Ali Abbasov
Aleksander Sładkowski
Tofig Babayev (Eds.)

Communications in Computer and Information Science 2767

Problems of Logistics, Management and Operation in the East-West Transport Corridor

4th International Conference, PLMO 2025
Baku, Azerbaijan, May 13–15, 2025
Proceedings

Springer

PLMO 2025

Communications in Computer and Information Science 2767

Ali Abbasov · Aleksander Sładkowski ·
Tofig Babayev
Editors

Problems of Logistics, Management and Operation in the East-West Transport Corridor

4th International Conference, PLMO 2025
Baku, Azerbaijan, May 13–15, 2025
Proceedings

Springer

Editors
Ali Abbasov
Institute of Control Systems
Baku, Azerbaijan

Tofig Babayev
Institute of Control Systems
Baku, Azerbaijan

Aleksander Sładkowski
Silesian University of Technology
Katowice, Poland

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-032-13671-8 ISBN 978-3-032-13672-5 (eBook)
https://doi.org/10.1007/978-3-032-13672-5

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Participant-Level Injury Outcome Prediction in Road Traffic Incidents Using Machine Learning: A Case Study in Poland

Artur Budzyński[1(✉)] and Aleksander Sładkowski[2]

[1] Krakow University of Economics, Rakowiecka 27, 31-510 Kraków, Poland
abudzyns@uek.krakow.pl
[2] Silesian University of Technology, Krasiński 8, Katowice, Poland

Abstract. This study investigates the application of supervised machine learning methods for predicting injury outcomes among participants involved in road traffic incidents. The analysis is based on detailed participant-level data collected in Poland between 2015 and 2022, covering over six million records. The dataset includes individual and incident-related characteristics, such as participant role, gender, driving license status, legal responsibility, and area type. A multi-class classification framework was developed to predict the injury status of participants, categorized as no injury, light injury, severe injury, or fatality. Three machine learning models - Random Forest, XGBoost, and LightGBM - were implemented and evaluated in terms of predictive performance. In addition, an analysis of feature importance was conducted to identify the most influential factors contributing to injury severity. The results demonstrated that ensemble learning models, particularly XGBoost, achieved the highest predictive performance. Participant role and legal responsibility were identified as the most critical factors influencing injury outcomes. The findings confirm the potential of machine learning techniques to improve the understanding of individual-level determinants of injury severity and to support data-driven road safety policies.

Keywords: Road Traffic Incidents · Injury Severity Prediction · Participant-Level Data · Supervised Machine Learning · Ensemble Learning · Feature Importance · Predictive Modelling · Road Safety · Poland

1 Introduction

Road traffic incidents are a major public safety concern, often resulting in injuries or fatalities and generating significant social and economic costs. Numerous studies have examined the factors influencing the severity of road traffic incidents, primarily focusing on incident-level characteristics such as roadway conditions, weather, and vehicle attributes. Early research in this area predominantly employed statistical modelling techniques. For example, Abdel-Aty and Keller applied logistic regression models to analyse crash severity levels at signalized intersections [1], while Yamamoto and Shankar used bivariate ordered probit models to investigate injury severity outcomes in fixed-object collisions [2]. Xie et al. further extended this approach by employing Bayesian ordered probit models to account for uncertainty in injury severity analysis [3].

A. Abbasov et al. (Eds.): PLMO 2025, CCIS 2767, pp. 82–93, 2026.
https://doi.org/10.1007/978-3-032-13672-5_8

In recent years, data-driven approaches based on machine learning techniques have gained increasing attention in road safety research. These methods offer greater flexibility in capturing complex, non-linear relationships between variables and do not require the strict statistical assumptions associated with traditional models. Chang and Chen demonstrated the effectiveness of tree-based models in analysing freeway accident severity [4]. Similarly, Zhang et al. applied various machine learning algorithms to investigate risk factors associated with traffic violations and injury severity in China [5]. Despite the growing popularity of such approaches, most prior studies have primarily focused on incident-level characteristics or environmental conditions, rather than on individual attributes of participants involved in traffic incidents.

