sustainability

MDPI

Article

Comparative Analysis of Parametric and Non-Parametric Data-Driven Models to Predict Road Crash Severity among Elderly Drivers Using Synthetic Resampling Techniques

Mubarak Alrumaidhi [1,2,*], Mohamed M. G. Farag [1,3] and Hesham A. Rakha [1,4]

[1] Center for Sustainable Mobility, Virginia Tech Transportation Institute, Blacksburg, VA 24061, USA
[2] Civil Engineering Department, College of Technological Studies, Public Authority for Applied Education and Training, Shuwaikh 70654, Kuwait
[3] College of Computing and Information Technology, Arab Academy for Science, Technology, and Maritime Transport, Alexandria 1029, Egypt
[4] Charles E. Via, Jr. Department of Civil and Environmental Engineering, Virginia Tech, Blacksburg, VA 24061, USA
* Correspondence: mubarak@vt.edu

Abstract: As the global elderly population continues to rise, the risk of severe crashes among elderly drivers has become a pressing concern. This study presents a comprehensive examination of crash severity among this demographic, employing machine learning models and data gathered from Virginia, United States of America, between 2014 and 2021. The analysis integrates parametric models, namely logistic regression and linear discriminant analysis (LDA), as well as non-parametric models like random forest (RF) and extreme gradient boosting (XGBoost). Central to this study is the application of resampling techniques, specifically, random over-sampling examples (ROSE) and the synthetic minority over-sampling technique (SMOTE), to address the dataset's inherent imbalance and enhance the models' predictive performance. Our findings reveal that the inclusion of these resampling techniques significantly improves the predictive power of parametric models, notably increasing the true positive rate for severe crash prediction from 6% to 60% and boosting the geometric mean from 25% to 69% in logistic regression. Likewise, employing SMOTE resulted in a notable improvement in the non-parametric models' performance, leading to a true positive rate increase from 8% to 36% in XGBoost. Moreover, the study established the superiority of parametric models over non-parametric counterparts when balanced resampling techniques are utilized. Beyond predictive modeling, the study delves into the effects of various contributing factors on crash severity, enhancing the understanding of how these factors influence elderly road safety. Ultimately, these findings underscore the immense potential of machine learning models in analyzing complex crash data, pinpointing factors that heighten crash severity, and informing targeted interventions to mitigate the risks of elderly driving.

Keywords: crash severity; machine learning; resampling techniques; imbalance data; road safety; elderly drivers; transportation safety

Citation: Alrumaidhi, M.; Farag, M.M.G.; Rakha, H.A. Comparative Analysis of Parametric and Non-Parametric Data-Driven Models to Predict Road Crash Severity among Elderly Drivers Using Synthetic Resampling Techniques. *Sustainability* **2023**, *15*, 9878. https://doi.org/10.3390/su15139878

Academic Editors: Juneyoung Park, Elżbieta Macioszek and Victoria Gitelman

Received: 5 April 2023
Revised: 1 June 2023
Accepted: 18 June 2023
Published: 21 June 2023

1. Introduction

1.1. Background

Road crashes are a major global public health problem that has far-reaching impacts on human life and economic growth. According to the World Health Organization (WHO), an estimated 1.35 million individuals die and another 50 million are injured annually in road crashes worldwide [1]. These crashes impose a significant burden on public health systems and economies, resulting in massive economic, social, and human capital losses. Furthermore, the WHO has predicted that road accidents will become the seventh leading cause of death worldwide by 2030 [1].

The severity of road crashes is a crucial factor in the loss of life and disability. A crash's severity is determined by the severity of the injuries sustained by the individuals involved, and this can lead to short- and long-term effects on their lives. The fatality rate in road crashes is a crucial measure of the impact of crashes on human life. The number of deaths and injuries resulting from road crashes places an enormous burden on healthcare systems worldwide, leading to a significant reduction in productivity and economic growth. The economic losses incurred due to crashes are substantial, leading to a reduction in the gross domestic product (GDP) of most nations [2].

The cost of road crashes to the world's economy is significant. The economic losses due to crashes are both direct and indirect. The direct losses include medical expenses, emergency services, insurance costs, and legal costs. Indirect losses include the impact on the economy, such as a reduction in productivity, absence from work, and the loss of life, leading to a decrease in the quality of life. According to the World Bank, road crashes cost most countries approximately 2–5% of their GDP [2].

1.2. Elderly Drivers

As the global population ages, the number of elderly drivers on the road continues to rise [3,4]. This demographic shift highlights the importance of road safety for seniors, particularly concerning the severity of crashes involving elderly drivers. Understanding the factors contributing to crash severity among elderly drivers is crucial for developing effective interventions to mitigate risks and enhance the safety of all road users.

The National Highway Traffic Safety Administration (NHTSA) indicates that senior drivers are at a higher risk of being involved in deadly crashes than their younger counterparts overall [5]. The age group of 65 and above demonstrates the most significant growth in fatal crash numbers when contrasted with other age groups. Moreover, the risk of injuries and fatalities among elderly drivers in crashes continues to escalate, with the death rates for these drivers and their passengers surpassing those of any other vehicular accident victims in terms of severity [5]. The Insurance Institute for Highway Safety (IIHS) states that drivers who are 70 years old or older exhibit elevated fatal crash rates per mile driven in comparison to middle-aged drivers [6]. This age group tends to face deteriorating physical, cognitive, and visual capabilities, which may contribute to a greater likelihood of severe injuries or death during crashes. The IIHS also emphasizes that the death rates for elderly drivers and their passengers surpass all other road-accident-related fatalities in terms of severity.

Moreover, the Korean Road Traffic Authority reported a significant increase in the number of traffic crash fatalities involving elderly drivers (aged over 64) in the Republic of Korea between 2011 and 2015 [7]. Specifically, the authority found a 34.7% increase in fatalities during this period, highlighting the growing importance of understanding the factors that contribute to crash severity among older drivers. This demographic shift has drawn attention to the issue of road safety for seniors, particularly in relation to the severity of crashes involving elderly drivers.

As individuals age, their risk factors for traffic crashes increase, including decreased ability to cope with complex traffic conditions, reduced driving stability, and slower reaction times, particularly at intersections [8–10]. These age-related declines in physical, cognitive, and visual abilities can pose significant challenges for elderly drivers on the road, increasing the likelihood of crashes and injuries [11].

In addition to the challenges posed by age-related deficits, the presence of comorbidities can exacerbate the risks associated with driving. For example, drivers with impaired vision, dementia, or Parkinson's disease may experience declines in their ability to make safe and appropriate driving decisions, placing themselves and other road users in danger [12,13]. Furthermore, the use of multiple medications can also affect driving performance by causing drowsiness or impairing reaction times [14].

Given the risks associated with elderly driving, it is important to understand the factors that contribute to crash severity in this population. Previous research has identified several

such factors, including driver-related factors, such as age, gender, medical history, and driving experience [15,16], as well as environmental factors, such as road conditions, traffic patterns, and weather conditions [17]. However, despite these efforts, the identification of the most influential factors remains an area of ongoing investigation.

1.3. Applications of Machine Learning Models in Crash Severity Prediction

Machine learning models have emerged as powerful tools for analyzing complex datasets and predicting outcomes across diverse sectors [18–21]. Their application to road safety, as demonstrated in this study, enables a comprehensive understanding of factors contributing to crash severity [16,22–29]. These models can discern patterns in large, complex datasets and assist in elucidating factors contributing to crash severity among elderly drivers. By analyzing factors such as driver demographics, vehicle type, road conditions, and traffic patterns, machine learning models can provide insights into the complex interactions that contribute to crash severity and help identify opportunities for intervention and prevention.

The application of machine learning models to the analysis of crash severity among elderly drivers is an area of active research, with studies using a variety of approaches to model the factors that contribute to crash severity. Several studies have examined the involvement of senior drivers in traffic accidents [16,22,28]. These studies suggest that older drivers are at a greater risk of being injured or killed in car crashes. For a more comprehensive understanding of the topic, readers are encouraged to refer to the detailed literature review presented in [30].

Despite the potential of machine learning models in predicting crash severity, there are limitations to their use in imbalanced data, particularly in the case of elderly drivers. One major challenge is the issue of class imbalance, where the number of instances in one class (e.g., severe crashes) is much smaller than the number of instances in another class (e.g., non-severe crashes). This imbalance can lead to biased predictions and poor model performance [16,23,31,32]. It is essential to take these limitations into account when employing machine learning models for crash severity prediction among elderly drivers and to implement suitable strategies to effectively address these challenges.

A limited number of studies have applied resampling techniques to address the issue of imbalanced datasets in crash severity prediction among elderly drivers [16,22]. In their research, ref. [16] utilized random undersampling of the majority class (RUMC) as a technique to balance the dataset, which subsequently improved the performance of multiple models, including multinomial and ordered random forests, as well as multinomial and ordered logistic regressions. The study emphasizes the importance of using resampling strategies to improve the accuracy of severe crash prediction with machine learning models, particularly in the context of crash severity prediction among elderly drivers. However, using traditional resampling techniques like RUMC can result in the loss of crucial information [23,32]. Consequently, it is vital to explore synthetic resampling strategies as an alternative approach to effectively tackle this issue [33–36]. Despite the potential benefits of synthetic resampling strategies, comprehensive comparative analyses assessing the predictive power of parametric and non-parametric machine learning techniques in conjunction with such strategies remain scarce.

Overall, using machine learning models to predict crash severity among elderly drivers offers valuable insights into the complex factors contributing to crashes in this population group. By identifying the most influential factors, these models can help develop effective interventions to mitigate risks and improve the safety of all road users. Additionally, the development of accurate models can inform policy decisions related to licensing requirements for elderly drivers and infrastructure design to accommodate older drivers' needs.

1.4. Research Objectives and Novelties

In this research, the primary objectives and contributions are as follows:

1. Compare the performance of parametric (logistic regression and LDA) and non-parametric (random forest and XGBoost) machine learning models in predicting crash severity among elderly drivers, utilizing crash data from the Commonwealth of Virginia (USA) between 2014 to 2021. We assess model performance employing various metrics such as accuracy, sensitivity, specificity, balanced accuracy, and geometric mean.
2. Investigate the impact of class imbalance on the predictive performance of these models and evaluate the potential benefits of employing synthetic resampling techniques, specifically random over-sampling examples (ROSE) and synthetic minority over-sampling technique (SMOTE), to address this issue.
3. Assess the impact of training the machine learning models on original, ROSE-balanced, and SMOTE-balanced datasets on their generalization capabilities when facing unseen data by comparing cross-validation and test dataset results.
4. Identify the most effective combination of machine learning models and resampling techniques that provides the best predictive performance in terms of sensitivity, specificity, balanced accuracy, and geometric mean.
5. Evaluate the effect of various contributing factors on crash severity among elderly drivers, providing guidance for risk mitigation and safety improvement strategies.
6. Provide insights and recommendations for future research and practical applications of machine learning models and resampling techniques in the field of crash severity prediction and traffic safety management.

The novelty of this work is primarily underscored by the distinctive application of advanced oversampling techniques, specifically synthetic minority over-sampling technique (SMOTE) and random over-sampling examples (ROSE), in the context of road crash severity prediction among elderly drivers. This approach effectively counters the significant issue of class imbalance inherent in crash datasets, a hurdle that has traditionally posed challenges in obtaining accurate and meaningful predictive results. The use of these techniques, in conjunction with powerful machine learning models, represents a novel contribution to the field, offering a robust methodology for enhanced prediction accuracy.

Furthermore, this study provides a unique perspective by targeting a critical but often underrepresented demographic in crash severity research—elderly drivers. It delves into a granular examination of various factors contributing to crash severity among this group, shedding light on the complex interplay of variables that culminate in severe crash outcomes. By focusing on this specific demographic, the research manages to provide a more nuanced understanding of crash severity determinants, thereby filling a crucial gap in the existing literature. This targeted approach, coupled with methodological innovations, is what sets this work apart in the domain of crash severity prediction.

The combined novelty of this research lies in its targeted exploration of elderly driver crash severity, along with the innovative application of advanced oversampling techniques for improving prediction accuracy, thereby enriching the field of crash severity studies with valuable insights and methodological advancements.

1.5. Structure of the Paper

The remainder of this paper is organized as follows: The "Methodology" section provides a detailed outline of the comprehensive research framework used in this study. It starts with a description of the dataset, followed by an explanation of the resampling techniques employed. This section also elaborates on the implementation of both parametric and non-parametric machine learning models, the application of K-fold cross-validation, and the evaluation metrics used to assess the predictive performance of these models.

The "Results and Discussion" section presents the obtained results and provides an in-depth discussion on them. This section begins with a comparison between cross-validation and test results, followed by the outcomes of the crash severity models. It further assesses the effectiveness of the applied resampling techniques on the performance of the predictive

models. A subsection is devoted to examining the effect of influential factors on crash severity levels.

Finally, the "Conclusions" and "Study Limitations and Future Directions" sections, respectively, summarize the study's key findings and their implications, and acknowledge the limitations of the current study, while suggesting potential avenues for future research in this field.

2. Methodology

2.1. Research Framework

In this study, the analysis and operational process, as shown in Figure 1, outlines the steps undertaken to address the research objectives. Initially, the original datasets were preprocessed to remove outliers and irrelevant cases. Categorical variables were transformed into indicator variables using one-hot encoding, while numeric variables were normalized to a range between zero and one. The dataset was subsequently partitioned into training (70%) and test sets (30%), employing stratified sampling to maintain the distribution of the outcome variable, as illustrated in Figure 1.

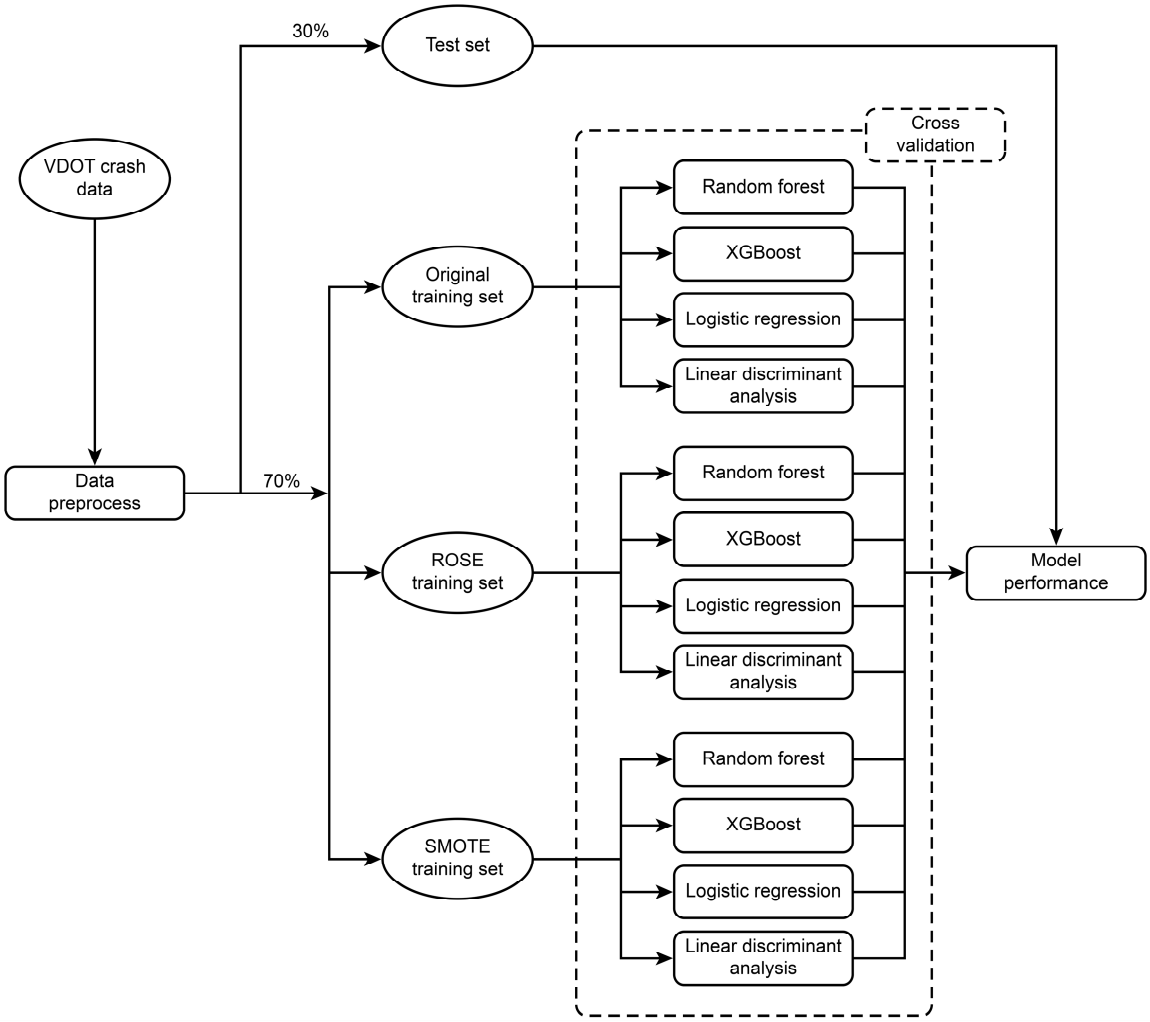

Figure 1. Research framework.

