Deep Learning, Machine Learning, or Statistical Models for Weather-related Crash Severity Prediction

Abimbola Ogungbire, M.S.
Hardik Gajera, Ph.D.

Panick Kalambay, Ph.D.
Srinivas Pulugurtha, Ph.D., P.E., F.ASCE
Founded in 1991, the Mineta Transportation Institute (MTI), an organized research and training unit in partnership with the Lucas College and Graduate School of Business at San José State University (SJSU), increases mobility for all by improving the safety, efficiency, accessibility, and convenience of our nation’s transportation system. Through research, education, workforce development, and technology transfer, we help create a connected world. MTI leads the Mineta Consortium for Transportation Mobility (MCTM) and the Mineta Consortium for Equitable, Efficient, and Sustainable Transportation (MCEET) funded by the U.S. Department of Transportation, the California State University Transportation Consortium (CSUTC) funded by the State of California through Senate Bill 1 and the Climate Change and Extreme Events Training and Research (CCEETR) Program funded by the Federal Railroad Administration. MTI focuses on three primary responsibilities:

Research
MTI conducts multi-disciplinary research focused on surface transportation that contributes to effective decision making. Research areas include: active transportation; planning and policy; security and counterterrorism; sustainable transportation and land use; transit and passenger rail; transportation engineering; transportation finance; transportation technology; and workforce and labor. MTI research publications undergo expert peer review to ensure the quality of the research.

Education and Workforce Development
To ensure the efficient movement of people and products, we must prepare a new cohort of transportation professionals who are ready to lead a more diverse, inclusive, and equitable transportation industry. To help achieve this, MTI sponsors a suite of workforce development and education opportunities. The Institute supports educational programs offered by the Lucas Graduate School of Business: a Master of Science in Transportation Management, plus graduate certificates that include High-Speed Rail and Intercity Rail Management and Transportation Security Management. These flexible programs offer live online classes so that working transportation professionals can pursue an advanced degree regardless of their location.

Information and Technology Transfer
MTI utilizes a diverse array of dissemination methods and media to ensure research results reach those responsible for managing change. These methods include publication, seminars, workshops, websites, social media, webinars, and other technology transfer mechanisms. Additionally, MTI promotes the availability of completed research to professional organizations and works to integrate the research findings into the graduate education program. MTI’s extensive collection of transportation-related publications is integrated into San José State University’s world-class Martin Luther King, Jr. Library.

Disclaimer
The contents of this report reflect the views of the authors, who are responsible for the facts and accuracy of the information presented herein. This document is disseminated in the interest of information exchange. MTI’s research is funded, partially or entirely, by grants from the U.S. Department of Transportation, the U.S. Department of Homeland Security, the California Department of Transportation, and the California State University Office of the Chancellor, whom assume no liability for the contents or use thereof. This report does not constitute a standard specification, design standard, or regulation.
Deep Learning, Machine Learning, or Statistical Models for Weather-related Crash Severity Prediction

Abimbola Ogungbire, M.S.
Panick Kalambay, Ph.D.
Hardik Gajera, Ph.D.
Srinivas Pulugurtha, Ph.D., P.E., F.ASCE

December 2023
Nearly 5,000 people are killed and more than 418,000 are injured in weather-related traffic incidents each year. Assessments of the effectiveness of statistical models applied to crash severity prediction compared to machine learning (ML) and deep learning techniques (DL) help researchers and practitioners know what models are most effective under specific conditions. Given the class imbalance in crash data, the synthetic minority over-sampling technique for nominal (SMOTE-N) data was employed to generate synthetic samples for the minority class. The ordered logit model (OLM) and the ordered probit model (OPM) were evaluated as statistical models, while random forest (RF) and XGBoost were evaluated as ML models. For DL, multi-layer perceptron (MLP) and TabNet were evaluated. The performance of these models varied across severity levels, with property damage only (PDO) predictions performing the best and severe injury predictions performing the worst. The TabNet model performed best in predicting severe injury and PDO crashes, while RF was the most effective in predicting moderate injury crashes. However, all models struggled with severe injury classification, indicating the potential need for model refinement and exploration of other techniques. Hence, the choice of model depends on the specific application and the relative costs of false negatives and false positives. This conclusion underscores the need for further research in this area to improve the prediction accuracy of severe and moderate injury incidents, ultimately improving available data that can be used to increase road safety.
ACKNOWLEDGMENTS

The authors sincerely thank the Highway Safety Information System (HSIS) for the help with the data required for this research. They also appreciate Dr. Sonu Matthew for assisting with problem formulation and preliminary data processing.
## CONTENTS

Acknowledgments ..................................................................................................................................................... vi

List of Figures.......................................................................................................................................................... ix

List of Tables........................................................................................................................................................ x

Executive Summary .................................................................................................................................................. 1

1. Introduction......................................................................................................................................................... 2

1.1 Organization of the Report .......................................................................................................................... 3

2. Literature Review................................................................................................................................................. 4

2.1 Overview of Weather-related Crashes ........................................................................................................ 4

2.2 Previous Research on Crash Severity Prediction ....................................................................................... 5

2.3 Methods in Crash Severity Analysis ......................................................................................................... 5

3. Methodology ..................................................................................................................................................... 8

3.1 Data Source ................................................................................................................................................... 8

3.2 Statistical Models ......................................................................................................................................... 11

3.3 Machine Learning (ML) Models ............................................................................................................... 13

3.4 Deep Learning (DL) Models ..................................................................................................................... 14

4. Results and Discussion ................................................................................................................................. 17

4.1. Imbalance Data Treatment ....................................................................................................................... 17

4.2. Model Estimation ....................................................................................................................................... 18

4.3. Model Performance Comparison and Analysis of Predictive Power .................................................. 21
5. Conclusion ......................................................................................................................... 28

5.1 Summary of Findings ....................................................................................................... 28

5.2 Limitations and Future Scope of Work ........................................................................... 29
Abbreviations and Acronyms ............................................................................................. 30
References ............................................................................................................................. 31
About the Authors .................................................................................................................. 38
LIST OF FIGURES

Figure 1. Distribution of Crash Severity before Data Treatment........................................ 17
Figure 2. Distribution of Crash Severity after Data Treatment........................................ 18
Figure 3. Permutation Feature Importance Plot................................................................. 21
Figure 4. Precision Comparison for Statistical, ML, and ML Models................................. 25
Figure 5. Recall Comparison for Statistical, ML, and DL Models........................................ 25
Figure 6. F1 Comparison for Statistical, ML, and DL Models............................................. 26
# List of Tables

Table 1. Summary Statistics ........................................................................................................... 9

Table 2. Confusion Matrix for the OLM ......................................................................................... 19

Table 3. Confusion Matrix for the OPM ......................................................................................... 19

Table 4. Confusion Matrix for the RF .............................................................................................. 20

Table 5. Confusion Matrix for the XGBoost .................................................................................... 21

Table 6. Model Performance Metrics .............................................................................................. 23
Executive Summary

The focus of this study is to evaluate and compare the performance of three different approaches for crash severity prediction: statistical, machine learning (ML), and deep learning (DL) models. The goal was to find the most effective model for each crash severity level and understand the strengths and limitations of each approach.

The performance of the ordered probit model (OPM), a statistical model, is similar to the ordered logit model (OLM), but both struggle to correctly identify severe-injury crashes and distinguish between moderate-injury and property-damage-only (PDO) crashes. In the realm of ML, models such as random forest (RF) and XGBoost are evaluated. For DL, models such as the multi-layer perceptron (MLP) and TabNet are assessed. The performance of these models varied across severity levels, with PDO predictions being the best and severe-injury predictions performing the worst. In the case of severe-injury classification, the best performing model is reported to be TabNet with a precision of 8.45%, a recall of 45.73%, and an F1 score of 28.78%. The model with the best performance for predicting moderate-injury crashes is RF, with a precision of 42.77%, a recall of 85.53%, and an F1 score of 57.03%. For PDO classification, TabNet again performed best with a precision of 79.62%, a recall of 82.72%, and an F1 score of 92.10%. However, all these models struggled with severe-injury classification due to factors such as class imbalance or the complexity of the problem.