While previous research has provided valuable insights into the environmental and incident-related determinants of injury severity, relatively little attention has been devoted to the analysis of participant-level characteristics. Factors such as the role of the participant, driving experience, gender, or legal responsibility may have a significant impact on the severity of injuries sustained in road traffic incidents. However, limited research has explicitly examined these participant-level factors in injury severity analysis, highlighting the need for further investigation in this area.

2 Problem Statement

The primary objective of this study is to investigate the relationship between participant-level characteristics and injury severity in road traffic incidents using machine learning techniques. Recent literature has demonstrated the effectiveness of various machine learning algorithms in predicting crash injury severity and highlighted their potential to improve road safety analysis [6]. The research presented in this paper focuses on analysing a dataset of incidents that occurred in Poland between 2015 and 2022, containing detailed information about the participants involved. The specific research problems addressed in this study are formulated as follows:

1) Can machine learning classification models accurately predict the injury status of participants based on individual and incident-related features?
2) Which participant-level and incident-related factors are the most influential in determining injury severity?
3) To what extent does the role of a participant (e.g., driver, passenger, pedestrian) affect the likelihood of sustaining different levels of injury?
4) How does the predictive performance differ across selected machine learning algorithms?

To answer these questions, the study applies supervised machine learning classification models and evaluates their predictive performance using appropriate metrics. Furthermore, the analysis includes an assessment of feature importance and additional exploratory analyses to better understand the relationship between participants' roles and injury outcomes.

3 Methods

The primary data source for this study is a dataset provided by the Polish Police. It contains detailed information on participants involved in road traffic incidents that occurred in Poland between 2015 and 2022. The dataset was obtained directly from the internal databases of the Polish Police and is not publicly available. Access to the data was granted exclusively for the purpose of this research. In total, the dataset includes over six million records, covering incidents reported across all administrative regions of Poland during the specified period.

The dataset consists of separate CSV files for each year between 2015 and 2022. Each annual file contains data exclusively about the participants involved in road traffic incidents reported in that particular year. All files share a consistent structure, enabling efficient consolidation and subsequent analysis across the entire study period. The CSV format was selected due to its simplicity, wide adoption, and suitability for systematic data analysis [7].

All annual CSV files were consolidated into a single dataset, resulting in over six million records in total. For efficient storage, faster loading times, and optimal performance during data analysis, the combined dataset was saved in the Apache Parquet format, which provides columnar storage and built-in compression. The selection of the Parquet format was motivated by its superior performance characteristics and optimized analytical query capabilities, especially when handling large datasets [8].

The data preparation stage involved transforming two original features, **Killed** and **Injured**, into a single categorical feature named **Injury_status**. Initially, the **Killed** feature included three unique values: “Killed at scene”, “Killed within 30 days”, and “Unspecified”. The **Injured** feature contained values: “Severely injured”, “Lightly injured”, and “Unspecified”. A custom Python function named **determine_injury_status()** was developed and applied conditional logic to each record, merging these two features into one with five mutually exclusive categories: “Killed at scene”, “Killed within 30 days”, “Severely injured”, “Lightly injured”, and “No injury”. The transformation was executed across the entire dataset using the **apply()** method, facilitated by the Pandas library, which provides efficient data manipulation capabilities for structured datasets [9, 10].

Data visualization methods were employed to facilitate exploratory analysis and clearly illustrate key patterns in the dataset. Specifically, categorical data distribution was visualized using bar plots. These visualizations were created utilizing the Python libraries Matplotlib and Seaborn. Matplotlib provides extensive functionality for generating publication-quality graphics in Python, supporting detailed customization of visual outputs [11]. Seaborn, built on top of Matplotlib, was employed due to its intuitive functions tailored specifically for statistical data visualization, enabling concise and informative graphical summaries [12].