The training set was further divided into a balanced sub-training set to facilitate the application of learning algorithms on balanced datasets. Both parametric and non-parametric machine learning models were trained using the original training set (imbalanced class)

and the balanced sub-training sets. Twelve classifiers were employed to predict crash severity on the test set, and their performance was evaluated and compared.

Alongside prediction, an important facet of the study focuses on the influence of various contributing factors on crash severity. For this purpose, the best-performing model was used to calculate the effect of these contributing factors on crash severity.

This comprehensive methodology ensures a robust investigation of crash severity prediction, taking into account data preprocessing, partitioning, and model evaluation. The results contribute to both predictive accuracy and a deeper understanding of the contributing factors affecting crash severity.

2.2. Data Description

The crash dataset utilized in this study was acquired from the Virginia Department of Transportation (VDOT) and spans an eight-year period from 2014 to 2021. It encompasses a total of 986,101 reported crashes within the Commonwealth of Virginia. The study's primary focus is to analyze crash severity involving motor vehicle accidents with senior drivers aged 65 years and older, as defined by the National Highway Traffic Safety Administration [37]. The sub-dataset consists of 157,800 crashes involving senior drivers, all of which were included in the analysis [17].

Crash severity levels were represented using the KABCO scale, which was further categorized into two groups: non-severe crashes (labeled as O, B, and C) and severe crashes (marked as K + A). Table 1 provides the descriptive statistics for the eighteen variables considered in the analysis. These variables comprise area type, alcohol use, animal involvement, seatbelt usage, bicycle involvement, crash type, distraction, drowsiness, drug usage, pedestrian involvement, posted speed limit, roadway alignment, roadway type, speed violation, time of the week (weekend or weekday), presence of a traffic signal, and weather conditions.

Table 1. Variables' descriptive statistics.

Variable	Category	Count	Percentage
Crash severity	Non-severe	148,473	94.09%
	Severe injury	9327	5.91%
Crash type	Fixed-object	13,399	8.49%
	Head-on	3813	2.42%
	Overturned	1156	0.73%
	Other	10,479	6.64%
	Rear-end	52,953	33.56%
	Sideswipe	17,471	11.07%
	Angle	58,529	37.09%
Traffic signal	Yes	40,998	25.98%
	No	116,802	74.02%
Weather condition	No adverse condition	137,196	86.94%
	Adverse condition	20,604	13.06%
Roadway alignment	Straight	142,472	90.29%
	Curve	15,328	9.71%
Roadway type	Two-way divided	91,375	57.91%
	Two-way undivided	62,206	39.42%
	One-way	4219	2.67%
Work zone	No	153,567	97.32%
	Yes	4233	2.68%
Alcohol	Yes	3483	2.21%
	No	154,317	97.79%
Belted	No	4271	2.71%
	Yes	153,529	97.29%

Table 1. *Cont.*

Variable	Category	Count	Percentage
Bike	Yes	915	0.58%
	No	156,885	99.42%
Distracted	Yes	28,054	17.78%
	No	129,746	82.22%
Drowsy	Yes	2733	1.73%
	No	155,067	98.27%
Drug	Yes	716	0.45%
	No	157,084	99.55%
Pedestrian	Yes	1605	1.02%
	No	156,195	98.98%
Speed violation	Yes	20,211	12.81%
	No	137,589	87.19%
Area type	Urban	121,884	77.24%
	Rural	35,916	22.76%
Animal	Yes	5060	3.21%
	No	152,740	96.79%
Posted speed (mph)	-	157,800	-
Weekend	Yes	32,622	20.67%
	No	125,178	79.33%

2.3. Resampling Techniques

Imbalanced datasets, where the number of instances in one class is significantly smaller than the number of instances in the other class(es), can lead to biased models that underperform on the minority class [16,23,32,33,38]. When the dataset is imbalanced, the model's focus may shift towards the majority class, leading to inaccurate predictions for the minority class. This can result in significant consequences in real-world applications, such as in medical diagnosis, fraud detection, or crash severity prediction [23,39].

Therefore, to address this issue, resampling strategies can be used to improve the model's performance on the minority class. Resampling strategies involve adjusting the dataset's class distribution by either over-sampling the minority class or under-sampling the majority class [16,38].

The over-sampling technique involves generating a balanced dataset by duplicating instances from the minority class randomly until the desired ratio is achieved [38]. The advantage of over-sampling is that it does not lead to any loss of information [32]. However, despite being widely used, over-sampling may not be effective in improving recognition of the minority class and can result in overfitting [23,40].

However, the undersampling method involves randomly removing instances from the majority class until the desired ratio between the classes is achieved [38]. This approach offers the advantage of reducing the size of the training data when dealing with large datasets. However, the removal of instances from the majority class may lead to a loss of valuable information and potentially result in a less representative sample [23,32]. Therefore, careful consideration is needed when using undersampling as a balancing strategy.

In this study, the focus was on overcoming the limitations of traditional resampling techniques by utilizing two synthetic resampling methods: the synthetic minority oversampling technique (SMOTE) and random over-sampling examples (ROSE). The goal was to enhance the model's ability to learn from a more balanced dataset and generate more accurate predictions for severe crashes. By adopting these techniques, it was possible to improve the model's performance and mitigate the risk of biased predictions. This facilitated better outcomes in real-world applications, particularly in the context of elderly driver crash severity prediction.

2.3.1. Synthetic Minority Over-Sampling Technique (SMOTE)

SMOTE (synthetic minority over-sampling technique) is a widely used data augmentation technique to address class imbalance in machine learning [41]. This technique generates synthetic samples in the feature space around the minority class by creating new instances between existing minority class samples. SMOTE works by selecting a minority class sample and computing its k nearest neighbors. It then randomly selects one of these k neighbors and generates a new sample at a point along the line connecting the original minority class sample and its chosen neighbor. This process is repeated until the desired level of over-sampling is achieved. In this study, the parameter k, which represents the number of nearest neighbors, was set to 5, and the amount of synthetic samples was determined such that the classes become balanced.

One of the significant benefits of SMOTE is that it can effectively address the issue of overfitting by generating synthetic samples that are close to the existing minority class samples, but not identical to them. SMOTE can help improve the performance of machine learning algorithms by increasing the representation of minority class samples in the training data. This approach has been shown to improve the accuracy and robustness of the models trained on imbalanced data [29,34,35].

2.3.2. Random Over-Sampling Examples (ROSE)

ROSE (random over-sampling examples) is a data augmentation technique used to address class imbalance in machine learning [31]. Like other resampling techniques, ROSE works by generating synthetic samples to increase the representation of the minority class in the training data.

ROSE is a three-step process. First, the majority class is undersampled using a bootstrap resampling technique, which removes instances from the majority class to create a more balanced dataset. Second, the minority class is over-sampled by generating synthetic samples in the feature space around the minority class. Finally, a new synthetic training dataset is created that is approximately the same size as the original dataset.

The generation of synthetic samples in ROSE is conducted by taking each minority class sample and identifying its k nearest neighbors in the feature space. The synthetic samples are then generated by randomly choosing one of the k neighbors and creating a new sample in the direction of that neighbor. The distance between the original sample and its new synthetic sample is determined by a function provided by the ROSE package in the R program [31].

One of the benefits of ROSE is that it can generate synthetic samples that are representative of the minority class but not identical to the original samples. This can help prevent overfitting and improve the model's generalization ability. Additionally, ROSE can help address the issue of class imbalance in a more effective way than traditional resampling techniques [31].

Research has indicated that generating synthetic data to balance an imbalanced dataset is a viable alternative to traditional resampling techniques, such as over-sampling and undersampling. This approach is believed to reduce the risk of overfitting and enhance the generalization ability that may be compromised by over-sampling methods [33,36].

Table 2 provides a detailed overview of the distribution of each class, including their respective shares and frequencies, for both the original training set and the augmented datasets generated using the SMOTE and ROSE techniques. Additionally, the table also includes the corresponding distributions for the test dataset. As can be observed, the original training set exhibited a highly imbalanced distribution of the classes, with only a small fraction of instances belonging to the severe crash class. However, after applying the SMOTE and ROSE techniques, the resulting training sets showed a more balanced distribution of classes. To ensure that the models' performance was evaluated under realistic conditions, the test dataset's distribution was kept the same as the original dataset. Overall, these augmented datasets enable the machine learning models to learn from a

more representative sample and potentially improve their ability to accurately predict the minority class [16,29,35,35,38].

Table 2. The distribution of crash severity classes in the training and test datasets.

Crash Severity Class	Training Data (Original)	Training Data (ROSE)	Training Data (SMOTE)	Test Data
Severe	6529 (5.9%)	55,197 (50%)	97,935 (48.5%)	2798 (5.9%)
Non-severe	103,932 (94.1%)	55,264 (50%)	103,932 (51.5%)	44,541 (94.1%)
Total	110,461	110,461	201,867	47,339

2.4. Parametric Machine Learning Models

2.4.1. Logistic Regression (LR)

Logistic regression is a well-known parametric machine learning algorithm used to analyze the relationship between a binary dependent variable and one or more independent variables [36]. It is a type of regression analysis that models the probability of an event occurring given a set of predictor variables.

The logistic regression model works by fitting a logistic function to the data, which produces an S-shaped curve. The logistic function transforms the linear combination of the predictor variables into a probability value between 0 and 1, which represents the likelihood of the event occurring.

One of the main advantages of logistic regression is its simplicity and interpretability. The output of logistic regression is a set of coefficients that represent the impact of each predictor variable on the outcome. These coefficients can be interpreted as the change in the log-odds of the outcome for a one-unit increase in the predictor variable, holding all other variables constant.

Logistic regression can handle both numerical and categorical predictor variables. Additionally, it can be used to model interactions between predictor variables and to perform variable selection to identify the most important predictors.

2.4.2. Linear Discriminant Analysis (LDA)

Linear discriminant analysis (LDA) is a parametric machine learning technique used for classification and dimensionality reduction [42–45]. It is a supervised learning algorithm that seeks to identify the underlying linear discriminants that separate different classes in the dataset.

The goal of LDA is to find a linear combination of the predictor variables that maximizes the separation between the classes. The linear combination is calculated by projecting the data onto a lower-dimensional space while maximizing the between-class variance and minimizing the within-class variance. This results in a set of linear discriminant functions that can be used to classify new data points.

LDA remains a popular choice for classification and dimensionality reduction due to its simplicity, interpretability, and computational efficiency. It has been successfully applied in various domains, including pattern recognition, computer vision, and bioinformatics [43–45].

2.5. Non-Parametric Machine Learning Models

2.5.1. Random Forest (RF)

Random forest is a non-parametric machine learning algorithm that is commonly used for classification and regression tasks [16,46,47]. It is an ensemble learning method that combines multiple decision trees to improve the accuracy and robustness of the model.

The random forest algorithm works by constructing a large number of decision trees on bootstrapped samples of the data, where each tree is trained on a random subset of the predictor variables. During training, at each node of each tree, the algorithm chooses the best split among a random subset of predictor variables, rather than considering

all variables. This randomness helps to avoid overfitting and can improve the model's generalization ability.

Once the individual decision trees are built, the random forest algorithm combines their predictions to make a final prediction. For classification tasks, the majority vote of the individual trees is used, while for regression tasks, the average of the individual tree predictions is used.

One of the advantages of random forest is that it can handle large datasets with a large number of predictor variables. It can also handle missing data and outliers, and it is relatively insensitive to the choice of hyperparameters.

2.5.2. eXtreme Gradient Boosting (XGBoost)

XGBoost (extreme gradient boosting) is a popular non-parametric machine learning algorithm used for classification and regression tasks [46,48]. It is an ensemble learning method that combines multiple weak learners (usually decision trees) to produce a more accurate and robust prediction.

In XGBoost, each decision tree is built sequentially, where each new tree is trained to correct the errors of the previous tree. During training, the algorithm assigns weights to each training instance to emphasize the samples that were incorrectly classified by the previous tree.

One of the advantages of XGBoost is that it can handle missing data, and it is relatively robust to outliers. Additionally, XGBoost has a built-in regularization parameter that helps to prevent overfitting. Another advantage of XGBoost is its ability to handle large datasets with a large number of predictor variables. The algorithm supports parallel processing, which can help speed up the training time significantly.

2.6. K-Fold Cross-Validation

K-fold cross-validation is a widely used technique in machine learning to evaluate the performance of a predictive model [49]. It involves dividing the dataset into multiple folds, typically between 5 and 10, and performing model training and testing iteratively. In each iteration, one-fold is used as a validation set to evaluate the model's performance, while the remaining folds are used as the training set. The process is then repeated for each fold, and the results are averaged to obtain an estimate of the model's performance on unseen data.

Cross-validation helps to avoid overfitting, which is when the model performs well on the training data but poorly on new, unseen data. By repeatedly testing the model on different subsets of the data, cross-validation provides a more reliable estimate of the model's performance on new data. It also helps to ensure that the model is not biased towards a specific subset of the data and that it can generalize well to new, unseen data.

Overall, cross-validation is a valuable tool in machine learning for evaluating the performance of models, selecting hyperparameters, and ensuring that the model can generalize well to new data.

2.7. Evaluation Metrics

Performance metrics are an essential component of machine learning (ML) models, providing a quantitative evaluation of how well the model is performing. The selection of appropriate performance metrics depends on the specific application and goals of the model.

