and a shallow multilayer perceptron (ANN-MLP). Tree-based ensembles are emphasized given their strong empirical record in crash-severity prediction (Zhang et al., 2018; Zhao et al., 2024) and benchmark evidence that they outperform deep networks on medium-sized tabular data with many categorical features (Grinsztajn et al., 2022). LR, linear SVM, and kNN serve as non-ensemble baselines; the shallow ANN-MLP provides a small-network reference distinct from the deeper architectures in Section 3.4.

Hyperparameters for each classifier, including tree depth, number of estimators, learning rate, regularization strength, and SVM penalty, are tuned using stratified 5-fold cross-validation on the training data, with macro-F1 score as the primary selection criterion. Grid or compact random search spaces are used to balance computational cost with sufficient exploration. Table 2 summarizes the final tuned configurations.

3.4 Deep-learning models.

Two deep learning architectures are evaluated as a sensitivity analysis: a deeper MLP with multiple hidden layers (ReLU activations, batch normalization, dropout), and a CNN

that maps the one-dimensional feature vector into a 2D grid (DeepInsight-style) so convolutional filters can exploit local feature structure. These models are trained under the sampling strategy identified as most effective in earlier experiments and evaluated on both multiclass and binary configurations, testing whether deeper architectures yield gains over tree-based ensembles for tabular crash data of this scale (Grinsztajn et al., 2022).

Table 2: Tuned hyperparameters for machine-learning models.

Model	*Key Tuned Hyperparameters*
LightGBM	num_leaves=127; n_estimators=400; min_child_samples=20; max_depth=6; learning_rate=0.03; colsample_bytree=0.8
XGBoost	n_estimators=300; min_child_weight=3; max_depth=4; learning_rate=0.1; colsample_bytree=0.8
Random Forest	n_estimators=600; min_samples_split=5; min_samples_leaf=5; max_features="sqrt"; max_depth=20
Logistic Regression	penalty="l2"; C=0.3
Gradient Boosting	n_estimators=100; max_depth=2; learning_rate=0.1
SVM	kernel="linear"; C=1.0; class_weight="balanced"
kNN	weights="distance"; p=1; n_neighbors=11
ANN-MLP	learning_rate_init=0.001; hidden_layer_sizes=(50,); alpha=1×10^{-5}

3.5 Class imbalance treatment.

In the primary pipeline, all models are trained on the original class distribution using class-weighting schemes that up-weight minority classes in the loss function (Zhang et al., 2018; Zhao et al., 2024). Class-weighted training has been shown to outperform synthetic resampling in several crash-severity studies (Santos et al., 2024), though over 70% of published studies do not explicitly address imbalance (Ali et al., 2024).

A separate set of experiments evaluates the impact of four resampling strategies on model performance: Random Under-Sampling (RUS), SMOTE-NC with a 0.5 ratio lift for the Severe class (SMOTENC-0.5), full parity oversampling for all classes (SMOTENC-1.0), and a combined SMOTE-NC plus under-sampling approach (SMOTENC + RUS). Each strategy modifies the class distribution at different intensities, ranging from mild augmentation of the minority class to fully balanced datasets. Resampled models are trained and assessed using the same metrics as the main pipeline.

3.6 Binary target reformulation.

To examine whether simpler targets improve performance, binary classifiers are estimated for three splits: Severe vs. Non-Severe, Moderate vs. Non-Moderate, and Minor/None vs. Non-Minor/None. The same classifier families and pipelines are reused, with class weights (and scale_pos_weight for XGBoost) tuned to the minority share. The Severe vs. Non-Severe task is the primary focus given its relevance for high-risk screening.

3.7 Evaluation metrics and interpretation.

Model performance is assessed on the held-out test set with confusion matrices and per-class precision, recall, and F1. Macro-F1 is the primary multiclass ranking metric because it weighs each severity level equally, better reflecting Moderate and Severe performance (Komol et al., 2021; Zhao et al., 2024). Supplementary metrics include weighted-F1, balanced accuracy, ROC–AUC (one-vs-rest), and the Brier score for calibration assessment, a metric rarely reported in crash-severity studies but critical for operational use (Ali et al., 2024).

Interpretation combines permutation importance (drop in macro-F1 when individual features are shuffled), partial-dependence plots for key numeric variables, SHAP values from the best-performing model for global rankings and local explanations (Lundberg and Lee, 2017), and OP/MNL coefficient signs for directional interpretation.

4. Results and discussion

4.1 *Multiclass model performance.*

Table 3 reports the test-set performance of all multiclass severity-prediction models. Overall accuracy ranges from 45% to 53%, reflecting the difficulty of three-class severity prediction given the overlap between Moderate and adjacent levels. Tree-based ensembles, Random Forest, XGBoost, and LightGBM, achieve the strongest results, with macro-F1 scores between 0.47 and 0.49 and ROC-AUC values near 0.65. Among these, LightGBM and Random Forest yield the highest macro-F1, indicating balanced performance across severity levels. kNN and the shallow MLP underperform, suggesting that high-dimensional categorical features favor models capable of capturing non-linear interactions.

Table 3: Performance of multiclass crash-severity models (test set).