Before building the predictive model, the dataset underwent further preprocessing. The feature **Accident_type** was excluded from the analysis to prevent data leakage, as this feature directly describes the nature of the incident (collision or accident), and therefore strongly correlates with participant injury. All remaining categorical features were encoded using Label Encoding, transforming text-based categories into numerical values suitable for modelling [13]. The target variable, **Injury_status**, was also numerically encoded. Subsequently, the dataset was randomly partitioned into training (70%) and testing (30%) subsets. Stratified sampling was used to ensure proportional representation of each injury category in both subsets, following best practices recommended in machine learning literature [14]. This preprocessing stage provided a robust basis for evaluating the predictive performance and generalization capability of the classification model.

A Random Forest classifier [15] was selected for predictive modelling due to its ability to handle large datasets, manage nonlinear relationships, and provide robustness against overfitting. The model was initialized with 100 decision trees (**n_estimators**=100), each limited to a maximum depth of 15 (**max_depth**=15) to control complexity and computational efficiency. The parameter **class_weight**='balanced' was applied to address class imbalance by automatically adjusting the weights of each class inversely proportional to their frequencies. Additionally, the model was trained using all available processor cores (**n_jobs**=−1) for computational efficiency, and the random state parameter (**random_state**=42) was fixed to ensure reproducibility of results. The trained model was subsequently evaluated on the independent test set, and performance metrics including accuracy, precision, recall, and F1-score, were computed and summarized using a classification report and confusion matrix [14].

The predictive performance of the classification model was evaluated using a confusion matrix. A confusion matrix is a widely used visualization tool in classification problems, summarizing prediction outcomes by displaying the number of correct and incorrect predictions for each class in the dataset. The confusion matrix was computed by comparing the model's predicted labels (**y_pred**) to the actual labels (**y_test**) from the test dataset. The resulting matrix was visualized as a heatmap using the **heatmap()** function from the Seaborn library, with explicitly defined class labels to clearly identify each injury category on both axes. This visualization approach facilitates the intuitive interpretation of the classifier's accuracy and helps identify specific classes that the model finds challenging to distinguish correctly [16].

To select the optimal predictive model, three widely-used classification algorithms were compared: Random Forest, XGBoost, and LightGBM. The Random Forest classifier was previously described. XGBoost [17] and LightGBM [18] are both ensemble-based algorithms utilizing gradient boosting techniques, which iteratively build decision trees to minimize prediction error by focusing on misclassified instances from previous iterations. Each model was initialized with consistent parameters: 100 decision trees (**n_estimators**=100) and a maximum depth of 15 (**max_depth**=15). Random states were fixed (**random_state**=42) to ensure reproducibility, and class imbalance was addressed by setting the **class_weight**='balanced' parameter in Random Forest and LightGBM.

Model performance was evaluated using three metrics: accuracy, macro-average F1-score, and weighted-average F1-score. Accuracy measures the overall proportion of correctly classified observations but can be misleading in datasets with significant class imbalance, as previously discussed. Therefore, the F1-score, defined as the harmonic mean of precision and recall, was also used to obtain a balanced measure of model performance. Specifically, the macro-average F1-score calculates the metric independently for each class and then averages these values equally, giving more importance to minority classes. In contrast, the weighted-average F1-score computes class-specific F1-scores and averages them weighted by class frequency, thus emphasizing the dominant classes in imbalanced datasets [19]. Finally, the results of the comparative analysis were visualized using Matplotlib and Seaborn libraries, facilitating clear interpretation of performance differences across evaluated models.

The relative importance of individual features was determined using the feature importance functionality implemented in the XGBoost algorithm. The importance score for each feature was calculated based on the gain criterion, which measures the average improvement in the model's predictive accuracy brought by a feature when used to split a decision node. Specifically, the gain reflects the reduction in the loss function achieved by including the feature in the tree-building process. The overall importance score assigned to each feature is the sum of gain values across all trees in the ensemble, normalized so that the total importance across all features equals one (gain-based feature importance). This approach enables the identification of variables that most significantly contribute to the predictive performance of the model.