2.7.1. Confusion Matrix

A confusion matrix, as shown in Table 3, is a fundamental tool used to assess the effectiveness of a classification model by comparing the predicted class labels with the actual class labels. This matrix is especially valuable in cases where there are two or more potential classes. Essentially, the matrix is presented as a table with the actual and predicted class labels represented in rows and columns, respectively. A confusion matrix comprises four fundamental components: true positives (TP), false positives (FP), true negatives

(TN), and false negatives (FN). TP refers to instances where the model accurately predicted the positive class, while FP denotes cases where the model inaccurately predicted the positive class. TN indicates instances where the model accurately predicted the negative class, while FN pertains to cases where the model inaccurately predicted the negative class. Using the confusion matrix, several metrics can be derived to evaluate the model's overall performance. The most commonly utilized metrics include accuracy, sensitivity, and specificity. Given that imbalanced data are present, it is recommended to employ additional metrics such as geometric mean and balanced accuracy to ensure a more comprehensive assessment of the model's performance.

Table 3. Confusion matrix for evaluating model's performance.

Predicted Class	**Actual Class**	
	Positive	Negative
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

2.7.2. Accuracy

Accuracy is a metric that measures the percentage of correct predictions made by the model. It is calculated by dividing the number of correct predictions by the total number of predictions. Accuracy can be a useful metric for balanced datasets, where the number of positive and negative examples is roughly equal. However, it can be misleading in cases where the dataset is imbalanced, and the model may achieve high accuracy by simply predicting the majority class.

$$\text{Accuracy } = \frac{TP + TN}{TP + TN + FP + FN}$$

2.7.3. Sensitivity

Sensitivity, also known as true positive rate, measures the percentage of positive examples that the model correctly identifies. It is calculated by dividing the number of true positives by the sum of true positives and false negatives. Sensitivity is a useful metric for datasets where the positive class is of particular interest, such as in medical diagnosis and crash severity.

$$\text{Sensitivity } = \frac{TP}{TP + FN}$$

2.7.4. Specificity

Specificity measures the percentage of negative examples that the model correctly identifies. It is calculated by dividing the number of true negatives by the sum of true negatives and false positives. Specificity is a useful metric for datasets where the negative class is of particular interest, such as in fraud detection.

$$\text{Specificity } = \frac{TN}{TN + FP}$$

2.7.5. Geometric Mean

Geometric mean is a metric that measures the geometric average of sensitivity and specificity. It is calculated by taking the square root of the product of sensitivity and specificity. The geometric mean is a useful metric for imbalanced datasets, where the positive and negative classes have different prevalences [50].

$$\text{Geometric Mean} = \sqrt{\text{Sensivity} \times \text{Specifity}}$$

2.7.6. Balanced Accuracy

Balanced accuracy is a metric that takes into account the balance between the positive and negative classes. It is calculated as the average of sensitivity and specificity, which is equivalent to the geometric mean when the dataset is balanced. Balanced accuracy is a useful metric for imbalanced datasets, where accuracy alone may be misleading.

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specifity}}{2}$$

It is important to note that accuracy may not be an appropriate metric for imbalanced datasets, as it can be heavily influenced by the majority class, leading to a falsely optimistic evaluation of the model's performance [16,27]. Therefore, using a combination of the above metrics is often necessary to provide a more comprehensive evaluation of the model's performance on imbalanced datasets.

3. Results and Discussion

3.1. Comparison between Cross-Validation and Test Results

Figures 2–13 present the confusion matrices for four different models: random forest, XGBoost, logistic regression, and LDA. These models were trained using the original dataset, a balanced dataset using the random over-sampling examples (ROSE) technique, and a balanced dataset using the synthetic minority over-sampling technique (SMOTE). The performance of the models was evaluated on both the cross-validation (CV) dataset and the test dataset. Comparing the CV results and test results provides insight into the models' generalization capability when facing unseen data.

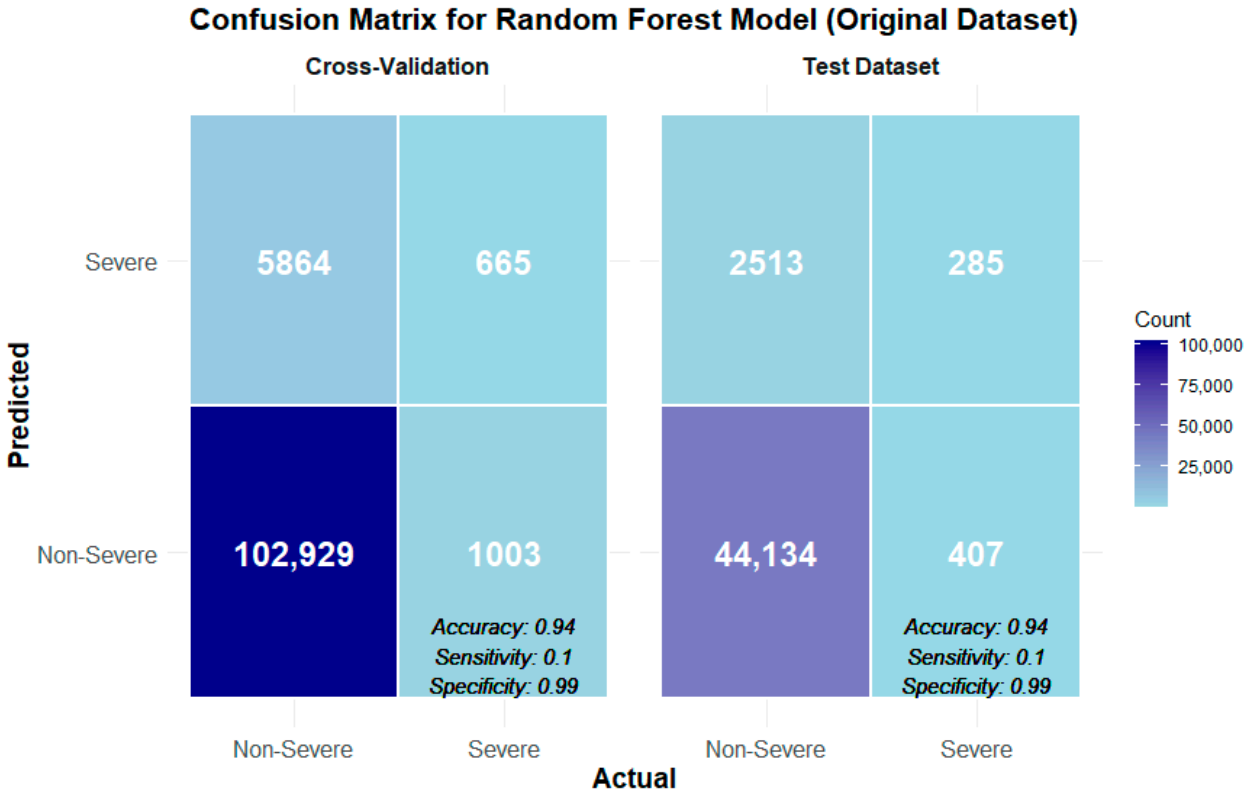

Figure 2. Confusion matrices for random forest model that used original dataset for training (training and testing data).

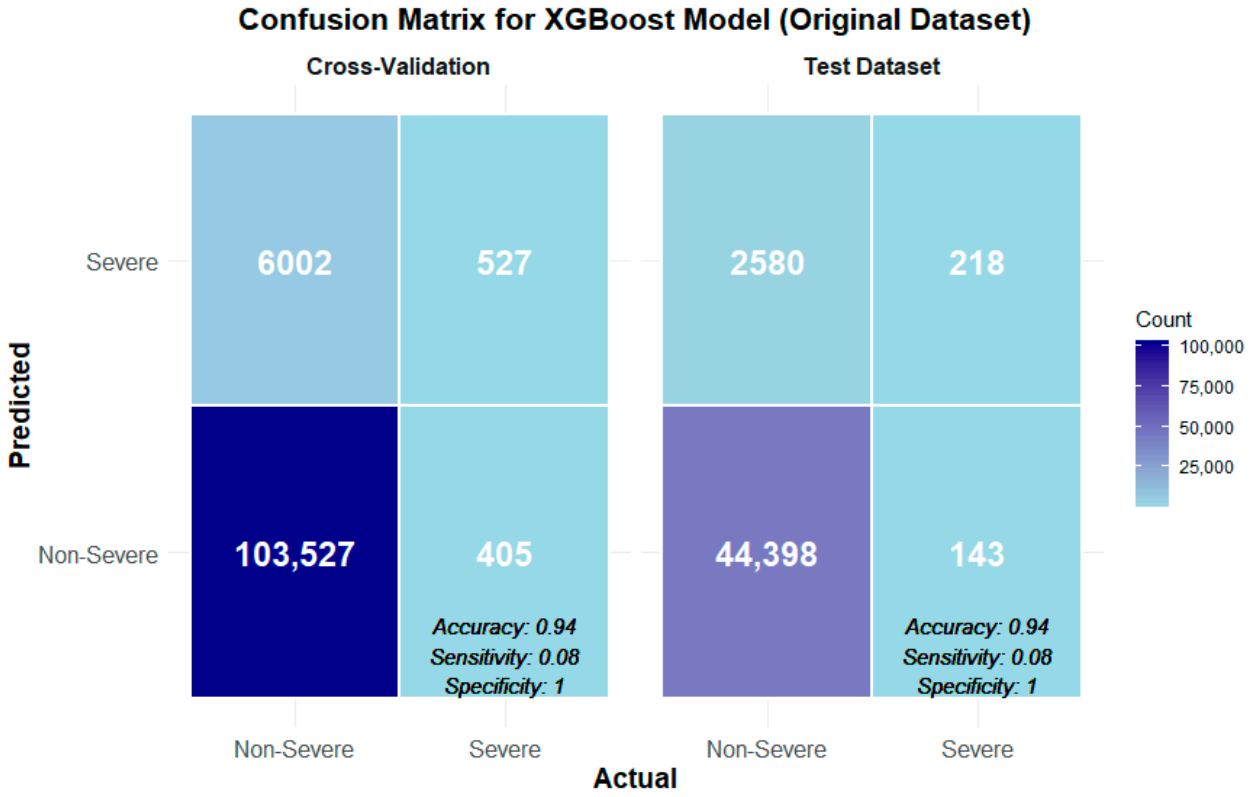

Figure 3. Confusion matrices for XGBoost model that used original dataset for training (training and testing data).

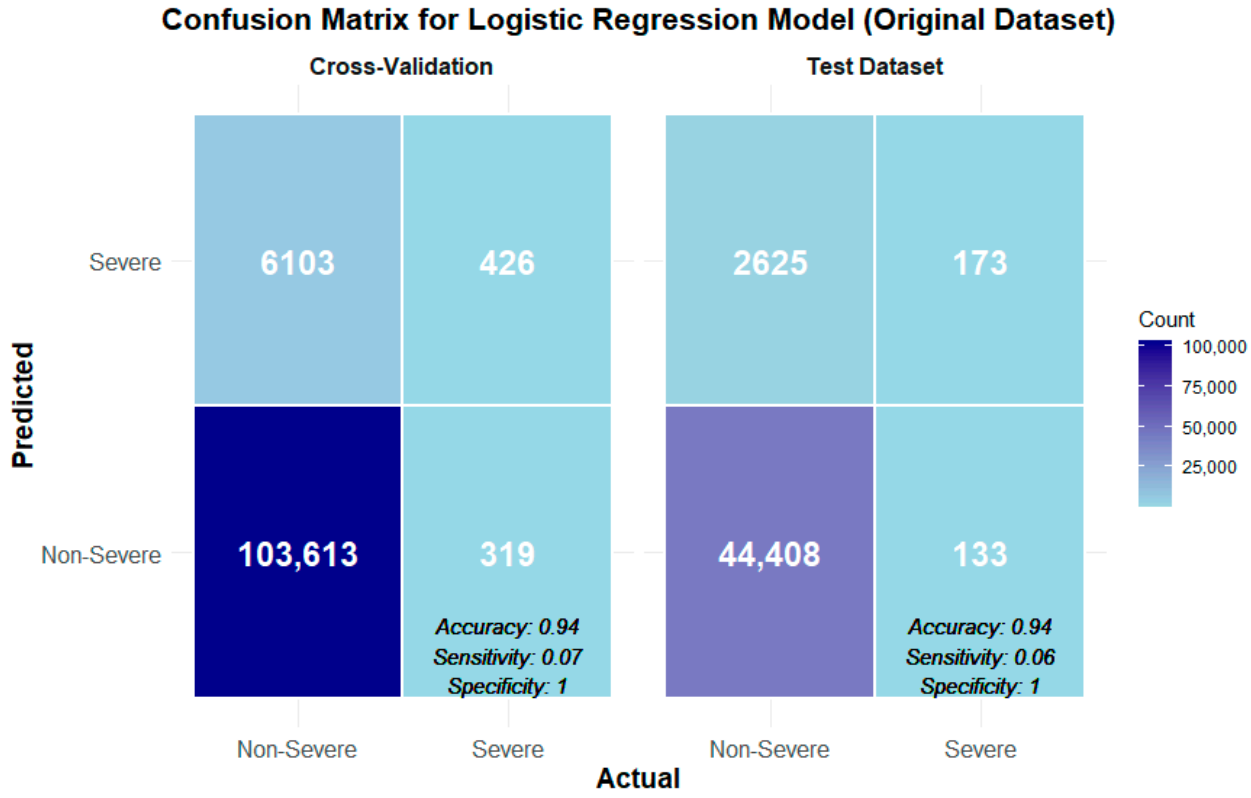

Figure 4. Confusion matrices for logistic regression model that used original dataset for training (training and testing data).

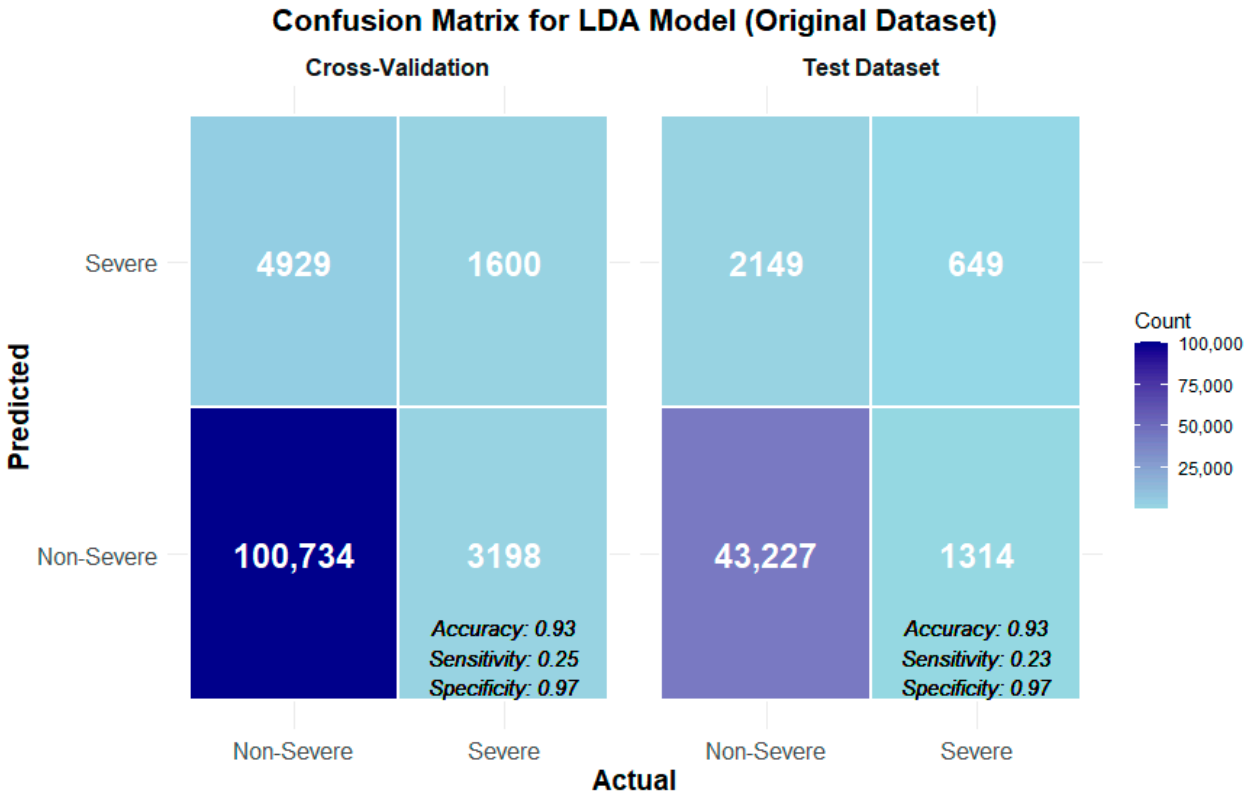

Figure 5. Confusion matrices for LDA model that used original dataset for training (training and testing data).