Model	*Accuracy*	*Prec. (macro)*	*Recall (macro)*	*Macro-F1*	*ROC-AUC (OvR)*	*Brier*
Logistic Regression	0.473	0.454	0.505	0.461	0.658	0.202
Gradient Boosting	0.528	0.530	0.451	0.447	0.659	0.187
SVM	0.516	0.478	0.493	0.476	0.657	0.192
Random Forest	0.509	0.479	0.506	0.487	0.653	0.234
XGBoost	0.507	0.504	0.455	0.467	0.654	0.193
LightGBM	0.506	0.496	0.477	0.485	0.630	0.268
Multinomial Logit	0.516	0.496	0.442	0.433	0.644	0.189
Ordered Probit	0.504	0.547	0.406	0.408	0.627	0.193
kNN	0.489	0.472	0.427	0.436	0.614	0.208
ANN-MLP	0.455	0.430	0.432	0.431	0.593	0.328

The confusion matrices for representative models (Figure 2) reveal a consistent structural pattern across algorithms. Minor/None is the easiest class to identify correctly, while Moderate shows the highest confusion and is frequently misclassified as either Minor/None or Severe. Severe achieves better separation than Moderate despite its lower prevalence, indicating that the circumstances surrounding severe injuries are more

distinguishable within the available feature space. These patterns are further reflected in the per-class F1, recall, and precision plots (Figure 3).

The ROC curves (Figure 4a) confirm that all multiclass models achieve similar overall discrimination (AUC ≈ 0.65), indicating moderate ability to rank-order severity outcomes despite the overlap among classes. The difficulty of separating Moderate cases motivates the binary reformulation examined next.

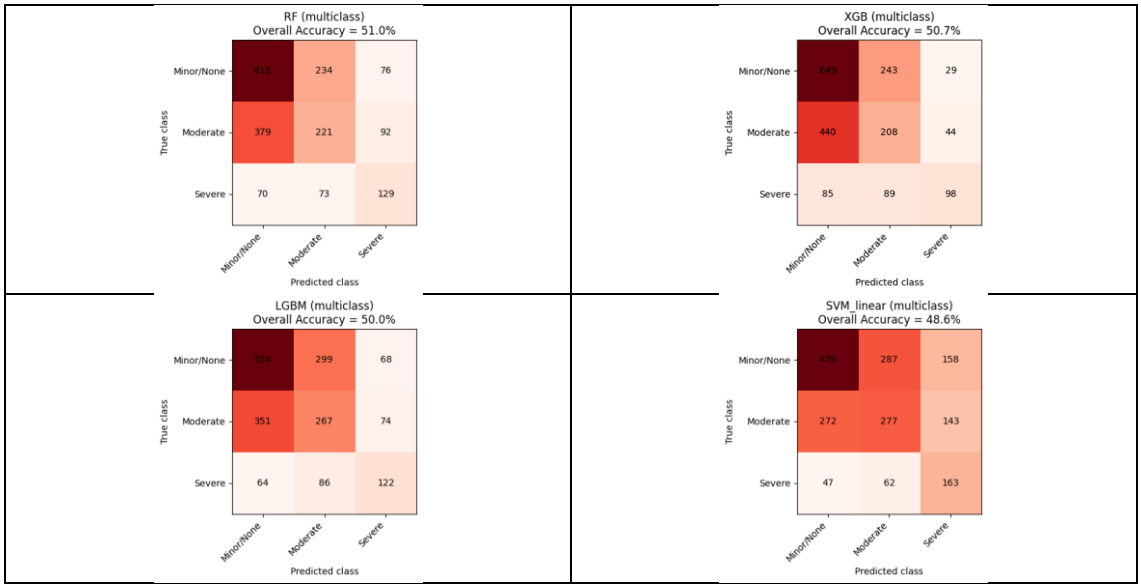

Figure 2: Confusion matrices for selected multiclass models: (a) Logistic Regression, (b) Gradient Boosting, (c) Random Forest, (d) XGBoost.

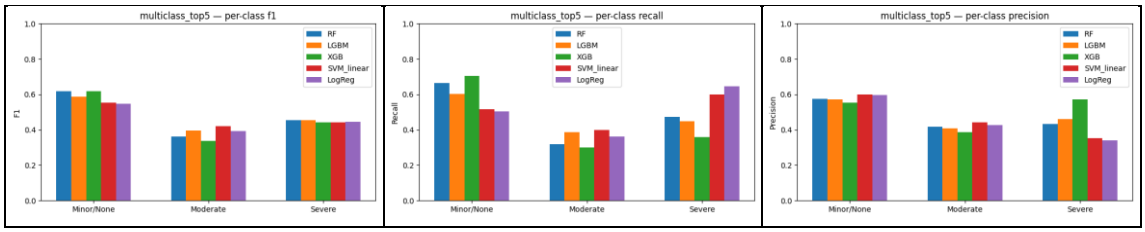

Figure 3: Per-class metrics for multiclass models: (a) Precision, (b) F1-score, (c) Recall.

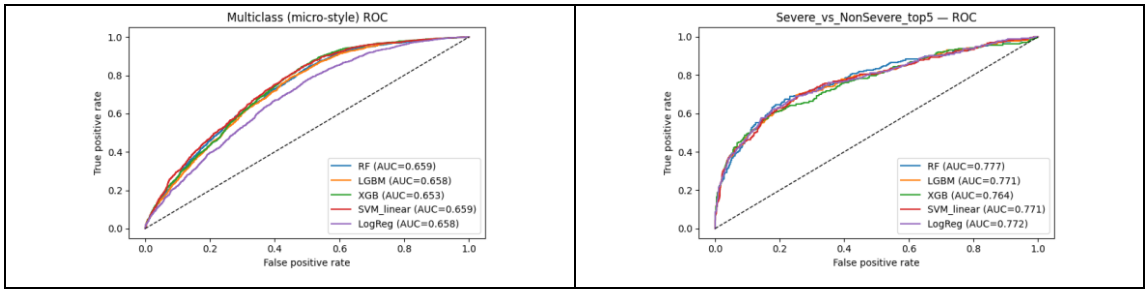

Figure 4: ROC curves for (a) multiclass models (OvR) and (b) Severe vs. Non-Severe binary models.

