PREDICTING PEDESTRIAN CRASH OCCURRENCE AND INJURY SEVERITY IN TEXAS USING TREE-BASED MACHINE LEARNING MODELS

Bo Zhao, M.Sc.
Graduate Research Assistant
Department of Plant Biology
The University of Texas at Austin
Austin, TX, 78712
bzhao@utexas.edu

Natalia Zuniga-Garcia, Ph.D.
Postdoctoral Fellow
Department of Civil, Architectural and Environmental Engineering
The University of Texas at Austin
Austin, TX, 78712
nzuniga@utexas.edu

Lu Xing, M.Sc.
Graduate Research Assistant
Department of Civil, Architectural and Environmental Engineering
The University of Texas at Austin
Austin, TX, 78712
xinglu@utexas.edu

Kara M. Kockelman, Ph.D., P.E.
(Corresponding Author)
Professor and Dewitt Greer Centennial Professor of Transportation Engineering
Department of Civil, Architectural and Environmental Engineering
The University of Texas at Austin – 6.9 E. Cockrell Jr. Hall
Austin, TX, 78712-1076
kkockelm@mail.utexas.edu

Under review for publication in *Traffic Injury Prevention*

ABSTRACT

This study investigates the frequency and injury severity of pedestrian crashes across Texas using tree-based machine learning models. Ten years of police records are used along with roadway inventory and other sources to map more than 78,000 pedestrian crashes over 700,000 road segments along with road design, land use, transit stops, and hospital location and weather information. Methods such as random forests (RF), gradient boosting (LightGBM and XGBoost), and Bayesian additive regression trees (XBART) are applied and compared. The crash frequency models indicate that highway design variables significantly positively impact crash frequencies.

Increments in total or fatal crash counts are related to a higher number of lanes, while higher speed and greater median and shoulder widths lead to fewer crash frequencies. Other variables such as proximity to schools, the number of transit stops, and population and job density increased pedestrian crash occurrences. Pedestrian severity models found that a high speed limit significantly increases the likelihood of pedestrian fatalities and severe injuries, and intoxicated drivers and pedestrians lead to more severe injuries. Also, pedestrian crashes are more likely to be severe and fatal at night and in areas with poor lighting conditions. An analysis of the vehicle type found that light-duty trucks (pickups, SUVs, and vans) also increase pedestrian severity. The comparison of the four models indicates that they performed similarly in predicting crash occurrences, with LightGBM showing significantly lower computational time. While for crash injury severity models, XBART obtained a higher precision value but with a significantly high computational time.

***Keywords**:* Pedestrian safety; Crash counts; Injury severity; Machine learning; Random forest; Gradient boosting; Bayesian Additive Regression Trees (BART).

INTRODUCTION

Walking is the most environmentally friendly form of transportation and has numerous health benefits [1], [2]. However, walking has become increasingly risky in recent years. During the ten years 2010-2019, the number of U.S. pedestrian fatalities increased by 46% [3], while the total walking miles traveled risen approximately 16% from 2009 to 2017 [4]. Walking trips accounted for 10% of the total trips taken in the nation in 2017, but pedestrians represented 16% of total traffic fatalities in the same year [3], [5]. This indicates that pedestrians experience a higher risk of fatality than motorists based on their exposure level. Furthermore, the risk of injury when walking is four times more than when driving a car [6]. There were 1.14 pedestrian fatalities per 100,000 people in Texas during 2019 [3], a rate 26% higher than the national average of 0.9. This study aims to investigate pedestrian-involved crashes in Texas using ten years of data from police records to determine which factors are critical to developing countermeasures for safer roadways.

Several studies have been conducted to examine the factors contributing to the frequency and severity of pedestrian crashes. Research findings suggest that pedestrian crash frequency can be influenced by roadway design characteristics, demographic, land use, and environmental, and weather conditions [7]–[11]. Several studies are based on macro-level information, with data aggregated at area levels like traffic analysis zones, census tracts, and zip code, while studies using micro-level aggregation, such as intersection and street segments, are more limited. Macro-level studies are useful to understand a city or state as a complex system. However, conducting micro-level studies can be crucial for the implementation of countermeasures for safety improvements. Using macro-level data at the Census tract, Wier et al. (2009) found that statistically significant predictors of vehicle-pedestrian collisions including traffic volume, employee and resident populations, arterial streets without public transit, proportions of the land area zoned for neighborhood commercial use, and residential-neighborhood commercial use, land area, the proportion of people living in poverty, and proportion of people aged 65 and over are statistically significant predictors of vehicle-pedestrian collisions. The micro-level analysis includes the work by Lee and Abdel-Aty (2005). They analyzed pedestrian crashes at intersections in Florida using negative binomial (NB) and log-linear models. They found that middle-aged male drivers and pedestrians were correlated to more pedestrian crashes than the other age and gender groups. With data aggregated at two different road segment levels, Vahedi Saheli and Effati (2021) used four count-based regression models (Poisson, NB, and their zero-inflated extensions) to show that residential, commercial, governmental, institutional, utility, and religious land uses have decisive impacts on the increase of pedestrian crash frequency.

Pedestrian injury severity analysis has also been widely used to the analyze factors leading to pedestrian injuries and fatalities. Tay et al. (2011) estimated a multinomial logit (MNL) model to identify the factors determining the severity of pedestrian-vehicle crashes. They found that fatal and serious crashes were associated with collisions involving heavy vehicles, drunk drivers, males or those under the age of 65, on high-speed roads, with inclement weather conditions, at night, on-road links, or on wider roads. Lee and Abdel-Aty (2005) used an ordered probit (OP) model and similarly found that pedestrian injury severity is closely related to pedestrians' physical condition (age and alcohol or drug use), the speed at the time of crashes, location of crashes, the presence of traffic control, weather and lighting conditions, and vehicle type. Kim et al. (2008) used heteroskedastic ordered probit (HOP) and MNL models to explore injury severity. The results show that pedestrian age induces heteroskedasticity, which affects the probability of fatal injury,

and the effect grows more pronounced with increasing age past 65. They found that the HOP model provides a better fit than the MNL model.

While extensive research has been developed using traditional frequency and severity models, studies using machine learning (ML) methods are limited. ML methods have been applied recently in the transportation safety literature to provide more accurate prediction models due to their ability to deal with more complex functions [16], [17]. Tree-based ML models are popular methods to make predictions. Ensemble tree models implementing bagging or boosting approaches usually outperform traditional statistically-based prediction models due to the informative and deliverable prediction. Examples of tree-based ML analysis include the use of random forest (FR) to evaluate the injury severity in pedestrian-bus crashes [18] and extreme gradient boosting (XGBoost) was used by Guo et al. (2021) to study older pedestrian crash severity. However, there is a lack of studies developing crash count models for pedestrian frequency prediction using tree-based ML models. There is also a need for model performance comparison across other models, such as light gradient boosting machine (LightGBM) and Bayesian additive regression tree (BART). In this study, four different tree-based machine learning models (RF, XGBoost, LightGBM, and accelerated BART) are applied to identify major factors contributing to pedestrian-vehicle crash occurrences and pedestrian injury severity.

DATA DESCRIPTION

Several data sources are used in this study. Crash records from 2010 to 2019 are obtained from the Texas Department of Transportation (TxDOT) Crash Records Information System (CRIS). The CRIS system comprises records of police reports generated in all 254 Texas counties and the hundreds of municipalities therein. Variables within the database characterize crashes according to time, location, severity, and road conditions. Also, the TxDOT Roadway Inventory database was used to obtain road-specific attributes.

The CRIS data was spatially matched with land use, population, job, rainfall, and other location features (schools, hospitals, transit stops) to examine the association between pedestrian crash counts and various contributing factors along Texas roads. Census tract-level population and job data were obtained from the 2010 US Census and Longitudinal Employer-Household Dynamics (LEHD) dataset respectively. Road segments were matched with the closest census tract centroid using the ArcGIS spatial join routine. Data were normalized by the area of census tracts. Other data sources include annual rainfall data (1981–2010) from the Texas Water Board, school locations from the Texas Education Agency, hospital locations from the Homeland Infrastructure Foundation, and transit stop locations from OpenStreetMap. ArcGIS Spatial Analysis tools were utilized to calculate numbers of transit stops and Euclidean distances from each road segment to the nearest schools and hospitals.

During the period between 2010 and 2019, a total of 5,631,223 crashes were recorded in the CRIS system. Among these crashes, 78,497 involved collisions or avoidance of pedestrians and a total of 72,243 pedestrians were involved. The distribution of the pedestrian crashes by fatality is described as follows: 5.9 % pedestrians were not injured, 33.0% presented a possible injury, 36.1% suffered a non-incapacitating injury, 15.8% were incapacitated, and 7.7% were killed, with 1.5% of unknown severity. Table 1 shows summary statistics of the variables at the road-segment level.

