

Article

Comparative Analysis of Machine Learning and Statistical Models for Railroad–Highway Grade Crossing Safety

Erickson Senkondo ^{1,*}, Deo Chimba ¹ , Masanja Madalo ¹ , Afia Yeboah ¹ and Shala Blue ²

¹ Department of Civil and Architectural Engineering, Tennessee state University, Nashville, TN 37209, USA; dchimba@tnstate.edu (D.C.); mmadalo@my.tnstate.edu (M.M.); ayeboah2@my.tnstate.edu (A.Y.)

² Federal Railroad Administration, Washington, DC 20590, USA; shala.blue@dot.gov

* Correspondence: esenkond@my.tnstate.edu; Tel.: +1-615-668-2126

Abstract

Railroad-highway grade crossings (RHGCs) are critical points of conflict between roadway and rail systems, contributing to over 2000 crashes and 250 fatalities annually in the United States. This study applied machine learning methods (ML) techniques to model and predict crash frequency at RHGCs, using a comprehensive dataset from the Federal Railroad Administration (FRA) and Tennessee Department of Transportation (TDOT). The dataset included 807 validated crossings, incorporating roadway geometry, traffic volumes, rail characteristics, and control features. Five ML models—Random Forest, XGBoost, PSO-Elastic Net, Transformer-CNN, and Autoencoder-MLP—were developed and compared to a traditional Negative Binomial (NB) regression model. Results showed that ML models significantly outperformed the NB model in predictive accuracy, with the Transformer-CNN achieving the lowest Mean Squared Error (21.4) and Mean Absolute Error (3.2). Feature importance analysis using SHAP values consistently identified Annual Average Daily Traffic (AADT), Truck Traffic Percentage, and Number of Lanes as the most influential predictors, findings that were underrepresented or statistically insignificant in the NB model. Notably, the NB model failed to detect the nonlinear relationships and interaction effects that ML algorithms captured effectively. While only three variables were statistically significant in the NB model, ML models revealed a broader spectrum of critical crash determinants, offering deeper interpretability and higher sensitivity. These findings emphasize the superiority of machine learning approaches in modeling RHGC safety and highlight their potential to support data-driven interventions and policy decisions for reducing crash risks at grade crossings.

Keywords: railroad highway grade crossings; machine learning; crash frequency prediction; Transformer-CNN model; feature importance



Academic Editor: Roberto Lot

Received: 30 October 2025

Revised: 1 December 2025

Accepted: 13 December 2025

Published: 17 December 2025

Citation: Senkondo, E.; Chimba, D.; Madalo, M.; Yeboah, A.; Blue, S. Comparative Analysis of Machine Learning and Statistical Models for Railroad–Highway Grade Crossing Safety. *Vehicles* **2025**, *7*, 163. <https://doi.org/10.3390/vehicles7040163>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Railroad-highway grade crossings remain a critical area of concern for traffic safety due to their unique operational characteristics and the high risk of vehicle–train collisions or vehicle-to-vehicle rear-end crashes. In the United States, thousands of crashes occur at these crossings annually, resulting in significant loss of life, injuries, and economic costs. Despite ongoing efforts by federal and state transportation agencies to mitigate this issue, crash frequencies at railroad crossings continue to present a substantial public safety challenge. Statistics show that railroad-highway grade crossing crashes are the second leading cause of rail-related deaths in the United States. Nationally, more than 2000 crashes and 200 fatalities

occur at grade crossings each year [1]. There are approximately 129,500 public at-grade crossings in the United States, with over 50% equipped with automatic warning systems [1]. Statistics also show that about 60% of collisions occur at crossings that are equipped with automatic warning systems [2].

Predicting crashes at railroad-highway grade crossings is a critical step toward enhancing transportation safety and optimizing resource allocation. Reliable crash prediction models help transportation agencies identify high-risk locations, prioritize interventions, and evaluate the potential impact of safety improvements. Accurate forecasting enables proactive decision-making, such as implementing enhanced warning devices, improving signage, or redesigning problematic crossings. Over the years, various modeling techniques have been employed to predict crash frequencies based on roadways, traffic, and environmental characteristics. However, the effectiveness of these models depends on their ability to capture the complex and often nonlinear dynamics present in crash data, making the selection of appropriate methodologies essential for meaningful and actionable insights. The accuracy and reliability of these predictions heavily depend on the robustness and accuracy of the modeling techniques used. Traditional statistical methods often face limitations, such as their inability to handle nonlinear relationships or over-dispersed crash data effectively. These shortcomings can result in biased predictions, inefficient allocation of resources, and reduced effectiveness of safety management strategies.

To address these modeling challenges, machine learning methods techniques offer a promising alternative [3]. Machine learning methods can model complex relationships between variables and account for site-specific factors such as crossing design, traffic mix, and train activity, and provide more reliable and precise crash frequency predictions. In this research, the traditional statistical methodology, specifically Negative Binomial regression was incorporated alongside five machine and deep learning models. These machine learning techniques were applied to identify and interpret the key factors influencing crash frequency at railroad-highway grade crossings (RHGCs). While the Negative Binomial model provides a foundational statistical baseline for detecting linear relationships, it often struggles to capture the complex, nonlinear interactions inherent in real-world crash data. In contrast, machine learning models enable a more flexible, data-driven framework that uncovers deeper patterns and variable dependencies by leveraging feature importance scoring methods and robust validation strategies such as cross-validation [4]. Through integrating traditional and modern approaches, this study offers a comprehensive analytical perspective on the underlying contributors to crash frequency, supporting more informed safety interventions and policy development at RHGCs [5]. This study makes three key contributions to RHGC safety research. First, it develops and evaluates advanced hybrid models, including Transformer CNN, PSO Elastic Net, and Autoencoder MLP, to capture complex nonlinear crash relationships that traditional methods cannot detect. Second, it applies SHAP-based interpretability to quantify the influence of individual roadway, traffic, and train-related features, providing engineering-relevant insight into crash mechanisms [6]. Third, it compares these deep-learning models with a conventional Negative Binomial baseline to demonstrate the added predictive value of hybrid approaches for supporting safety planning and policy development.

2. Literature Review

According to Washington and Nam [7], crashes at railroad-highway grade crossings (RHGCs) represent a significant transportation safety concern due to their severe consequences for both motorists and train operators. To mitigate these incidents, the Federal Railroad Administration (FRA) collaborates with railroads, state and local governments, and various non-governmental organizations. The FRA promotes technologies aimed at

improving crossing safety and provides essential data and resources to the public and stakeholders [2]. Historically, numerous statistical models have explored correlations between crossing features, crash predictions, and crash severity; however, the primary objective of this research is to identify critical factors influencing collision frequency. This study employs analytical methods to assess collision frequency by examining specific geographic locations, typically road segments or intersections, over defined time periods (weeks, months, years). The approach incorporates both spatial and temporal aspects, ensuring sufficient data for robust statistical modeling [2]. Previous studies have demonstrated the effectiveness of machine learning (ML) techniques in analyzing factors contributing to RHGC crashes. For instance, a decision tree-based analysis of North Dakota crash data from 1996 to 2014 found that increased train volume, higher annual average daily traffic (AADT), and elevated train speeds significantly increased crash likelihood, whereas train detection systems and advance warning devices reduced crash risk [7]. Similarly, research comparing ML-based approaches with traditional statistical models showed that ML techniques consistently achieve superior predictive performance and identified roadway and environmental factors as key contributors to crash occurrences [5]. The advantages of ML in traffic safety analysis extend beyond rail-road crossings; several studies have shown that ML models capture complex and nonlinear relationships that conventional regression techniques struggle to represent [8]. For example, an ensemble framework integrating gradient boosting regression trees with LASSO regression achieved higher prediction accuracy than individual models under both normal and abnormal traffic conditions, demonstrating the robustness of ML methods in handling heterogeneous traffic patterns [9].

