Comparative Analysis of Machine Learning Based Algorithms for Predicting Injury Severity in Road Accidents

Soumaya AMRI[1], Mohammed AL ACHHAB[2], Mohamed LAZAAR[3]
Faculty of Sciences in Tetuan, Abdelmalek Essaadi University, Tetuan, Morocco[1, 2]
ENSIAS, Mohammed V University in Rabat, Rabat, Morocco[3]

Abstract—**Road crash injury severity prediction is essential for intelligent transportation systems, yet challenged by severe class imbalance, rigid 4-class severity schemes (unhurt/slight/hospitalized/fatal), and optimal methodological selection. This study proposes a structured framework systematically evaluating four machine learning models—CatBoost, HistGradientBoosting, Random Forest, and SVM—across multiclass (native 4-class and ordinal wrapper), binary reduction (non-severe vs. severe), and oversampling techniques using crash data. Multiclass approaches reveal tree ensemble dominance but persistent rare severe class prediction difficulties. Binary class reduction substantially improves severe injury detection performance on this dataset by simplifying decision boundaries, while SMOTE oversampling provides algorithm-specific imbalance mitigation. Random Forest demonstrates the most stable binary performance across evaluation metrics, independent of oversampling strategies. This performance gain comes at the cost of reduced severity granularity compared to the original multiclass formulation. Overall, under imbalance-sensitive evaluation metrics, binary class reduction provides a pragmatic and operationally effective alternative to complex multiclass strategies for severe injury detection.**

Keywords—Machine learning; imbalanced data; road accident; multiclass classification; binary classification; injury severity prediction

I. Introduction

Injuries, particularly among road users aged 15 to 64, result in significant mortality and long-term disability, ultimately reducing the workforce and weakening the economy. According to the International Road Assessment Program (iRAP), road traffic injuries contribute an estimated $1.85 trillion burden to the global economy each year [1].

In a world of rapidly advancing human mobility and a growing need for transport infrastructure, road safety has become a critical component of government initiatives aimed at reducing road user fatality rates.

The availability of large datasets and advancements in machine learning techniques provide powerful tools for analyzing traffic data and supporting the development of intelligent transportation systems before the implementation of new road safety strategies. However, predicting the severity of injuries resulting from road crashes is a complex challenge. Data collection, often conducted manually by police officers, frequently results in substantial inaccuracies and incomplete records. PIARC, the World Road association noted that a balance between the data collection burden and practical constraints is necessary to ensure police compliance in completing accident reports [2]. Furthermore, the contextual circumstances of each crash blur the line between rare events and standard classification tasks, requiring specialized techniques to manage data imbalance [3]. The factors influencing accident occurrence are highly intricate, further increasing the complexity and rendering it a non-linear classification problem [4]. Combined, these elements make the application of machine learning techniques in road safety research and intelligent transportation systems challenging at every stage of the prediction process.

Despite the substantial body of research on machine learning–based injury severity prediction, an important gap persists. Specifically, the combined influence of data imbalance, preprocessing strategies, and algorithm selection remains insufficiently examined. This limitation is particularly evident in large-scale, real-world crash datasets. While some studies focus on improving model architectures or evaluating different algorithms, others address data imbalance in isolation. However, few investigations examine the interaction between robust data preprocessing techniques and machine learning methods. Such integrated analysis is essential to enhance the accurate prediction of rare yet critical cases, including severe injuries in road accidents. This gap limits our understanding of how to effectively optimize the entire machine learning pipeline for imbalanced classification problems.

To address the challenges associated with road crash injury severity prediction, this study presents a critical review of related research. It also proposes a step-by-step prediction framework based on machine learning techniques and a real-world crash dataset. The objective is to establish a rigorous methodology for identifying a robust predictive model for injury severity.

The study further examines the challenges of training machine learning models on large, imbalanced datasets. It addresses multiclass injury severity prediction and systematically evaluates binary class reduction, nominal multiclass learning, and ordinal wrapper formulations. The goal is to identify the most suitable approach for similar applications.

The structure of the study is as follows: The second section reviews related works on predicting the severity of road crash injuries, with a particular focus on the application of machine learning techniques in similar contexts. The third section describes the proposed prediction framework, detailing its architecture and key interactions. The fourth section explains the experimental evaluation of the proposed model. The fifth section discusses the performance results obtained from the experiments. Finally, the last section concludes the study, summarizing the main findings and contributions.

II. Related Work

Numerous significant studies have investigated injury severity in road crashes, outlining various methods to predict injury severity levels. These prediction methodologies include classical statistical models [5], Poisson regression [6], logistic regression analysis [7] and discriminant analysis [8]. They also encompass advanced artificial intelligence approaches, including neural networks [9], decision trees [10], random forest models [11], and ensemble learning techniques [3].

The selection of appropriate machine learning techniques during the prediction process is influenced by data characteristics (data size and quality), as well as the specific objectives of each case study.

To provide a systematic evaluation of prior work, Table I synthesizes the methodological and empirical characteristics of related injury severity prediction studies.