It can be concluded that the statistical models have high precision but low recall for severe-injury crashes, suggesting they are confident but not comprehensive in identifying severe-injury crashes. The ML models offer a more balanced performance with moderate precision and recall. In contrast, the DL models have high recall but low precision, indicating they may incorrectly classify incidents as severe- or moderate-injury crashes. The choice of a model may depend on the specific application and the relative costs of false negatives and false positives. For example, a model with a high recall (such as the DL models) might be preferable in preliminary screening tools where the cost of a false negative is much higher than the cost of a false positive.
1. Introduction

Road traffic crashes are a significant public health concern (Gopalakrishnan, 2012; CDC, 2023). Rain, snow, ice, and poor visibility often create unsafe road conditions for drivers (Druta et al., 2020; Greaves et al., 2023). Recent reports published by the Federal Highway Administration (FHWA) indicate that nearly 5,000 people are killed and more than 418,000 people are injured in weather-related crashes each year (FHWA, 2020). The consequences of a driver's failure to stop when confronted by either poor visible conditions or wet pavement frequently result in severe crashes. In order to reduce the intensity or frequency of these weather-related crashes, it is important to understand their contributing factors and to select and implement relevant countermeasures. In addition, providing an assessment of the effectiveness of statistical models, machine learning (ML) models and deep learning (DL) models applied to crash severity prediction will help researchers and practitioners know which models are most effective under specific conditions.

Researchers in the past adopted statistical methods such as binary logistic regression models, multinomial logit models, OPMs, and random parameter logit models to evaluate crash severity (Penmetsa & Pulugurtha, 2017, 2019; Penmetsa et al., 2018a; Chen et al., 2019; Tagar & Pulugurtha, 2021; Fanyu et al., 2021; Hou et al., 2022). Historically, statistical models have been widely utilized for crash severity prediction based on weather variables. Researchers employed regression techniques to establish linear relationships between weather conditions and crash outcomes (Mao et al., 2019; Abdulhafedh, 2022). Although these models provided initial insights, they often failed to capture intricate nonlinear interactions and complex patterns in the data. The emergence of ML techniques revolutionized crash severity prediction by enabling the development of more sophisticated models. Decision trees, random forests (RFs), and support vector machines (SVMs) were among the first algorithms employed in the context of crash severity analysis (Theofilatos et al., 2019; Hadjidimitriou et al., 2019; Mokhtarimousavi et al., 2020). These models introduced the concept of feature importance, enabling researchers to identify the most influential weather-related variables affecting crash severity. DL, a subset of ML, has gained substantial popularity in recent years due to its ability to automatically learn hierarchical representations from raw data. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been leveraged to extract intricate patterns from weather data, empowering more accurate crash severity predictions. These models excel at handling complex temporal and spatial dependencies within weather variables, capturing subtle nuances that were previously overlooked. A few studies were conducted to evaluate the performance of statistical models, ML models, and DL models in weather-severity crash severity prediction. These studies often employed large datasets encompassing diverse weather conditions and incorporating a range of evaluation metrics such as accuracy, precision, recall, and F1 score. The results varied across these studies, with each modeling technique demonstrating its strengths and limitations.
A diverse set of models were chosen to comprehensively assess their effectiveness in predicting weather-related crash severity. OLM and OPM were chosen due to their established use in similar research and their interpretability, which is well-suited for ordinal crash severity data. ML techniques such as RF and XGBoost were chosen for their robustness in handling noisy data, capturing intricate interactions, and providing feature importance insights. These models strike a balance between accuracy and interpretability. In the realm of DL, the multi-layer perceptron neural network (MLP NN) and TabNet were chosen, leveraging their capacity to extract complex non-linear patterns from weather data, particularly adept at handling the temporal and spatial dependencies within weather variables. TabNet, designed for tabular data, offers feature selection capabilities. By including this array of models, the study aims to offer a comprehensive understanding of their strengths and limitations, thereby providing valuable insights for predicting crash severity under various weather conditions. These models are designed to forecast the level of traffic crash injuries, specifically factoring in prevailing weather conditions as one of their inputs. In other words, these models aim to estimate the severity of injuries resulting from a traffic crash, and they consider the weather conditions at the time of the crash as a significant factor in making these predictions.

1.1 Organization of the Report

The remainder of the report comprises four chapters. Chapter 2 presents and analyzes the literature of weather-related crashes, including the methods. Chapter 3 describes the data, study area, and method used for crash severity analyses in this study. Chapter 4 presents and analyzes the results from this study, and Chapter 5 summarizes the conclusions and scope for future work.
2. Literature Review

2.1 Overview of Weather-related Crashes

Weather-related crashes pose significant risk to road users and have been a subject of extensive research in the transportation safety domain. Numerous studies were conducted to investigate the relationship between weather conditions and crash occurrence (Zhao et al., 2019; Al-Mistarehi et al., 2022; Wei et al., 2022). Adverse weather factors such as rain, snow, fog, and icy roads have been found to significantly increase the likelihood of crashes (Hammad et al., 2019; Robins & Fotio, 2020; Das et al., 2020; Zhang et al., 2021). For instance, studies have reported higher crash rates during rainfall or heavy precipitation events (Das et al., 2020). Additionally, reduced visibility due to fog or mist has been linked to an increased crash risk (Robins & Fotio, 2020). Understanding the specific weather variables and their impact on crash occurrence is essential for developing targeted interventions.

Weather conditions play a critical role in determining crash severity. Several studies have explored the association between weather variables and the severity of crashes (Xing et al., 2019; Abohassan et al., 2021; Yang et al., 2022). Harsh weather conditions, including heavy rain, snowstorm, and high wind, have been linked to an increased likelihood of severe crashes (Xing et al., 2019). Poor road surface conditions during adverse weather, such as slippery roads or reduced tire grip, contribute to more severe crashes (Druta et al., 2020). Investigating the relationship between weather conditions and crash severity is crucial for allocating resources and implementing appropriate safety measures.

Studies on weather-related crashes utilize diverse data sources, including police reports, crash databases, weather station data, and advanced sensor technologies (Das et al., 2020; Duddu et al., 2020; Mathew & Pulugurtha, 2022). These datasets provide valuable information for conducting analyses. However, they may have limitations such as underreporting crashes or incomplete weather data. Addressing data limitations and improving data collection methods remain critical challenges in weather-related crash research. Understanding the impact of weather conditions on crash occurrence and severity has practical implications for road safety management.

While significant effort has been made in crash severity analysis, there is a clear need for more comprehensive studies that build on identifying the best method in analyzing weather-related crash severity. This study compares a broader range of statistical, ML, and DL models on the same representative crash dataset. This research could provide greater understanding of the intricate relationship between weather conditions and crash severity, and ultimately contribute to the development of more effective strategies for enhancing road safety during adverse weather conditions.
2.2 Previous Research on Crash Severity Prediction

Crash severity prediction is a critical area of research that aims to identify factors and develop models to accurately predict the severity of traffic crashes. Accurate crash severity prediction can aid in emergency response planning, resource allocation, and the development of effective countermeasures to reduce crash severity. Numerous studies were conducted to investigate the factors influencing crash severity (Zhao et al., 2019; Das et al., 2020; Al-Mistarehi et al., 2022; Wei et al., 2022; Yang et al., 2022). These factors can be broadly categorized into three main groups: driver-related, vehicle-related, and environmental-related factors (Robins & Fotio, 2020; Hou et al., 2022; Yang et al., 2022). Driver-related factors include driver age, gender, impairment (e.g., alcohol or drug use), distraction, and fatigue (Dingus et al., 2016). Vehicle-related factors encompass vehicle type, size, and safety features. Environmental-related factors consist of road conditions, weather conditions, lighting, and traffic characteristics. Understanding the impact of these factors is crucial for developing effective crash-severity prediction models. The selection of predictor variables significantly impacts the accuracy and interpretability of crash-severity prediction models (Sattar et al., 2023). Previous research has identified a wide range of potential predictor variables, including driver characteristics (e.g., age and gender), roadway attributes (e.g., speed limit and road type), environmental conditions (e.g., weather and lighting), and crash-specific variables (e.g., crash type and time of day) (Duddu et al., 2019; Shi et al, 2019; Yuan et al., 2019; Islam & Mannering, 2020). Feature selection techniques such as stepwise regression, principal component analysis, or recursive feature elimination have been employed to identify the most influential variables for crash-severity prediction.