To investigate the relationship between participants' roles and injury severity, a bivariate analysis was performed using cross-tabulation. Two contingency tables were constructed: the first captured the distribution of injury status within each participant type, while the second reflected the distribution of participant types within each injury severity category. Both tables were normalized to express the results as relative frequencies, facilitating interpretation of proportional differences. This analytical approach is commonly used in categorical data analysis to explore potential associations between variables [20]. The normalized tables were subsequently visualized using heatmaps to provide a clear and interpretable graphical representation of the observed patterns.

All data preprocessing, exploratory analysis, feature engineering, model training, and evaluation were conducted using the Python programming language within a JupyterLab environment. JupyterLab is an interactive development environment designed for reproducible computational workflows, integrating code, narrative text, and visual outputs in a single document [21].

To ensure full transparency and reproducibility of the analytical process, the complete Jupyter notebook used in this study has been made publicly available on GitHub. GitHub is a widely adopted platform that facilitates collaborative development, version control, and open dissemination of scientific code [22]. The notebook includes all steps of data processing, model implementation, and visualization described in this paper. Due to data privacy restrictions, the dataset itself is not included in the repository. The notebook is available at [23].

4 Results

Figure 1 illustrates the distribution of participants' injury status based on road traffic incidents reported in Poland between 2015 and 2022. The vast majority of participants (over 6 million) did not sustain any injury. Among the injured, the number of lightly injured participants was significantly higher than severely injured ones. Fatalities were notably fewer, with participants killed at the scene slightly more numerous than those who succumbed to their injuries within 30 days after the incident.

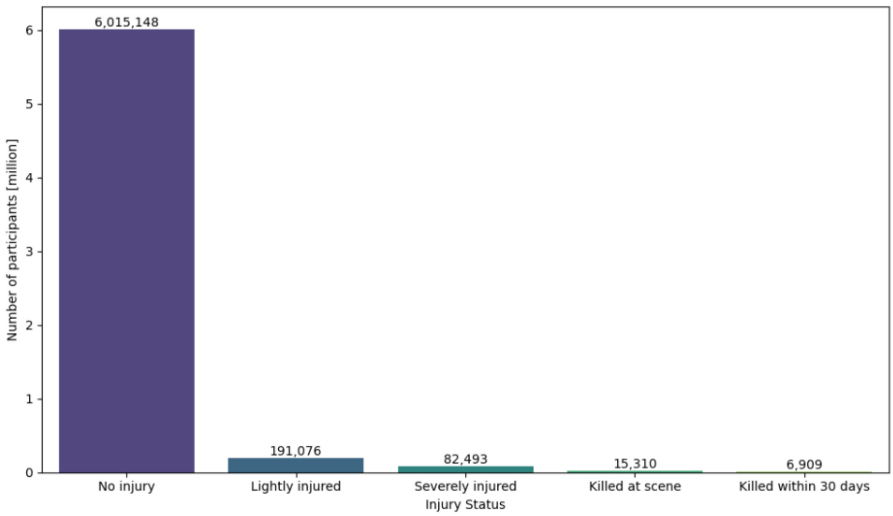

Fig. 1. Distribution of injury status among participants (2015–2022)

Figure 2 presents the confusion matrix illustrating the performance of the Random Forest classifier in predicting participants' injury status. The diagonal values indicate correctly classified observations, while off-diagonal values represent misclassifications. The model demonstrated exceptional accuracy in predicting the No injury category, correctly classifying a vast majority of these instances. However, the classifier faced significant challenges in accurately predicting minority classes, particularly those involving fatalities (Killed at scene, Killed within 30 days) and severe injuries (Severely injured). The relatively high number of misclassifications among these critical categories highlights the inherent difficulty of accurately predicting rare yet significant injury outcomes. This imbalance suggests further exploration into more specialized modeling techniques or additional balancing strategies to improve predictive performance for minority classes.