4.2 Binary classification performance.

Binary reformulation yields substantially clearer discrimination than the multiclass task. Table 4 summarizes results for three binary splits. The Severe vs. Non-Severe formulation, the primary focus of this study, produces the strongest performance across all model families. LightGBM and Random Forest both attain accuracies near 0.85 and macro-F1 scores around 0.68, while XGBoost exhibits similar overall metrics but lower Severe-class recall. SVM achieves the highest Severe recall (0.65) at the cost of more false positives. These trade-offs are visible in the confusion matrices (Figure 5).

Table 4: Performance of binary classification models (test set).

Model	*Acc.*	*Prec.*	*Recall*	*Macro-F1*	*W-F1*	*AUC*	*Brier*
RF Severe vs. Non-Severe	0.854	0.519	0.401	0.677	0.847	0.777	0.130
RF Moderate vs. Non-Moderate	0.581	0.426	0.322	0.522	0.567	0.567	0.238
RF Minor vs. Non-Minor	0.603	0.586	0.651	0.602	0.602	0.645	0.232
LGBM Severe vs. Non-Severe	0.855	0.488	0.449	0.687	0.850	0.771	0.113
LGBM Moderate vs. Non-Moderate	0.545	0.384	0.400	0.513	0.546	0.542	0.252
LGBM Minor vs. Non-Minor	0.594	0.574	0.609	0.594	0.594	0.631	0.242
XGB Severe vs. Non-Severe	0.839	0.667	0.287	0.682	0.841	0.763	0.125
XGB Moderate vs. Non-Moderate	0.548	0.407	0.191	0.518	0.550	0.544	0.256
XGB Minor vs. Non-Minor	0.583	0.567	0.598	0.583	0.583	0.624	0.247
SVM Severe vs. Non-Severe	0.751	0.317	0.654	0.638	0.782	0.771	0.101
SVM Moderate vs. Non-Moderate	0.515	0.392	0.597	0.512	0.523	0.558	0.230
SVM Minor vs. Non-Minor	0.598	0.574	0.645	0.598	0.597	0.640	0.234

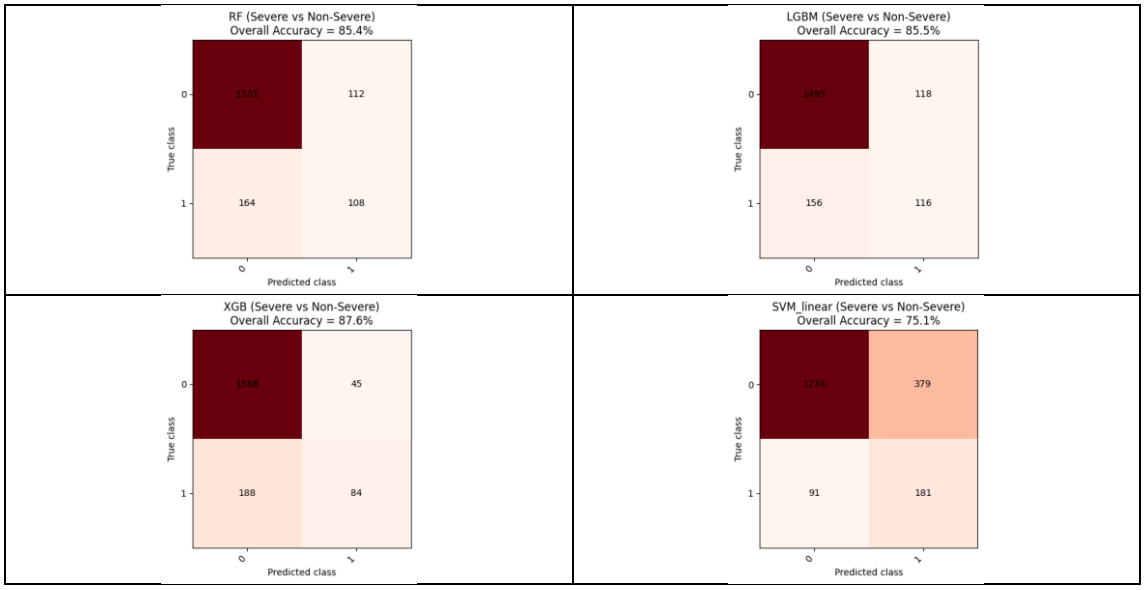

Figure 5: Confusion matrices for Severe vs. Non-Severe binary models: (a) Random Forest, (b) LightGBM, (c) XGBoost, (d) SVM.

The two additional binary tasks, Moderate vs. Non-Moderate and Minor/None vs. Non-Minor/None, exhibit weaker separation. Moderate injuries remain difficult to isolate, with recall frequently below 0.40, while the Minor/None split performs slightly better with

more consistent recall above 0.60. ROC curves (Figure 4b) show AUC values between 0.76 and 0.78 for all Severe vs. Non-Severe models, indicating stable ranking ability across classifiers. The per-class metric plots for binary models (Figure 6) confirm that tree-based models provide the most balanced precision–recall trade-off for the Severe class.