Evaluation metrics play a crucial role in assessing the performance of crash-severity prediction models. Commonly used metrics include accuracy, precision, recall, F1 score, and area under the curve—receiver operating characteristic curve (AUC-ROC) (Yuan et al., 2019). Additionally, confusion matrix analysis provides insights into model performance across different severity levels. Studies have compared the performance of different models and techniques, highlighting the strengths and limitations of each approach. Higher accuracy and AUC-ROC values indicate better model performance (Ke et al., 2017). Accurate crash-severity prediction models have practical implications for road safety management. For instance, emergency response systems can use predicted severity levels to dispatch appropriate medical personnel and resources. Transportation agencies can prioritize road safety improvements and allocate funding based on predicted crash severity hotspots (Jamal et al., 2021). Furthermore, crash-severity-prediction models can aid in the development of intelligent transportation systems and advanced driver-assistance systems to prevent or mitigate crashes.

2.3 Methods in Crash Severity Analysis

Researchers have widely used statistical models in the past to determine the effect of factors affecting crashes on injury severity levels. Amongst statistical models, discrete choice models are
widely employed in the existing literature considering the discrete nature of variables in crash datasets (Penmetsa & Pulugurtha, 2018b; Yan et al., 2021). Logistic regression techniques are widely used in cases when the dependent variable is discrete or ordinal. Several researchers used different models such as the OLM (Ayuso and Santolino, 2007), OPM (Abdel-Aty 2003; Gray et al., 2008; Garrido et al., 2014; Penmetsa et al., 2017), Bayesian OPM (Xie et al., 2009), probit or logit models with random parameters (Behnoood and Mannering 2015; Behnoood and Mannering 2016; Behnoood and Mannering 2017), hierarchical OLMs, and OPMs (Chen et al., 2016; Fountas and Anastasopoulos, 2018). Advanced statistical models provide flexibility to account for heterogeneity due to unobserved variables as well as account for random or hierarchical components (Duddu et al., 2018; Duvvuri et al., 2022).

Traditional OLMs and OPMs are used in this study since they are computationally efficient and do not account for any random component in modeling. Traditional models also provide flexibility to compare the accuracy matrix with an ML or DL model, with results employed on the same dataset.

With the advancement of computational power and data availability, ML models have started to gain attention in traffic crash-severity prediction. These models include decision trees, RFs, SVMs, and gradient boosting machines, among others. Zeng et al. (2019) employed a DL approach, specifically a long short-term memory (LSTM) model, to predict crash severity. Their study demonstrated that LSTM, a type of recurrent neural network, is well suited to handling time-series data and can effectively model the temporal dependencies of various factors contributing to crash severity. Chen et al. (2016) used a decision tree model and found it to outperform logistic regression in terms of prediction accuracy. However, the model might overfit the data if not properly tuned. Wang et al. (2019) showed that SVMs, combined with an effective feature selection strategy, provided highly accurate predictions and outperformed traditional logistic regression models. Santos et al. (2020) compared various ML models, including decision trees, RFs, and XGBoost, for crash severity prediction and found XGBoost to deliver the best performance. Tang et al. (2020) reviewed the use of RF models for crash severity prediction and highlighted their superior performance over traditional statistical models, particularly in handling high-dimensional and non-linear data. Nevertheless, the interpretability of RF models remains a challenge.

Deep learning (DL) models for crash severity prediction have been used in a limited number of studies (Abdelwahab & Abdel-Aty, 2001; Alkheder et al., 2017; Das et al., 2018; Zheng et al., 2019; Ma et al., 2021; Rahim & Hassan, 2021; Khan & Ahmed, 2022). Abdelwahab & Abdel-Aty (2001) employed multi-layer perceptron (MLP) and fuzzy adaptive resonance theory (ART) to develop driver-injury severity models. Their DL models outperformed OLMs, achieving an accuracy of 65.6% and 60.14% on the training and testing datasets, respectively. Alkheder et al. (2017) also demonstrated the superiority of DL models, specifically artificial neural networks (ANNs), over the OPM. Their ANN-based model achieved an accuracy of 81.6% and 74.6% on
the training and testing datasets, respectively. Das et al. (2018) introduced "DeepScooter," a deep learning framework based on MLP, for predicting crash severities involving at-fault motorcycle riders. Remarkably, their framework achieved unprecedented accuracies of 100% and 94% on the training and testing datasets, respectively. Ma et al. (2021) proposed an analytic framework utilizing a DL model. Their model achieved a recall value of 0.82–0.85 for geographical clusters of fatal and serious injury crashes. In contrast to the aforementioned studies, Zheng et al. (2019), Rahim & Hassan (2021), and Khan & Ahmed (2022) utilized CNN-based models to predict crash severity levels. Zheng et al. (2019) developed a novel CNN-based model named "TASP-CNN" for crash-severity prediction. Their model achieved a recall of 6.3% on fatal crashes (16.7% and 93.2% for slight and serious injury crashes, respectively) and an overall F1 score of 87%. Rahim & Hassan (2021) proposed a CNN-based framework for predicting crash severity in highway work zones, utilizing the EfficientNet model (Tan et al., 2019). They achieved the highest recall of 67% for fatal crashes. More recently, Khan & Ahmed (2022) developed a crash-severity-prediction model for rural mountainous freeways, employing the ResNet18 algorithm. This model achieved a recall of 99.3% for fatal crashes. It is worth noting that, unlike the other mentioned studies, their study focused solely on predicting crash severity in adverse weather conditions.

Despite the extensive literature on traffic crash-severity prediction using statistical, ML, and DL models, there is still a lack of studies that thoroughly compare these models in the same context, using the same dataset. Therefore, a comprehensive comparison study is necessary. This study aims to fill these gaps by conducting a comprehensive comparison of different models in terms of various performance metrics. This study will provide a more complete picture of the strengths and weaknesses of different models—potentially contribute to the development of more effective and efficient traffic safety strategies.
3. Methodology

3.1 Data Source

The data used in this study were obtained from the Highway Safety Information System (HSIS) database. Crashes that took place between January 1, 2015 and December 31, 2017 were extracted. Crashes are reported using case numbers and observations with the same number indicating that the vehicles involved are part of the same crash incident. To gain a thorough understanding of the crash occurrence process, Washington and Haque (2013) argued that crashes due to different causes should be modeled separately. Hence, only weather-related crashes were extracted, while crashes occurring under clear weather conditions were excluded to obtain a final dataset. The final dataset only included crashes that happened under non-clear weather conditions (i.e., with cloudy, rain, fog/smog, sleet/hail/freezing rain/drizzle, severe crosswinds, or blowing sand conditions described in the crash reports).

Crash severity is defined in the HSIS database at five different levels: fatal crashes, injury type class A, injury type class B, injury type class C, and no injury/property damage only (PDO). For this analysis, the crash severity was re-categorized into three levels, i.e., severe injury (fatal and injury type class A), moderate injury (injury type class B and injury type class C), and PDO/no injury. A total of 238,252 weather-related crashes were recorded within the study period with 2,952 severe crashes, 71,688 moderate crashes, and 163,612 PDO crashes.

Table 1 presents the summary statistics providing information about the various categories and variables associated with weather-related crashes. It shows the count and percentage of severe, moderate, and PDO crashes for each category, such as weather condition, contributing factor of the crash, road surface condition, functional class of the road, location type, light condition, road characteristic, driver gender, driver age, speed limit class, crash type, work zone area, vehicle type, seasonal factors, road terrain, time of day, day of the week, locality, and more.