Fig. 2. Confusion matrix for injury status prediction

Figure 3 presents a comparative analysis of the classification models: Random Forest, XGBoost, and LightGBM. The evaluation was based on three performance metrics: accuracy, macro-average F1-score, and weighted-average F1-score. The XGBoost classifier achieved the highest accuracy (0.96) and weighted F1-score (0.96), clearly outperforming the other models. Additionally, XGBoost obtained the highest macro-average F1-score (0.40), indicating better classification performance across all injury categories, including minority classes. The Random Forest model yielded lower scores, with an accuracy of 0.91 and macro-average F1-score of 0.34, while the LightGBM model demonstrated the weakest performance among the three classifiers. These results confirm the superior predictive capacity of XGBoost in the analysed dataset.

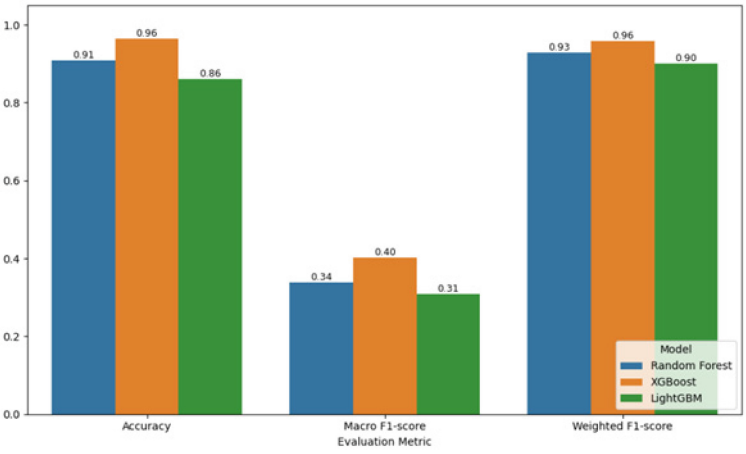

Fig. 3. Comparison of classification models

Figure 4 presents the ranking of the fifteen most important features identified by the XGBoost classifier. The feature importance was determined based on the contribution of each variable to the model's predictive performance. The analysis revealed that the variable **Participant_type** had by far the highest importance score (0.70), indicating its dominant role in predicting participants' injury status. Other features, such as **Driving_license_status** (0.06) and **Legal_decision** (0.03), also demonstrated some predictive relevance, albeit substantially lower. The remaining features contributed marginally to the model's decisions. These results suggest that the type of participant involved in the incident is the most influential factor in determining injury severity, while the predictive value of other attributes is comparatively limited.