Similarly, research on COVID-19 outbreak predictions highlighted the superior accuracy of the XGBoost algorithm compared to Random Forest Regression, Support Vector Regression, and Linear Regression. Specifically, in the context of transportation safety, ML methods have effectively predicted crash frequency at RHGCs. One study developed an advanced deep learning model with both unsupervised feature-learning and supervised fine-tuning modules, significantly outperforming traditional Negative Binomial Regression in predicting crash occurrences [9]. Another investigation employed artificial neural networks and support vector machines to classify traffic flow data and identify hazardous conditions, leveraging actual accident and vehicle trajectory data rather than simulations for enhanced realism [10]. Abdel-Aty et al. [11] conducted statistical modeling of crashes at signalized intersections to analyze spatial correlations and identify critical risk factors. Utilizing generalized estimating equations (GEE) instead of standard negative binomial models, the authors accounted for spatial correlation among intersections along various corridors. Their study included 476 signalized intersections along 41 corridors in Orange, Brevard, and Miami counties. The spatially correlated crash frequency data were modeled using GEE models with negative binomial link functions, revealing higher crash frequencies at intersections with multiple lanes, heavy traffic, shorter signal spacing, numerous phases per cycle, and higher speed limits. Conversely, lower crash frequencies were associated with three-legged intersections located in residential areas featuring exclusive right-turn lanes and protected left-turn phases. A 2017 FRA report [12] emphasized that highway-rail grade crossing incidents remain a significant safety issue in the United States, causing extensive harm to motorists and imposing substantial economic costs. Between 2010 and 2014, approximately 2100 accidents occurred annually at these crossings, resulting in over 250 fatalities per year. The FRA maintains a comprehensive database documenting key incident details, such as vehicle types, train speeds, and environmental conditions, which facilitates detailed crash analyses. Additionally, the FRA maintains an inventory of 211,631 operational highway-rail grade crossings, documenting crucial details such as types of warning devices and daily traffic volumes of trains and vehicles.

Integration of these databases enables detailed examination of the factors influencing crash rates, providing a robust dataset for analyzing crash frequency at railroad grade crossings. This extensive dataset is utilized in this study for applying advanced ML techniques to identify significant predictors of crash occurrences.

Recent studies on RHGC safety prediction demonstrate a gradual shift from traditional statistical models toward more flexible machine-learning and deep-learning techniques; however, each modeling family still exhibits limitations that restrict its applicability to RHGC environments. Linear models such as Negative Binomial regression offer interpretability but fail to represent nonlinear or interaction effects. Tree-based methods, including decision trees and Random Forests, identify interactions but lack the depth required for high-dimensional feature learning. Advanced algorithms such as XGBoost, Gradient Boosting, and various hybrid ML frameworks achieve higher predictive accuracy, yet most are tested outside the RHGC domain or rely on limited geographic or modal inputs. Deep neural networks—including DNNs, LSTMs, and CNNs—excel in capturing complex nonlinearities and temporal or spatial structures, but their application to multimodal roadway–rail datasets remains scarce. Even interpretability techniques such as SHAP have been applied with restricted model diversity. Overall, prior work reveals strong methodological advances but a persistent gap in applying integrated, multimodal, and hybrid deep-learning approaches specifically tailored to RHGC crash prediction. Despite substantial efforts to analyze safety at railroad highway grade crossings, prior studies have largely relied on linear, statistical, or tree-based modeling techniques that struggle to capture the nonlinear interactions and heterogeneous roadway railway characteristics that influence crash occurrences. Many earlier works use limited predictor sets, focus on narrow geographic contexts, or lack interpretability frameworks capable of explaining how roadway, traffic, and train-related factors jointly shape crash risk. Additionally, hybrid deep-learning architectures and optimization-assisted modeling strategies have not been fully explored in this domain, leaving important gaps in understanding complex crash-related patterns. To address these limitations, the present study evaluates a wide range of machine-learning and deep-learning models including Random Forest, XGBoost, Transformer-assisted CNN, PSO-Elastic Net, and Autoencoder-MLP to more effectively represent nonlinear crash relationships and identify the most influential predictors at RHGCs [13]. Leveraging a multi-source dataset from the FRA and TDOT, this study aims to produce more accurate crash-frequency predictions and provide insights that support improved safety strategies and policy decisions at grade crossings.

3. Methodology

To provide greater clarity on the modeling framework, this study includes a brief explanation of how each machine-learning and deep-learning model functions. Random Forest and XGBoost use ensembles of decision trees to learn nonlinear crash patterns by aggregating multiple tree-based predictions [14]. The PSO Elastic Net hybrid integrates a metaheuristic optimizer with a regularized linear model to enhance coefficient stability when predictors are correlated. The Autoencoder–MLP hybrid reduces feature dimensionality before classification, allowing deeper pattern extraction in the crash data. The Transformer CNN model combines convolutional layers for local feature learning with Transformer encoders that capture longer-range dependencies among variables [15]. Together, these models address both linear and complex nonlinear relationships influencing RHGC crash frequency.

3.1. Data Gathering and Preprocessing

The dataset used in this study was obtained from FRA crash and inventory records and the Enhanced Tennessee Roadway Information Management System (eTRIMS), covering a 13-year period from 2010 to 2023. Table 1 shows the summary of continuous variables. As shown in Table 1 and subsequent figures, various variables were used during model training including the percentage of trucks in traffic, AADT that denotes the Average Annual Daily Traffic, the absence of illumination at the grade crossing, rolling or mountainous terrains, the number of roadway lanes, non-asphalt surface pavement, train speed, posted speed limit greater than 35, and the daily train volume at the crossing. The combined dataset included crash history, roadway geometry, traffic characteristics, warning device information, and additional inventory attributes relevant to RHGC safety analysis. The initial crash dataset contained 5401 crossing sections, which was refined to 929 unique crossings with recorded crashes. Further filtering removed crossings with zero AADT, private crossings, and non-grade crossings, yielding a final dataset of 807 crossings. These records were merged using the FRA crossing number as the unique identifier. During preprocessing, incomplete or inconsistent entries were removed while maintaining the overall distribution of key variables. Numerical features were normalized using MinMax scaling to support efficient learning, and categorical variables were one-hot encoded to ensure interpretability across models [16]. The processed dataset was then split into training and testing subsets, resulting in a clean and consistent foundation for the machine-learning and deep-learning crash-frequency analysis.

Table 1. Summary of modeling Variables.