The summary statistics give information about the counts and percentages of different types of weather conditions for severe, moderate, and PDO injury cases. Cloudy weather is the most common weather condition in all three categories, with 63.4% of severe crashes, 57.6% of moderate-injury crashes, and 56.6% of no-injury crashes occurring under cloudy conditions. Rain is the second most common weather condition, followed by snow, fog, smog, and smoke. Sleet, hail, and freezing rain/drizzle are less common, with blowing sand and dirt being the least common weather condition for all three severity types.
Table 1. Summary Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Category</th>
<th>Description</th>
<th>Severe Injury</th>
<th>Moderate Injury</th>
<th>PDO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Count</td>
<td>%</td>
<td>Count</td>
</tr>
<tr>
<td>Weather condition</td>
<td>1</td>
<td>Cloudy</td>
<td>1871</td>
<td>63.4</td>
<td>41293</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Rain</td>
<td>893</td>
<td>30.3</td>
<td>26828</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Snow</td>
<td>41</td>
<td>1.4</td>
<td>1085</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Fog, smog, smoke</td>
<td>104</td>
<td>3.5</td>
<td>1247</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Sleet, hail, freezing rain/drizzle</td>
<td>39</td>
<td>1.3</td>
<td>1196</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>Severe crosswinds</td>
<td>3</td>
<td>0.1</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>Blowing sand, dirt</td>
<td>1</td>
<td>0.03</td>
<td>8</td>
</tr>
<tr>
<td>Contributing factor of the crash</td>
<td></td>
<td>No contributing factors</td>
<td>1191</td>
<td>40.3</td>
<td>32095</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Disregarding signs or signals</td>
<td>98</td>
<td>3.3</td>
<td>2124</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Exceeded safe speed/speed limit or failed to reduce speed</td>
<td>548</td>
<td>18.6</td>
<td>17302</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Improper turn or right turn on red</td>
<td>17</td>
<td>0.6</td>
<td>730</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Crossed centerline, improper lane change, or use of an improper lane</td>
<td>222</td>
<td>7.5</td>
<td>2135</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>Overcorrected, oversteered, improper passing, or improper backing</td>
<td>77</td>
<td>2.6</td>
<td>1503</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>Failing to yield to the right-of-way, or driver inattention</td>
<td>267</td>
<td>9.0</td>
<td>9147</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>Operating too closely, aggressive driving, or alcohol use</td>
<td>429</td>
<td>14.5</td>
<td>3768</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>Visibility obstruction, or defective equipment</td>
<td>16</td>
<td>0.5</td>
<td>471</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>Other/unable to determine</td>
<td>87</td>
<td>2.9</td>
<td>2413</td>
</tr>
<tr>
<td>Road surface condition</td>
<td>1</td>
<td>Dry</td>
<td>1290</td>
<td>43.7</td>
<td>27329</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Wet, presence of water (standing/moving)</td>
<td>1577</td>
<td>53.4</td>
<td>41617</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Ice, snow, slush</td>
<td>81</td>
<td>2.7</td>
<td>2708</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Sand, mud, dirt, gravel, fuel, or oil</td>
<td>4</td>
<td>0.1</td>
<td>34</td>
</tr>
<tr>
<td>Functional class of road</td>
<td>1</td>
<td>Principal arterial: interstate, freeways and expressways</td>
<td>402</td>
<td>13.6</td>
<td>12617</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Principal arterial: other</td>
<td>682</td>
<td>23.1</td>
<td>22955</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Minor arterial</td>
<td>693</td>
<td>23.5</td>
<td>18303</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Major collector</td>
<td>789</td>
<td>26.7</td>
<td>11795</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Local</td>
<td>386</td>
<td>13.1</td>
<td>6018</td>
</tr>
<tr>
<td>Location type</td>
<td>0</td>
<td>Non-intersection</td>
<td>2516</td>
<td>85.2</td>
<td>57486</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Intersection</td>
<td>436</td>
<td>14.8</td>
<td>14202</td>
</tr>
<tr>
<td>Light condition</td>
<td>1</td>
<td>Daylight</td>
<td>1727</td>
<td>58.5</td>
<td>51358</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Dusk, and dawn</td>
<td>163</td>
<td>5.5</td>
<td>3582</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Dark lighted roadway/unknown lighting</td>
<td>244</td>
<td>8.3</td>
<td>7146</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Roadway not lighted</td>
<td>818</td>
<td>27.7</td>
<td>9402</td>
</tr>
<tr>
<td>Variable</td>
<td>Category</td>
<td>Description</td>
<td>Severe Injury</td>
<td>Moderate Injury</td>
<td>PDO</td>
</tr>
<tr>
<td>------------------------</td>
<td>-----------------------------------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>--------------</td>
<td>----------------</td>
<td>-----</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Count</td>
<td>%</td>
<td>Count</td>
<td>%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Road characteristic</td>
<td>1</td>
<td>1675</td>
<td>56.7</td>
<td>50437</td>
<td>70.4</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>485</td>
<td>16.4</td>
<td>12319</td>
<td>17.2</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>788</td>
<td>26.7</td>
<td>8876</td>
<td>12.4</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>4</td>
<td>0.1</td>
<td>56</td>
<td>0.1</td>
</tr>
<tr>
<td>Driver gender</td>
<td>1</td>
<td>1006</td>
<td>68.0</td>
<td>38058</td>
<td>53.1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>946</td>
<td>32.0</td>
<td>33630</td>
<td>46.9</td>
</tr>
<tr>
<td>Driver age</td>
<td>1</td>
<td>262</td>
<td>8.9</td>
<td>7105</td>
<td>9.9</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2505</td>
<td>84.9</td>
<td>60725</td>
<td>84.7</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>185</td>
<td>6.3</td>
<td>3858</td>
<td>5.4</td>
</tr>
<tr>
<td>Speed limit class</td>
<td>1</td>
<td>5</td>
<td>0.2</td>
<td>415</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>17</td>
<td>0.6</td>
<td>1064</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>302</td>
<td>10.2</td>
<td>15354</td>
<td>21.4</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>860</td>
<td>29.1</td>
<td>28690</td>
<td>40.0</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>1495</td>
<td>50.6</td>
<td>20053</td>
<td>28.0</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>273</td>
<td>9.2</td>
<td>6112</td>
<td>8.5</td>
</tr>
<tr>
<td>Crash type</td>
<td>1</td>
<td>113</td>
<td>3.8</td>
<td>2524</td>
<td>3.5</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>124</td>
<td>4.2</td>
<td>1191</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>186</td>
<td>6.3</td>
<td>484</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>26</td>
<td>0.9</td>
<td>727</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>664</td>
<td>22.5</td>
<td>9296</td>
<td>13.0</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>395</td>
<td>13.4</td>
<td>29527</td>
<td>41.2</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>340</td>
<td>11.5</td>
<td>9243</td>
<td>12.9</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>416</td>
<td>14.1</td>
<td>1534</td>
<td>2.1</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>599</td>
<td>20.3</td>
<td>15879</td>
<td>22.2</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>89</td>
<td>3.0</td>
<td>1283</td>
<td>1.8</td>
</tr>
<tr>
<td>Work zone area</td>
<td>0</td>
<td>2880</td>
<td>97.6</td>
<td>69872</td>
<td>97.5</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>72</td>
<td>2.4</td>
<td>1816</td>
<td>2.5</td>
</tr>
<tr>
<td>Vehicle type</td>
<td>1</td>
<td>1331</td>
<td>45.1</td>
<td>40185</td>
<td>56.1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1228</td>
<td>41.6</td>
<td>28305</td>
<td>39.5</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>13</td>
<td>0.4</td>
<td>255</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>192</td>
<td>6.5</td>
<td>1847</td>
<td>2.6</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>179</td>
<td>6.1</td>
<td>876</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>9</td>
<td>0.3</td>
<td>220</td>
<td>0.3</td>
</tr>
<tr>
<td>Seasonal factors</td>
<td>1</td>
<td>623</td>
<td>21.1</td>
<td>18168</td>
<td>25.3</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>762</td>
<td>25.8</td>
<td>17669</td>
<td>24.6</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>717</td>
<td>24.3</td>
<td>15408</td>
<td>21.5</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>850</td>
<td>28.8</td>
<td>20443</td>
<td>28.5</td>
</tr>
<tr>
<td>Road terrain</td>
<td>1</td>
<td>748</td>
<td>25.3</td>
<td>13405</td>
<td>18.7</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1975</td>
<td>66.9</td>
<td>53124</td>
<td>74.1</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>229</td>
<td>7.8</td>
<td>5159</td>
<td>7.2</td>
</tr>
</tbody>
</table>
In this study, the statistical models were setup to preserve the ordered nature of the severity class. Two ML models with different classification logics were applied, i.e., bagging and boosting techniques. The dataset was divided into training and test sets to evaluate the model’s performance on unseen data. The training set was used to train the model, while the test set was used to assess its performance. The split ratio used in this study is 75:25 for the training and testing dataset, respectively. Since the dataset suffers from class imbalance, where one class has significantly fewer samples than the others, the Synthetic Minority Over-Sampling Technique for Nominal (SMOTE-N) data was applied to balance the classes. These methods help prevent the model from being biased towards the majority class and improve its ability to learn from the minority class.