TABLE 1: Summary statistics of variables for road segments across Texas

Variables	Mean	Std. Dev	Min	Median	Max
Number of pedestrian crashes	**0.08**	**0.65**	**0.00**	**0.00**	**115**
Number of fatal pedestrian crashes	**0.01**	**0.10**	**0.00**	**0.00**	**10.00**
Segment length (in miles)	0.43	0.81	0.00	0.19	44.24
Number of lanes	2.23	0.78	1.00	2.00	14.00
Median width (in feet)	1.74	11.79	0.00	0.00	519.00
Average shoulder width (in feet)	1.41	3.62	0.00	0.00	42.00
On-system road	0.23	0.42	0.00	0.00	1.00
Indicator of curvature	0.11	0.31	0.00	0.00	1.00
Curve length (in meters)	21.68	125.77	0.00	0.00	9,630.57
Curve angle (degrees)	3.54	12.95	0.00	0.00	331.80
Average daily traffic (ADT) per lane	888	2,366	0.00	165	92,090
Percentage of truck ADT	5.96	7.22	0.00	3.20	95.80
Daily VMT (DVMT)	1,035	7,319	0.00	54	793,942
Speed limit (mph)	56.97	28.69	10.00	0.00	85.00
Rural (pop. < 5000)	0.41	0.49	0.00	0.00	1.00
Small urban (pop: 5000–49999)	0.10	0.30	0.00	0.00	1.00
Urbanized (pop: 50000–199999)	0.09	0.29	0.00	0.00	1.00
Large urbanized (pop: 200000+)	0.40	0.49	0.00	0.00	1.00
Population density (per sq. mile)	1,672	2,275	0.00	636	55,240
Job density (per sq. mile)	805	3,285	0.00	140	130,011
Average yearly precipitation (1981–2010) (inches)	36.48	11.52	8.00	37.00	61.00
Distance to nearest hospital (miles)	6.82	7.28	0.00	3.97	98.21
Distance to nearest school (miles)	2.08	3.09	0.01	0.74	53.95
Presence of transit stop within 100-meter buffer	0.01	0.08	0.00	0.00	1.00
Number of transit stops within 100-meter buffer	0.01	0.20	0.00	0.00	27.00

METHODOLOGY

This study investigates the performance of tree-based ensemble machine learning models in predicting pedestrian occurrence and severity in roadway segments in Texas. An outline of the proposed model training and sensitivity analysis process is presented in Figure 1. First, the pedestrian crash dataset is split into training and test sets. After fitting the models on the training set, the performance of the models is evaluated based on several metrics, such as the root mean squared error (RMSE) and R-squared in crash occurrence models. For crash severity prediction models, the metrics used include F1 score, area under the curve (AUC)—specifically under the Receiver Operating Characteristic (ROC) curve—a widely used measure of performance of classification, and model margins. Subsequently, the Bayesian optimization algorithm is applied to obtain the hyperparameter settings to achieve optimal model performance. Finally, the sensitivity analysis is carried out to identify the importance of the features. The models are developed using Python programming language.

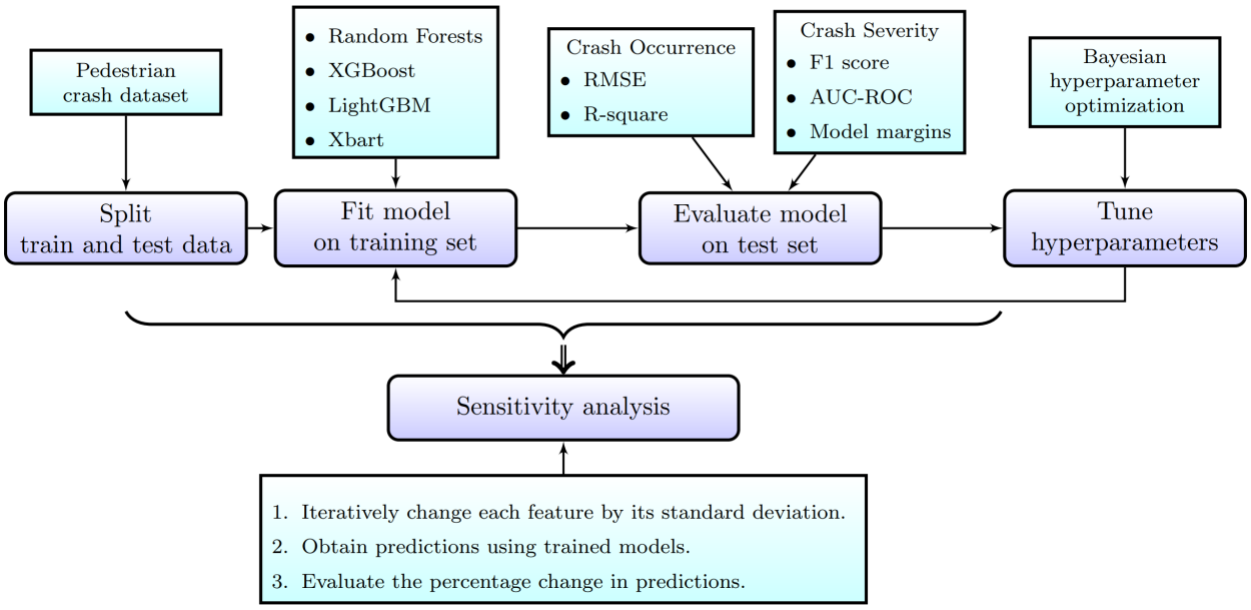


FIGURE 1: The process of model building, testing, and sensitivity analysis

Tree-Based Ensample Machine Learning Models

Tree-based models use a series of if-then rules to generate predictions from one or more decision trees. Various methods combining a set of tree models, i.e., ensemble methods, have attracted much attention and have been widely used for supervised learning tasks. These include random forests [20], [21], gradient boosting [22], [23], and Bayesian additive regression trees [24], [25], each of which uses different techniques to fit a linear combination of trees. This section investigates the performance of different tree-based ML models in predicting pedestrian crash occurrence and injury severity in Texas. The following subsections briefly introduce the investigated tree-based models.

Random Forest (RF)

An RF model comprises decision trees constructed by splitting each node using the best among a subset of predictors randomly chosen at that node with a different bootstrap sample of the data [20]. Running an RF algorithm can be described as follows [21]: (1) draw n_{tree} bootstrap samples from the original data; (2) for each bootstrap sample, grow an unpruned tree using the following procedure: at each node, randomly sample m_{try} of the predictors and choose the best split from among those variables; and, (3) predict new data by aggregating the prediction of the n_{tree} trees, i.e., majority votes for classification, average for regression. With the two layers of randomness, i.e., random feature selection and bootstrap/bagging, RF is powerful at handling complex and non-linear relationships. RF can also be trained quickly since the trees do not rely on each other and thus can be trained in parallel. However, RF is known to be less accurate for regression problems as it tends to overfit.

Extreme Gradient Boosting (XGBoost)

XGBoost is a scalable ML system for gradient tree boosting, which gives state-of-the-art results on a wide range of problems [22]. Boosting is an ensemble tree method that builds consecutive small trees with each tree focused on correcting the net error from the previous trees. For example, the first tree is split on the most predictive feature, and then the weights are updated to ensure that

the subsequent tree splits on whichever feature allows it to correctly classify the data points that were misclassified in the initial tree. The next tree will then focus on correctly classifying errors from that tree, and so on. The final prediction is a weighted sum of all individual predictions. Gradient boosting is the most popular extension of boosting and uses the gradient descent algorithm for optimization.

Efficient Gradient Boosting Decision Tree (LightGBM)

LightGBM is a popular gradient boosting decision tree model. Compared with XGBoost, LightGBM incorporates gradient-based one-side sampling (GOSS) to improve computational efficiency [23]. The basic assumption behind GOSS is that those samples with larger gradients, i.e., under-trained instances, will contribute more to the information gain. Therefore, to retain the accuracy of information gain estimation, GOSS keeps all the instances with large gradients (e.g., larger than a pre-defined threshold or among the top percentiles) and only randomly drops those instances with small gradients. It was shown that LightGBM could lead to a more accurate gain estimation than uniformly random sampling, with the same target sampling rate, especially when the value of information gain has a large range.