Variable	Mean	Min	Max
Number of Crashes	1.84	0	64
AADT	4875	70	30,559
Percentage of Truck in Traffic	3.46	0	24
Lanes	2.38	1	6
Number of single-unit trucks in traffic	3.44	0	24
Trains Per Day	5.99	1	78

3.2. Crash Modeling

This study employs a dual approach to crash modeling by combining traditional statistical techniques with modern data-driven methods to examine factors influencing crash frequency at railroad-highway grade crossings (RHGCs). The inclusion of the Negative Binomial regression model establishes a statistical baseline [17], while advanced models enable the exploration of complex relationships among variables. This integrated framework enhances both the interpretability and depth of the analysis, supporting a more robust understanding of crash determinants.

3.3. Statistical Approach (Negative Binomial Regression)

The Negative Binomial Regression model was initially applied to identify the relationship between key variables, such as traffic volume, crossing type, and vehicle classification, with crash frequency. While effective for count data with overdispersion, The Negative Binomial Regression model serves as a baseline for comparison with more machine learning methods and optimization algorithms. While traditional models like Negative Binomial provide foundational insights, they often fall short in addressing complex patterns and interactions present in the data. Therefore, advanced techniques, such as Random Forest, XG-Boost, and hybrid machine learning models, are employed to explore these complexities more accurately. The negative binomial regression is a commonly applied alternative

statistical model to deal with over dispersed data. The negative binomial model takes the relationship between the expected number of accidents occurring at the i -th element and the M parameters, it introduces a parameter, usually called α to relax the strict mean-variance equality assumption in this model, [18,19] the variance is modeled as:

$$\text{var}(y) = \mu + \alpha\mu^2 \quad (1)$$

where $\text{var}(y)$ = Variance of crash count (y), μ = Mean crash frequency and α = Dispersion parameter (controls overdispersion; when $\alpha = 0$, reduces to Poisson)

$$\mu = E(y) = \exp(X\beta) \quad (2)$$

$$\text{Mean of Crashes } \mu = E(y) = e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_n x_n)} \quad (3)$$

where μ = Expected value (mean) of crash count, $E(y)$ = Expected number of crashes

X = Matrix of predictor variables (AADT, truck percentage, lanes, etc.), β = Vector of regression coefficients

Probability Mass Function (PMF):

$$p(y) = \frac{(y + \alpha^{-1})}{(\alpha^{-1})!} \left(\frac{1}{1 + \alpha\mu} \right)^{1/\alpha} \left(\frac{\alpha\mu^1}{1 + \alpha\mu} \right)^y \quad (4)$$

Log-likelihood

$$\log L = \sum_{i=1}^n \left[\ln(y_i + r) - \ln y_i! - \ln(r) + y_i \ln \left(\frac{\mu_i}{\mu_i + r} \right) + r \ln \left(\frac{r}{\mu_i + r} \right) \right] \quad (5)$$

where $\log L$ = Log-likelihood function (to be maximized), n = Total number of observations (crossings), y_i = Observed crash count at crossing i , r = Alternative notation for $1/\alpha$ (inverse dispersion parameter), μ_i = Predicted mean crash count at crossing i , Σ = Summation operator

To obtain α and β for crash Y prediction involve Iteratively updating β coefficients and α until the log-likelihood converges from:

$$\ln(\beta, \alpha) = \sum_{i=1}^n \left[\ln(y_i + r) - \ln y_i! - \ln(r) + y_i \ln \left(\frac{\mu_i}{\mu_i + r} \right) + r \ln \left(\frac{r}{\mu_i + r} \right) \right] \quad (6)$$

3.4. Machine Learning Models

The machine learning portion of this study employed a diverse set of advanced models tailored for predictive accuracy and the capacity to uncover nonlinear relationships inherent in transportation crash data. These models were selected to address the limitations of traditional statistical techniques and to enhance the model's ability to generalize across varied roadway and operational contexts. The ensemble models Random Forest and XGBoost were included due to their robustness and interpretability in high-dimensional spaces. Additionally, PSO-Elastic Net was applied for its unique integration of Particle Swarm Optimization with Elastic Net regularization to enhance feature selection. For modeling complex temporal and spatial dependencies, a hybrid Transformer-CNN architecture was utilized, while Autoencoder-MLP was employed for its strength in unsupervised feature learning followed by multilayer perceptron-based prediction. The following subsections describe the structure, functions, and formulation of each machine learning model used in this study. Random Forest is an ensemble learning method that aggregates predictions

from multiple decision trees to improve accuracy and reduce overfitting, making it effective for handling complex datasets [20–22].

$$\hat{y}_{tree} = \frac{1}{n_{leaf}} \sum_{i=1}^{n_{leaf}} y_i \quad (7)$$

where $\hat{y}_{tree}$ = Predicted crash count from a single decision tree, n_{leaf} = Number of observations in the terminal leaf node, y_i = Crash count of observation i in the leaf.

XG-Boost is a powerful gradient boosting algorithm known for its speed and efficiency. It builds an ensemble of trees sequentially, each one focusing on correcting the errors of the previous one, making it particularly effective for regression tasks with complex relationships [23]. With loss function.

$$\hat{L}_{(\theta)} = \sum_{i=1}^{n_1} l(y_i, \hat{y}_i) + \sum_{k=1}^{K_1} \varphi(f_k) \quad (8)$$

where $\hat{L}_{(\theta)}$ = Total loss function to minimize, θ = Model parameters, n = Number of training samples, $l(y_i, \hat{y}_i)$ = Loss function comparing actual (y_i) vs. predicted ($\hat{y}_i$) crashes, K = Number of trees in ensemble, $\varphi(f_k)$ = Regularization term for tree k (penalizes model complexity), f_k = k -th tree function.

PSO-Elastic Net is a hybrid model that optimizes hyperparameters using Particle Swarm Optimization and applies Elastic Net regression for feature selection and regularization, helping to manage high-dimensional data. This hybrid model combines the long-range dependency learning of Transformers with the feature extraction power of CNNs, making it effective for handling complex temporal and spatial patterns in the data [24].

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) V \quad (9)$$

where Q = Query matrix (what information to look for), K = Key matrix (what information is available), V = Value matrix (actual feature values), K^T = Transpose of key matrix, d_k = Dimension of key vectors (scaling factor), SoftMax activation function (normalizes attention weights).

The Autoencoder-MLP model uses Autoencoders to extract features and MLPs for predictions, capturing nonlinear relationships and improving feature learning for complex data.

$$z_j = \sum_{i=1}^n w_{ji} x_i + b_j \quad (10)$$

where z_j = j -th encoded feature in latent space, n = Number of input features, w_{ji} = Weight connecting input feature i to encoded feature j , x_i = i -th input feature (e.g., AADT, truck %, lanes).

b_j = Bias term for encoded feature j the autoencoder compresses input features into a lower-dimensional representation.