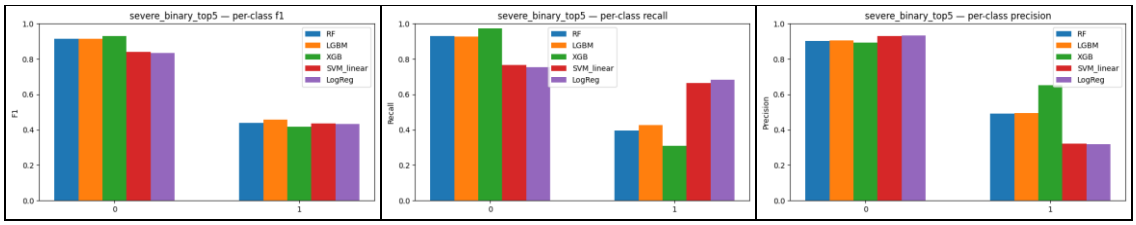

Figure 6: Per-class metrics for Severe vs. Non-Severe binary models: (a) Precision, (b) F1-score, (c) Recall.

4.3 Impact of resampling strategies.

Table 5 reports the multiclass performance under alternative resampling strategies for all four main model families. A consistent pattern emerges: the original class-weighted models (NoResampling) achieve the highest or near-highest macro-F1 for every algorithm, typically in the range of 0.47-0.48. Mild SMOTENC oversampling produces marginal changes, while full balancing and combined SMOTE-plus-under-sampling reduce macro-F1 and degrade calibration. Random under-sampling underperforms the class-weighted baseline. This pattern aligns with observations by Santos et al. (2024), who reported that under-sampling dramatically improved sensitivity for the fatal class in motorcycle crashes but at the cost of reduced overall precision, and with the broader finding from Ali et al. (2024) that synthetic resampling often overgeneralizes minority-class boundaries. The results indicate that the natural class proportions, when combined with model-specific class weighting, provide a more effective learning signal than synthetic resampling for this dataset.

Table 5: Multiclass resampling comparison (selected configurations).

Model	*Sampling*	*Accuracy*	*Macro-F1*	*W-F1*	*AUC (OvR)*	*Brier*
LGBM	NoResampling	0.492	0.477	0.491	0.646	0.200
LGBM	RUS	0.438	0.431	0.441	0.630	0.225
LGBM	SMOTENC_0.5	0.500	0.470	0.493	0.643	0.200
LGBM	SMOTE_RUS	0.488	0.453	0.479	0.635	0.205
RF	NoResampling	0.513	0.481	0.501	0.662	0.194
RF	RUS	0.470	0.453	0.470	0.647	0.204
RF	SMOTENC_0.5	0.486	0.452	0.474	0.635	0.197
RF	SMOTE_RUS	0.477	0.441	0.463	0.628	0.198
XGB	NoResampling	0.520	0.479	0.505	0.654	0.193
XGB	RUS	0.438	0.431	0.441	0.631	0.223
XGB	SMOTENC_0.5	0.499	0.470	0.490	0.645	0.198
XGB	SMOTE_RUS	0.491	0.458	0.482	0.638	0.203
SVM	NoResampling	0.486	0.472	0.489	0.661	0.187
SVM	RUS	0.473	0.462	0.477	0.642	0.203
SVM	SMOTENC_0.5	0.476	0.455	0.475	0.624	0.201

SVM	SMOTE_RUS	0.476	0.453	0.474	0.620	0.201

Note: All models in this table are trained on the training partition only (70%) to enable controlled comparison across resampling conditions. Results differ slightly from Table 3, where models are trained on the combined training and validation sets (85%) for final evaluation.

4.4 Deep-learning model performance.

Table 6 summarizes the performance of the MLP and CNN under both task formulations. In the multiclass setting, the MLP achieves a macro-F1 of 0.457, marginally below the top tree-based models, while the CNN attains only 0.285, reflecting difficulty in learning from the feature-image representation at this dataset size. Both architectures show weaker probability calibration than RF and LightGBM. In the binary task, the MLP performs more competitively (macro-F1 = 0.615), approaching the best ensemble results.

Table 6: Performance of deep-learning models.

Model	*Task*	*Accuracy*	*Macro-F1*	*W-F1*	*AUC (OvR)*	*Brier*
MLP	Multiclass	0.525	0.457	0.490	0.663	0.185
MLP	Severe vs. Non-Severe	0.875	0.615	0.840	0.779	0.099
CNN	Multiclass	0.497	0.285	0.368	0.616	0.197
CNN	Severe vs. Non-Severe	0.859	0.487	0.798	0.745	0.109

Figure 7 visualizes macro-F1 across model families. Tree-based ensembles match or exceed both deep-learning architectures in the multiclass setting, and the MLP closes the gap in the binary task without exceeding the best ensembles. The result agrees with Grinsztajn et al. (2022) and the limited deep-learning gains reported in crash-severity work (Ali et al., 2024).