Accelerated Bayesian Additive Regression Trees (XBART)

XBART is a variant of the Bayesian additive regression tree (BART) model with improved computational efficiency [25]. Conceptually, BART is a Bayesian nonparametric approach that fits a parameter-rich model using a strongly influential prior distribution [24]. BART is similar to GBT models, i.e., XGBoost and LightGBM, in that they all sum the contribution of sequential weak learners. However, BART weakens the individual trees using a prior, instead of multiplying each sequential tree by a small constant, i.e., the learning rate, as in GBT models. Additionally, BART performs the iterative fitting by using the back-fitting Monte Carlo Markov Chain (MCMC) algorithm rather than using gradient descent algorithms. The Bayesian perspective yields a number of practical advantages of BART, including the robustness to hyperparameter settings, more accurate predictions, and the inherent Bayesian measure of uncertainties. On the other side, the incorporation of the MCMC algorithm also imposes severe computational demands, especially in the application of high-dimensional large datasets. XBART improves the computational efficiency by adopting the novel stochastic hill-climbing algorithms, which follow the Gibbs update framework in BART but replace the Metropolis-Hasting updates of each tree with a novel grown-from-root back-fitting strategy [25]. XBART is shown to yield very similar results to BART, but with much higher computational efficiency [25].

Hyperparameter Tuning

The aim of hyperparameter optimization is to find the hyperparameters of a given ML algorithm that return the best performance as measured by the specified evaluation metric. The optimization of hyperparameters (θ) can be represented in equation form as:

$$\theta^* = argmin_{\theta \in \Theta} f(\mathcal{M}, \theta) \quad (1)$$

where, $\mathcal{M}$ is the ML model; *f(x)* represents an objective function to minimize, such as RMSE for regression models or F1 score for classification models, evaluated on the validation set; θ^* is the set of hyperparameters that yields the lowest value of the score; and θ can take on any value in the domain Θ. Bayesian hyperparameter optimization methods build a probability model of the objective function, i.e., $P(f(\mathcal{M}, \theta)|\theta)$, by tracking the past evaluation results and using them to

select the most promising hyperparameters to evaluate in the true objective function [26]. Specifically, the process of Bayesian hyperparameter tuning can be described as follows: (1) build a surrogate probability model of the objective function; (2) find the hyperparameters that perform best on the surrogate; (3) apply these hyperparameters to the true objective function; (4) update the surrogate model incorporating the new results; and, (5) repeat steps 2–4 until the maximum number of iterations or specified time is reached.

Sensitivity Analysis

ML models excel at capturing complex relationships between input independent and output response variables. However, they can be less intuitive in explaining how and why such relationships are captured. Several sensitivity analysis methods were developed to mitigate the interpretability deficiency, aiming to unveil the cause-and-effect relationship between the input and output variables. Sensitivity analysis is a simple yet powerful way to understand an ML model by examining what impact each feature has on the model's prediction. The feature value was changed to calculate feature sensitivity, while all the other features stay constant, and the output of the model was recorded. If the model's outcome has been altered drastically by changing the feature value, it means that this feature significantly impacts the prediction.

Specifically, given a test set X, the process of evaluating the sensitivity of feature X_i can be described as follows: (1) train the baseline model on X and denote the prediction vector as y; (2) create a new set X^* where a transformation was applied, such as reshuffling or dropping, over feature X_i; (3) perform prediction on X^* and denote the prediction vector as y^*; (4) measure the change in the outcome using the percentage change in the prediction mean, i.e., $\frac{\bar{y^*}-\bar{y}}{\bar{y}} \times 100\%$. In pedestrian crash occurrence prediction, the transformation is defined as in Li and Kockelman (2020): an increase of one standard deviation for continuous input variables and binary (0 to 1) change for binary input variables. Specifically, for each input variable, one standard deviation or binary change is applied to each data point. The modified variables are passed to the model to calculate the prediction, i.e., permuted prediction. Then, the difference between the mean of original prediction and permuted prediction is calculated to represent the contribution of that feature. In injury severity prediction, the probability of each class was obtained for every single data point. The same imputation and computation approach was used to analyze the marginal effects, as in pedestrian crash occurrence prediction, except that each class's probability is used instead of class values.

Ordinal Classification

In classification problems, the class labels can be ordered, e.g. injury severity (0,1,2,3,4) and cancer stage (I,II,III,IV) [28]. It is clear that there is an order among these labels and in terms of the severity, where 4>3>2>1>0. The ordered probit or ordered logit models are widely used in this case [8], [29]. Standard machine learning methods for multiclass classification commonly assume the class labels are not ordered. Frank and Hall [30] proposed a simple method to implement ordinal classification with standard binary classification, such as RandomForest classifier. Take injury severity as an example, the process is summarized as: the injury class that is higher than 0 is encoded as 1, resulting in new injury classes (0,1,1,1,1). After applying the standard binary classification, we have the probability for new classes, $P(y^* = 0)$ and $P(y^* = 1)$ where $P(y^* = 1)$ is actually equivalent to $P(y > 0)$. In second time iteration, the injury class that is higher than 1 is encoded as 1 and that is lower than 1 is encoded as 0, resulting in new classes

(0,0,1,1,1). Similarly, the probability $P(y > 1)$ is obtained. After iterating all injury class values, the probability $P(y > i), i = 0, 1,2,3$ is obtained. In general, for class value i:

$$P(y = i) = \begin{cases} 1 - P(y > i), i = 0 \\ P(y > i - 1) - P(y > i), i = 1, 2, 3 \\ P(y > i - 1), i = 4 \end{cases} \quad (2)$$

RESULTS

Pedestrian Crash Occurrence Prediction

Four tree-based ensemble ML models were developed to predict pedestrian crash occurrence: RF, XGBoost, LightGBM, and XBART. For each model configuration, two models were trained—one for total pedestrian crashes and another for fatal pedestrian crashes. The optimal hyperparameters for each model were obtained using Bayesian optimization.

Model Performance Evaluation

Table 2 summarizes the model performance measured by R-square and RMSE on the testing set to predict total and fatal pedestrian crash occurrence. For the total pedestrian crash occurrence prediction model, LightGBM achieves the best performance in terms of both R-square and RMSE, while for predicting the fatal pedestrian crash occurrence, RF yields the best performance. The R-square values for the fatal pedestrian crashes are lower than the values for the total pedestrian crashes, which can be related to the significantly low number of fatal pedestrian crashes. The computation times, including training and testing after the optimal hyperparameters are obtained, are also compared among the models. LightGBM is the most computationally efficient model due to the efficient GOSS optimization algorithm. XBART is the most computationally expensive model, which can be explained by the expensive MCMC connection between the trees.

TABLE 2: Comparison of model performance and computation time

Model	Total pedestrian crash occurrence			Fatal pedestrian crash occurrence		
	R-square	RMSE	Time (s)	R-square	RMSE	Time (s)
RF	0.359	0.242	216	**0.148**	**0.008**	278
XGBoost	0.318	0.258	126	0.070	0.009	133
LightGBM	**0.363**	**0.241**	**43**	0.133	0.009	**25**
XBART	0.351	0.245	354	-0.001	0.010	5110

Feature Sensitivity Analysis

The practical importance of input variables can be estimated using the proposed sensitivity analysis approach. The value of continuous features is increased by one standard deviation, and binary changes are made on binary features for each data point in the dataset. Then the percentage change in the mean of the model prediction is estimated. The estimated feature importance for total and fatal pedestrian crash occurrence are shown in Figure 2. The y-axis shows the name of the input variables. The x-axis represents the percentage change in the mean of model prediction, i.e., total or fatal pedestrian crash occurrence, after applying the proposed transformation on the corresponding input features. Different colors represent different ML models: blue for RF, orange for LightGBM, green for XGBoost, and red for XBART.

As shown in Figure 2, the vehicle miles traveled (VMT) have the most significant impact on total and fatal pedestrian crash occurrence. One standard deviation increase on VMT can lead to around a 270% and more than 300% increase in the total number and number of fatal pedestrian crash occurrences per roadway segment, respectively. However, it should be noted that one standard deviation increase of VMT (7,319) on all roadway segments is not practical, considering the capacity limit of segments. Therefore, the process was repeated considering a double VMT on each segment. The results indicate that the total and fatal pedestrian occurrence will increase by 50%, which is still a significant impact. These results, consistent with literature findings [31], represent the higher crash risk faced by pedestrians with increasing VMT, which is consistent with the expectation that crash frequencies increase with the increase in pedestrian exposure to motorized vehicles.

The number of transit stops within a buffer of 100 meters is a relevant variable in predicting total and fatal pedestrian crashes in roadway segments, according to the results of the LightGBM and XBART models. This variable is an indirect measure of pedestrian exposure as pedestrian activity surrounding transit stops is high. Similarly, variables for the distance to nearest school and distance to nearest hospital offer a practical significance. The number of pedestrian crashes increases in areas near schools and decreases as distance from schools increases, consistent with literature findings [32]. Interestingly, the hospital proximity is particularly significant for fatal crashes, where the frequency increases as the distance to the hospital increases, possibly related to the response time of emergency services. Although relevant, these variables are rarely considered in pedestrian safety literature [8].