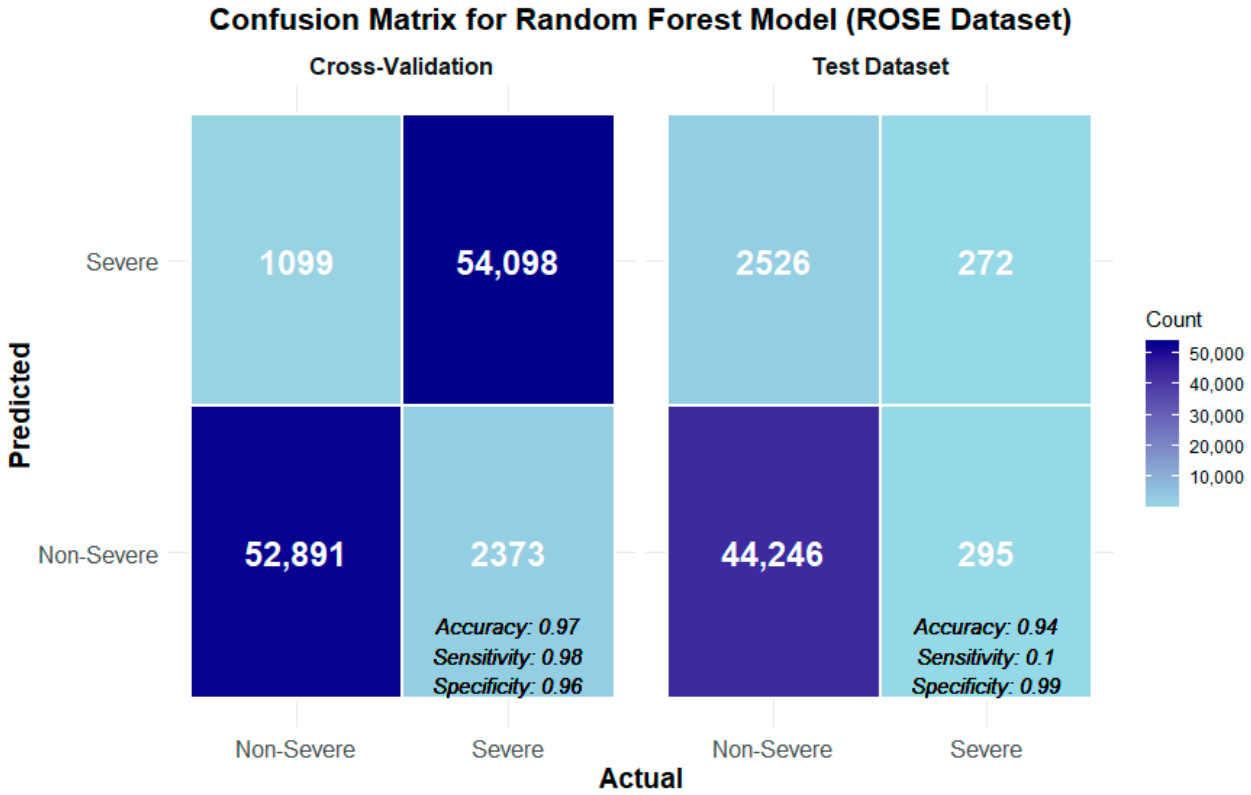

Figure 6. Confusion matrices for random forest model that used ROSE dataset for training (training and testing data).

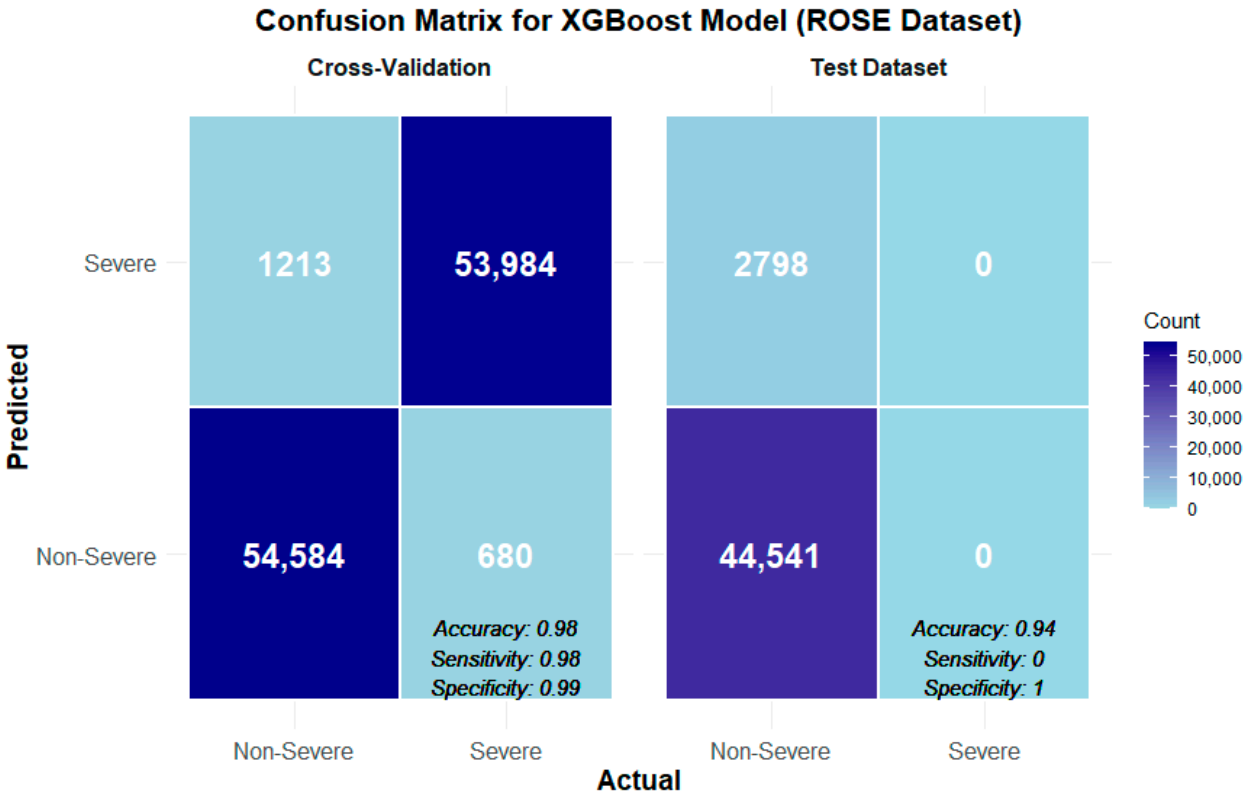

Figure 7. Confusion matrices for XGBoost model that used ROSE dataset for training (training and testing data).

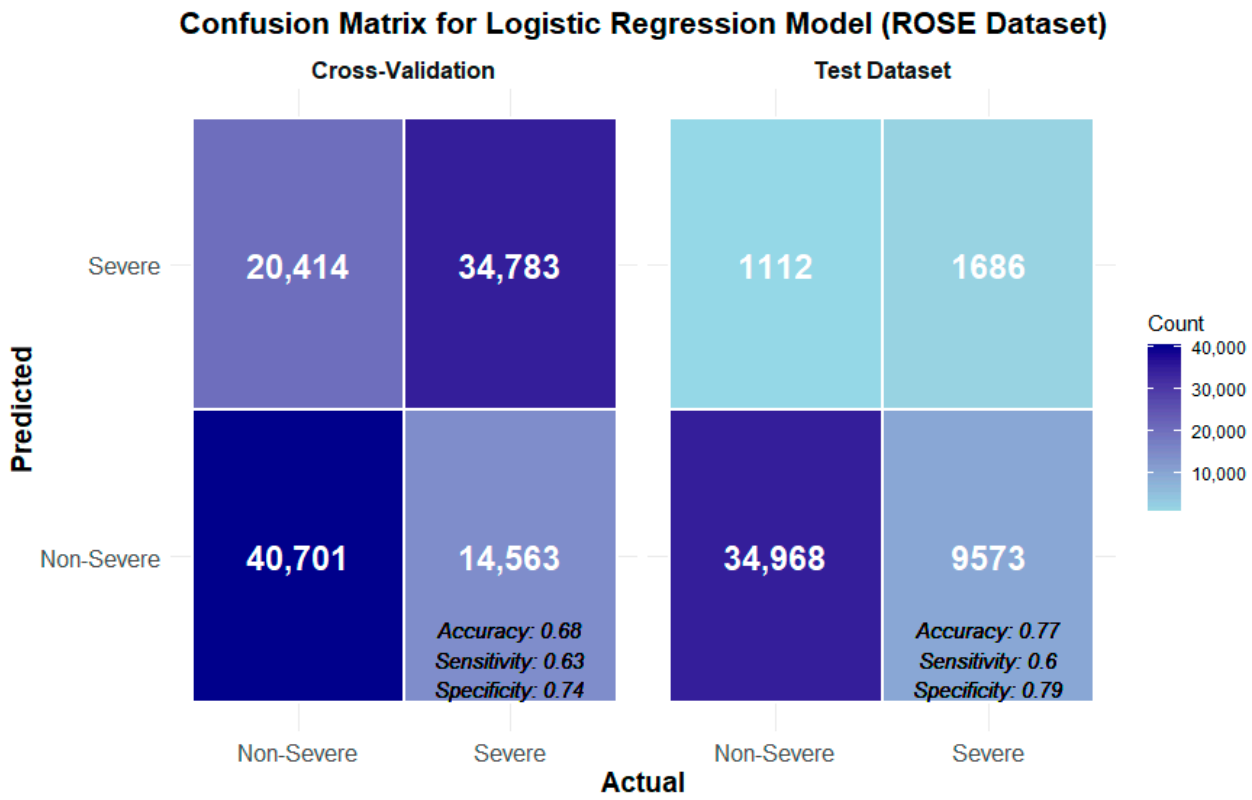

Figure 8. Confusion matrices for logistic regression model that used ROSE dataset for training (training and testing data).

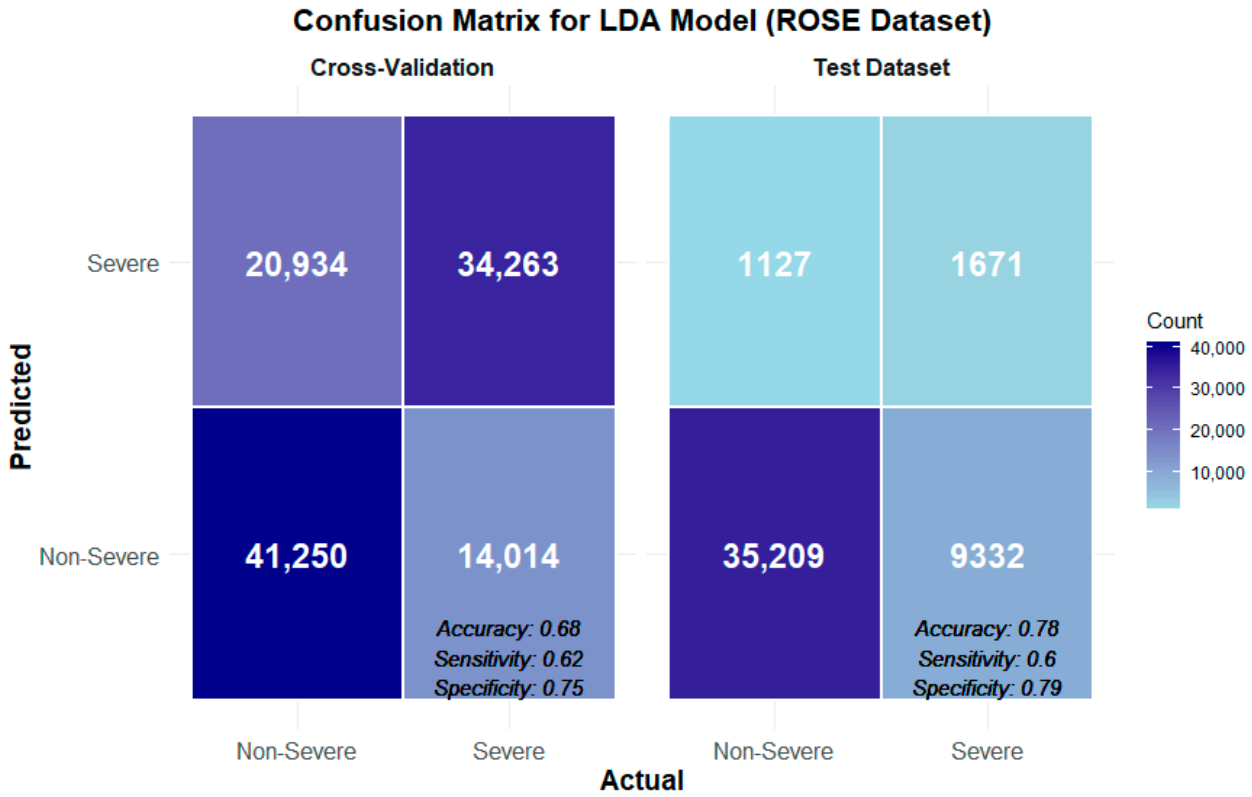

Figure 9. Confusion matrices for LDA model that used ROSE dataset for training (training and testing data).