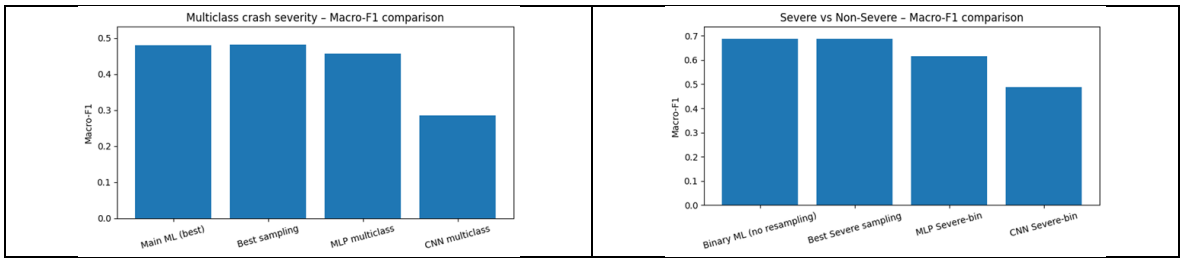

Figure 7: Macro-F1 comparison across (a) multiclass and (b) Severe vs. Non-Severe models.

4.5 Interpretation of severity predictors.

The interpretability analyses converge on a stable set of severity predictors. The SHAP summary plots (Figure 8) show that non-motorist age exerts a strong, non-linear effect: older non-motorists sharply increase predicted severe-crash probability, in line with the elevated physical vulnerability of elderly pedestrians and bicyclists. The pattern aligns with results from multiple jurisdictions: pedestrians aged 65+ were 3.22 times more likely to experience KSI in Hong Kong (Zeng et al., 2023), cyclists aged 75+ in Great Britain showed a 73.79% rise in fatal crash probability (Scarano et al., 2023), and elderly VRUs were the most vulnerable group in Queensland (Komol et al., 2021). Driver age has a milder declining effect. Other contributors include violation counts, darker lighting, high-

speed roadway types, and straight-through vehicle motions, with bicyclist involvement frequent among high-impact predictors.

The partial-dependence plots (Figure 9) clarify the functional form of key relationships. The predicted probability of severe injury rises steadily with non-motorist age, particularly beyond 60 years, and remains elevated for drivers with repeated violations. Driver age shows a more nuanced pattern, with moderate risk for younger drivers that gradually declines. Permutation-importance results agree, ranking non-motorist violations, road type, time-of-day encodings, non-motorist type, vehicle movement, and lighting as the strongest contributors; season, weekday, and several driver demographic attributes play minimal roles.

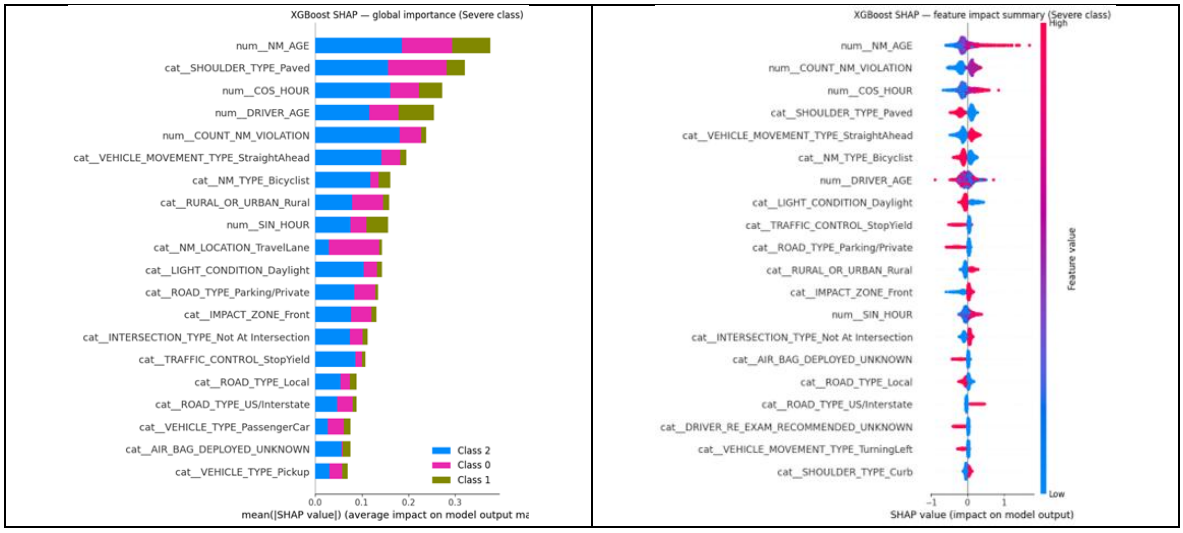

Figure 8: SHAP interpretation for Severe predictions: (a) global feature importance and (b) feature-wise impact summary.