Highway design variables such as on-system roads (or state-maintained arterials), number of lanes, curve angle, curvature indicator, and curvature length have a significant positive impact on pedestrian crash frequencies. One standard deviation increment on the number of lanes can lead to more than a 25% increment in total or fatal crash counts. On-system roads are found to be strongly correlated to the number of total crashes. The speed limit is found to be negatively correlated to the number of crashes. This can be related to the reduced exposure of pedestrians to high-speed roadway segments. However, high speed limits lead to more severe injuries, as discussed in the following section. Variables such as median and shoulder widths show diverse variations across the different models, limiting the conclusions for these variables.

Land use characteristics are described by variables such as population, job density, and types of urban areas. These metrics are directly related to pedestrian exposure. For example, dense urban areas with high job density usually have higher traffic volumes and pedestrian activity. Changes in one standard deviation led to a positive, significant increment in pedestrian crash frequencies, as expected. The number of pedestrian crash occurrences increases by approximately 50% for total counts and 30% for fatal counts when the population and job density are increased by one standard deviation. Large-urbanized, urbanized, and small urban locations have more conservative increments of 10% for total pedestrian crashes. However, for the fatal count model, the effect differs significantly across models, possibly related to the low count number within the different categories.

The four different models come to a similar conclusion about the significance of some features, such as distance to the nearest school, job density, population density, and VMT. However, the

results diverge on other features, such as the number of transit stops within a 100-meter buffer. XBART and LightGBM consider the number of transit stops a very important feature in predicting the total pedestrian crash occurrence. One standard deviation increase on the transit stop variable can lead to 150% and 300% increase on total pedestrian crash occurrence, respectively. However, results from LightGBM and XGBoost show that the number of transit stops has little impact on the total pedestrian crash occurrence. This observation indicates that different ML models interpret the significance of the input features differently. It might make more sense to look only at the model that yields the best performance, i.e., prediction accuracy. Noticeably, the discrepancies in the results from different models are even more obvious in fatal pedestrian occurrence prediction as compared with total pedestrian occurrence. This again stresses the importance of choosing the best performing model when comparing the evaluation metrics and then analyzing the feature importance using the chosen optimal model.

Pedestrian Crash Injury Severity Prediction

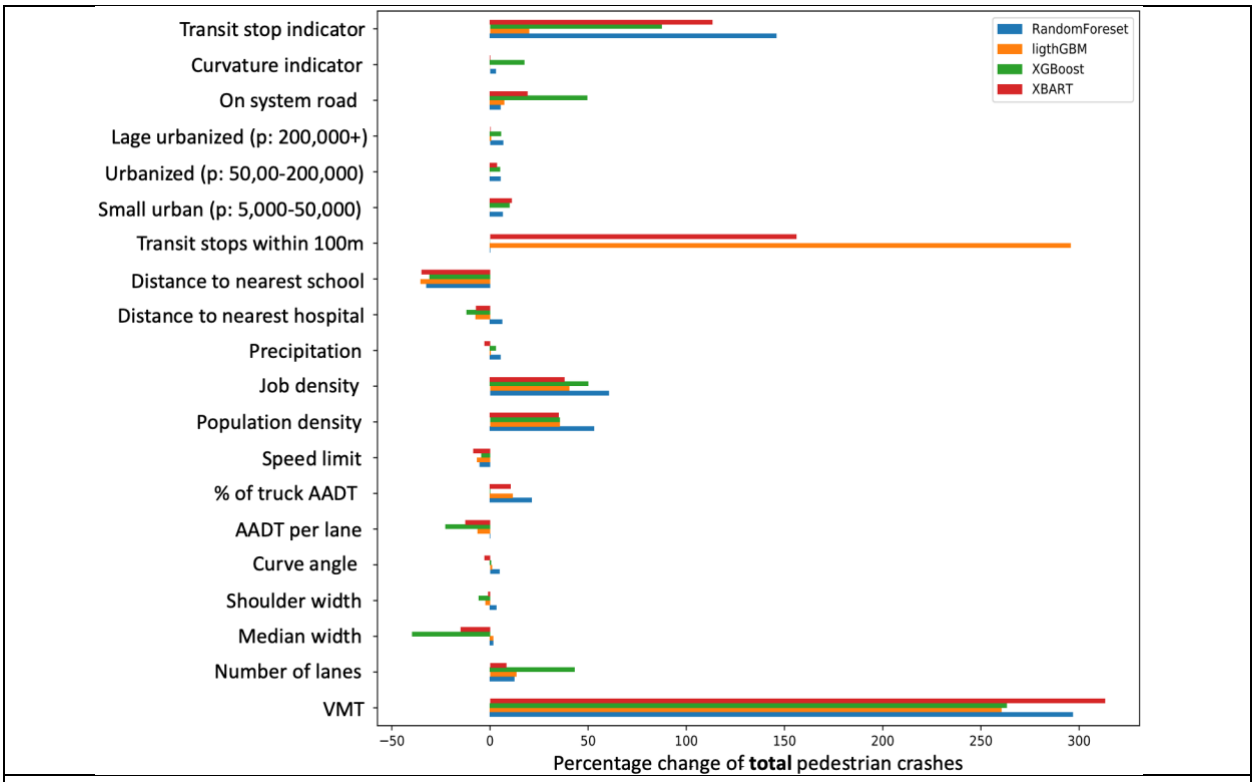
To estimate the crash injury severity models, first, the complete dataset was randomly split into training and testing datasets to predict crash injury severity. Then, four models (RF, XGBoost, LightGBM, and XBART) were fitted on the training dataset and tested on the testing dataset. Before model fitting, parameters for each model were tuned with the Bayesian optimization method to obtain optimal parameters. The simulation was repeated ten times.

Model Performance Evaluation

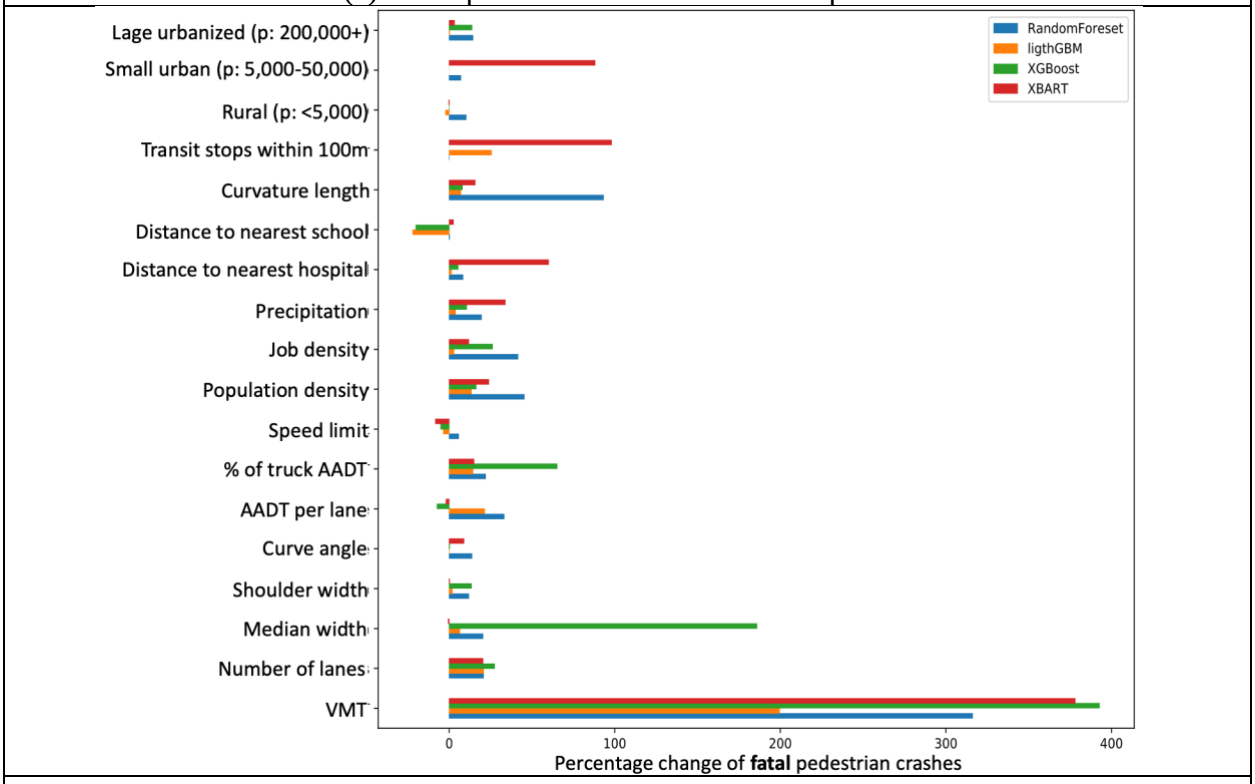
Table 3 summarizes model performance on model running time, accuracy, precision, recall, F1 score, and GM. In general, RF, XGBoost, and LightGBM behave similarly in terms of these classification metrics, but LightGBM runs much faster than the other two GBT models. Even with higher precision, XBART shows lower recall and F1, and it is heavily time-consuming. The crash injury data is imbalanced with 7%, 33%, 36%, 17%, and 7% of class 0 (not injured), 1 (possibly injury), 2 (non-incapacitating), 3 (incapacitating) and 4 (killed), respectively. The geometric mean (GM) metric is less sensitive to an imbalanced dataset [33]. Results show that RF, XGBoost, and LightGBM achieve the same GM value of 0.53, which is higher than XBART with 0.51. This indicates that XBART may be more sensitive to imbalanced data.