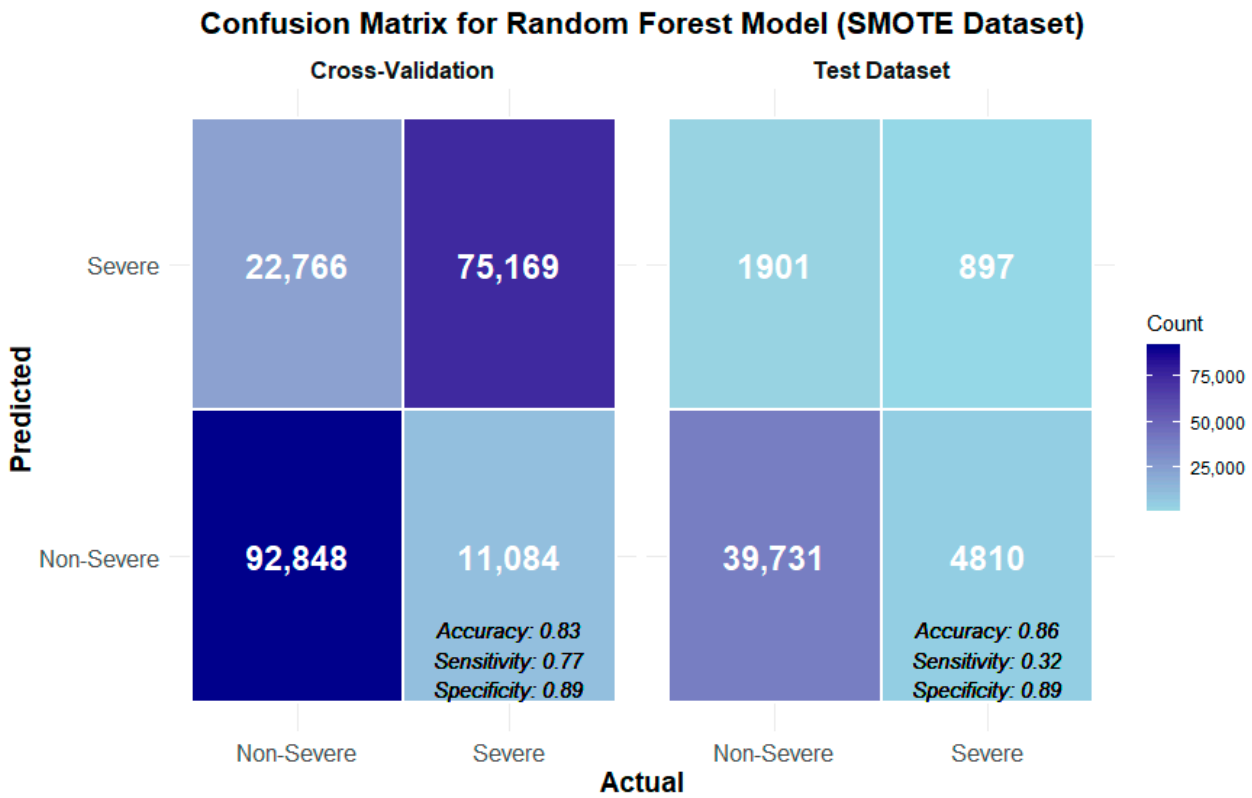

Figure 10. Confusion matrices for random forest model that used SMOTE dataset for training (training and testing data).

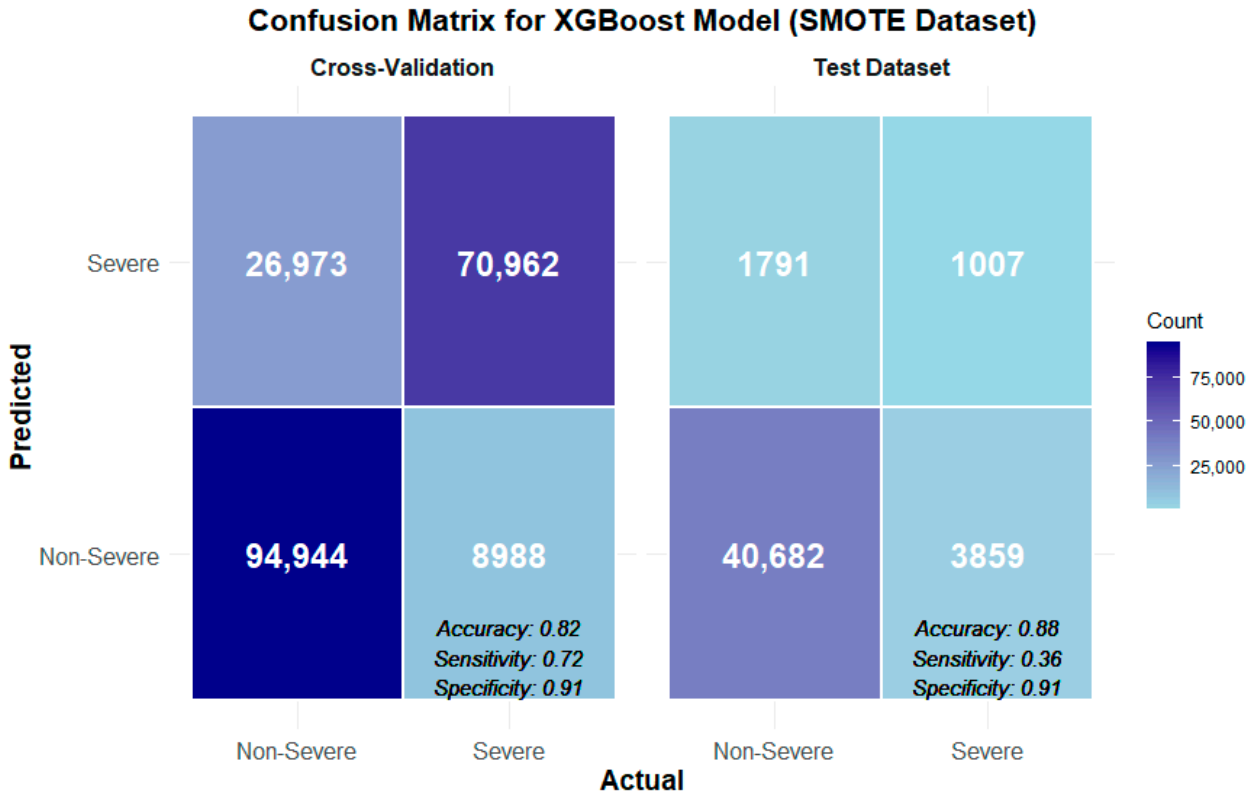

Figure 11. Confusion matrices for XGBoost model that used SMOTE dataset for training (training and testing data).

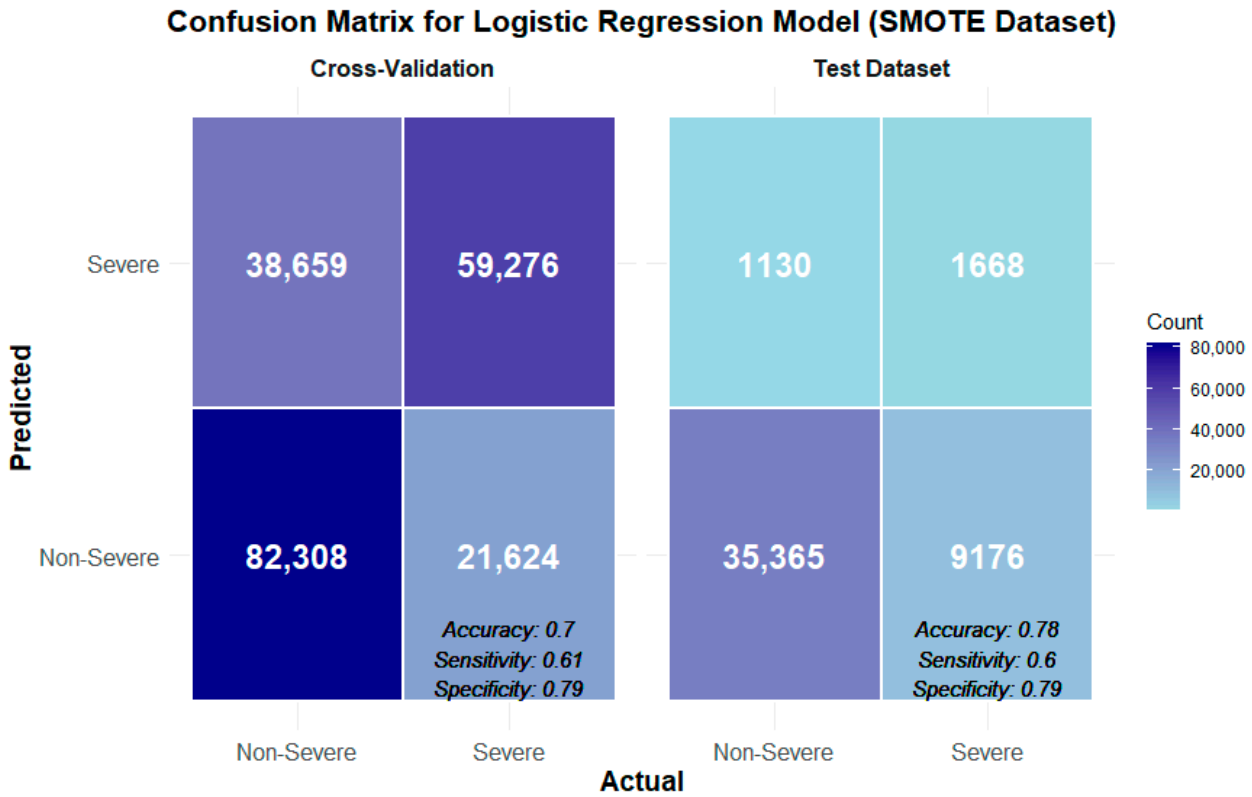

Figure 12. Confusion matrices for logistic regression model that used SMOTE dataset for training (training and testing data).

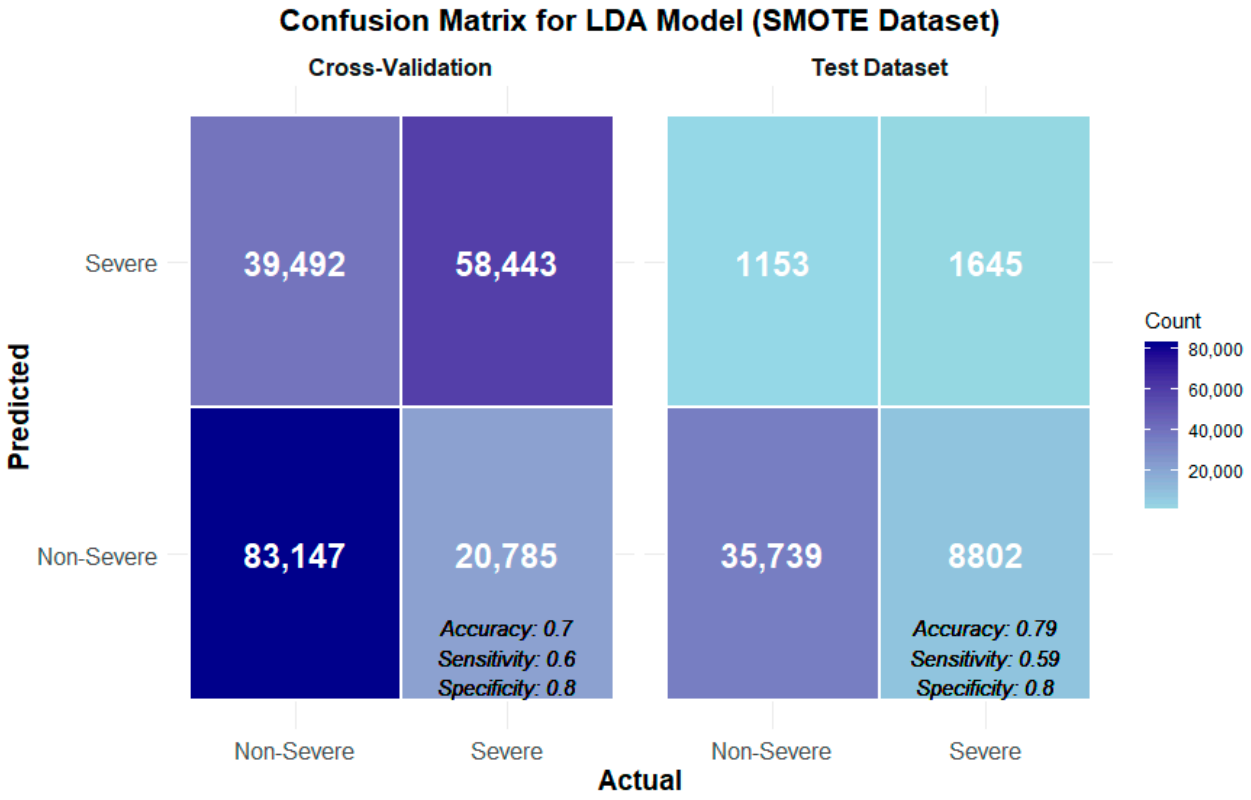

Figure 13. Confusion matrices for LDA model that used SMOTE dataset for training (training and testing data).

3.1.1. Original Dataset

When trained on the original dataset, the models exhibit similar performance in both CV and test datasets in terms of accuracy, sensitivity, and specificity. The slight differences in the metrics between the CV and test datasets indicate that the models have a good generalization ability when trained on the original dataset. However, it is important to note that the low sensitivity values in both CV and test datasets indicate poor performance in identifying the 'Severe' class.

3.1.2. ROSE Dataset

For the models trained on the ROSE dataset, there was a more significant discrepancy between the CV and test dataset results. The non-parametric models, namely, the random forest and XGBoost achieved high accuracy and specificity on the CV dataset, but the sensitivity was lower on the test dataset. This difference suggests that the models may have overfitted the training data, leading to reduced generalization capabilities.

However, the parametric models, namely, the logistic regression and LDA trained on the ROSE dataset showed a less pronounced difference between the CV and test datasets, with the accuracy, sensitivity, and specificity being relatively consistent. This consistency indicates better generalization capabilities for these models when trained on the ROSE dataset.

3.1.3. SMOTE Dataset

When trained on the SMOTE dataset, the models displayed a relatively consistent performance between the CV and test datasets. Although the non-parametric models, namely, random forest and XGBoost, exhibited higher sensitivity in the CV dataset compared to the test dataset. The discrepancy was not as substantial as with the ROSE dataset. The parametric models, namely, logistic regression and LDA, demonstrated similar performance in both the CV and test datasets in terms of accuracy, sensitivity, and specificity.

The results demonstrate that the models trained on the ROSE and SMOTE datasets exhibit different levels of generalization capabilities. The ROSE-trained models, especially the non-parametric models random forest and XGBoost, showed a more significant discrepancy between the CV and test datasets, indicating potential overfitting. On the other hand, the SMOTE-trained models demonstrated more consistent performance across the CV and test datasets, suggesting better generalization capabilities.

3.2. Results of the Crash Severity Models

3.2.1. Models Trained on Original Training Set

In this section, the performance of four different machine learning models is discussed: random forest (RF), XGBoost, logistic regression (LR), and linear discriminant analysis (LDA), on the test dataset. The models were trained on the original dataset, and their performance is summarized in Figure 14. The performance measures used to evaluate these models include accuracy, sensitivity, specificity, balanced accuracy, and geometric mean.