Unlike earlier crash-modeling studies that rely on single-model deep-learning architectures or conventional machine-learning techniques, the Transformer CNN framework used in this study incorporates the Transformer as an attention-weighting and optimization layer rather than as a temporal encoder. This structure enhances the CNN's ability to prioritize informative roadway, traffic, and train-related features while reducing noise, making it more suitable for RHGC crash prediction. Similarly, the PSO Elastic Net approach differs from standard Elastic Net applications by using particle swarm optimization to identify optimal regularization parameters across heterogeneous RHGC variables, improving

feature-selection stability [25,26]. These two hybrid frameworks have not been previously applied to RHGC crash modeling, highlighting their novelty and suitability for capturing complex, nonlinear crash relationships.

3.5. Model Development

The data used in this analysis includes a variety of features, such as traffic volume, road configuration, crossing type, and other site-specific variables. These features were preprocessed, normalized, and split into training and testing sets using an 80/20 ratio, which is appropriate given the dataset size of more than 800 crossings. Although cross-validation is commonly applied in machine-learning studies, it was not implemented here due to the high computational cost associated with training hybrid deep-learning models such as the Transformer-assisted CNN and Autoencoder-MLP. Instead, model robustness was supported through consistent performance patterns across multiple model types and SHAP-based sensitivity analysis. The models employed in this study include Random Forest, XG-Boost, PSO-Elastic Net Hybrid, Transformer-CNN, and Autoencoder-MLP, all developed using Python libraries such as Scikit-learn and PyTorch version 2.9.1 [27]. To ensure fairness and comparability across models, all machine-learning and deep-learning frameworks were trained using consistent optimization procedures. Traditional models such as Random Forest and XGBoost were tuned using grid search with cross-validation to identify optimal combinations of tree depth, number of estimators, learning rate, and regularization strength. Deep-learning models (Autoencoder-MLP and Transformer CNN) were trained using the Adam optimizer, ReLU activations, mini-batch training, and early stopping to prevent overfitting. Batch sizes, learning rates, and training epochs were iteratively adjusted based on validation performance. These settings ensure that each model was trained in stable and reproducible conditions. Table 2 summarizes the key characteristics of each model, including algorithmic foundation, regularization strategy, and core structure.

Table 2. Summary of Model Structures and Development Techniques.

Model	Framework	Architecture	Hyperparameters	Regularization/Techniques
Random Forest (RF)	Scikit-learn	Ensemble of Decision Trees	n_estimators = 100, max_depth = 10	Feature Bagging, Grid Search
XG-Boost	XG-Boost	Gradient Boosted Trees	n_estimators = 200, learning_rate = 0.05, max_depth = 6	Regularization (L1 and L2), Grid Search
PSO + Elastic Net	Scikit-learn	Elastic Net Regression	alpha = 0.01, l1_ratio = 0.5	Particle Swarm Optimization (PSO), Elastic Net Regularization
Transformers + CNN	PyTorch	Transformer + Convolutional Neural Network	n_heads = 8, num_layers = 3, cnn_filters = 32	Early Stopping, Dropout
Autoencoder + MLP	PyTorch	Autoencoder + Multi-layer Perceptron	hidden_layers = [128,64], dropout = 0.3, learning_rate = 0.001	Early Stopping, Dropout

Figure 1 illustrates the overall workflow adopted for modeling crash frequency using machine learning algorithms. The process begins with data acquisition from the Enhanced Tennessee Roadway Information Management System, followed by data cleaning, feature selection, and preprocessing steps such as normalization and train-test splitting. Subsequently, various machine learning algorithms including Random Forest, XGBoost, PSO-Elastic Net, Transformer-CNN, and Autoencoder-MLP are trained on the prepared dataset. The models are evaluated using standard performance metrics, and feature importance is extracted to interpret key predictors. This structured approach ensures consistency, repro-

ducibility, and accuracy in identifying the critical factors contributing to crash frequency at railroad-highway grade crossings. Overall, as shown in Figure 1, the process begins with importing RHGC crash datasets followed by preprocessing steps such as cleaning missing values, removing duplicates, and encoding categorical variables. Feature selection using an Extra Trees Classifier identified the most relevant predictors for modeling. The dataset was then processed through handling class imbalance using Imbalanced-Learn techniques and split into training and testing subsets. Model development involved training regression and deep-learning models with hyperparameter tuning to optimize performance. The trained models were evaluated using multiple fitness metrics, and the final outputs included variable importance rankings, stored models, and graphical visualization of results [28–31].

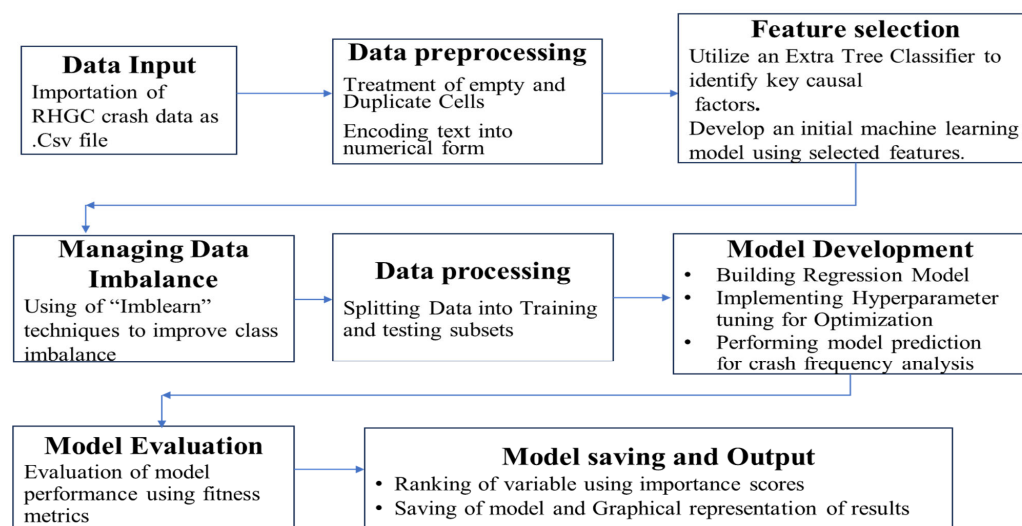


Figure 1. Workflow in Machine Learning.

4. Results and Discussion

Evaluations of various machine learning models used to analyze factors influencing crash frequency at railroad-highway grade crossings were performed. Model performance is assessed using metrics like Mean Squared Error (MSE) and Mean Absolute Error (MAE). Additionally, key variables contributing to crash frequency are identified and discussed through feature importance analysis.

4.1. Model Performance Evaluation

Among the models evaluated (Table 3), Transformer-CNN, PSO-Elastic Net, and Random Forest emerged as the top performers in predicting crash frequency at railroad-highway grade crossings. The Transformer-CNN hybrid model performed better than others due to its ability to capture both temporal and spatial patterns in the data. The Transformer component allows the model to learn long-range dependencies in the data, which is critical for understanding complex relationships. Meanwhile, the CNN component helps in extracting key features from the data, which further enhances the model's ability to identify patterns and improve generalization. This combination between Transformer and CNN allowed the model to achieve the lowest error metrics. The PSO-Elastic Net hybrid model also performed well by effectively minimizing large prediction discrepancies. PSO was used to optimize the hyperparameters, including the regularization strength of Elastic Net, which combines both Lasso (L1) and Ridge (L2) regularization. This combination allows the model to perform both feature selection and shrinkage, providing better generalization and reducing overfitting, especially when handling noisy or high-dimensional data. This made PSO-Elastic Net a reliable model for managing complex data while mini-

mizing large errors. The Random Forest model, while reliable, was slightly less efficient in minimizing prediction errors compared to the other two. On the other hand, XG-Boost and Autoencoder-MLP models, though competitive, demonstrated higher error rates, with XG-Boost being more prone to large discrepancies and Autoencoder-MLP showing challenges in feature extraction and generalization. Overall, the top three models proved to be superior in predicting crash frequency at grade crossings, offering a combination of accuracy, robustness, and reliability.