### 3.2 Statistical Models

A class of logistic models known as ordered probability models, such as the OLM or OPM, proves useful for regression analysis when dealing with a modeled variable containing three or more categories, with crucial consideration of the order among these categories (Sasidharan & Menéndez, 2014). OLMs and OPMs are widely used to analyze and predict the relationship between ordinal dependent variables and a set of independent variables. OLMs and OPMs have been widely used to predict crash severity (Abdel-Aty, 2003; Ayuso & Santolino, 2007).

#### 3.2.1 Ordered Logit Model (OLM)

The OLM was developed for weather-related crash severity analysis with severity levels defined as severe, moderate, and PDO.
Let $Y_i$ be the ordinal response variable representing the crash severity of the $i$th observation. The severity levels are coded as follows: severe ($Y_i = 2$), moderate ($Y_i = 1$), and PDO ($Y_i = 0$). The OLM can be represented as follows:

\[
\text{logit}(P(Y_i \leq 1 | X_i)) = \alpha_1 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}
\]

\[
\text{logit}(P(Y_i \leq 2 | X_i)) = \alpha_2 + \beta_1 X_{i2} + \ldots + \beta_p X_{ip}
\]

where

$P(Y_i \leq 1 | X_i)$ represents the probability that the crash severity of the $i$-th observation is less than or equal to moderate (level 1);

$P(Y_i \leq 2 | X_i)$ represents the probability that the crash severity of the $i$-th observation is less than or equal to severe (level 2);

$X_i$ denotes the vector of explanatory variables for the $i$-th observation;

$\alpha_1$ and $\alpha_2$ are the intercepts specific to moderate and severe levels of crash severity, respectively; and

$\beta_1, \ldots, \beta_p$ are the coefficients associated with the explanatory variables $X_{i1}, \ldots, X_{ip}$, respectively.

Note that the reference category for the ordinal response variable is PDO (level 0), which does not have a corresponding logit equation. Its probability can be derived as $1 - P(Y_i \leq 1 | X_i)$ since the sum of probabilities for all levels should equal 1.

3.2.2 Ordered Probit Model (OPM)

The OPM assumes that the relationship between a set of independent variables and the probability of each category of the dependent variable follows a standard normal distribution (Long, 1997). The OPM can be represented mathematically as follows:

\[
P\left(Y \leq \frac{j}{X}\right) = \Phi(\beta'X - \varphi_j)
\]

where

$P\left(Y \leq \frac{j}{X}\right)$ represents the probability that the dependent variable $Y$ takes a value less than or equal to $j$, given the value of the independent variables $X$;

$\Phi$ represents the cumulative distribution function of the standard normal distribution;
\( \beta \) is a vector of coefficients;

\( X \) is a vector of independent variables; and

\( \varphi_j \) represents the threshold parameters delineating different categories of dependent variables.

3.3 Machine Learning (ML) Models

3.3.1 Random Forest Model (RF)

RF is an ensemble learning method that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Suppose an RF model consists of \( N \) decision trees (Biau, 2012). Each tree gives a classification, and the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest). Each tree is grown as follows:

- If the number of cases in the training set is \( N \), then \( N \) cases are sampled at random—but with replacement from the original data. This sample will be the training set for growing the tree.

- If there are \( M \) input variables, a number \( m \) is specified such that at each node, \( m \) variables are selected at random out of the \( M \) and the best split on these \( m \) is used to split the node. The value of \( m \) is held constant during the forest growing.

- Each tree is grown to the largest extent possible and there is no pruning.

For a given test record, each tree in the forest gives a classification. The forest chooses the classification having the most votes (over all the trees in the forest), and in case of regression, it takes the average of outputs by different trees.

Mathematically, the prediction of an RF model for an input \( x \) can be written as

\[
\hat{y} = \frac{1}{N} \sum_{i=1}^{N} f_i(x)
\]

where \( f_i(x) \) is the prediction of the \( i \)-th decision tree.

The model would take as input features of a traffic incident (such as speed, weather condition, time of day, etc.), and output a severity class (severe injury, moderate injury, PDO). The model would be trained on a labeled dataset, and the aim would be to minimize the discrepancy between
the predicted and actual labels. An RF's ability to combine multiple decision trees helps it avoid overfitting and generally results in a robust prediction performance.

3.3.2 Extreme Gradient Boosting Model (XGBoost)

XGBoost model is an ensemble of decision trees, where each subsequent tree tries to correct the errors made by the previous ones. It is an iterative process that aims to minimize a loss function (Ke et al., 2017; Charm et al., 2023). Mathematically, the prediction $\hat{y}$ of an XGBoost model with $K$ trees for an input $x$ can be written as follows:

$$\hat{y} = \sum_{k=1}^{K} f_k(x)$$

where $f_i(x)$ is the prediction of the $k$-th tree.

The objective function that XGBoost tries to minimize is represented as follows:

$$\text{obj}(\theta) = L(\theta) + \vartheta(\theta)$$

where $\theta$ represents the parameters of the model, $L(\theta)$ is the training loss function, and $\vartheta(\theta)$ is a regularization term that controls the complexity of the model.

The "gradient boosting" part of XGBoost comes from the fact that it trains each new tree to predict the negative gradient (or "residual") of the loss function with respect to the current predictions (Ke et al., 2017). This is why it is called "gradient boosting," as it uses gradient information to boost the performance of the ensemble. The model would learn to predict the severity of a crash based on input features such as weather condition, time of day, lighting condition, etc., by minimizing the discrepancy between its predictions and the actual labels, while also controlling the complexity of the model to prevent overfitting.

3.4 Deep Learning (DL) Models

3.4.1 Multi-layer Perceptron

MLP is a class of ANN composed of multiple layers of nodes (or "neurons") in a directed graph. Each layer is fully connected to the next one, meaning that each node in a given layer is connected to all nodes in the adjacent layers. Consider an input vector $x$ of dimension $d$, which corresponds to the features of a given traffic incident. This includes factors such as lighting condition, weather condition, time of day, etc. The input is then passed through one or more hidden layers. Each node in a hidden layer computes a weighted sum of its inputs, adds a bias term, and applies an activation function. Mathematically, this can be expressed as follows.
\[ h_i^{(l)} = f\left( \sum_j w_{ij}^{(l)} h_j^{(l-1)} + b_i^{(l)} \right) \]

Here, \( h_i^{(l)} \) is the output of the \( i \)-th node in the \( l \)-th layer, \( f \) is the activation function, \( w_{ij}^{(l)} \) is the weight connecting the \( j \)-th node in the \((l - 1)\)-th layer to the \( i \)-th node in the \( l \)-th layer, and \( b_i^{(l)} \) is the bias term for the \( i \)-th node in the \( l \)-th layer. The final hidden layer is fully connected to the output layer. In a classification context such as traffic crash severity prediction, the output layer would have one node for each class (severe injury, moderate injury, and PDO) and would use a SoftMax function to convert the outputs into probabilities summing to 1 (Sattar et al., 2023). The class with the highest probability is chosen as the prediction.

3.4.2 TabNet

The TabNet model is designed to learn and capture the complex relationships between weather-related features and the severity levels of crashes. It leverages various components to process and analyze the tabular data effectively. It consists of shared feature transformers, sparse attention mechanisms, sequential feature selection, and adversarial learning components. Shared feature transformers process the input features and extract meaningful representations. They learn patterns and relationships within the weather-related crash data, enabling the model to capture the important features that contribute to crash severity prediction.

Sparse attention mechanisms are incorporated in TabNet to selectively focus on relevant features while disregarding irrelevant ones. This attention mechanism identifies important patterns and interactions among the weather-related features. By attending to the most relevant features, TabNet can effectively learn the relationships between weather conditions and crash severity (Sattar et al., 2023). Sequential feature selection is a critical aspect of TabNet. It adaptively chooses the most informative features at each decision step, allowing the model to make accurate predictions (Jiang et al., 2022). This sequential selection process ensures that the relevant features are utilized for classifying the severity levels of weather-related crashes. To enhance model robustness and interpretability, TabNet incorporates adversarial learning. It includes an adversarial loss component that encourages the model to capture useful information from the dropped features while maintaining privacy and interpretability.