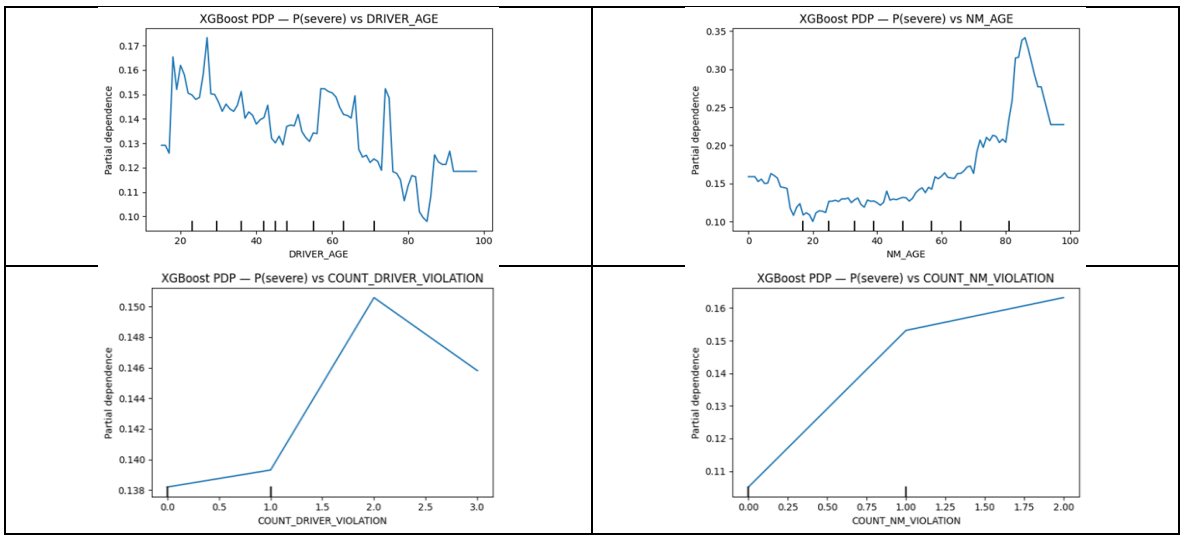

Figure 9: Partial-dependence plots for P(Severe): (a) Driver age, (b) Non-motorist age, (c) Driver violation count, (d) Non-motorist violation count.

The role of lighting as a severity predictor is well-established. Darkness raised fatal cyclist crash probability by 14.47% in Great Britain (Scarano et al., 2023), approximately

80% of pedestrian fatalities in Texas occurred between 8 PM and 7 AM (Zhao et al., 2024), and early morning hours (1–6 AM) showed elevated two-wheeler crash severity in France (Kashifi, 2023). The present SHAP results align with these patterns, indicating that lighting deficiency is a stable, cross-jurisdictional predictor of severe VRU injury.

Severe outcomes in vehicle-non-motorist crashes arise from a convergence of human vulnerability (age), risky behaviors (violations, impairment), adverse exposure (poor lighting, high-speed corridors), and roadway design. No single factor dominates: combinations of risk conditions most reliably predict severe injury, aligning with the interaction-effect findings of Sun et al. (2023).

4.6 Comparison with existing studies.

To contextualize the performance achieved in this study, Table 7 compares the present results against representative crash-severity prediction studies that share similar methodological characteristics (ML-based classification, VRU focus, comparable metrics reported). This comparison is necessarily approximate, as differences in datasets, severity definitions, target formulations, sample sizes, and class distributions limit strict cross-study equivalence.

Table 7: Comparison of the present study with existing crash-severity prediction studies.

Study	*Location*	*Target*	*N*	*Best Model*	*Macro-F1*	*AUC*	*Imbalance*
Iranitalab & Khattak (2017)	Nebraska, USA	4-level	68,448	NNC+KC	-	-	K-means
Zhang et al. (2018)	Florida, USA	5-level	5,538	RF	-	-	None
Komol et al. (2021)	Queensland, AUS	Binary	21,158	RF	0.75	0.70	Binary agg.
Zahid et al. (2024)	Rawalpindi, PAK	4-level	5,144	CatBoost	0.66	0.86	SMOTE
Santos et al. (2024)	Portugal	3-level	37,728	RF (bal.)	-	0.81	Under-samp.
Zhao et al. (2024)	Texas, USA	5-level	78,497	RF/LGBM	-	0.90*	None
Present study (multi)	Florida, USA	3-level	12,563	RF/LGBM	0.48	0.65	Class wt.
Present study (binary)	Florida, USA	Binary	12,563	LightGBM	0.68	0.77	Class wt.

Note: * indicates AUC for the fatal class only. Dashes indicate metrics not reported or not directly comparable.

The multiclass macro-F1 of 0.48 sits within the range reported for studies with three or more severity classes; the Moderate category introduces well-documented classification noise (Iranitalab and Khattak, 2017; Zhang et al., 2018). Binary reformulation (macro-F1 = 0.68, AUC = 0.77) closes most of the gap to studies that use binary or heavily aggregated severity targets. The systematic class-imbalance comparison adds value relative to studies that apply a single technique without controlled comparison.

The performance levels here align with the pattern noted by Ali et al. (2024): moderate accuracy is the norm for crash-severity prediction, and high-accuracy claims often reflect evaluation on majority classes rather than genuine discrimination of rare severe outcomes.

The emphasis on macro-F1, per-class metrics, and calibration provides a more honest assessment.

4.7 Model calibration assessment.

Beyond discriminative performance, the reliability of predicted probabilities matters for operational use such as risk scoring and threshold-based screening. The Brier score (Tables 3, 4, 6) summarizes calibration, with lower values indicating closer agreement between predicted and observed outcomes. In the multiclass setting (Table 3), Gradient Boosting (0.187) and SVM (0.192) achieve the best calibration, followed by XGBoost (0.193). Random Forest (0.234), LightGBM (0.268), and the ANN-MLP (0.328) calibrate worse despite some achieving strong macro-F1 scores, reflecting the known tendency of bagging-based and gradient-boosted ensembles to produce overconfident or underconfident probabilities without explicit recalibration. The statistical baselines (MNL = 0.189, OP = 0.193) calibrate competitively, likely owing to their parametric smoothing.

For the binary Severe vs. Non-Severe task (Table 4), all models achieve lower Brier scores, with SVM (0.101), LightGBM (0.113), and XGBoost (0.125) leading. In the multiclass setting, the deeper MLP reaches a Brier of 0.185 (Table 6), and in the binary task it reaches 0.099, suggesting that the simpler binary target supports more reliable probability estimation. All values fall below approximately 0.165 expected from a naive predictor using marginal class frequencies, indicating meaningful probability discrimination.