TABLE 3: Summary of model performance on injury severity prediction

Models	Time (s)	Accuracy	Precision	Recall	F1 score	GM
RandomForest	36.55	0.42	0.49	0.33	0.34	0.53
XGBoost	75.17	0.42	0.43	0.33	0.34	0.53
LightGBM	18.59	0.42	0.45	0.34	0.34	0.53
XBART	1447.56	0.42	0.53	0.31	0.32	0.51



(a) Total pedestrian crash occurrence prediction



(b) Fatal pedestrian crash count prediction

FIGURE 2: Sensitivity analysis for crash occurrence predictions

The model margins show the confidence of a classifier making a correct classification. A *positive margin value* means the classifier votes for the right classification [20], and a *negative margin value* indicates the classifier voted incorrectly. Figure 12 shows ten-time repeated results plotted in a bar graph; model margins are affected by the injury class. For example, all four models achieve higher margins on class 2 and class 4 (around 0.2), while the margin values for class 0 and class 3 are negative (-0.4 and -0.6), indicating a high discrepancy between true class and predicted class. This discrepancy is not necessarily related to data imbalance as class 3 makes up a greater proportion of the data than class 4, which has a higher margin. Also, XBART shows a much greater capacity for correctly classifying class 2, while for other classes XBART is limited to a weak classification capacity.

The ROC curve is one of the most important metrics to visualize the performance of multiclass classification models. It quantifies the extent to which the model is able to distinguish between classes (Narkhede, 2018). The AUC is the quantitative measure of ROC. The point (0,1) in the ROC curve represents the prefect classifier, meaning no false-positive error happens (Fawcett, 2006). The ROC curve of one simulated result is presented in Figure 4 to analyze how those models behave in the different classes. Based on the results, RF, XGBoost, and LightGBM are capable of classifying on class 4 while achieving an AUC around 0.9. However, XBART shows less ability to vote for the right classification on class 4 (AUC = 0.72). For other injury classes, RF, XGBoost, and LightGBM obtain an AUC value ranging from 0.59 to 0.67, which exceeds that of XBART.

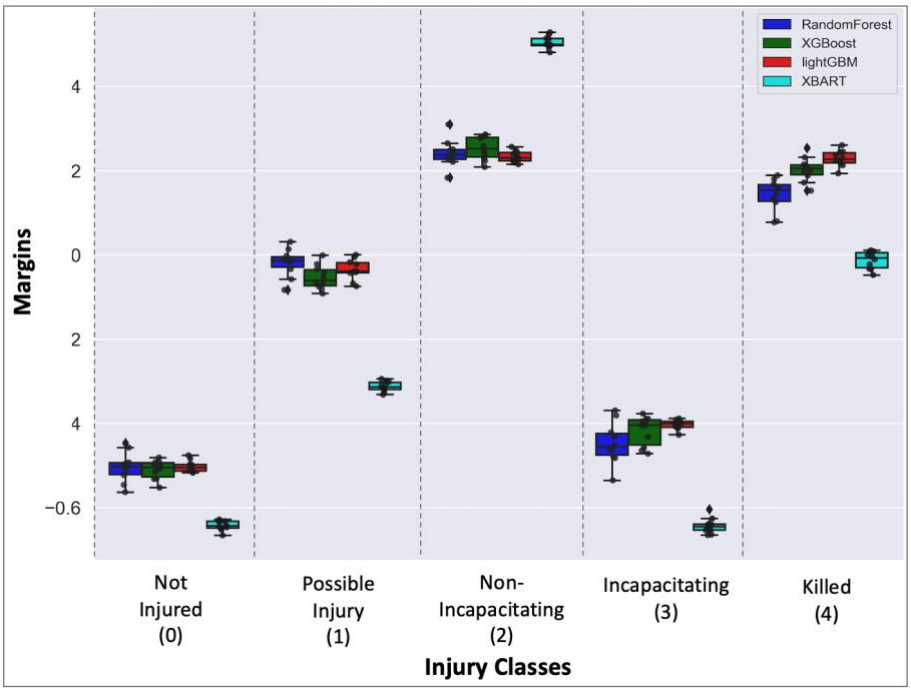


FIGURE 3: Margins of classification models on different classes

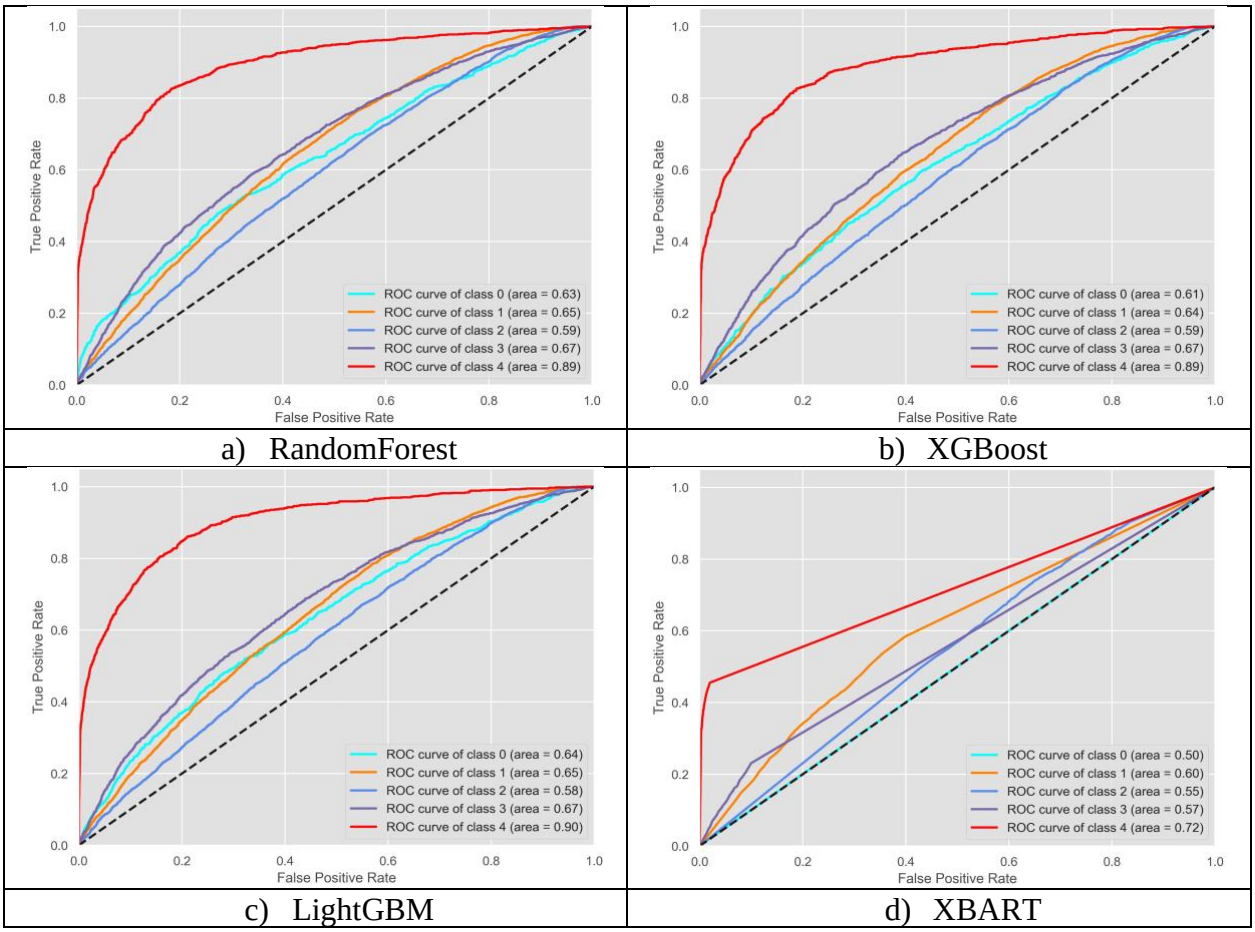


FIGURE 4: ROC curve of different classification models

Marginal Effects

The marginal effects of each variable are analyzed by data imputation to identify the most important factors that contribute to fatal pedestrian injury. First, the data is randomly split for training and testing data. Then, the hyperparameters are optimized on the training data, and models are trained on training data with optimal hyperparameters. Finally, the model is used for prediction on the testing dataset before and after one variable is imputed, and the probability difference between two times of prediction is defined as the marginal effect for that imputed variable. This process is repeated ten times, and the results are summarized in Figure 5.