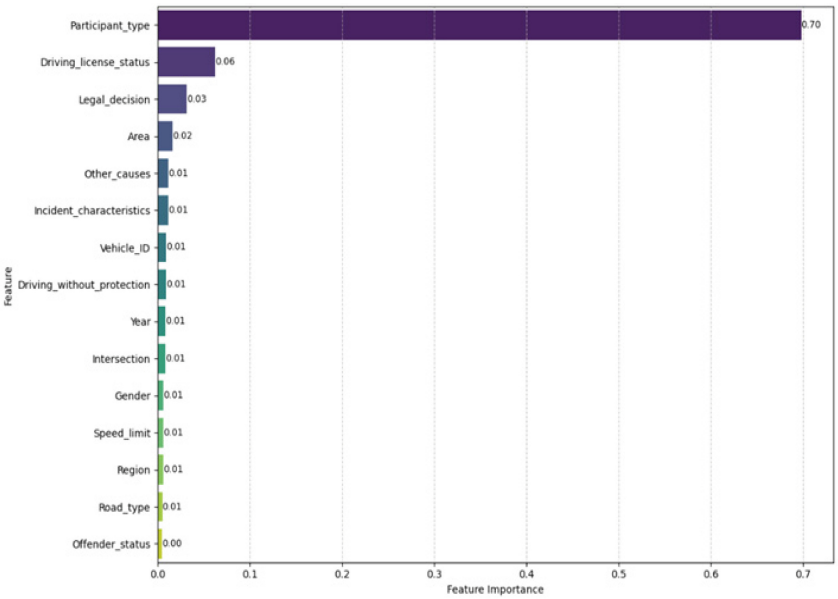

Fig. 4. Top 15 most important features (XGBoost)

A clear association was observed between participants' roles in road traffic incidents and the severity of their injuries. The distribution of injury outcomes, presented in Fig. 5, indicates that drivers and individuals with an unknown participant type were the least affected, with the vast majority (97% and 100%, respectively) classified as uninjured. In contrast, passengers and pedestrians exhibited a substantially higher risk of injury. Among passengers, 66% sustained light injuries and 22% were severely injured. Pedestrians demonstrated the most unfavorable distribution, with nearly half (49%) of cases involving injuries, including 17% classified as severe. Although fatalities were relatively rare, they were more likely to occur among pedestrians and passengers. These findings confirm the strong relationship between participant type and injury severity, consistent with the results of the feature importance analysis.

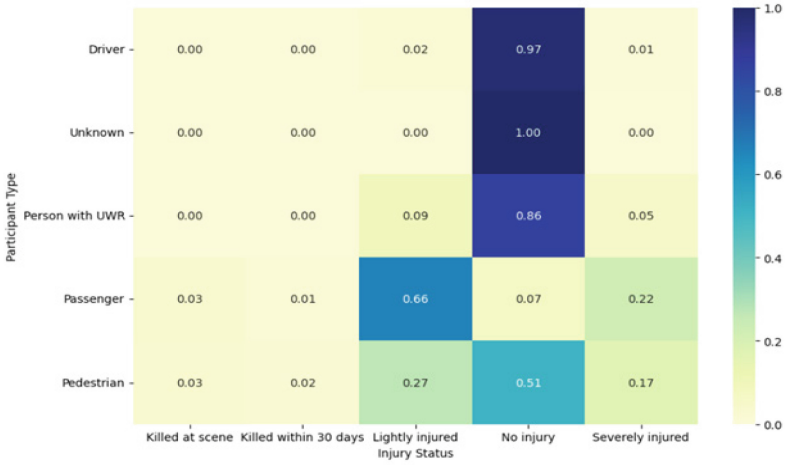

Fig. 5. Distribution of injury status by participant type

An additional analysis was conducted to examine the composition of participant types within each injury severity category. The results, Fig. 6, reveal substantial differences in the distribution of participant roles across injury outcomes. Uninjured participants were predominantly drivers, who accounted for 86% of cases in this category. Conversely, the majority of fatalities occurred among pedestrians and passengers. Specifically, pedestrians represented 51% of participants killed at the scene and 43% of those who died within 30 days of the incident. Passengers also constituted a considerable proportion of fatalities and severe injuries. The category of light injuries was dominated by passengers, who accounted for 67% of cases, followed by pedestrians. These findings further emphasize the disproportionate vulnerability of pedestrians and passengers in road traffic incidents compared to drivers.