3.5 Model Evaluation and Comparison

Performance metrics are essential for assessing the effectiveness and reliability of predictive models in various contexts. The models are evaluated using metrics such as precision, recall, and F1 score. Precision measures the proportion of true positive predictions among all positive predictions, while recall measures the proportion of true positives among all actual positives. The F1 score is a balanced measure of precision and recall, and accuracy measures the proportion of correct predictions among all predictions. The model comparison was carried out by assessing the performance of different models to determine which model performs best. The best model is often
the one that balances the trade-off between bias and variance and performs well on unseen data, demonstrating good generalization capabilities. The specific criteria for choosing the best model can depend on the context, including the relative costs of different types of prediction errors.
4. Results and Discussion

4.1. Imbalance Data Treatment

Wen et al. (2021) reviewed the current status on the use of ML in traffic crash prediction and identified data imbalance as a major issue. In this study, the SMOTE-N algorithm was employed in addressing class imbalance by generating synthetic samples for the minority class. The SMOTE-N algorithm is specifically designed for datasets with nominal predictor features. Figures 1 and 2 show the visual representation of the dataset before and after data treatment.

Figure 1. Distribution of Crash Severity before Data Treatment

![Graph showing distribution before data treatment](image)

Figure 1 illustrates the data distribution wherein the minority class represents severe injury or class 1. Training an ML or DL model using such imbalanced datasets can lead to biased training samples. Consequently, the SMOTE-N technique was employed as a remedy for the data imbalance issue. Figure 2 visually depicts the distribution of the dataset subsequent to the application of the SMOTE-N data-treatment technique.
4.2. Model Estimation

The main objective was to evaluate the performance of three different approaches (i.e., statistical, ML, and DL) to classify crash severity. For the ML and DL models, the SMOTE-N data treatment was used to resolve the data imbalance issue.

4.2.1. Statistical Models

4.2.1.1. Ordered Logit Model (OLM)

The confusion matrix from the OLM is as presented in Table 2. For severe crashes, the model did not correctly predict any instances. All severe crashes were misclassified as either moderate (215 instances) or PDO (525 instances). This indicates a major issue as the model is unable to correctly identify any of the severe crashes, which could be the most critical to predict accurately in a real-world setting. For moderate crashes, the model correctly predicted 1,101 instances. However, a significant number (16,776) were misclassified as PDO. This suggests that while the model has some ability to identify moderate crashes, it is largely confusing them with PDO crashes. The model performs best in predicting PDO crashes. It correctly predicted 40,195 PDO crashes and misclassified 751 as moderate. None of the PDO crashes were misclassified as severe. Hence, the model's performance in classifying crash severity is suboptimal. It never predicts severe crashes, which could be the most important to predict in practice. Additionally, it often confuses moderate and PDO crashes. The model may benefit from additional training, better feature selection, or a different modeling strategy.
Table 2. Confusion Matrix for the OLM

<table>
<thead>
<tr>
<th>Reference</th>
<th>Severe Injury</th>
<th>Moderate Injury</th>
<th>PDO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Severe Injury</td>
<td>0</td>
<td>215</td>
<td>525</td>
</tr>
<tr>
<td>Moderate Injury</td>
<td>0</td>
<td>1101</td>
<td>16776</td>
</tr>
<tr>
<td>PDO</td>
<td>0</td>
<td>751</td>
<td>40195</td>
</tr>
</tbody>
</table>

4.2.1.2. Ordered Probit Model (OPM)

The confusion matrix from the OPM is as presented in Table 3. The model correctly predicted two moderate-injury crashes but also misclassified many crashes. Although the performance of both models is poor in this category, the OPM marginally outperformed the OLM. In addition, the model correctly predicted fewer moderate-injury crashes (1,022) and misclassified many as PDO. Therefore, the OLM slightly outperformed the OPM in this category. The model correctly predicted a slightly higher number of PDO crashes (40,410) and misclassified fewer as moderate-injury crashes (536). Thus, the OPM performs better in predicting PDO crashes. The performance of the OLM and OPM is relatively similar in this case, with each outperforming the other in different crash severity categories. However, both models struggle a great deal in correctly identifying moderate-injury crashes and distinguishing between moderate-injury and PDO crashes. These results suggest that further model refinement and exploration of other modeling techniques could be beneficial.

Table 3. Confusion Matrix for the OPM

<table>
<thead>
<tr>
<th>Reference</th>
<th>Severe Injury</th>
<th>Moderate Injury</th>
<th>PDO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Severe Injury</td>
<td>0</td>
<td>236</td>
<td>502</td>
</tr>
<tr>
<td>Moderate Injury</td>
<td>0</td>
<td>1022</td>
<td>16855</td>
</tr>
<tr>
<td>PDO</td>
<td>0</td>
<td>536</td>
<td>40410</td>
</tr>
</tbody>
</table>

4.2.2. Machine Learning (ML) Models

4.2.2.1. Random Forest (RF)

Table 4 shows the confusion matrix from the RF model. Out of the actual severe-injury cases, the model correctly predicted 77 of them (true positives). However, it also misclassified 321 as moderate injury and 342 as PDO (false negatives). For the actual moderate injury cases, the model correctly classified 7,002 of them. On the other hand, it incorrectly classified 109 cases as severe injury and 10,766 cases as PDO. The model performed best in predicting the PDO crashes,
correctly classifying 31,745 cases. However, it misclassified 154 cases as severe injury and 9,047 as moderate injury. While the model offers some value in identifying and categorizing crash severity, its current performance highlights several areas of potential concern in practical applications. These primarily revolve around the misclassification of cases, which could lead to inadequate responses, misallocation of resources, misguided policy decisions, and potential financial implications in an insurance context.

Table 4. Confusion Matrix for the RF

<table>
<thead>
<tr>
<th></th>
<th>Severe Injury</th>
<th>Moderate Injury</th>
<th>PDO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Severe Injury</td>
<td>77</td>
<td>321</td>
<td>342</td>
</tr>
<tr>
<td>Moderate Injury</td>
<td>109</td>
<td>7002</td>
<td>10766</td>
</tr>
<tr>
<td>PDO</td>
<td>154</td>
<td>9047</td>
<td>31745</td>
</tr>
</tbody>
</table>

Figure 3 represents the permutation feature importance as determined by the RF model. The permutation feature importance is a technique for estimating the contributions of individual features to the predictive power of a model by observing the effect on model performance of randomly permuting the values of each feature, one at a time (Altmann et al., 2010). In this study, the RF model was trained on a subset of the dataset. A representative dataset was used for this process because this technique is computationally demanding. In the plot, each bar corresponds to a specific feature in the dataset, and the length of the bar corresponds to the importance of that feature. Positive values indicate that the performance of the model decreases when the feature is shuffled, suggesting that the model relies on the feature to make accurate predictions. Conversely, negative values indicate that the performance of the model improves when the feature is shuffled, suggesting that the model might be overfitting to noise in the feature.

The top three features according to this analysis are crash types, vehicle types, and locality, suggesting that these features are the most important for predicting the severity of a crash according to the trained model. However, it's important to note a few caveats. First, while permutation feature importance provides a useful way to rank the importance of features, it does not provide any information about the nature of the relationship between each feature and the target variable (Altmann et al., 2010). Second, it is possible that important features might appear unimportant if they are highly correlated with other features (Altmann et al., 2010). Finally, because this analysis was performed on a sample of the original dataset, the results might differ if the analysis were performed on the full dataset or a different sample.
4.2.2.1. Extreme Gradient Boosting (XGBoost)

Of the severe-injury cases, the model correctly predicted 139 of them (true positives). However, it misclassified 322 as moderate injury and 279 as PDO (false negatives). This model appears to be better at predicting severe-injury cases than the RF model, with more actual severe-injury cases correctly classified. For the moderate-injury cases, the model correctly classified 8,485 of them. However, it incorrectly classified 297 crashes as severe injury and 9,095 crashes as PDO.