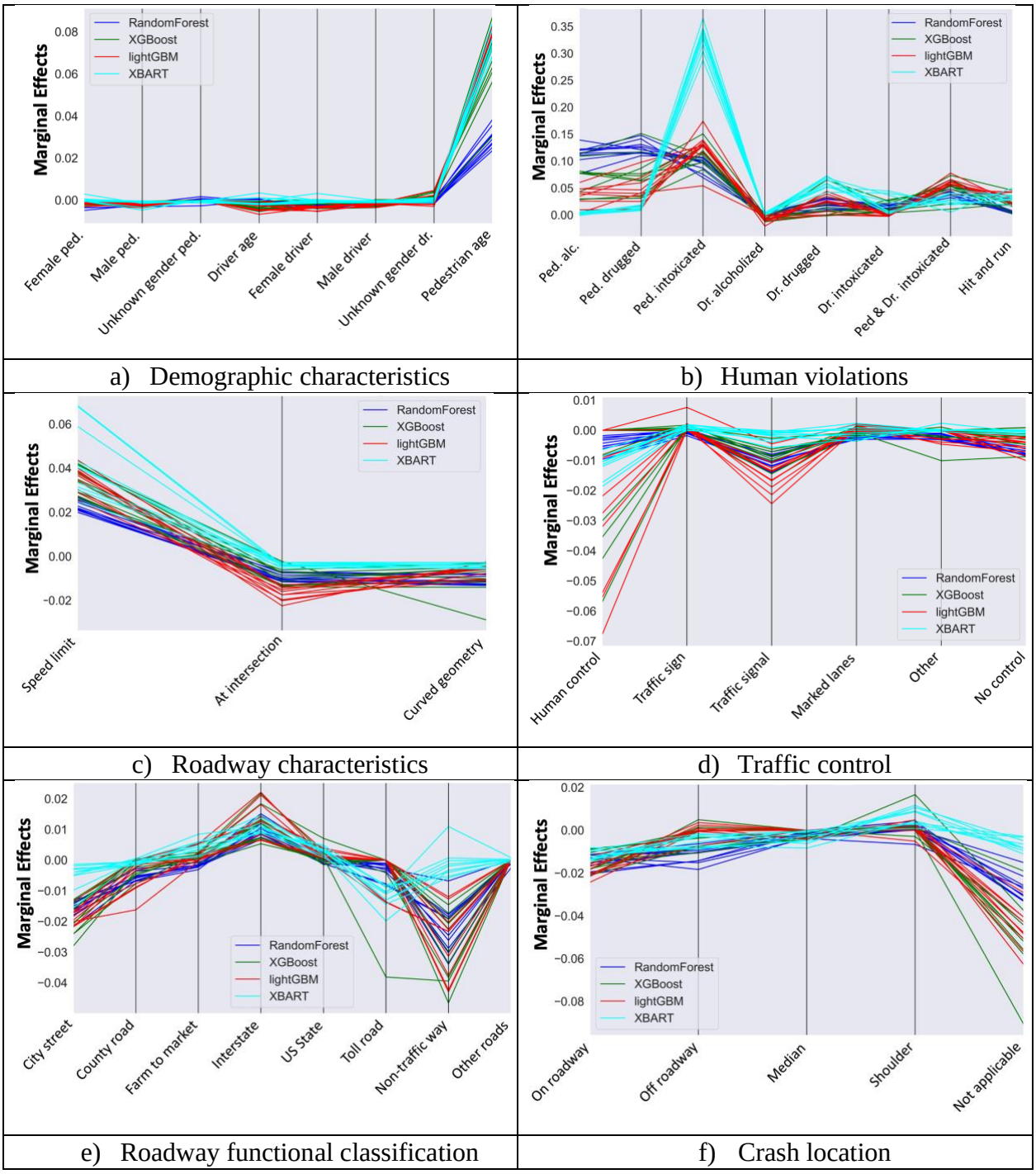
The variables are classified into ten categories. In terms of driver and pedestrian characteristics, driver age seems to have both positive and negative marginal effects. This can be related to the impact of driver age observed in some literature, with the involvement of younger drivers increasing the risk of high severity as compared to the presence of middle-age drivers [34], [35]. One study found that drivers aged 65 and older also increase the risk for pedestrian injury severity [36]. However, some researchers have found that this is not always the case, as older drivers may also be more experienced [37]. Furthermore, a high value for pedestrian age has a high likelihood of increasing injury severity, which might be due to the greater physical vulnerability of older people.

Several human violation variables were tested to analyze the effect of pedestrian and driver intoxication. Figure 5b shows that the intoxicated pedestrian variable is the most important factor that contributes to fatal pedestrian injury among those variables. The probability of a pedestrian fatality shows significant positive changes after data imputation, indicating that pedestrian intoxication greatly increases the risks of pedestrian death in a crash. When pedestrian intoxication is imputed, the probability of pedestrian death increases by 15% on average in RF, XGBoost, and LightGBM models, and XBART shows a probability change as great as 30%. Driver intoxication also leads to an increase in pedestrian death probability change. Intoxication is more likely to cause pedestrian fatality in a crash, and this result is also supported by the CRIS dataset, where intoxication is involved in 38% of fatal crashes. The simulation result is consistent with the previous report that intoxication has the strongest effect on pedestrian death [8]. Hit-and-run accidents are also related to a high pedestrian severity level. In the CRIS data, 19% of pedestrian deaths are hit-and-run cases, highlighting the relevance of this variable. The high severity is largely due to the time delay incurred when the driver leaves the crash location, which delays emergency services and prompt attention to the pedestrian.

Speed limit contributes significantly to the probability of pedestrian death, and the marginal effect ranges from 2% to 6% among these classification models. As expected, roads with higher speed limits led to more pedestrian injuries, consistent with previous studies [38]. However, crash frequency is reduced as speed limit increases, as found in the analysis of the previous section. Approximately 21% of the crashes were located at intersections. The results indicate that those crashes have a lesser risk of pedestrian fatalities than mid-segment crashes, likely due to reduced speed at these locations. Traffic control, including traffic signs and traffic signals, can help to reduce the probability of a crash and thus pedestrian death. Results suggest that traffic control is predicted to reduce pedestrian death probability on average by 3% in LightGBM, 2% in XGBoost, 1% in XBART, and 0.5% in RF. Also, data imputation on traffic signals decreases pedestrian death probability by as much as 2% on average in LightGBM.

Roadway functional classification is also an important factor. Different road types (such as country roads, city streets, and interstate roads) also play distinct roles in fatal pedestrian injury. For example, city streets and non-trafficways help reduce the marginal effect in RF, XGBoost, and LightGBM models, while the interstate seems to increase the marginal impact on all models. In Texas, interstate highways account for 6% of pedestrian crashes but 21% of pedestrian fatalities. This outcome is likely related to the speed of the crash. As analyzed previously, high-speed roadway segments tend to have fewer pedestrian crashes, but the severity is higher due to the speed of impact. The crash location analysis seems to indicate that crashes occurring on the roadway shoulder have a higher risk of causing pedestrian fatalities compared to crashes on the roadway and in the median area.