Table 3. Machine Learning Model Performance.

Model	Mean Squared Error (MSE)	Mean Absolute Error (MAE)
Random Forest	41.8	3.5
XG-Boost	57.2	3.5
PSO + Elastic Net	27.0	3.7
Autoencoder and MLP	35.2	4.4
Transformers + CNN	21.4	3.2

4.2. Feature Importance Analysis for Machine Learning

Feature importance analysis plays a pivotal role in identifying and interpreting the variables that most significantly influence crash frequency at railroad-highway grade crossings (RHGCs). This analysis helps not only with model transparency but also in uncovering actionable insights for safety interventions. The importance score, derived from machine learning algorithms, ranks input features based on their contribution to model prediction accuracy. In this study, SHAP (SHapley Additive exPlanations) was employed for both global and instance-level interpretability. SHAP values assign consistent, quantitative contributions to each feature for every prediction, making it an effective tool for understanding complex, nonlinear model behaviors. The analysis revealed that Percentage of Truck Traffic, Annual Average Daily Traffic (AADT), and Number of Lanes are the top predictors of crash frequency across multiple models. Figures 2 and 3 illustrate the relative feature importance as determined by the PSO-Elastic Net hybrid model and the Random Forest model, respectively. In both models, Percentage of Truck Traffic and AADT consistently emerged as dominant contributors. This result aligns well with engineering expectations: a higher percentage of trucks increases crash risk due to longer stopping distances, slower acceleration, and larger turning radii, all of which complicate maneuvering through crossings [32]. Likewise, higher AADT elevates crash exposure by increasing the number of potential vehicle-train conflict points, especially at crossings that lack gates or active warning systems.

The SHAP summary plot in Figure 4 visually reinforces this pattern, showing a strong positive relationship between AADT and predicted crash frequency. Higher AADT values (shown in red) correspond to larger SHAP contributions, indicating that increased traffic volumes directly elevate crash risk. This heightened exposure underscores the need to prioritize high-volume crossings for enhanced protection, such as active warning devices, four-quadrant gates, or predictive signal preemption. Beyond traffic volume, Truck Percentage and Train Speed also demonstrate notable SHAP impacts. Crossings with a higher share of trucks exhibit increased crash likelihood due to longer stopping distances, slower acceleration, and greater queue buildup near tracks, suggesting the need for improved signage visibility, extended clearance intervals, or truck-specific signal timing strategies. Train speed shows a nonlinear influence: moderate speeds intensify risk due to gap misjudgment by drivers, while very high speeds often correspond to crossings equipped with advanced warning systems that mitigate risk. These SHAP-derived insights not only confirm that exposure-related and operational variables are the

strongest determinants of crash frequency at RHGCs but also provide direct guidance for safety countermeasures, including speed harmonization, enhanced sight distance, and targeted warning device upgrades. Overall, the feature-importance patterns help agencies allocate resources more efficiently by identifying high-risk crossings and matching them with the most appropriate engineering interventions. Figure 4. SHAP dependence plot illustrating the marginal effect of AADT on predicted crash frequency. Each point represents a roadway–rail crossing, with the x-axis showing the normalized AADT class range and the y-axis representing the corresponding SHAP value for AADT. Positive SHAP values indicate that higher AADT contributes to an increase in predicted crash frequency, while negative values indicate a decreasing effect. The red trend line and shaded confidence band highlight the overall positive relationship, confirming that crossings with higher traffic volumes are associated with elevated crash risk.

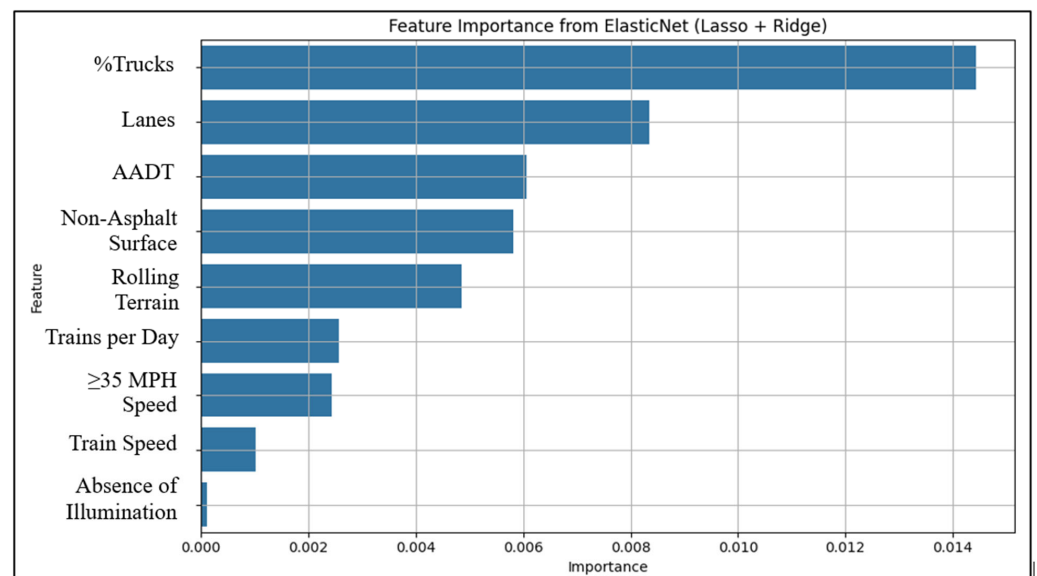


Figure 2. Important Features Ranked in Descending Scores for PSO-Elastic Net Hybrid model.