Table 5. Confusion Matrix for the XGBoost

<table>
<thead>
<tr>
<th>Test</th>
<th>Predicted</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Severe Injury</td>
<td>Moderate Injury</td>
<td>PDO</td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>139</td>
<td>322</td>
<td>279</td>
<td></td>
</tr>
<tr>
<td>Severe Injury</td>
<td>297</td>
<td>8485</td>
<td>9095</td>
<td></td>
</tr>
<tr>
<td>PDO</td>
<td>350</td>
<td>11999</td>
<td>28597</td>
<td></td>
</tr>
</tbody>
</table>

4.3. Model Performance Comparison and Analysis of Predictive Power

Table 6 presents the performance metrics of all the developed models. Evaluating model performance is a critical part of the ML process (Ke et al., 2017). By quantifying how well a model
performs on a given dataset, one can understand its strengths and weaknesses, guide the selection of models, tune hyperparameters, and assess the effectiveness of feature selection or engineering (Ke et al., 2017). Without this step, one would not know if our model is effective, if it generalizes well to unseen data, or how it might be improved. When dealing with imbalanced datasets, the choice of evaluation metric becomes particularly important. In this study, the F1 score, alongside precision and recall, was used.
Table 6. Model Performance Metrics

<table>
<thead>
<tr>
<th>Method</th>
<th>Model</th>
<th>Level of Severity</th>
<th>Precision %</th>
<th>Recall %</th>
<th>F1 Score %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OLM</td>
<td>Severe Injury</td>
<td>53.27</td>
<td>6.16</td>
<td>11.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Moderate Injury</td>
<td>69.91</td>
<td>98.17</td>
<td>81.66</td>
</tr>
<tr>
<td></td>
<td>OPM</td>
<td>Severe Injury</td>
<td>100</td>
<td>0.27</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Moderate Injury</td>
<td>56.97</td>
<td>5.72</td>
<td>10.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PDO</td>
<td>69.95</td>
<td>98.69</td>
<td>81.87</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>Severe Injury</td>
<td>22.65</td>
<td>10.41</td>
<td>14.26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Moderate Injury</td>
<td>42.77</td>
<td>85.53</td>
<td>57.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PDO</td>
<td>95.72</td>
<td>77.53</td>
<td>85.67</td>
</tr>
<tr>
<td></td>
<td>XGBoost</td>
<td>Severe Injury</td>
<td>17.68</td>
<td>18.78</td>
<td>18.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Moderate Injury</td>
<td>40.78</td>
<td>47.46</td>
<td>43.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PDO</td>
<td>75.71</td>
<td>69.84</td>
<td>72.47</td>
</tr>
<tr>
<td></td>
<td>MLP</td>
<td>Severe Injury</td>
<td>11.17</td>
<td>36.91</td>
<td>17.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Moderate Injury</td>
<td>5.08</td>
<td>60.44</td>
<td>9.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PDO</td>
<td>98.80</td>
<td>70.02</td>
<td>81.75</td>
</tr>
<tr>
<td></td>
<td>TabNet</td>
<td>Severe Injury</td>
<td>8.45</td>
<td>45.73</td>
<td>28.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Moderate Injury</td>
<td>3.73</td>
<td>72.19</td>
<td>14.54</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PDO</td>
<td>79.62</td>
<td>82.72</td>
<td>92.10</td>
</tr>
</tbody>
</table>
4.3.1. Severe-injury classification

The most effective model for predicting severe-injury crashes is reportedly the TabNet model, which presumably is a form of DL model. This model has a precision of 8.45, which indicates that out of all instances predicted as severe-injury crashes, only 8.45% were severe-injury crashes. The recall is 45.73, meaning that the model correctly identified 45.73% of the actual severe-injury crashes. The F1 score, which is a harmonic mean of precision and recall, and is often used as a single metric to compare models, is 28.78. This relatively low score indicates that the model struggles with the severe-injury classification, which could be due to several reasons such as class imbalance or the complexity of the problem. However, when it comes to predicting crash severity, other studies have found XGBoost to perform better (Guo et al., 2021; Jamal et al., 2021). For instance, a comparative study by Jamal et al. (2021) found that an XGBoost model outperformed other ML models in predicting crash severity. In another study, Pensantez-Narvaez et al. (2019) compared the XGBoost model and logistic regression in crash severity analysis and found XGBoost to perform better. They attributed its success to its ability to handle non-linear relationships and interactions between features, which are common in crash data. While other studies might have found XGBoost to be a better model, the choice of model should be guided by the specific characteristics of the dataset and the problem at hand, and no single model is likely to be the best choice for all tasks (Li et al., 2023).

4.3.2. Moderate-injury classification

The model with the best performance for predicting moderate-injury crashes is RF. However, it has a precision of 42.77, which means that 42.77% of the instances it predicted as moderate-injury crashes were indeed moderate-injury crashes. It has a recall of 85.53, signifying that it correctly identified 85.53% of all actual moderate-injury crashes. The F1 score is 57.03, which is considerably higher than the severe-injury model, indicating a better balance of precision and recall. In previous research on crash severity prediction (Theofilatos et al., 2019), achieving a balance between precision and recall is a common challenge, as the relative importance of false positives and false negatives can vary depending on the specific application and costs associated with prediction errors. The RF model’s high recall indicates its ability to capture a substantial proportion of actual moderate injuries, a crucial consideration in traffic safety applications.

4.3.3. Property Damage Only (PDO) Classification

The model with the highest performance in predicting PDO crashes is TabNet. Its precision is 79.62, implying that 79.62% of instances it predicted as PDO were PDO. The model has a recall of 82.72, implying it correctly identified 82.72% of actual PDO cases. The F1 score is notably high at 92.1, indicating a superior balance of precision and recall for this class.

Hence, the model performances vary significantly across the severity levels, with the PDO predictions faring the best and the severe-injury predictions performing the worst. This
discrepancy in performance may be attributed to various factors, including but not limited to, class imbalance, differing complexities in predicting each class, or nuances in the data distribution for each class.

Figure 4. Precision Comparison for Statistical, ML, and ML Models

![Precision Comparison for Statistical, ML, and ML Models](image1)

Figure 5. Recall Comparison for Statistical, ML, and DL Models

![Recall Comparison for Statistical, ML, and DL Models](image2)
Figure 6 shows the graphical comparison of the performance metrics of different models. The statistical modes show high precision for severe-injury crashes but very low recall, suggesting they are confident but not comprehensive in identifying severe-injury crashes. The F1 score, which balances precision and recall, is also low for severe-injury crashes. For moderate-injury crashes, precision is relatively high, but recall is low, leading to a low F1 score. However, these models perform better for PDO crashes, with a relatively high F1 score. The ML models, on the other hand, provide a more balanced performance across all severity levels. For severe-injury crashes, both precision and recall are lower than for other categories, resulting in a lower F1 score. For moderate injury and PDO crashes, the performance is better with higher F1 scores. The DL models, however, have lower precision for moderate- and severe-injury crashes, suggesting they may incorrectly classify incidents as these categories. However, they have high recall, particularly for PDO and moderate-injury crashes, indicating that they are good at identifying these categories when present. The F1 scores are relatively low for severe- and moderate-injury crashes, but higher for PDO crashes.

While the statistical models are very confident in their predictions of severe-injury crashes (high precision), they miss a lot of actual severe-injury crashes (low recall). This might be acceptable if the goal is to avoid false alarms at the cost of missing some real cases. However, in many applications, such as emergency response or healthcare, missing severe-injury crashes could have serious consequences, making these models less effective for real-world application. For predicting PDO crashes, statistical models perform very well with a high recall and good precision. The ML models offer a more balanced performance, making fewer mistakes in both directions (moderate precision and recall). They may be useful in situations where both false positives and false negatives carry significant costs. For example, if resources are being allocated based on these predictions, it
would be important to both catch as many real cases as possible (high recall) and to avoid wasting resources on false alarms (high precision). The DL models, however, are very sensitive, catching a large proportion of actual incidents (high recall), but they also produce a significant number of false alarms (low precision). This could be useful in a situation where the cost of a false negative is much higher than the cost of a false positive. For example, in a preliminary screening tool, it might be acceptable to have more false alarms to ensure that all or most real cases are identified for further investigation.
5. Conclusion

5.1 Summary of Findings

An extensive evaluation of various statistical, ML, and DL models to predict crash severity was conducted in this study. The analysis began with data preprocessing using SMOTE-N to address imbalances in the crash dataset. The findings revealed that each model category presented unique strengths and weaknesses. Statistical models, including OLM and OPM, demonstrated high precision but low recall in predicting severe-injury crashes. These models excelled in their confidence but lacked comprehensive identification of severe incidents. They were most effective in predicting PDO crashes.