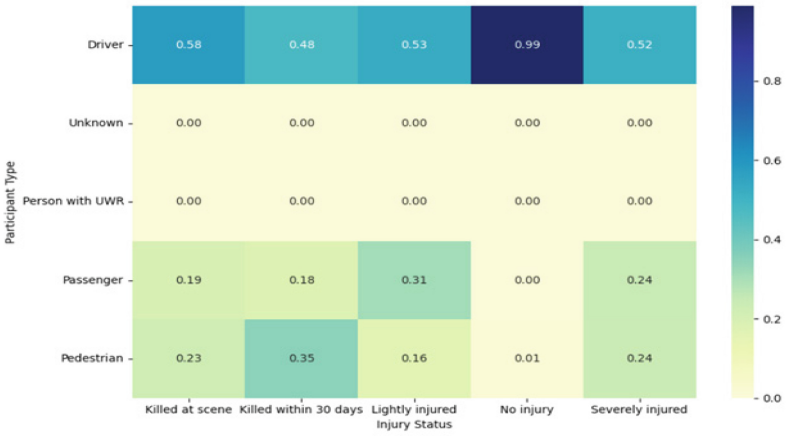

Fig. 6. Distribution of participant type by injury status

5 Discussion and Conclusions

The results of this study demonstrate the effectiveness of machine learning methods in predicting injury severity among participants involved in road traffic incidents. Among the evaluated models, XGBoost achieved the highest predictive performance across all considered metrics, including accuracy, macro-average F1-score, and weighted-average F1-score. This finding aligns with recent research indicating the superior performance of XGBoost in traffic crash severity prediction tasks. For instance, a study by Jiang et al. developed an improved XGBoost model to predict injury severity, achieving notable accuracy improvements over traditional models [24]. Such models are capable of handling complex, non-linear relationships between variables and addressing the challenges posed by imbalanced datasets, which are common in injury severity analysis.

The analysis of feature importance revealed that the role of the participant was the most influential factor in determining injury severity. This observation is consistent with the exploratory analyses conducted in this study, which demonstrated substantial differences in injury outcomes depending on whether the participant was a driver, passenger, or pedestrian. Participants classified as pedestrians were particularly vulnerable, with a significantly higher likelihood of sustaining severe injuries or fatalities. Passengers also exhibited an increased risk of injury compared to drivers, reflecting their passive role in the incident and limited ability to react or avoid danger. Previous studies have highlighted this increased vulnerability among passengers in various traffic scenarios [13]. In addition to participant roles, other individual-level and incident-related variables, such as legal responsibility, driving license status, and area type, were found to moderately influence injury outcomes. These findings confirm the relevance of participant-specific information in injury severity prediction and highlight the importance of considering individual characteristics in road safety analyses.

Despite the overall high predictive accuracy of the developed models, the analysis identified persistent challenges in correctly classifying minority classes, particularly fatalities and severe injuries. The confusion matrix revealed that most misclassifications occurred in these categories, primarily due to the class imbalance inherent in the dataset. This limitation is commonly observed in classification tasks involving rare but critical outcomes and reflects the well-known difficulties associated with learning from imbalanced data. Further research should explore advanced resampling techniques, alternative modelling approaches, or the inclusion of additional explanatory variables to improve the predictive performance for minority classes. Moreover, the availability of more comprehensive participant-level data, including behavioural and contextual factors, could enhance the understanding of the underlying determinants of injury severity. Previous studies have demonstrated that such variables may significantly improve the accuracy of injury severity prediction models. The findings of this study may support policymakers and road safety authorities in developing targeted preventive measures based on individual-level risk factors.