These calibration results have practical implications. For agencies seeking to use predicted probabilities as risk scores, for example, to flag locations where predicted P(Severe) exceeds a decision threshold, models with better calibration (Gradient Boosting, SVM, XGBoost) are preferable to those with higher macro-F1 but weaker calibration (LightGBM, Random Forest), unless post-hoc calibration methods such as Platt scaling or isotonic regression are applied. This dimension of model assessment is rarely reported in crash-severity studies (Ali et al., 2024) but is essential for translating predictions into operational safety tools.

5. Conclusions, practical implications, and limitations

This study developed a structured framework for predicting injury severity in vehicle-non-motorist crashes using Florida's Signal4 database (12,563 crashes). Gradient-boosted tree models and Random Forest achieved the strongest multiclass performance (macro-F1 near 0.48, ROC-AUC of 0.65), and binary reformulation to Severe vs. Non-Severe improved discrimination substantially (macro-F1 of 0.68, ROC-AUC of 0.77). Deep learning models trained on tabular features did not exceed tree-based results, in line with evidence favoring tree-based methods for structured data of this scale (Grinsztajn et al., 2022). Model interpretation identified non-motorist age, violation history, roadway classification, lighting, and vehicle movement type as the most influential severity predictors, echoing findings from comparable studies in multiple jurisdictions (see Section 4.5).

These findings carry several practical implications for transportation agencies pursuing Vision Zero or Safe System approaches. The strong influence of non-motorist age on severity, with predicted P(Severe) rising sharply beyond age 60 in partial-dependence

analysis, supports prioritized infrastructure investments in corridors with high elderly pedestrian exposure, including refuge islands, pedestrian-scale lighting, and speed management. The binary severity model can serve as an operational screening tool. As an illustration, at a threshold of P(Severe) > 0.25, the LightGBM model captures approximately 45% of Severe crashes while maintaining 85% overall accuracy, though the optimal threshold would depend on agency-specific cost-sensitivity considerations. The importance of high-speed roadway types and vehicle movement patterns points to opportunities for speed management on arterials with significant VRU exposure and protected pedestrian signal phases at high-conflict intersections. The finding that resampling yields minimal benefit over class-weighted training simplifies deployment of severity prediction tools in other jurisdictions.

Several limitations temper the generalizability of these findings. Severity labels derived from police-reported KABCO classifications are subject to well-documented reporting inconsistencies, and the Moderate category is particularly susceptible to subjective officer judgment (Ali et al., 2024). The analysis relies on a single state's data, and findings may not transfer to regions with different road designs, traffic laws, or crash-reporting practices. The study also does not account for spatial autocorrelation or temporal trends, and variables such as real-time traffic conditions, detailed roadway geometry, and built-environment features are not available in the Signal4 database (Zeng et al., 2023; Sadeghi et al., 2024).

Future work should incorporate spatial structure through conditional autoregressive priors, test temporal instability across pre- and post-pandemic periods, and integrate richer contextual variables such as real-time traffic conditions and detailed roadway geometry. Cross-jurisdictional validation using crash databases from other states would test model transferability and the stability of identified risk factors. Hybrid modeling approaches combining ML predictive power with random-parameters econometric frameworks also represent a promising direction (Mannering et al., 2020; Sun et al., 2023; Sadeghi et al., 2024). Translating model predictions into operational decision-support tools, including threshold-calibrated risk-flagging systems for high-severity locations, offers the clearest path from research to proactive safety management.

References

Abdel-Aty, M.A., Radwan, A. (2000) "Modeling traffic accident occurrence and involvement", *Accident Analysis and Prevention*, 32(5), pp. 633-642.

Ali, Y., Hussain, F., Haque, M.M. (2024) "Advances, challenges, and future research needs in machine learning-based crash prediction models: A systematic review", *Accident Analysis and Prevention*, 194, 107378.

Anowar, P., Haque, N., Raihan, M.A., Hadiuzzaman, M. (2025) "Trajectory-based real-time pedestrian crash prediction at intersections: A novel non-linear link function for block maxima led Bayesian GEV framework addressing heterogeneous traffic condition", *arXiv preprint, arXiv:2510.12963.*

Du, B., Zhang, C., Sarkar, A., Shen, J., Telikani, A., Hu, H. (2024) "Identifying factors related to pedestrian and cyclist crashes in ACT, Australia with an extended crash dataset", *Accident Analysis and Prevention*, 207, 107742.

Grinsztajn, L., Oyallon, E., Varoquaux, G. (2022) "Why do tree-based models still outperform deep learning on typical tabular data?", *Advances in Neural Information Processing Systems (NeurIPS) 2022 Datasets and Benchmarks Track.*

Hoque, I., Ananda, T.N., Anowar, P., Naz, A., Murshed, M.N. (2025) "Machine learning approach in calibrating VISSIM microsimulation model for mixed traffic conditions", *Journal of Engineering Science*, 16(1), pp. 21-30.

Hossain, A., Sun, X., Das, S., Jafari, M., Rahman, A. (2024) "Investigating pedestrian-vehicle crashes on interstate highways: Applying random parameter binary logit model with heterogeneity in means", *Accident Analysis and Prevention*, 199, 107503.