The area type is also an important factor for injury severity. The results suggest that rural and small urban areas present a higher risk for pedestrian fatalities. Factors such as distance to hospitals and speed limits can influence this finding. Rural and small urban areas tend to be less dense, and the emergency response time is slower than in urban areas. However, results from the previous section indicate that pedestrian activity is lower in these areas compared to large urban areas.



a) Demographic characteristics

b) Human violations

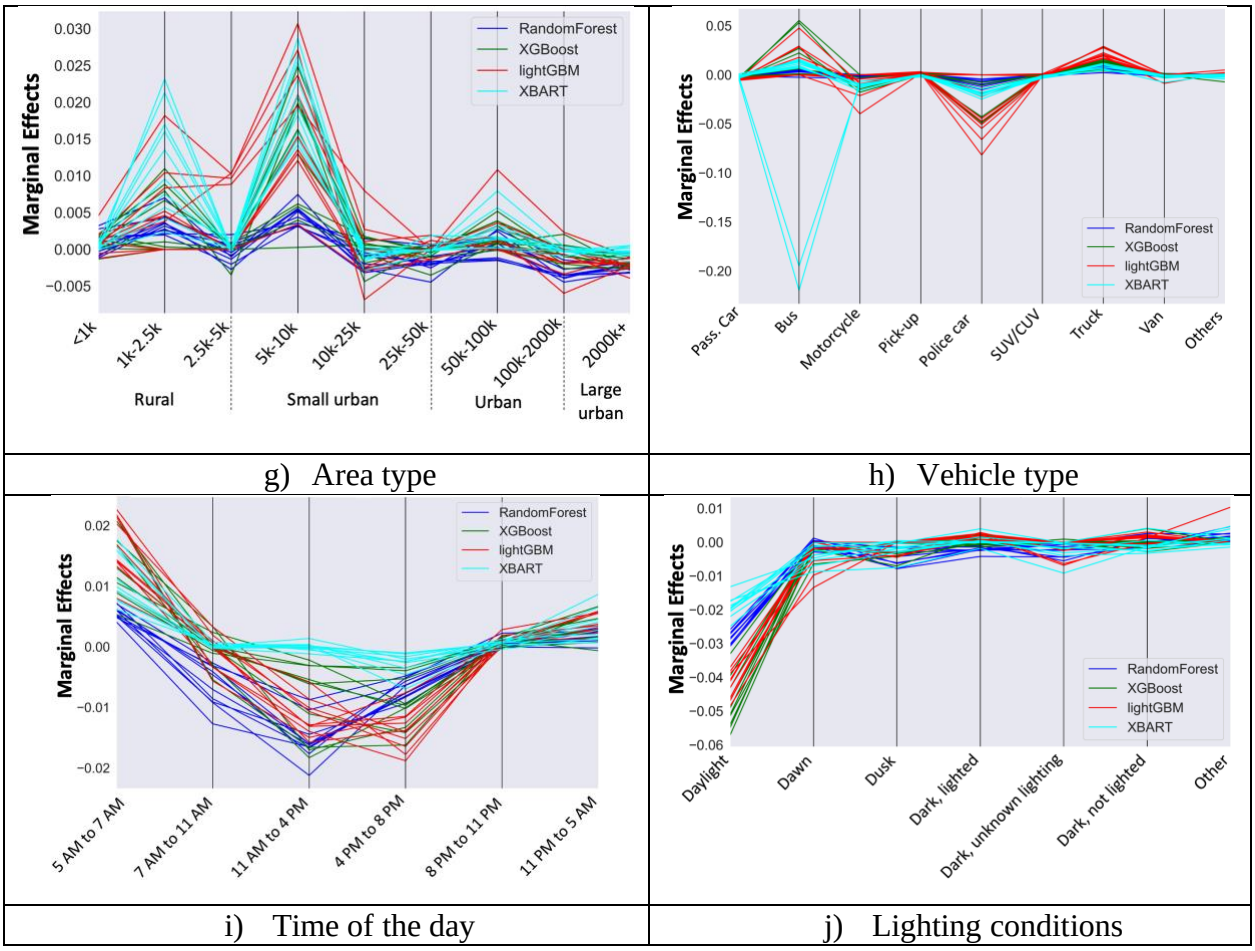
c) Roadway characteristics

d) Traffic control

e) Roadway functional classification

f) Crash location

FIGURE 5: Parallel coordinate plots of marginal effects for fatal crashes



g) Area type | h) Vehicle type

i) Time of the day | j) Lighting conditions

FIGURE 5: Parallel coordinate plots of marginal effects for fatal crashes (Cont.)

In terms of vehicle types, research in the field indicates that high injury severity is associated with light-duty vehicles, such as SUV/CUVs, pickup trucks, and vans, due to the heavy mass involved in the collision [8], [35], [39], [40]. However, this study shows that trucks involved in a crash are more likely to cause the death of pedestrians, but the effects of vans and SUV/CUVs are not significant. Busses also have a significant effect, but it is important to mention that the number of crashes involving buses is low compared to other vehicle body types.

Environmental factors such as crash time and lighting condition strongly affect the pedestrian injury severity [35], [41]. The time of day is found to influence its marginal effect. Specifically, the period after 8 PM and before 7 AM (under dark conditions) has a positive effect on pedestrian deaths. Approximately 80% of pedestrian deaths occurred at this time. In contrast, in the daytime, the probability of pedestrian death is reduced in all models. Similarly, the brighter daytime conditions significantly help to lower the likelihood of pedestrian death in all models, but not the other light types. This finding highlights the importance of streetlight improvements to reduce pedestrian crashes.

CONCLUSIONS

Tree-based machine learning models were applied to identify major factors contributing to pedestrian-vehicle crash occurrences and pedestrian injury severity. The analysis showed that

pedestrian and driver intoxication levels, speed limit, and lighting conditions significantly affect the severity of crashes. Variables such as crash location, traffic control, and roadway characteristics were also analyzed. The principal findings suggest that crashes located at intersections have a lesser risk of pedestrian fatalities than mid-segment crashes, likely due to reduced speeds at these locations. Also, traffic control, including traffic signs and traffic signals, can help to reduce the probability of a crash and pedestrian death. In terms of pedestrian crash frequencies, VMT was one of the most significant variables, and the number of transit stops within a buffer of 100 meters was relevant in predicting total and fatal pedestrian crashes in roadway segments. Results also found that highway design variables significantly positively impact pedestrian crash frequencies. Finally, the number of pedestrian crash occurrences increases by more than one third when the population and job density increase by one standard deviation.

This study also showed a comparison across four different tree-based models. In the pedestrian crash occurrence prediction, the principal results showed that all four models perform similarly, with close root mean square error (RMSE) and R-square for total crash occurrence. Still, LightGBM exceeds the other three models in terms of computational efficiency. For fatal crash occurrence, LightGBM and RF have comparable performance. However, XGBoost and XBART showed significantly lower goodness of fit values. Also, XBART is more sensitive to imbalanced data than the other models are. In the injury severity prediction, RF, XGBoost, and LightGBM achieved similar goodness of fit performance, evaluated by the metrics' accuracy, precision, recall, F1, and geometric mean. XBART obtained a higher precision value, but the other metrics were lower, with a significantly high computational time.

Findings from this study underscore the importance of campaigns against driving and walking while intoxicated, installation of streetlights in pedestrian-active areas, improved roadway design, and enforcement of safety countermeasures in areas where pedestrians are more vulnerable (such as near bus stops and schools). It also highlights the importance of detailed police reports to develop analyses of this type that can be used to improve pedestrian safety.

AUTHOR CONTRIBUTION STATEMENT

The authors confirm contribution to the paper as follows: writing-original draft preparation: B. Zhao, N. Zuniga-Garcia, L. Xing; conceptualization and design: K. Kockelman, B. Zhao, L. Xing; methodology: K. Kockelman, B. Zhao, L. Xing; data assembly and analysis: B. Zhao, L. Xing, N. Zuniga-Garcia; writing-reviewing and editing: N. Zuniga-Garcia, B. Zhao, L. Xing, K. Kockelman. All authors have reviewed the results and approved the final version of the manuscript.

REFERENCES

[1] H. Christian *et al.*, "Encouraging Dog Walking for Health Promotion and Disease Prevention," *Am. J. Lifestyle Med.*, vol. 12, no. 3, pp. 233–243, May 2018, doi: 10.1177/1559827616643686.

[2] P. Kelly, M. Murphy, and N. Mutrie, "The Health Benefits of Walking," in *Walking*, vol. 9, Emerald Publishing Limited, 2017, pp. 61–79. doi: 10.1108/S2044-994120170000009004.

[3] G. H. S. A. GHSA, "Pedestrian Traffic Fatalities by State," 2020.

[4] U. States. D. O. Transportation. B. O. T. S. BTS, "National Transportation Statistics: U.S. Passenger-Miles," 2019.

https://rosap.ntl.bts.gov/gsearch?collection=dot:35533&type1=mods.title&fedora_terms1=National+Transportation+Statistics (accessed May 16, 2021).

[5] F. H. A. FHWA, "Summary of Travel Trends 2017 National Household Travel Survey," ORNL/TM-2004/297, 885762, 2018. doi: 10.2172/885762.

[6] R. Elvik, "The non-linearity of risk and the promotion of environmentally sustainable transport," *Accid. Anal. Prev.*, vol. 41, no. 4, pp. 849–855, Jul. 2009, doi: 10.1016/j.aap.2009.04.009.

[7] C. Lee and M. Abdel-Aty, "Comprehensive analysis of vehicle–pedestrian crashes at intersections in Florida," *Accid. Anal. Prev.*, vol. 37, no. 4, pp. 775–786, Jul. 2005, doi: 10.1016/j.aap.2005.03.019.

[8] Rahman, M. and Kockelman, K.M., 2020. Predicting Pedestrian Crash Occurrences and Injury Severity in Texas Presented at the 100th Annual Meeting of the Transportation Research Board. Under review for publication in Traffic Injury and Prevention.