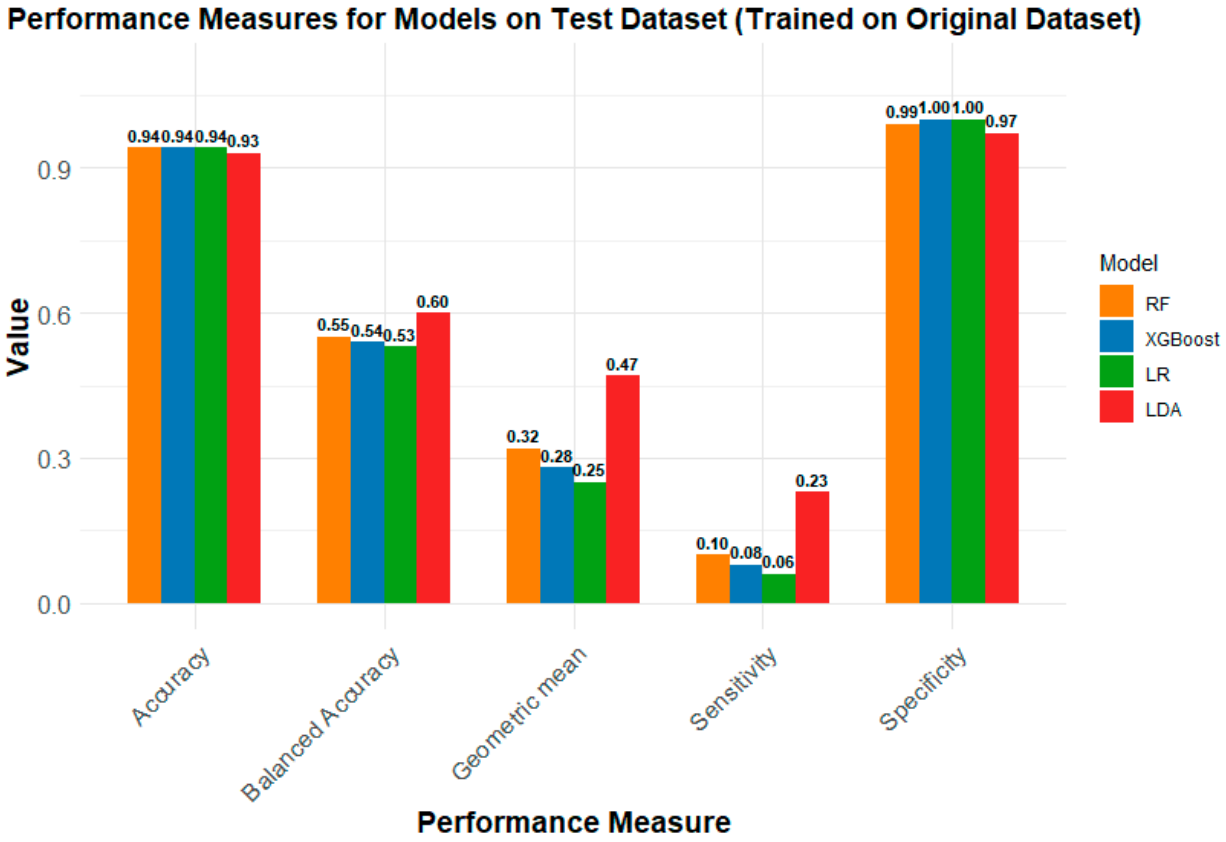

Figure 14. Performance measures for parametric and non-parametric models on the test dataset, trained on original dataset.

From Figure 14, it can be observed that all four models exhibit a high level of accuracy, ranging from 0.93 for LDA to 0.94 for RF, XGBoost, and LR. However, accuracy alone does not provide a comprehensive assessment of the models' performance, as it may be misleading in cases of imbalanced datasets [27].

When evaluating the sensitivity, which measures the proportion of true positive cases among the actual positive cases, the values for all models were notably low, with the LDA model achieving the highest sensitivity at 0.23, while the other models demonstrated much lower values. This indicates that all models struggled to identify the 'Severe' class correctly, which might be due to the imbalanced nature of the original dataset [16,27].

Specificity, which assesses the proportion of true negative cases among the actual negative cases, demonstrates high scores for all models, with XGBoost and LR achieving

perfect scores of 1.0, RF scoring 0.99, and LDA scoring 0.97. This indicates that all models perform well in identifying the 'Non-Severe' class. It is crucial to maintain a balance between sensitivity and specificity, as the trade-off between the two metrics determines the overall performance of the models.

Balanced accuracy provides a more balanced view of the models' performance, taking into account both sensitivity and specificity. In this case, the LDA model achieved the highest balanced accuracy at 0.60, while the other models scored lower values, which highlights the LDA model's relatively better performance in terms of the balance between sensitivity and specificity.

The geometric mean is another metric that accounts for both sensitivity and specificity. It reflects the models' ability to identify both classes equally well. The LDA model again outperformed the other models with a geometric mean of 0.47, indicating a better balance between sensitivity and specificity compared to the other models.

The results demonstrate that, despite high accuracy values, the models' performance in identifying the 'Severe' class was relatively poor. This issue could be attributed to the imbalanced nature of the original dataset, which might have led the models to favor the majority class ('Non-Severe') [16,27,50].

3.2.2. Models Trained on ROSE Training Set

Figure 15 presents the performance measures of parametric (logistic regression and LDA) and non-parametric (random forest and XGBoost) models on the test dataset when trained on the ROSE dataset, which employs resampling techniques to address class imbalance.

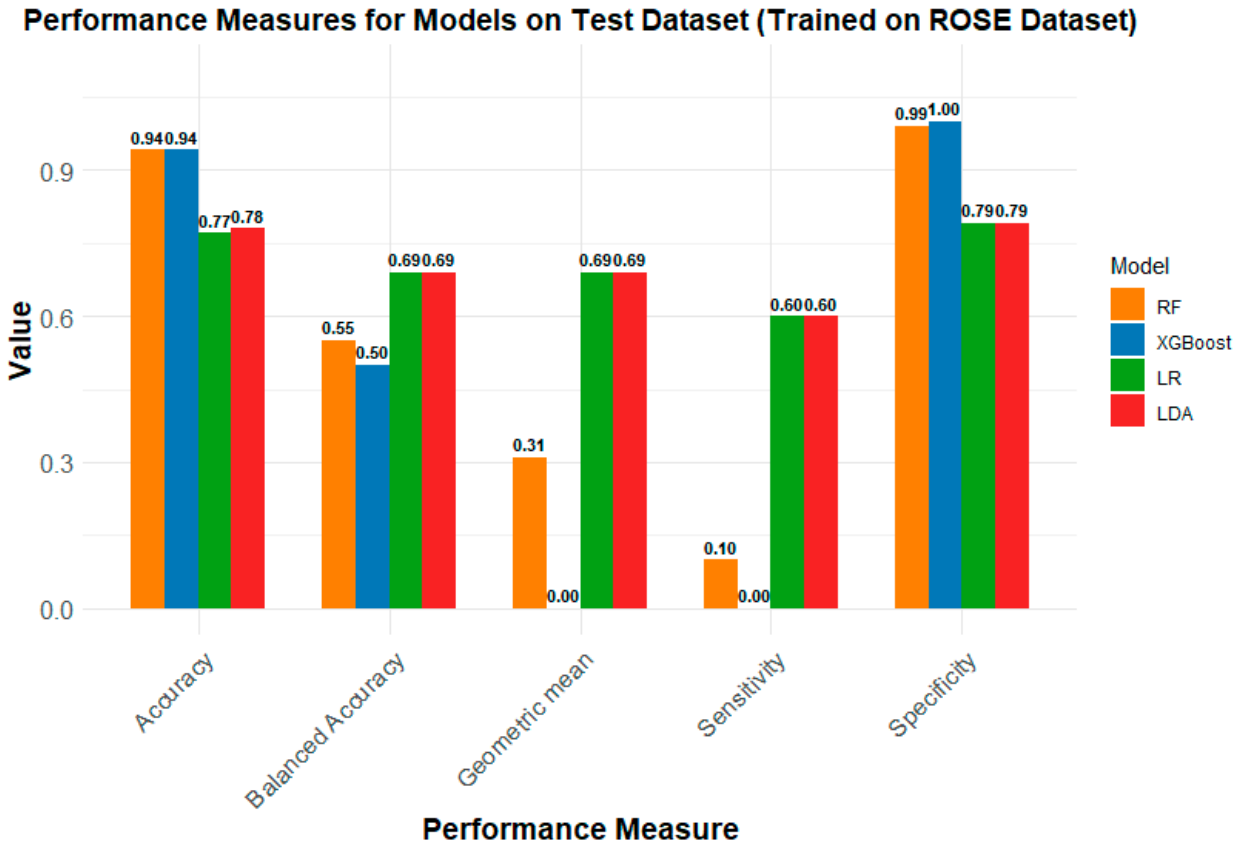

Figure 15. Performance measures for parametric and non-parametric models on the test dataset, trained on ROSE dataset.

The accuracy values for the non-parametric models (RF and XGBoost) remained relatively high at 0.94 for both, while the parametric models (LR and LDA) experienced a decrease in accuracy, falling to 0.77 and 0.78, respectively.

After training on the ROSE dataset, the sensitivity of the logistic regression and LDA models saw a significant improvement, with values of 0.60 for both, indicating better performance in identifying the 'Severe' class. However, the sensitivity for RF slightly increased to 0.10, while the XGBoost model's sensitivity dropped to 0, implying that it failed to identify any 'Severe' cases correctly.

All models maintained high specificity values, with the XGBoost model achieving perfect specificity (1). However, logistic regression and LDA experienced a slight decrease in specificity to 0.79 for both.

The balanced accuracy and geometric mean provide a more comprehensive evaluation of the models' performance. After training on the ROSE dataset, both the parametric models logistic regression and LDA showed a notable improvement in balanced accuracy (0.69 for both) and geometric mean (0.69 for both). In contrast, the RF model had minimal change in balanced accuracy and a slight increase in the geometric mean, while the XGBoost model's balanced accuracy and geometric mean decreased to 0.5 and 0.0, respectively.

The results in Figure 15 highlight the importance of addressing class imbalance when evaluating model performance. The ROSE dataset led to significant improvements in sensitivity, balanced accuracy, and geometric mean for the parametric models (i.e., logistic regression and LDA). However, the XGBoost model's performance deteriorated, and the random forest model experienced only minimal changes.

3.2.3. Models Trained on SMOTE Training Set

Figure 16 presents the performance measures of parametric (logistic regression and LDA) and non-parametric (random forest and XGBoost) models on the test dataset when trained on the SMOTE dataset.

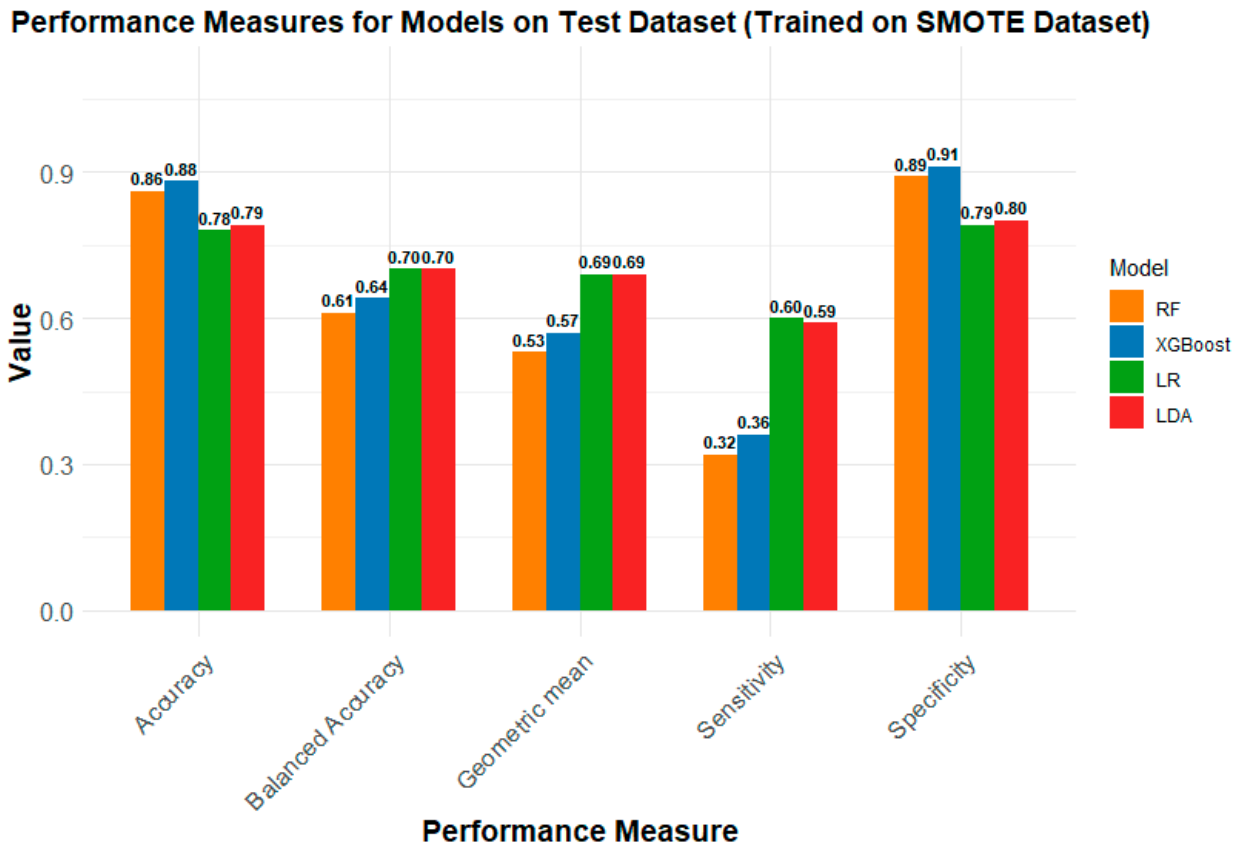

Figure 16. Performance measures for parametric and non-parametric models on the test dataset, trained on SMOTE dataset.

The results indicate that the XGBoost model has the highest accuracy (0.88), followed by RF (0.86), LDA (0.79), and LR (0.78). Although accuracy is an essential performance measure, it might not be sufficient when dealing with imbalanced datasets. Therefore, other performance measures such as sensitivity, specificity, balanced accuracy, and geometric mean are considered to provide a comprehensive evaluation of the models.

Sensitivity measures the proportion of true positive cases among the actual positive cases. In this aspect, LR has the highest sensitivity (0.60), closely followed by LDA (0.59). Meanwhile, XGBoost and RF exhibit lower sensitivity values, 0.36 and 0.32, respectively. This indicates that the parametric models (LR and LDA) perform better in identifying severe class crashes.

The specificity values for all models decreased compared to their performance on the original dataset, with RF and XGBoost models achieving specificity values of 0.89 and 0.91, respectively. The specificity of the parametric models (LR and LDA) also decreased, with values of 0.79 and 0.80, respectively.

After training on the SMOTE dataset, all models showed improvement in their balanced accuracy and geometric mean values. The RF and XGBoost models had balanced accuracy values of 0.61 and 0.64, respectively, and geometric mean values of 0.53 and 0.57, respectively. The parametric models (LR and LDA) had balanced accuracy values of 0.70 for both and geometric mean values of 0.69 for both.

In summary, although XGBoost has the highest accuracy, it exhibits relatively lower sensitivity and balanced accuracy compared to LR and LDA. In contrast, LR and LDA demonstrate a better balance between sensitivity and specificity, making them more suitable for this imbalanced dataset.

3.3. Effectiveness of Resampling Techniques on Predictive Models

Given the higher involvement of elderly drivers in severe crashes, this study aimed to improve the prediction of such crashes by utilizing synthetic resampling techniques. The following section will demonstrate how these techniques can improve the performance of machine learning models.

3.3.1. ROSE

Figure 17 presents the performance measures for four machine learning models after applying the ROSE balancing strategy. The results show that the implementation of this strategy leads to an improvement in predictive performance in some measures, but not all. The accuracy for the RF and XGBoost models slightly improved by 0.21% and decreased by 0.16%, respectively, while the accuracy for LR and LDA models decreased significantly by 16.74% and 14.77%, respectively.