ML models, exemplified by RF and XGBoost, delivered a more balanced performance. They provided moderate precision and recall across different severity levels, making them suitable for situations where the costs of false positives and false negatives are significant. DL models, MLP, and TabNet offered high recall rates but lower precision. These models were particularly effective in identifying severe- and moderate-injury crashes, despite a higher rate of false positives. TabNet stood out as the best performer for predicting both severe-injury and PDO crashes. It can be deduced from this study that the choice of model heavily depends on the specific application and the trade-off between false positives and false negatives. While statistical models provide high precision, ML models offer balanced performance, and DL models yield high recall. Therefore, to optimally predict crash severity, practitioners should carefully consider the context and requirements of their specific application.

The models developed in this study hold substantial promise for enhancing traffic safety in real-world applications. Specifically, the RF model exhibits notable performance in predicting moderate-injury crashes, offering a valuable tool for traffic safety authorities and organizations. Its balanced precision and recall make it suitable for identifying crashes with a moderate level of injury severity, allowing for more targeted allocation of resources, such as emergency response teams and medical personnel, to the crash scenes. This can lead to quicker and more effective responses, potentially reducing the severity of injuries and saving lives. Furthermore, the TabNet model’s exceptional precision and recall in predicting PDO crashes make it a valuable asset for traffic management and insurance agencies. By accurately identifying PDO crashes, these organizations can streamline their claims processes, minimize fraud, and allocate resources more efficiently.

Moreover, the deployment of these models can significantly contribute to proactive traffic safety measures. By leveraging the insights gained from predictive modeling, traffic safety authorities can implement targeted interventions and educational campaigns in areas prone to severe crashes. Additionally, these models can inform the design and implementation of infrastructure improvements, such as road modifications and signage enhancements, to mitigate crash risks effectively. As traffic safety continues to be a paramount concern worldwide, the integration of
predictive models such as RF and TabNet into decision-making processes can lead to more effective and data-driven approaches to reducing injuries and fatalities on our roadways. Ultimately, the utilization of these models empowers stakeholders with the tools they need to make more informed, proactive, and lifesaving decisions in the field of traffic safety.

5.2 Limitations and Future Scope of Work

The study provides valuable insights into the performance of various models for predicting crash severity. The following are limitations and should be further explored:

- Imbalanced data were dealt with using the SMOTE-N technique. Although this method helps balance the classes, it also created synthetic data, which might not represent real-world scenarios accurately. There is a need to further explore the sensitivity of different data-treatment methods to crash data.

- There is a need for further refinement of the developed models, especially in correctly identifying moderate-injury crashes and distinguishing between moderate injury and PDO crashes.

- ML techniques have shown promise in improving prediction accuracy, while statistical models provide interpretability. Future research should focus on addressing data limitations, integrating real-time data sources, and exploring emerging technologies such as sensor data fusion and DL algorithms to enhance crash severity prediction and improve road safety outcomes.
## Abbreviations and Acronyms

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN</td>
<td>Artificial Neural Network</td>
</tr>
<tr>
<td>AUC-ROC</td>
<td>Area Under Curve- Receiver Operating Characteristic</td>
</tr>
<tr>
<td>CNN</td>
<td>Convolutional Neural Network</td>
</tr>
<tr>
<td>DL</td>
<td>Deep Learning</td>
</tr>
<tr>
<td>HSIS</td>
<td>Highway Safety Information System</td>
</tr>
<tr>
<td>ML</td>
<td>Machine Learning</td>
</tr>
<tr>
<td>MLP</td>
<td>Multi-Layer Perceptron</td>
</tr>
<tr>
<td>OLM</td>
<td>Ordered Logit Model</td>
</tr>
<tr>
<td>OPM</td>
<td>Ordered Probit Model</td>
</tr>
<tr>
<td>PDO</td>
<td>Property Damage Only</td>
</tr>
<tr>
<td>ResNet</td>
<td>Residual Neural Network</td>
</tr>
<tr>
<td>RF</td>
<td>Random Forest</td>
</tr>
<tr>
<td>RNN</td>
<td>Recurrent Neural Network</td>
</tr>
<tr>
<td>XGBoost</td>
<td>Extreme Gradient Boosting</td>
</tr>
</tbody>
</table>
References


About the Authors

Abimbola Ogungbire, M.S.

Mr. Ogungbire is pursuing his Ph.D. in the Infrastructure and Environmental Systems (INES) program at the University of North Carolina at Charlotte. His areas of interest are traffic data analytics and visualization.

Panick Kalambay, Ph.D.

Dr. Kalambay recently graduated with his doctoral degree in Civil Engineering at the University of North Carolina at Charlotte. His areas of interest are traffic operations & safety, and transportation planning.

Hardik Gajera, Ph.D.

Dr. Gajera recently graduated with his doctoral degree in Civil Engineering at the University of North Carolina at Charlotte. His areas of interests include traffic operations & safety, traffic simulations, and connected & automated vehicles.

Srinivas S. Pulugurtha, Ph.D., P.E., F.ASCE

Dr. Pulugurtha is currently working as a Professor & Research Director of the Department of Civil and Environmental Engineering at the University of North Carolina at Charlotte.
MINETA TRANSPORTATION INSTITUTE

Founded in 1991, the Mineta Transportation Institute (MTI), an organized research and training unit in partnership with the Lucas College and Graduate School of Business at San José State University (SJSU), increases mobility for all by improving the safety, efficiency, accessibility, and convenience of our nation’s transportation system. Through research, education, workforce development, and technology transfer, we help create a connected world. MTI leads the Mineta Consortium for Transportation Mobility (MCTM) and the Mineta Consortium for Equitable, Efficient, and Sustainable Transportation (MCEEST) funded by the U.S. Department of Transportation, the California State University Transportation Consortium (CSUTC) funded by the State of California through Senate Bill 1 and the Climate Change and Extreme Events Training and Research (CCEETR) Program funded by the Federal Railroad Administration. MTI focuses on three primary responsibilities:

Research
MTI conducts multi-disciplinary research focused on surface transportation that contributes to effective decision making. Research areas include: active transportation; planning and policy; security and counterterrorism; sustainable transportation and land use; transit and passenger rail; transportation engineering; transportation finance; transportation technology; and workforce development. MTI’s research publications undergo expert peer review to ensure the quality of the research.

Education and Workforce Development
To ensure the efficient movement of people and products, we must prepare a new cohort of transportation professionals who are ready to lead a more diverse, inclusive, and equitable transportation industry. To help achieve this, MTI sponsors a suite of workforce development and education opportunities. The Institute supports educational programs offered by the Lucas Graduate School of Business—a Master of Science in Transportation Management, plus graduate certificates that include High-Speed and Intercity Rail Management and Transportation Security Management. These flexible programs offer live online classes so that working transportation professionals can pursue an advanced degree regardless of their location.

Information and Technology Transfer
MTI utilizes a diverse array of dissemination methods and media to ensure research results reach those responsible for managing change. These methods include publication, seminars, workshops, websites, social media, webinars, and other technology transfer mechanisms. Additionally, MTI promotes the availability of completed research to professional organizations and works to integrate the research findings into the graduate education program. MTI’s extensive collection of transportation-related publications is integrated into San José State University’s world-class Martin Luther King, Jr. Library.

MTI FOUNDER
Hon. Norman Y. Mineta

MTI BOARD OF TRUSTEES

Founders, Honorable Norman Mineta***
Founder, Retired Transportation Executive

Karen Philbrick, PhD
Executive Director

Hilary Nixon, PhD
Deputy Executive Director

Asha Weinstein Agrawal, PhD
Education Director

National Transportation Finance Center Director

Brian Michael Jenkins
National Transportation Security Center Director

MTI’s extensive collection of transportation-related publications is integrated into San José State University’s world-class Martin Luther King, Jr. Library.

Disclaimer
The contents of this report reflect the views of the authors, who are responsible for the facts and accuracy of the information presented herein. This document is disseminated in the interest of information exchange. MTI’s research is funded, partially or entirely, by grants from the U.S. Department of Transportation, the U.S. Department of Homeland Security, the California Department of Transportation, and the California State University Office of the Chancellor, whom assume no liability for the contents or use thereof. This report does not constitute a standard specification, design standard, or regulation.