Ibrahim, S., Sandt, A., Al-Deek, H., McCombs, J., Uddin, N. (2026) "Corridor-level and approach-level features associated with arterial wrong-way driving crashes and hotspots in South Florida", *Journal of Transportation Safety and Security.*

Ijaz, M., Lan, L., Zahid, M., Jamal, A. (2021) "A comparative study of machine learning classifiers for injury severity prediction", *Accident Analysis and Prevention*, 154, 106094.

Iranitalab, A., Khattak, A. (2017) "Comparison of four statistical and machine learning methods for crash severity prediction", *Accident Analysis and Prevention*, 108, pp. 27-36.

Kashifi, M.T. (2023) "Investigating two-wheelers risk factors for severe crashes using an interpretable machine learning approach and SHAP analysis", *IATSS Research*, 47, pp. 357-371.

Komol, M.M.R., Hasan, M.M., Elhenawy, M., Yasmin, S., Masoud, M., Rakotonirainy, A. (2021) "Crash severity analysis of vulnerable road users using machine learning", *PLoS ONE*, 16(8), e0255828.

Lee, C., Abdel-Aty, M. (2005) "Comprehensive analysis of vehicle-pedestrian crashes at intersections in Florida", *Accident Analysis and Prevention*, 37(4), pp. 775-786.

Li, W., Kockelman, K.M. (2022) "How does machine learning compare to conventional econometrics for transport data sets?", *Growth and Change*, 53(1), pp. 342-376.

Lundberg, S.M., Lee, S.-I. (2017) "A unified approach to interpreting model predictions", *Advances in Neural Information Processing Systems*, 30.

Mannering, F., Bhat, C.R., Shankar, V., Abdel-Aty, M. (2020) "Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis", *Analytic Methods in Accident Research*, 25, 100113.

Sadeghi, M., Aghabayk, K., Quddus, M. (2024) "A hybrid machine learning and statistical modeling approach for analyzing the crash severity of mobility scooter users considering temporal instability", *Accident Analysis and Prevention*, 206, 107696.

Santos, K., Firme, B., Dias, J.P., Amado, C. (2024) "Analysis of motorcycle accident injury severity and performance comparison of machine learning algorithms", *Transportation Research Record*, 2678(1), pp. 736-748.

Scarano, A., Rella Riccardi, M., Mauriello, F., D'Agostino, C., Pasquino, N., Montella, A. (2023) "Injury severity prediction of cyclist crashes using random forests and random parameters logit models", *Accident Analysis and Prevention*, 192, 107275.

Scarano, A., Sadeghi, M., Mauriello, F., Rella Riccardi, M., Aghabayk, K., Montella, A. (2025) "Cyclist crash severity modeling: A hybrid approach of XGBoost-SHAP and random parameters logit with heterogeneity in means and variances", *Journal of Safety Research*, 93, pp. 373-398.

Sun, Z., Wang, D., Gu, X., Abdel-Aty, M., Xing, Y., Wang, J., Lu, H., Chen, Y. (2023) "A hybrid approach of random forest and random parameters logit model of injury

severity modeling of vulnerable road users involved crashes", *Accident Analysis and Prevention*, 192, 107235.
Tahmid, M.M., Islam, Z., Abdel-Aty, M., Wang, C., Ahsan, M.J. (2025) "Estimating lane control signs (LCS) and variable speed limit (VSL) effects on expressway incident duration: A double machine learning causal forest approach", *SSRN preprint.*
Uddin, A.S.M.N., Abdel-Aty, M., Wang, C. (2025) "Joint modeling of segment-level crash risk and frequency on arterials: A copula-based framework with temporal instability", *SSRN preprint.*
Wang, C., Abdel-Aty, M., et al. (2026) "From prediction to explanation: A machine learning and causal mediation framework for roadway crash risk with connected vehicle data", *Transportation Research Part C*, 173, 105479.
Wen, X., Xie, Y., Jiang, L., Pu, Z., Ge, T. (2021) "Applications of machine learning methods in traffic crash severity modelling: Current status and future directions", *Transport Reviews*, 41(6), pp. 855-879.
WHO (2019) Global Status Report on Road Safety 2018, World Health Organization, Geneva.
Yue, L., Abdel-Aty, M., Wu, Y., Zheng, O., Yuan, J. (2020) "In-depth approach for identifying crash causation patterns", *Journal of Safety Research*, 73, pp. 119-132.
Zahid, M., Habib, M.F., Ijaz, M., Ameer, I., Ullah, I., Ahmed, T., He, Z. (2024) "Factors affecting injury severity in motorcycle crashes: Different age groups analysis using CatBoost and SHAP", *Traffic Injury Prevention*, 25(3), pp. 472-481.
Zeng, Q., Wang, Q., Zhang, K., Wong, S.C., Xu, P. (2023) "Analysis of the injury severity of motor vehicle-pedestrian crashes at urban intersections using spatiotemporal logistic regression models", *Accident Analysis and Prevention*, 189, 107119.
Zhang, J., Li, Z., Pu, Z., Xu, C. (2018) "Comparing prediction performance for crash injury severity among various machine learning and statistical methods", *IEEE Access*, 6, pp. 60079-60087.
Zhao, B., Zuniga-Garcia, N., Xing, L., Kockelman, K.M. (2024) "Predicting pedestrian crash occurrence and injury severity in Texas using tree-based machine learning models", *Transportation Planning and Technology*, 47(8), pp. 1205-1226.

Acknowledgements

The author thanks the Signal4 Analytics system for the data.

Conflict of interest

The author declares no conflict of interest.