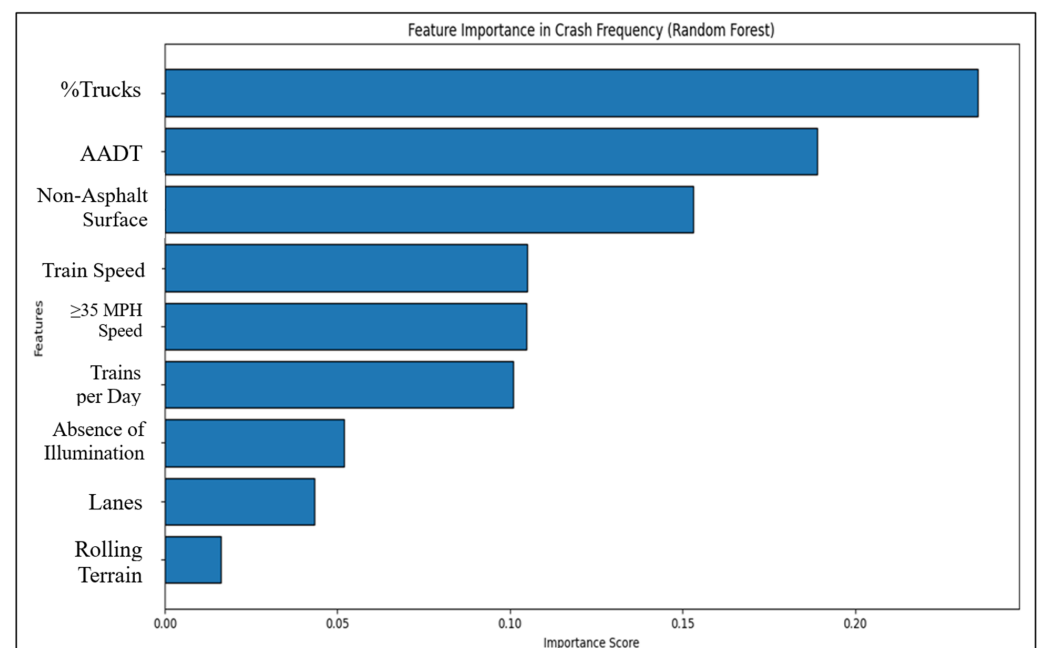


Figure 3. Various Important Features Ranked in Descending Scores for RF Model.

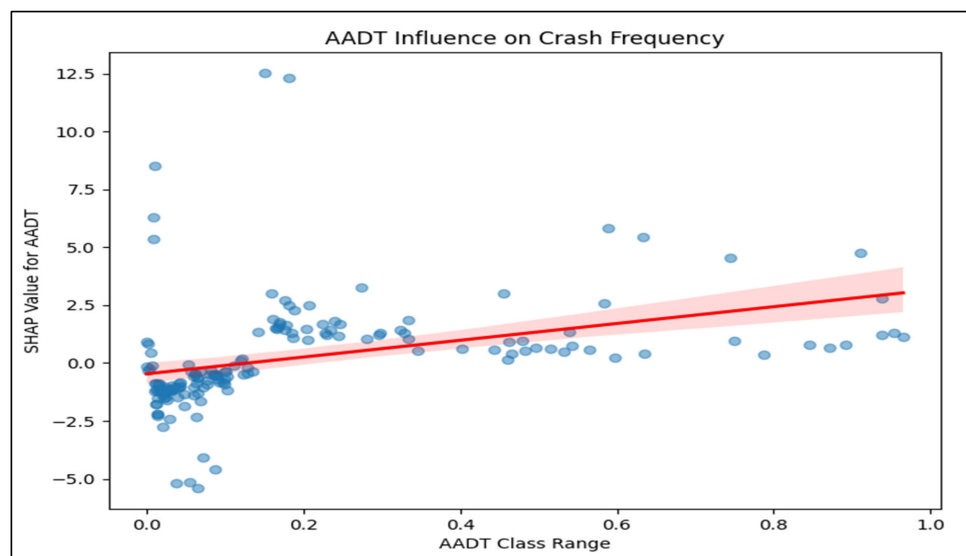


Figure 4. SHAP Graphical Representation of Influence of AADT on Crash Frequency.

4.3. Negative Binomial Regression Results and Comparison with Machine Learning Insights

The Negative Binomial (NB) regression model was employed as a foundational statistical approach to explore relationships between key roadway and operational variables and crash frequency at railroad-highway grade crossings (RHGCs), with results shown in Table 4. While NB modeling provides interpretable coefficients and significance tests, its capacity to capture complex, nonlinear interactions is limited. Therefore, findings from NB were compared against machine learning (ML) model outcomes, which are assumed to offer more accurate and comprehensive predictive performance. In the NB model The Negative Binomial model results were expanded to include coefficient estimates, p-values, z-statistics, and 95% confidence intervals to provide a more complete statistical interpretation [33]. Several variables, including AADT, truck percentage, and train speed, were found to be statistically significant predictors of crash frequency, with positive coefficients indicating increased crash likelihood as these variables rise. The dispersion parameter confirmed overdispersion in the crash data, validating the use of the NB model over Poisson regression. While the NB model captured general exposure-related trends, its linear functional form limited its ability to represent nonlinear and interaction effects that were more effectively learned by the machine-learning and deep-learning models. In Table 4, AADT had a positive coefficient with marginal statistical significance. This result aligns with the machine learning models, where AADT emerged as one of the most influential predictors across all algorithms, as higher traffic volumes inherently increase exposure to train-vehicle conflict points. The consistency of AADT as a top contributor in both models reinforces its critical role in crash risk assessment at RHGCs. Percentage of Truck Traffic also had a positive coefficient, suggesting that greater truck presence increases crash frequency, likely due to longer braking distances and turning limitations. However, with a z-value of 1.23, the variable is not statistically significant in the NB model. The SHAP results provide engineering insight into how major variables shape RHGC crash predictions. AADT remains the strongest contributor, confirming that higher traffic exposure elevates crash likelihood. Truck percentage also shows substantial influence due to the limited maneuverability and longer stopping distances of heavy vehicles. These findings contrast with the Negative Binomial model, which did not detect truck percentage as significant, highlighting NB's limitations in capturing nonlinear effects and variable interactions that ML models learn more effectively. Train speed exhibits a nonlinear pattern in the SHAP

analysis, with moderate speeds increasing risk through reduced driver reaction time, while very high speeds often correspond to crossings equipped with enhanced warning devices. This differs from the NB model, where Train Speed > 45 mph showed a significant negative coefficient, likely to reflect operational design differences rather than true safety effects. Several categorical variables, such as absence of warning signs and crossing surface type, appeared strong in the NB model but were less dominant in ML models, which prioritize continuous variables with larger marginal contributions. Likewise, the NB model showed insignificant effects for number of lanes and highway speed, whereas SHAP values suggested moderate relevance due to their influence on driver behavior, maneuver complexity, and speed profiles. It can be observed that, while the NB model provided valuable initial insights, its limited ability to model complex relationships and interactions resulted in several key predictors being statistically insignificant despite their clear importance in ML models. Machine learning methods, especially those incorporating ensemble and deep learning approaches, consistently identified AADT, Truck Percentage, Train Speed, and Number of Lanes as major contributors. The contrast underscores the greater robustness of ML models in identifying critical predictors of crash frequency at RHGCs and supports their use in guiding data-driven safety interventions and policy development [34]. As shown in Table 4, analysis also determined sensitivity of the variables in terms of marginal effects. It shows what the possible change in crash frequency at the crossings is if the variable is changed by one unit while others are kept at the mean. Some of the strong positive marginal effects are seen on the Absence of Warning Signs ($dy/dx = 3.17$), and higher truck percentages ($dy/dx = 0.64$) are consistent with expected safety risks, as inadequate warning devices and heavy truck presence are well-known contributors to increased crash exposure.

Table 4. Negative Binomial Regression Results.