[9] S. Ukkusuri, L. F. Miranda-Moreno, G. Ramadurai, and J. Isa-Tavarez, "The role of built environment on pedestrian crash frequency," *Saf. Sci.*, vol. 50, no. 4, pp. 1141–1151, Apr. 2012, doi: 10.1016/j.ssci.2011.09.012.

[10] Y. Wang and K. M. Kockelman, "A Poisson-lognormal conditional-autoregressive model for multivariate spatial analysis of pedestrian crash counts across neighborhoods," *Accid. Anal. Prev.*, vol. 60, pp. 71–84, Nov. 2013, doi: 10.1016/j.aap.2013.07.030.

[11] L. Yue, M. Abdel-Aty, Y. Wu, O. Zheng, and J. Yuan, "In-depth approach for identifying crash causation patterns and its implications for pedestrian crash prevention," *J. Safety Res.*, vol. 73, pp. 119–132, Jun. 2020, doi: 10.1016/j.jsr.2020.02.020.

[12] M. Wier, J. Weintraub, E. H. Humphreys, E. Seto, and R. Bhatia, "An area-level model of vehicle-pedestrian injury collisions with implications for land use and transportation planning," *Accid. Anal. Prev.*, vol. 41, no. 1, pp. 137–145, Jan. 2009, doi: 10.1016/j.aap.2008.10.001.

[13] M. Vahedi Saheli and M. Effati, "Segment-Based Count Regression Geospatial Modeling of the Effect of Roadside Land Uses on Pedestrian Crash Frequency in Rural Roads," *Int. J. Intell. Transp. Syst. Res.*, vol. 19, no. 2, pp. 347–365, Jun. 2021, doi: 10.1007/s13177-020-00250-1.

[14] R. Tay, J. Choi, L. Kattan, and A. Khan, "A multinomial logit model of pedestrian–vehicle crash severity," *Int. J. Sustain. Transp.*, vol. 5, no. 4, pp. 233–249, 2011.

[15] J.-K. Kim, G. F. Ulfarsson, V. N. Shankar, and S. Kim, "Age and pedestrian injury severity in motor-vehicle crashes: A heteroskedastic logit analysis," *Accid. Anal. Prev.*, vol. 40, no. 5, pp. 1695–1702, Sep. 2008, doi: 10.1016/j.aap.2008.06.005.

[16] M. Effati and M. Vahedi Saheli, "Examining the influence of rural land uses and accessibility-related factors to estimate pedestrian safety: The use of GIS and machine learning techniques," *Int. J. Transp. Sci. Technol.*, Apr. 2021, doi: 10.1016/j.ijtst.2021.03.005.

[17] S. Mokhtarimousavi, J. C. Anderson, A. Azizinamini, and M. Hadi, "Factors affecting injury severity in vehicle-pedestrian crashes: A day-of-week analysis using random parameter ordered response models and Artificial Neural Networks," *Int. J. Transp. Sci. Technol.*, vol. 9, no. 2, pp. 100–115, Jun. 2020, doi: 10.1016/j.ijtst.2020.01.001.

[18] S. K. Sivasankaran and V. Balasubramanian, "Severity of Pedestrians in Pedestrian - Bus Crashes: An Investigation of Pedestrian, Driver and Environmental Characteristics Using Random Forest Approach," in *Proceedings of the 21st Congress of the International*

Ergonomics Association (IEA 2021), Cham, 2021, pp. 825–833. doi: 10.1007/978-3-030-74608-7_101.

[19] M. Guo, Z. Yuan, B. Janson, Y. Peng, Y. Yang, and W. Wang, "Older Pedestrian Traffic Crashes Severity Analysis Based on an Emerging Machine Learning XGBoost," *Sustainability*, vol. 13, no. 2, Art. no. 2, Jan. 2021, doi: 10.3390/su13020926.

[20] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[21] A. Liaw and M. Wiener, "Classification and Regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[22] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785--794. doi: 10.1145/2939672.2939785.

[23] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," 2017.

[24] H. A. Chipman, E. I. George, and R. E. McCulloch, "BART: Bayesian additive regression trees," *Ann. Appl. Stat.*, 2012, doi: 10.1214/09-AOAS285.

[25] J. He, S. Yalov, and P. R. Hahn, "XBART: Accelerated Bayesian additive regression trees," *arXiv*. 2018.

[26] A. Klein, S. Falkner, S. Bartels, P. Hennig, and F. Hutter, "Fast Bayesian optimization of machine learning hyperparameters on large datasets," in *Artificial Intelligence and Statistics*, 2017, pp. 528--536.

[27] W. Li and K. M. Kockelman, "How does machine learning compare to conventional econometrics for transport data sets? A test of ML vs MLE," *Transp. Res. Rec.*, 2020.

[28] B. Misganaw and M. Vidyasagar, "Exploiting Ordinal Class Structure in Multiclass Classification: Application to Ovarian Cancer," *IEEE Life Sci. Lett.*, 2015, doi: 10.1109/LLS.2015.2451291.

[29] X. Wang and K. M. Kockelman, "Occupant injury severity using a heteroscedastic ordered logit model: distinguishing the effects of vehicle weight and type," *Transp. Res. Rec.*, vol. 1908, pp. 195–204, 2005.

[30] E. Frank and M. Hall, "A simple approach to ordinal classification," in *European Conference on Machine Learning*, 2001, pp. 145--156. doi: 10.1007/3-540-44795-4_13.

[31] T. Nashad, S. Yasmin, N. Eluru, J. Lee, and M. A. Abdel-Aty, "Joint Modeling of Pedestrian and Bicycle Crashes: Copula-Based Approach," *Transp. Res. Rec.*, vol. 2601, no. 1, pp. 119–127, Jan. 2016, doi: 10.3141/2601-14.

[32] J. Warsh, L. Rothman, M. Slater, C. Steverango, and A. Howard, "Are school zones effective? An examination of motor vehicle versus child pedestrian crashes near schools," *Inj. Prev. J. Int. Soc. Child Adolesc. Inj. Prev.*, vol. 15, no. 4, pp. 226–229, Aug. 2009, doi: 10.1136/ip.2008.020446.

[33] A. Tharwat, "Classification assessment methods," *Appl. Comput. Inform.*, 2018, doi: 10.1016/j.aci.2018.08.003.

[34] J.-K. Kim, G. F. Ulfarsson, V. N. Shankar, and F. L. Mannering, "A note on modeling pedestrian-injury severity in motor-vehicle crashes with the mixed logit model," *Accid. Anal. Prev.*, vol. 42, no. 6, pp. 1751–1758, Nov. 2010, doi: 10.1016/j.aap.2010.04.016.

[35] M. Pour-Rouholamin and H. Zhou, "Investigating the risk factors associated with pedestrian injury severity in Illinois," *J. Safety Res.*, vol. 57, pp. 9–17, Jun. 2016, doi: 10.1016/j.jsr.2016.03.004.

[36] M. G. Mohamed, N. Saunier, L. F. Miranda-Moreno, and S. V. Ukkusuri, "A clustering regression approach: A comprehensive injury severity analysis of pedestrian–vehicle crashes in New York, US and Montreal, Canada," *Saf. Sci.*, vol. 54, pp. 27–37, Apr. 2013, doi: 10.1016/j.ssci.2012.11.001.

[37] J. M. Wood, P. Lacherez, and R. A. Tyrrell, "Seeing pedestrians at night: effect of driver age and visual abilities," *Ophthalmic Physiol. Opt.*, vol. 34, no. 4, pp. 452–458, 2014.

[38] Z. Chen and W. (David) Fan, "A multinomial logit model of pedestrian-vehicle crash severity in North Carolina," *Int. J. Transp. Sci. Technol.*, vol. 8, no. 1, pp. 43–52, Mar. 2019, doi: 10.1016/j.ijtst.2018.10.001.

[39] A. J. Anarkooli, M. Hosseinpour, and A. Kardar, "Investigation of factors affecting the injury severity of single-vehicle rollover crashes: A random-effects generalized ordered probit model," *Accid. Anal. Prev.*, vol. 106, pp. 399–410, Sep. 2017, doi: 10.1016/j.aap.2017.07.008.

[40] J. Liu, A. Hainen, X. Li, Q. Nie, and S. Nambisan, "Pedestrian injury severity in motor vehicle crashes: An integrated spatio-temporal modeling approach," *Accid. Anal. Prev.*, vol. 132, p. 105272, Nov. 2019, doi: 10.1016/j.aap.2019.105272.

[41] H. M. A. Aziz, S. V. Ukkusuri, and S. Hasan, "Exploring the determinants of pedestrian–vehicle crash severity in New York City," *Accid. Anal. Prev.*, vol. 50, pp. 1298–1309, Jan. 2013, doi: 10.1016/j.aap.2012.09.034.