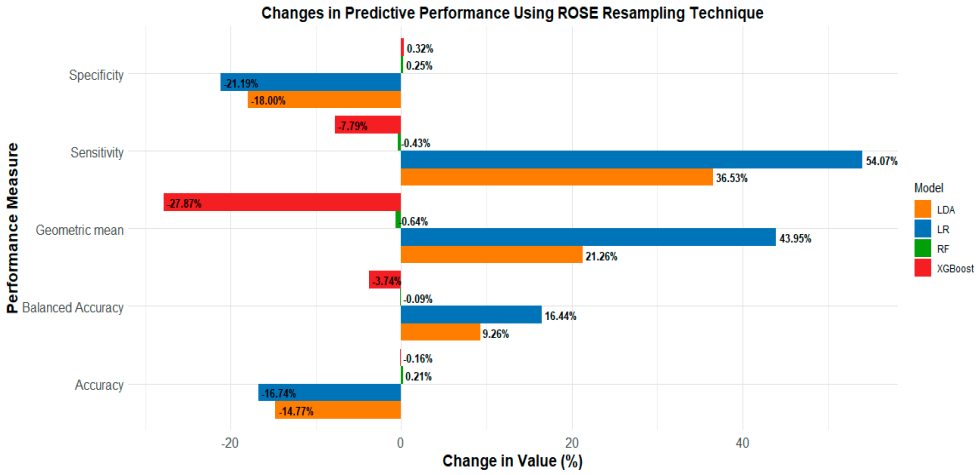

Figure 17. The implementation of the ROSE balancing strategy results in enhanced predictive performance.

In terms of sensitivity, the implementation of the ROSE balancing strategy resulted in a decrease of 0.43% for the RF model, while it was reduced by 7.79% for the XGBoost model. On the other hand, the LR and LDA models showed a significant increase in sensitivity by 54.07% and 36.53%, respectively. In terms of specificity, the RF and XGBoost models showed slight improvements of 0.25% and 0.32%, respectively. However, the implementation of the ROSE balancing strategy resulted in a significant decrease in specificity of 21.19% and 18.00% for the LR and LDA models, respectively.

The balanced accuracy for the RF and XGBoost models showed a slight decrease of 0.09% and 3.74%, respectively, while the LR and LDA models showed a significant increase of 16.44% and 9.26%, respectively. Finally, in terms of geometric mean, the implementation of the ROSE balancing strategy resulted in a decrease of 0.64% and 27.87% for the RF and XGBoost models, respectively. However, the LR and LDA models showed a significant increase of 43.95% and 21.26%, respectively.

3.3.2. SMOTE

Figure 18 demonstrates the impact of implementing the synthetic minority oversampling technique (SMOTE) balancing strategy on the predictive performance of four machine learning models—random forest (RF), extreme gradient boosting (XGBoost), logistic regression (LR), and linear discriminant analysis (LDA)—for crash severity prediction. The results indicate that the application of the SMOTE strategy leads to enhanced predictive performance across various performance measures.

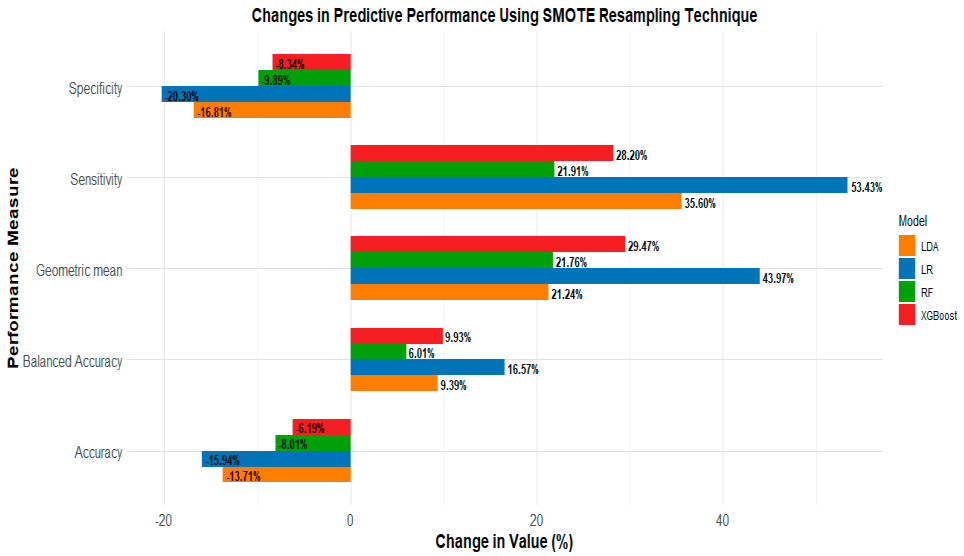

Figure 18. The implementation of the SMOTE balancing strategy results in enhanced predictive performance.

For accuracy, the implementation of the SMOTE strategy resulted in decreases for all models: RF (−8.01%), XGBoost (−6.19%), LR (−15.94%), and LDA (−13.71%). However, the sensitivity significantly improved for all models: RF (21.91%), XGBoost (28.20%), LR (53.43%), and LDA (35.60%).

In terms of specificity, the SMOTE strategy led to a slight decrease for all models: RF (−9.89%), XGBoost (−8.34%), LR (−20.30%), and LDA (−16.81%). On the other hand, balanced accuracy exhibited improvements for all models following the implementation of the SMOTE strategy: RF (6.01%), XGBoost (9.93%), LR (16.57%), and LDA (9.39%).

Lastly, the geometric mean also exhibited improvements for all models with the implementation of the SMOTE strategy: RF (21.76%), XGBoost (29.47%), LR (43.97%), and LDA (21.24%)

In summary, the findings indicate that employing the SMOTE balancing strategy can substantially improve the predictive performance of machine learning models when

predicting the minority class (i.e., 'Severe'), as well as enhance both balanced accuracy and geometric mean.

3.4. The Effect of Influential Factors on Crash Severity

In Table 4, the results of the logistic regression model employing a synthetic minority over-sampling technique (SMOTE) dataset are outlined, detailing the parameter estimates and odds ratios for various factors contributing to crash severity. This methodology, a combination of logistic regression and SMOTE, was selected as it yielded the most robust and insightful results in our study, thereby providing a solid foundation for the ensuing discussion on the influential determinants of crash severity.

Table 4. Logistic regression model results.

Variable	Category	Estimate	SE	*p*-Value	Odds Ratio
Crash type	Angle	1.10874	0.021	<0.001	3.031
	Fixed object	1.60342	0.024	<0.001	4.967
	Head-on	2.21490	0.032	<0.001	9.160
	Overturned	1.88370	0.052	<0.001	6.578
	Rear end	0.29230	0.021	<0.001	1.340
	Sideswipe *				
Traffic signal	Yes	0.05584	0.013	<0.001	1.057
	No *				
Weather condition	No adverse condition	0.44410	0.017	<0.001	1.559
	Adverse condition *				
Roadway alignment	Curve	0.13316	0.019	<0.001	1.142
	Straight *				
Roadway type	One-way	−0.70767	0.041	<0.001	0.493
	Two-way divided	−0.13259	0.012	<0.001	0.876
	Two-way undivided *				
Work zone	Yes	−0.39516	0.035	<0.001	0.674
	No *				
Alcohol	Yes	0.60794	0.032	<0.001	1.837
	No *				
Belted	No	1.94371	0.026	<0.001	6.985
	Yes *				
Bike	Yes	2.27529	0.055	<0.001	9.731
	No *				
Distracted	Yes	0.11692	0.014	<0.001	1.124
	No *				
Pedestrian	Yes	2.72422	0.049	<0.001	15.245
	No *				
Speed violation	Yes	0.47653	0.014	<0.001	1.610
	No *				
Area type	Rural	0.44437	0.014	<0.001	1.560
	Urban *				
Animal	Yes	−1.89159	0.052	<0.001	0.151
	No *				
Weekend	Yes	0.09184	0.012	<0.001	1.096
	No *				
Intercept	Non-severe Injury | Severe Injury	3.2344	0.146	<0.001	25.391
Log-likelihood at convergence	−115,577				
Log-likelihood at zero	−139,834				
Likelihood ratio test	48,514				

* Reference category.

The results of the logistic regression analysis, including model coefficients and corresponding odds ratios (ORs), substantiate the relative impact of various factors on crash severity when compared to a designated baseline or reference category. An odds ratio exceeding 1, aligned with a positive coefficient estimate, signifies a greater influence on crash severity compared to the reference category, and vice versa for an OR less than 1. These odds ratios serve as valuable interpretive tools, providing critical insights into the increased or decreased likelihood of a severe crash outcome given a specific predictor or condition, assuming all other factors are constant.

The model indicates that the type of crash is an essential determinant of crash severity. When compared to sideswipe crashes (reference category), the odds of severe outcomes are increased for angle crashes (OR = 3.031), fixed object crashes (OR = 4.967), head-on crashes (OR = 9.160), and overturned crashes (OR = 6.578). These crash types typically involve a greater force of impact, which may explain the heightened severity.

Considering environmental and situational factors, crashes that occur at locations with traffic signals (OR = 1.057) have slightly increased odds of severe outcomes compared to those where no signals are present. This may be due to the complexity of such intersections, which can lead to more severe crashes when errors occur. Additionally, a higher likelihood of severe crashes was found under no adverse weather conditions as opposed to adverse conditions (OR = 1.559). A possible explanation for this could be that drivers tend to exercise greater caution and lower speeds in the presence of adverse weather conditions. This phenomenon can be attributed to the concept of risk compensation, where drivers adjust their behavior in response to perceived levels of risk [51].

The nature of the roadway also influences crash severity. Crashes on curved roads, compared to those on straight roads, have higher odds of severity (OR = 1.142). Curved roads may require more complex maneuvering and judgment from drivers, possibly leading to more severe crashes when errors are made. Furthermore, one-way and two-way divided roads tend to have less severe crashes compared to two-way undivided roads (ORs of 0.493 and 0.876, respectively).

Crashes occurring in work zones show lower odds of resulting in severe outcomes (OR = 0.674) compared to those outside of work zones. Work zones typically have lower speed limits and increased enforcement, which may contribute to the reduced severity.

The non-use of seatbelts markedly increases the odds of severe crashes (OR = 6.985) compared to crashes where seatbelts are used. Seatbelts are known to significantly reduce the risk of injury by preventing ejection from the vehicle during a crash, hence their absence may contribute to more severe injuries.

The study further reveals that crashes involving cyclists pose dramatically higher odds of severity (OR = 9.731) when compared to their counterparts excluding bikes. Cyclists' vulnerability, given their lack of physical protection compared to vehicle occupants, potentially accounts for this increased severity. Similarly, pedestrian involvement in crashes considerably increases the odds of severity (OR = 15.245) compared to crashes with no pedestrians involved. Pedestrians, like cyclists, lack the physical protection that a vehicle provides, which makes them particularly vulnerable in crashes.

Crashes where distractions, such as texting on mobile devices, are a contributing factor demonstrate higher odds of severity (OR = 1.124) in contrast to scenarios free of such distractions. The subsequent delayed reaction times and impaired decision-making due to these distractions indeed magnify the potential severity of the crashes. Similarly, crashes involving alcohol, particularly those associated with intoxicated drivers, considerably elevate the odds of severe outcomes (OR = 1.837) relative to crashes that occur without the influence of alcohol. This increased risk can be primarily attributed to alcohol's deleterious impacts on crucial cognitive abilities such as reaction time and decision-making capacity.

Speed violations further compound the odds of severe crashes (OR = 1.610), reaffirming the dangerous consequences of high-speed impacts. Interestingly, crashes transpiring over weekends demonstrate marginally increased odds of severity (OR = 1.096), which may be

attributed to altered traffic patterns, increased alcohol consumption, or escalated travel speeds during leisurely periods.

The geographical context also influences crash severity, with rural crashes associated with higher severity odds (OR = 1.560) compared to urban occurrences. This divergence could reflect disparities in speed limits, emergency response times, and access to trauma care between rural and urban settings. In contrast, crashes involving animals are less likely to be severe (OR = 0.151) compared to those without animal involvement.

3.5. Operational and Management Implications

The findings of this research provide valuable insights with direct implications for operational and management strategies in traffic safety. Understanding the critical factors contributing to crash severity could serve as a fundamental resource for traffic safety managers, city planners, and policymakers in their endeavor to enhance road safety, particularly for elderly drivers. This study also demonstrates the utility of machine learning models in developing targeted interventions for preventing such crashes.

Regarding speed violations, this study identified them as significant determinants of crash severity. Speed management remains a key aspect of road safety because it directly influences both the occurrence and severity of crashes. Therefore, enhancing law enforcement measures against speed violations becomes a necessary strategy. Moreover, the implementation of intelligent speed assistance systems and designing roads to naturally limit speed could be potential strategies for reducing the instances and impact of speed violations.

Similarly, the effectiveness of seat belts in reducing the severity of crash injuries was reaffirmed in this study. This underscores the urgency for measures that encourage seat belt usage. Strategies could include public awareness campaigns highlighting the importance of seat belts, the strict enforcement of seat belt laws, and incorporating seat-belt reminder systems in vehicles. These initiatives could considerably enhance compliance with seat belt usage, consequently lowering crash severity.

This study further revealed that the involvement of pedestrians and bicyclists in crashes significantly increases crash severity. As such, ensuring the safety of these vulnerable road users becomes an urgent priority. Infrastructure improvements could include creating dedicated bike lanes and designing pedestrian-friendly intersections. The implementation of effective traffic calming measures could also play a crucial role in reducing the severity of crashes involving pedestrians and bicyclists.

Moreover, the role of alcohol consumption and driver distraction in contributing to severe crashes was highlighted in our findings. This puts forth a strong case for robust strategies aimed at tackling drunk and distracted driving. Measures could range from strict law enforcement to technological solutions such as ignition interlocks for DUI offenders. Concurrently, public awareness campaigns stressing the dangers of drunk and distracted driving could promote safer driving habits.

Lastly, the machine learning models developed and tested in this study could serve as invaluable tools for real-time crash prediction systems. These systems could evaluate current driving conditions, predict high-risk situations based on the identified influential factors, and initiate appropriate safety measures, such as adjusting speed limits or delivering warning messages to drivers. This represents a proactive, data-driven approach to traffic safety management.

In conclusion, the insights derived from this study offer a solid foundation for the development of targeted interventions, focusing on the key factors contributing to crash severity. These interventions can supplement and enhance the effectiveness of existing traffic safety management strategies, ultimately paving the way for safer road conditions for all users, particularly elderly drivers.

4. Conclusions

This study explored the potential of both parametric and non-parametric machine learning models in predicting crash severity involving elderly drivers, utilizing crash data

from the Commonwealth of Virginia (USA) spanning from 2014 to 2021. A thorough comparison of performance metrics revealed that resampling techniques, specifically ROSE and SMOTE, effectively tackled class imbalance, resulting in enhanced sensitivity, balanced accuracy, and geometric mean for parametric models such as logistic regression and LDA. Notably, the application of the SMOTE balancing technique substantially improved the predictive performance of all evaluated models.

Study findings highlight that incorporating resampling techniques can significantly boost the performance of parametric models, leading to an impressive 54% increase in the true positive rate for severe crash prediction and a 44% improvement in geometric mean for logistic regression. Furthermore, the use of SMOTE enhances the prediction of severe crashes in non-parametric models, yielding a 28% increase in the true positive rate and a 29% enhancement in geometric mean for XGBoost. The results also suggest that parametric models outperform non-parametric models when employing balancing resampling techniques, which can be critical for developing effective interventions and improving traffic safety for elderly drivers.

Furthermore, the study highlighted a broad range of factors that contribute significantly to crash severity among elderly drivers. The findings revealed that crash types, environmental conditions, roadway characteristics, driver behaviors, the involvement of vulnerable road users, and geographical context are all key determinants of crash severity. This comprehensive understanding underscores the necessity for multifaceted interventions, addressing not only individual behaviors but also environmental factors and road infrastructure.