Variable	Coefficient	Std. Error	Z-Value	p-Value	Marginal Effect (dy/dx)
AADT	5.7×10^{-5}	3.1×10^{-5}	1.83	0.07	0.00
Lanes	0.18	0.24	0.90	0.37	0.40
Percentage of Trucks in Traffic	0.16	0.13	1.23	0.22	0.64
Absence of Pavement Markings	−0.13	0.36	−0.37	0.71	−1.71
Absence of Warning Signs	1.08	0.32	3.37	0.00	3.17
Terrain: Rolling and Hills	0.27	0.53	0.66	0.51	2.37
Train Speed $\geq 30 \leq 45$ mph	0.66	1.02	1.24	0.22	5.49
Train Speed > 45 mph	−2.52	0.29	−6.98	0.00	−2.42
Urban Areas	−0.50	0.24	−1.24	0.21	−0.85
Government Control (Municipal and State)	0.25	0.27	1.15	0.25	0.93
Crossing Surface (Non-Asphalt)	−1.05	1.00	−3.66	0.00	−2.94
Highway Speed ≥ 35 mph	−0.25	0.28	−0.91	0.36	−0.94
Constant	−2.98	0.90	−3.32	0.00	0.00

The NB model fitness is as shown in Table 5. The log-likelihood of −1821.42 and AIC of 3648.83 reflect the overall adequacy of the model, while the LR chi-square value of 133.39 ($p < 0.001$) confirms that the included predictors collectively improve model performance relative to a null model. The dispersion parameter ($\alpha = 9.77$), along with the highly significant likelihood-ratio test of $\alpha = 0$ ($\chi^2 = 1.2 \times 10^5$, $p < 0.001$), demonstrates substantial overdispersion in the data and strongly justifies the use of a Negative Binomial specification instead of a Poisson model.

Table 5. Goodness-of-Fit Measures for NB Crash Frequency.

Model Fit Statistics	Value
Log-Likelihood	−1821
LR Chi-Square	133
Prob > Chi ²	0.000
Pseudo R ²	0.0353
AIC	3648.83
Dispersion Parameter (α)	9.77
Likelihood-Ratio Test of $\alpha = 0$	$\chi^2 = 1.2 \times 10^5, p = 0.000$

5. Conclusions

This study presented a comprehensive and data-driven evaluation of crash frequency at railroad-highway grade crossings (RHGCs), leveraging machine learning methods and comparing them with traditional statistical methods. The goal was to identify critical factors influencing crash frequency and develop more accurate and interpretable prediction models to support proactive safety management strategies. The machine learning framework, which included Random Forest, XGBoost, PSO-Elastic Net, Transformer-CNN, and Autoencoder-MLP models, outperformed the traditional Negative Binomial (NB) regression in both prediction accuracy and depth of insights. Among these, the Transformer-CNN hybrid model demonstrated superior performance due to its ability to capture complex spatial and temporal dependencies in the data. PSO-Elastic Net also performed robustly by integrating optimization and regularization strategies for enhanced generalization. Random Forest offered high interpretability, while XGBoost and Autoencoder-MLP provided competitive yet less stable performance. Feature importance analysis, supported by SHAP (SHapley Additive exPlanations), consistently identified Annual Average Daily Traffic (AADT), Percentage of Truck Traffic, Number of Lanes, and Train Speed as the most influential variables contributing to crash frequency. These findings align with engineering judgment and offer actionable insights: higher traffic volumes and truck percentages elevate the probability of crashes due to increased exposure and maneuvering challenges, while higher train and highway speeds amplify risk due to reduced driver reaction time and increased severity of collisions. In contrast, the NB model identified only a few statistically significant variables, namely, absence of warning signs, train speed over 45 mph, and crossing surface type, highlighting its limited ability to capture nonlinear and interaction effects. Variables like AADT and truck percentage, while highly influential in ML models, were underrepresented in the NB model, demonstrating the limitations of traditional statistical approaches in handling high-dimensional, complex transportation data.

This study therefore confirms the superiority of machine learning approaches in modeling RHGC crash frequency, not only in predictive performance but also in their ability to uncover nuanced relationships between variables. Through the provision of deeper insights into the underlying contributors to crashes, ML models were found to be able to support more targeted interventions, such as prioritizing high-risk crossings for upgrades, implementing dynamic warning systems, or enforcing truck-specific policies. It is important to clarify that the machine-learning and deep-learning models developed in this study were not designed for real-time operational deployment at railroad highway grade crossings. Instead, their primary purpose is to support crash-risk prediction, long-term safety analysis, and the identification of key predictors that influence crash occurrences. Given the computational requirements of hybrid deep-learning architectures such as the Transformer CNN and Autoencoder MLP, real-time implementation was not a design objective. Rather, these models serve as analytical tools that can guide safety planning, infrastructure prioritization, and policy development.

This study has several limitations that should be acknowledged. Although the dataset covered a 13-year period and included diverse roadway, traffic, and train-related attributes, some relevant contextual factors such as driver behavior, land-use characteristics, and real-time train operations were not available. In particular, certain safety treatments associated with higher-speed train corridors (e.g., constant-warning-time devices, active gates, and enhanced signal systems) were not fully captured in the dataset, which may influence the interpretation of the nonlinear train-speed effects observed in the SHAP analysis. Additionally, the hybrid deep-learning models are computationally intensive and were trained on tabular data without explicit spatial or temporal dependencies, which may limit their ability to capture cluster effects or year-to-year variability. Finally, model comparison did not include formal significance testing, which may affect the interpretation of performance differences across models. Future research can build on this work by integrating spatial and temporal modeling frameworks, such as graph-based neural networks or recurrent deep-learning architectures, to better reflect clustering and dynamic crash patterns. Incorporating supplementary data sources such as driver behavior indicators, land-use context, or near-miss events derived from sensor and video analytics could further enhance predictive capability. Upcoming studies will also include statistical significance testing and explore model compression techniques to evaluate the feasibility of lightweight or real-time prediction models for operational use at grade crossings.

Author Contributions: The research was conceptualized and led by E.S.; Data acquisition, preprocessing, and modeling were conducted by E.S. and M.M.; D.C. and A.Y. contributed to the technical oversight and methodological guidance throughout the project; S.B. served as the project coordinator, ensuring alignment with project goals and facilitating communication with the funding agency; The manuscript was drafted and revised by E.S., with critical review and feedback from all co-authors. All authors have read and agreed to the published version of the manuscript.

Funding: This research was sponsored by the Federal Railroad Administration (FRA), Contract Number 693JJ622C000035.

Institutional Review Board Statement: The study did not require ethical approval as study does not involve humans or animals. It is just publicly available data.