References

1. Abdel-Aty, M., Keller, J.: Exploring the overall and specific crash severity levels at signalized intersections. Accid. Anal. Prev. **37**(3), 417–425 (2005). https://doi.org/10.1016/j.aap.2004.11.002
2. Zhang, Y., Fu, C., Cheng, S.: Exploring driver injury severity at intersection: an ordered probit analysis. Adv. Mech. Eng. **7**(2), 567124 (2015). https://doi.org/10.1155/2014/567124
3. Xie, Y., Zhang, Y., Liang, F.: Crash injury severity analysis using Bayesian ordered probit models. J. Transp. Eng. **135**(1), 18–25 (2009). https://doi.org/10.1061/(ASCE)0733-947X(2009)135:1(18)
4. Chang, L.-Y., Chen, W.-C.: Data mining of tree-based models to analyze freeway accident frequency. J. Saf. Res. **36**(4), 365–375 (2005). https://doi.org/10.1016/j.jsr.2005.06.013
5. Zhang, G., Yau, K.K.W., Chen, G.: Risk factors associated with traffic violations and accident severity in China. Accid. Anal. Prev. **59**, 18–25 (2013). https://doi.org/10.1016/j.aap.2013.05.004
6. Santos, K., Dias, J.P., Amado, C.: A literature review of machine learning algorithms for crash injury severity prediction. J. Saf. Res. **80**, 254–269 (2022). https://doi.org/10.1016/j.jsr.2021.12.007
7. Van Den Burg, G.J.J., Nazábal, A., Sutton, C.: Wrangling messy CSV files by detecting row and type patterns. Data Min. Knowl. Disc. **33**(6), 1799–1820 (2019). https://doi.org/10.1007/s10618-019-00646-y
8. Zeng, X., Hui, Y., Shen, J., Pavlo, A., McKinney, W., Zhang, H.: An empirical evaluation of columnar storage formats (2023). https://arxiv.org/abs/2304.05028
9. McKinney, W.: Data structures for statistical computing in Python. In: Proceedings of the Python in Science Conference, Austin, Texas, pp. 56–61 (2010). https://doi.org/10.25080/Majora-92bf1922-00a
10. Hunter, J.D.: Matplotlib: a 2D graphics environment. Comput. Sci. Eng. **9**(3), 90–95 (2007). https://doi.org/10.1109/MCSE.2007.55
11. Waskom, M.: Seaborn: statistical data visualization. J. Open Source Softw. **6**(60), 3021 (2021). https://doi.org/10.21105/joss.03021
12. Géron, A.: Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd edn. O'Reilly Media, Sebastopol (2019)
13. Kuhn, M., Johnson, K.: Applied Predictive Modeling. Springer, New York (2013). https://doi.org/10.1007/978-1-4614-6849-3
14. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324
15. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. ACM, San Francisco (2016). https://doi.org/10.1145/2939672.2939785
16. Ke, G., et al.: LightGBM: a highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems, pp. 3146–3154 (2017)
17. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Inf. Process. Manag. **45**(4), 427–437 (2009). https://doi.org/10.1016/j.ipm.2009.03.002
18. Agresti, A.: Categorical Data Analysis. Wiley, Hoboken (2002). https://doi.org/10.1002/0471249688
19. Kluyver, T., et al.: Jupyter Notebooks - a publishing format for reproducible computational workflows. In: International Conference on Electronic Publishing, Göttingen (2016). https://api.semanticscholar.org/CorpusID:36928206

20. Tsay, J., Dabbish, L., Herbsleb, J.: Influence of social and technical factors for evaluating contribution in GitHub. In: Proceedings of the 36th International Conference on Software Engineering, pp. 356–366. ACM, Hyderabad (2014). https://doi.org/10.1145/2568225.2568315
21. Budzyński, A.: Injury Severity Analysis. https://github.com/BudzynskiA/injury-severity-analysis. Accessed 20 July 2025
22. Vadhwani, D., Thakor, D.: An improved XGBoost model to predict the injury severity of person in road crash. Int. J. Crashworthiness **30**(2), 115–124 (2025). https://doi.org/10.1080/13588265.2024.2366567
23. Yang, T., Fan, W.D., Song, L.: Modeling pedestrian injury severity in pedestrian-vehicle crashes considering different land use patterns: mixed logit approach. Traffic Inj. Prev. **24**(2), 114–120 (2023). https://doi.org/10.1080/15389588.2022.2156789
24. Lee, C., Abdel-Aty, M.: Comprehensive analysis of vehicle–pedestrian crashes at intersections in Florida. Accid. Anal. Prev. **37**(4), 775–786 (2005). https://doi.org/10.1016/j.aap.2005.03.019