The findings ultimately underscore the need for further research to refine preventative strategies, ensuring safer road conditions for elderly drivers. Moreover, the findings underscore the potential of machine learning models to effectively analyze complex crash data, identify factors contributing to crash severity, and inform targeted interventions to mitigate risks associated with elderly drivers.

These results provide valuable insights for policymakers and transportation safety professionals to develop data-driven strategies that can enhance road safety and reduce the number of severe crashes involving elderly drivers.

5. Study Limitations and Future Directions

Despite these promising results, this study has some limitations that warrant acknowledgment. First, the data used in this study is specific to Virginia and the United States, which may limit the generalizability of the findings to other regions or countries with different road infrastructure, traffic rules, or driving behaviors. Second, while the resampling techniques employed in this study have improved the models' performance, other resampling methods or alternative strategies for handling imbalanced datasets should be considered and compared to assess their impact on predictive accuracy and generalization.

Future research could build upon the findings of this study by applying the resampling strategies discussed herein to a broader range of modeling approaches, such as deep neural networks (e.g., convolutional neural networks). This would facilitate a more comprehensive understanding of the effectiveness of these strategies across various machine learning techniques and offer insights into the most suitable methods for handling imbalanced datasets in the context of crash severity prediction.

Additionally, future studies could consider incorporating other relevant factors related to vehicle characteristics, such as vehicle size, type, and age. Additionally, the distinct circumstances of single-vehicle crashes and multi-vehicle crashes could be considered separately, as these might be influenced by different factors. Furthermore, road geometric variables such as lane width can be incorporated. This could further enhance the analysis by providing a more holistic view of the factors that contribute to crash severity among elderly drivers.

Lastly, future work could extend the current research by conducting comparative studies in different geographic regions and among various driver populations, which

would contribute to a better understanding of the generalizability of the proposed modeling approaches and resampling strategies in diverse contexts and settings.

By incorporating these suggestions and ideas, future research can continue to advance the field of transportation safety and machine learning, ultimately contributing to more effective interventions and strategies for reducing crash severity among elderly drivers and other vulnerable road users.

Author Contributions: Conceptualization, M.A., M.M.G.F. and H.A.R.; methodology, M.A.; software, M.A.; analysis and interpretation of the results, M.A.; data curation, M.A.; writing—original draft preparation, M.A.; writing—review and editing, M.M.G.F. and H.A.R.; visualization, M.A.; supervision, M.M.G.F. and H.A.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study are publicly available at: https://www.virginiaroads.org/maps/1a96a2f31b4f4d77991471b6cabb38ba/about (accessed on 1 February 2022).

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

References

1. World Health Organization. *Global Status Report on Road Safety 2018*; World Health Organization: Geneva, Switzerland, 2018; ISBN 978-92-4-156568-4.
2. Road Crashes Have More Impact on Poverty than You Probably Thought. Available online: https://blogs.worldbank.org/transport/road-crashes-have-more-impact-poverty-you-probably-thought (accessed on 3 March 2023).
3. U.S. Census Bureau. 2017 National Population Projections Tables: Main Series. Available online: https://www.census.gov/data/tables/2017/demo/popproj/2017-summary-tables.html (accessed on 11 March 2023).
4. The Myth of an "Ageing Society". Available online: https://www.weforum.org/agenda/2018/05/the-myth-of-the-aging-society/ (accessed on 11 March 2023).
5. Traffic Deaths Decreased in 2018, but Still 36,560 People Died | NHTSA. Available online: https://www.nhtsa.gov/traffic-deaths-decreased-2018-still-36560-people-died (accessed on 10 March 2023).
6. Older Drivers. Available online: https://www.iihs.org/topics/older-drivers (accessed on 10 March 2023).
7. Lee, J.; Gim, T.-H.T. Analysing the Injury Severity Characteristics of Urban Elderly Drivers' Traffic Accidents through the Generalised Ordered Logit Model: A Case of Seoul, South Korea. *J. Transp. Saf. Secur.* **2022**, *14*, 1139–1164. [CrossRef]
8. Cobb, R.W.; Coughlin, J.F. Are Elderly Drivers a Road Hazard? Problem Definition and Political Impact. *J. Aging Stud.* **1998**, *12*, 411–427. [CrossRef]
9. Hakamies-Blomqvist, L. *Elderly Drivers, Results from a Nordic in-Depth Study on Elderly Car Drivers. Comments on Im Bernhoft's Paper*; VTI Rapport; Swedish National Road and Transport Research Institute: Linkoping, Sweden, 1991.
10. Mathias, J.L.; Lucas, L.K. Cognitive Predictors of Unsafe Driving in Older Drivers: A Meta-Analysis. *Int. Psychogeriatr.* **2009**, *21*, 637–653. [CrossRef] [PubMed]
11. Bélanger, A.; Gagnon, S.; Yamin, S. Capturing the Serial Nature of Older Drivers' Responses towards Challenging Events: A Simulator Study. *Accid. Anal. Prev.* **2010**, *42*, 809–817. [CrossRef]
12. Andrews, E.C.; Westerman, S.J. Age Differences in Simulated Driving Performance: Compensatory Processes. *Accid. Anal. Prev.* **2012**, *45*, 660–668. [CrossRef]
13. Rao, P.; Munoz, B.; Turano, K.; Munro, C.; West, S.K. The Decline in Attentional Visual Fields over Time among Older Participants in the Salisbury Eye Evaluation Driving Study. *Investig. Opthalmology Vis. Sci.* **2013**, *54*, 1839–1844. [CrossRef]
14. de Wit, H. Impulsivity as a Determinant and Consequence of Drug Use: A Review of Underlying Processes. *Addict. Biol.* **2009**, *14*, 22–31. [CrossRef]
15. Hanrahan, R.B.; Layde, P.M.; Zhu, S.; Guse, C.E.; Hargarten, S.W. The Association of Driver Age with Traffic Injury Severity in Wisconsin. *Traffic Inj. Prev.* **2009**, *10*, 361–367. [CrossRef]
16. Kim, S.; Lym, Y.; Kim, K.-J. Developing Crash Severity Model Handling Class Imbalance and Implementing Ordered Nature: Focusing on Elderly Drivers. *Int. J. Environ. Res. Public Health* **2021**, *18*, 1966. [CrossRef]
17. Alrumaidhi, M.; Rakha, H.A. Factors Affecting Crash Severity among Elderly Drivers: A Multilevel Ordinal Logistic Regression Approach. *Sustainability* **2022**, *14*, 11543. [CrossRef]
18. Wang, X.; Xia, G.; Zhao, J.; Wang, J.; Yang, Z.; Loughney, S.; Fang, S.; Zhang, S.; Xing, Y.; Liu, Z. A Novel Method for the Risk Assessment of Human Evacuation from Cruise Ships in Maritime Transportation. *Reliab. Eng. Syst. Saf.* **2023**, *230*, 108887. [CrossRef]

19. Hellton, K.H.; Tveten, M.; Stakkeland, M.; Engebretsen, S.; Haug, O.; Aldrin, M. Real-Time Prediction of Propulsion Motor Overheating Using Machine Learning. *J. Mar. Eng. Technol.* **2022**, *21*, 334–342. [CrossRef]
20. Babichev, S.; Yasinska-Damri, L.; Liakh, I. A Hybrid Model of Cancer Diseases Diagnosis Based on Gene Expression Data with Joint Use of Data Mining Methods and Machine Learning Techniques. *Appl. Sci.* **2023**, *13*, 6022. [CrossRef]
21. Almasoudi, F.M. Enhancing Power Grid Resilience through Real-Time Fault Detection and Remediation Using Advanced Hybrid Machine Learning Models. *Sustainability* **2023**, *15*, 8348. [CrossRef]
22. Al Mamlook, R.E.; Abdulhameed, T.Z.; Hasan, R.; Al-Shaikhli, H.I.; Mohammed, I.; Tabatabai, S. Utilizing Machine Learning Models to Predict the Car Crash Injury Severity among Elderly Drivers. In Proceedings of the 2020 IEEE International Con-ference on Electro Information Technology (EIT), Naperville, IL, USA, 31 July–1 August 2020; pp. 105–111.
23. Aldhari, I.; Almoshaogeh, M.; Jamal, A.; Alharbi, F.; Alinizzi, M.; Haider, H. Severity Prediction of Highway Crashes in Saudi Arabia Using Machine Learning Techniques. *Appl. Sci.* **2022**, *13*, 233. [CrossRef]
24. Alhomaidat, F.; Abushattal, M.; Morgan Kwayu, K.; Kwigizile, V. Investigating the Interaction between Age and Liability for Crashes at Stop-Sign-Controlled Intersections. *Transp. Res. Interdiscip. Perspect.* **2022**, *14*, 100612. [CrossRef]
25. Amin, S. Backpropagation-Artificial Neural Network (BP-ANN): Understanding Gender Characteristics of Older Driver Accidents in West Midlands of United Kingdom. *Saf. Sci.* **2020**, *122*, 104539. [CrossRef]
26. Amiri, A.M.; Sadri, A.; Nadimi, N.; Shams, M. A Comparison between Artificial Neural Network and Hybrid Intelligent Genetic Algorithm in Predicting the Severity of Fixed Object Crashes among Elderly Drivers. *Accid. Anal. Prev.* **2020**, *138*, 105468. [CrossRef]
27. Fiorentini, N.; Losa, M. Handling Imbalanced Data in Road Crash Severity Prediction by Machine Learning Algorithms. *Infrastructures* **2020**, *5*, 61. [CrossRef]
28. Mafi, S.; AbdelRazig, Y.; Doczy, R. Machine Learning Methods to Analyze Injury Severity of Drivers from Different Age and Gender Groups. *Transp. Res. Rec. J. Transp. Res. Board* **2018**, *2672*, 171–183. [CrossRef]
29. Taghipour, H.; Parsa, A.B.; Chauhan, R.S.; Derrible, S.; Mohammadian, A. (Kouros) A Novel Deep Ensemble Based Approach to Detect Crashes Using Sequential Traffic Data. *IATSS Res.* **2022**, *46*, 122–129. [CrossRef]
30. Gu, X.; Lu, X.; Jin, X.; Guo, Y.; Zhou, Y.; Chen, Y. Analysis of Studies on Traffic Crashes Involving the Elderly. *Int. Rev. Spat. Plan. Sustain. Dev.* **2023**, *11*, 4–23. [CrossRef]
31. Lunardon, N.; Menardi, G.; Torelli, N. ROSE: A Package for Binary Imbalanced Learning. *R J.* **2014**, *6*, 79–89. [CrossRef]
32. Tantithamthavorn, C.; Hassan, A.E.; Matsumoto, K. The Impact of Class Rebalancing Techniques on the Performance and Interpretation of Defect Prediction Models. *IEEE Trans. Softw. Eng.* **2020**, *46*, 1200–1219. [CrossRef]
33. Menardi, G.; Torelli, N. Training and Assessing Classification Rules with Imbalanced Data. *Data Min. Knowl. Discov.* **2012**, *28*, 92–122. [CrossRef]
34. Gupta, R.; Asgari, H.; Azimi, G.; Rahimi, A.; Jin, X. Analysis of Fatal Truck-Involved Work Zone Crashes in Florida: Application of Tree-Based Models. *Transp. Res. Rec. J. Transp. Res. Board* **2021**, *2675*, 1272–1290. [CrossRef]
35. Rendón, E.; Alejo, R.; Castorena, C.; Isidro-Ortega, F.J.; Granda-Gutiérrez, E.E. Data Sampling Methods to Deal with the Big Data Multi-Class Imbalance Problem. *Appl. Sci.* **2020**, *10*, 1276. [CrossRef]
36. Vilaça, M.; Macedo, E.; Coelho, M.C. A Rare Event Modelling Approach to Assess Injury Severity Risk of Vulnerable Road Users. *Safety* **2019**, *5*, 29. [CrossRef]
37. Older Drivers | NHTSA. Available online: https://www.nhtsa.gov/road-safety/older-drivers (accessed on 15 March 2023).
38. Haixiang, G.; Yijing, L.; Shang, J.; Mingyun, G.; Yuanyue, H.; Bing, G. Learning from Class-Imbalanced Data: Review of Methods and Applications. *Expert Syst. Appl.* **2017**, *73*, 220–239. [CrossRef]
39. Thabtah, F.; Hammoud, S.; Kamalov, F.; Gonsalves, A. Data Imbalance in Classification: Experimental Evaluation. *Inf. Sci.* **2020**, *513*, 429–441. [CrossRef]
40. Pei, X.; Sze, N.N.; Wong, S.C.; Yao, D. Bootstrap Resampling Approach to Disaggregate Analysis of Road Crashes in Hong Kong. *Accid. Anal. Prev.* **2016**, *95*, 512–520. [CrossRef] [PubMed]
41. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
42. Karacasu, M.; Ergül, B.; Altın, A. Estimating the Causes of Traffic Accidents Using Logistic Regression and Discriminant Analysis. *Int. J. Inj. Control Saf. Promot.* **2013**, *21*, 305–313. [CrossRef] [PubMed]
43. Zhang, D.; Zhao, X.; Han, J.; Zhao, Y. A Comparative Study on PCA and LDA Based EMG Pattern Recognition for Anthropomorphic Robotic Hand. In Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, 31 May—7 June 2014; pp. 4850–4855.
44. Yang, L.; Gao, H.; Wu, K.; Zhang, H.; Li, C.; Tang, L. Identification of Cancerlectins by Using Cascade Linear Discriminant Analysis and Optimal G-Gap Tripeptide Composition. *Curr. Bioinform.* **2020**, *15*, 528–537. [CrossRef]
45. Mothwa, L.; Tapamo, J.-R.; Mapati, T. Conceptual Model of the Smart Attendance Monitoring System Using Computer Vision. In Proceedings of the 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Las Palmas de Gran Canaria, Spain, 26–29 November 2018; pp. 229–234.
46. Yan, X.; He, J.; Zhang, C.; Liu, Z.; Qiao, B.; Zhang, H. Single-Vehicle Crash Severity Outcome Prediction and Determinant Extraction Using Tree-Based and Other Non-Parametric Models. *Accid. Anal. Prev.* **2021**, *153*, 106034. [CrossRef]

47. Dimitrijevic, B.; Khales, S.D.; Asadi, R.; Lee, J. Short-Term Segment-Level Crash Risk Prediction Using Advanced Data Modeling with Proactive and Reactive Crash Data. *Appl. Sci.* **2022**, *12*, 856. [CrossRef]
48. Guo, M.; Yuan, Z.; Janson, B.; Peng, Y.; Yang, Y.; Wang, W. Older Pedestrian Traffic Crashes Severity Analysis Based on an Emerging Machine Learning XGBoost. *Sustainability* **2021**, *13*, 926. [CrossRef]
49. Islam, M.K.; Reza, I.; Gazder, U.; Akter, R.; Arifuzzaman, M.; Rahman, M.M. Predicting Road Crash Severity Using Classifier Models and Crash Hotspots. *Appl. Sci.* **2022**, *12*, 11354. [CrossRef]
50. Jeong, H.; Jang, Y.; Bowman, P.J.; Masoud, N. Classification of Motor Vehicle Crash Injury Severity: A Hybrid Approach for Imbalanced Data. *Accid. Anal. Prev.* **2018**, *120*, 250–261. [CrossRef]
51. Adams, J.; Hillman, M. The Risk Compensation Theory and Bicycle Helmets. *Inj. Prev.* **2001**, *7*, 89–91. [CrossRef]