Data Availability Statement: The data used in this study were sourced from two primary platforms. The Federal Railroad Administration (FRA) crash and inventory datasets are publicly available and can be accessed through the FRA's official website. Additional roadway and intersection data were obtained through the state of Tennessee database which is managed by the Tennessee Department of Transportation (TDOT).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Federal Railroad Administration (FRA). *Highway–Rail Grade Crossing Safety*; U.S. Department of Transportation: Washington, DC, USA, 2025. Available online: <https://railroads.dot.gov/railroad-safety/divisions/crossing-safety-and-trespass-prevention/railroad-crossing-safety> (accessed on 15 August 2025).
2. National Highway Traffic Safety Administration (NHTSA). *Early Estimate of Traffic Fatalities*; NHTSA: Washington, DC, USA, 2021.
3. Amin, A.; Chimba, D.; Hasan, K.; Emmanuel, S. Intelligent Railroad Grade Crossing: Leveraging semantic segmentation and object detection for enhanced safety. *arXiv* **2024**, arXiv:2403.11060. [CrossRef]
4. Banyong, C.; Hantanong, N.; Wisutwattanasak, P.; Champahom, T.; Theerathitichaipa, K.; Ratanavaraha, V.; Seefong, M.; Jomnonkwao, S. A machine learning comparison of transportation mode changes from high-speed railway promotion in Thailand. *Results Eng.* **2024**, *24*, 103110. [CrossRef]
5. Silva, B.; Andrade, M.; Ferreira, S. Machine learning applied to road safety modeling: A systematic literature review. *J. Traffic Transp. Eng.* **2020**, *7*, 775–790. [CrossRef]
6. Zhang, L.; Jiang, L. Game-theoretic SHAP-driven interpretable forecasting of air cargo demand using Bayesian-optimized random forests. *Front. Phys.* **2025**, *13*, 1705687. [CrossRef]

7. Oh, J.; Washington, S.P.; Nam, D. Accident prediction model for railway–highway interfaces. *Accid. Anal. Prev.* **2006**, *38*, 346–356. [\[CrossRef\]](#)
8. Ardabili, F.; Mosavi, A.; Ghamisi, P.; Ferdinand, F.; Varkonyi-Koczy, A.R.; Reuter, U.; Rabczuk, T.; Atkinson, P.M. COVID-19 outbreak prediction with machine learning. *Algorithms* **2020**, *13*, 249. [\[CrossRef\]](#)
9. Dong, C.; Shao, C.; Li, J.; Xiong, Z. An improved deep learning model for traffic crash prediction. *J. Adv. Transp.* **2018**, *2018*, 3869106. [\[CrossRef\]](#)
10. Chong, M.; Abraham, A.; Paprzycki, M. Traffic accident analysis using machine learning paradigms. *Informatica* **2005**, *29*, 89–98.
11. Mohamed, A.; Wang, X. Crash estimation at signalized intersections along corridors: Analyzing spatial effects. *Transp. Res. Rec.* **2006**, *1953*, 98–111.
12. Brabb, C.; Vithani, A.; Martin, K. *In-Depth Data Analysis of Grade Crossing Accidents Resulting in Injuries and Fatalities*; U.S. Department of Transportation: Washington, DC, USA, 2017.
13. Chengula, J.; Kutela, B.; Novat, N.; Shita, H.; Kinerio, A.; Tamakloe, R.; Kasomi, S. Spatial instability of crash prediction models: A case of scooter crashes. *Mach. Learn. Appl.* **2024**, *17*, 100574. [\[CrossRef\]](#)
14. Ding, H.; Wang, R.; Chen, T.; Chung, H.; Dong, N.; Sze, N.N. A hybrid approach for modeling bicycle crash frequencies using SHAP and random-parameter NB models. *Accid. Anal. Prev.* **2024**, *208*, 107778. [\[CrossRef\]](#)
15. Long, H. Hybrid design of CNN and vision transformer: A review. In Proceedings of the 2024 7th International Conference on Computer Information Science and Artificial Intelligence, Shaoxing, China, 13–15 September 2024; pp. 121–127. [\[CrossRef\]](#)
16. Koramati, S.; Majumdar, B.; Sahu, P. Investigation of risk factors for midblock road traffic crashes using negative binomial model. In *Lecture Notes in Civil Engineering*; Springer: Berlin/Heidelberg, Germany, 2025; p. 117. [\[CrossRef\]](#)
17. Washington, S.; Karlaftis, M.; Mannering, F. *Statistical and Econometric Methods for Transportation Data Analysis*; CRC Press: Boca Raton, FL, USA, 2003. [\[CrossRef\]](#)
18. Chimba, D.; Sando, T.; Kwigizile, V.; Kutela, B. Modeling School Bus Crashes Using Zero Inflated Model. *J. Transp. Stat.* **2014**, *10*, N1.
19. Zheng, Z.; Lu, P.; Tolliver, D. Decision tree approach to accident prediction for highway–rail grade crossings. *Transp. Res. Rec.* **2016**, *2545*, 115–122. [\[CrossRef\]](#)
20. Wang, Y.; Zhai, H.; Cao, X.; Geng, X. Cause analysis and accident classification using complex networks. *Appl. Sci.* **2023**, *13*, 23.
21. Chen, Y.; Chen, H.; Ye, P.; Lv, Y. Acting as a decision maker: Traffic-condition-aware ensemble learning for traffic flow prediction. *IEEE Trans. Intell. Transp. Syst.* **2020**, *23*, 3190–3200. [\[CrossRef\]](#)
22. Ma, R.; Chen, R.; Liang, B.; Li, X. An XGBoost-based prediction method for livestock transport stress. *Sensors* **2024**, *24*, 7826. [\[CrossRef\]](#)
23. Ahern, Z.; Corry, P.; Rabbani, W.; Paz, A. Extensive hypothesis testing for crash frequency models. *Heliyon* **2024**, *10*, e26634. [\[CrossRef\]](#)
24. Mitchell, T.M. *Machine Learning*; McGraw-Hill: Boston, MA, USA, 1997.
25. Xu, Y.; Goodacre, R. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *J. Anal. Test.* **2018**, *2*, 249–262. [\[CrossRef\]](#)
26. Amin, A.; Chimba, D.; Hasan, K. Integrating AI and edge computing for advanced safety at railroad grade crossings. *J. Rail Transp. Plan. Manag.* **2025**, *33*, 100501. [\[CrossRef\]](#)
27. Hilbe, J. *Negative Binomial Regression*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2012.
28. Tan, H.; Xie, Y. Financial text analysis and credit risk assessment using GPT-4 and BERT fusion. *PLoS ONE* **2025**, *20*, e0336217. [\[CrossRef\]](#)
29. Abualigah, L. Particle swarm optimization: Advances and applications. *Comput. Mater. Contin.* **2025**, *82*, 1539–1592. [\[CrossRef\]](#)
30. Devarajan, K.; Ponnann, S.; Perumal, S. Hybrid CNN–transformer architecture for EEG-based emotion recognition. *Discover Comput.* **2025**, *28*, 87. [\[CrossRef\]](#)
31. Muraina, O. Ideal dataset splitting ratios in ML algorithms. In Proceedings of the 7th International Mardin Artuklu Scientific Research Conference, Mardin, Turkey, 10–12 December 2022.
32. Dzinyela, R.; Shirazi, M.; Das, S.; Lord, D. Negative binomial–Lindley model with time-dependent parameters. *Accid. Anal. Prev.* **2024**, *207*, 107711. [\[CrossRef\]](#)
33. Tang, L.; Tang, C.; Fu, Q.; Ma, C. Predicting travel mode choice with neural networks and SHAP. *IET Intell. Transp. Syst.* **2024**, *18*, 1339–1354. [\[CrossRef\]](#)
34. Wang, M.; Lee, W.-C.; Liu, N.; Fu, Q.; Wan, F.; Yu, G. A deep learning framework for predicting traffic crashes at intersections. *Appl. Sci.* **2025**, *15*, 752. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.