TABLE I. Synthesis of Research Papers on Crash Severity Modeling and Predictive Performance

Study	Dataset Size	Classes	Ratio of severe classes	Main ML Model	Best Global Model Results	Severe Injury Results
[3]	308641	4	0.8% (Severe injury and fatal)	Gradient Boosting	Accuracy= 0.82	F1 score = 0.21, Precision = 0.31, Recall = 0.16
[10]	2238	3 (Slight, Serious, Fatal)	41,91% (Serious Injuries =39.83% Fatalities= 2.08%)	Ordered Forest	Gmean = 0.47	Precision =0,50
[12]	770561	3 levels initially later integrated into 2 classes	14.5% (13.3% serious and 1.2% fatal injuries)	SSAE	Precision=0.80, Recall=0.85, F1=0.84	Precision=0.42, Recall=0.21, F1=0.28
[13]	54364	3 (Slight, Severe, Fatal)	2.49% (2.34% severe and 0.15% fatal)	2D-CNN	F1= 0.95	F1-score = 0.004 (Fatal)
[15]	5740	4 levels (Slight, Medium, Severe, Fatal)	9.69% (6.74% severe and 2.94% fatal)	Bayesian Network	Accuracy= 0.66	Not evaluated
[16]	900690	Binary	Not described	Random Forest , LightGBM	ROC-AUC scores = 0.93	Precision = 0.65
[17]	1620	3 (Fatal, Injury, No-injury)	4,75%	CART	Accuracy =0.67	Precision = 0.45 (Fatal)
[18]	8516	4 (Fatal, Hospitalized, Injured, and Damage-only)	65% (22.9% Fatal and 42.1% Hospitalized)	Random Forest	Accuracy = 0.73	Precision = 0.85 Recall = 0.44
[19]	3834	4 (Minor, Moderate, Severe, Fatal)	13% (7.6% Severe Injuries, 5.4%Fatal Injuries)	Adaboost, Random Forest	Accuracy = 0.87	Precision =0.27 Recall =0.48

A comparative analysis of model performance in cited studies [3],[10], [12], [13], [15]-[19] reveals substantial heterogeneity in dataset size, modeling strategies, and evaluation metrics, with no standardized methodological framework emerging. Moreover, severe injury prediction remains insufficiently examined, as reported performance often emphasizes global accuracy while overlooking minority-class detection.

The implementation of a generic method is complex, as it requires careful examination of several parameters to determine the most appropriate machine learning methodology.

The first criterion relates to the dataset size and the number of features; both the algorithm's performance and the available computational resources should be assessed based on these data characteristics. Additionally, techniques such as scaling and distributed computing can be considered to facilitate training on large datasets.

Another critical criterion involves the distribution of injury severity levels, as a significant challenge in road crash analysis is the presence of highly imbalanced data. In most case studies, severe injuries are less frequent than minor injuries, compelling researchers to adopt supplementary techniques to address imbalanced data challenges and improve road crash injury analysis.

In the crash-injury severity literature, three modeling strategies are commonly used: binary reduction, nominal multiclass learning, and ordinal multiclass modeling. Binary reduction is often applied to mitigate severe class imbalance and stabilize model training [20], [21]. However, it reduces the granularity required for prioritizing safety interventions. Recent studies emphasize preserving richer severity levels whenever possible [22].

Nominal multiclass methods, particularly tree ensembles and gradient boosting models such as XGBoost, LightGBM, and CatBoost, often achieve strong overall performance. When combined with resampling techniques like SMOTE-NC or class weighting, their results improve [23]. However, these models may confuse adjacent severity levels because they do not consider label order.

To address this limitation, ordinal approaches explicitly encode the inherent rank structure from minor to fatal injuries. These methods include cumulative-threshold decomposition and parametric ordinal regression models. They reduce distant misclassifications and enhance policy and clinical interpretability. Empirical studies report improved performance over nominal classifiers, especially when combined with imbalance handling and modern learners [24]. Recent advances in dynamic ensemble selection and interpretable ensemble methods further improve robustness in imbalanced multiclass settings. Overall, evidence supports retaining multiclass severity and adopting ordinal-aware learning with appropriate rebalancing to preserve information and respect outcome ordering [23].

The following provides a comparative analysis of the machine learning models employed in research related to predicting the severity of road crash injuries. It outlines their advantages and limitations.

Neural network models present numerous advantages, particularly their ability to capture complex nonlinear relationships between independent and dependent variables and their effectiveness in scaling to large datasets. This capability makes them particularly suitable for extensive road crash databases. However, a significant drawback is the potential for overfitting, which can compromise prediction accuracy. A study on traffic accident severity in Madrid employed a framework based on one- and two-dimensional convolutional neural networks [13]. The findings highlighted the framework's scalability and its rapid prediction capabilities, making it suitable for real-time applications. Nonetheless, the precision scores for fatal and serious injuries were notably low, primarily due to the influence of imbalanced data, which adversely affected the prediction outcomes for these injury categories.

The Bayesian network, evaluated on a medium-sized dataset of 5740 instances and four injury classes, achieved the highest overall accuracy on both the training and testing sets. Its performance was superior to that of the Exhaustive Chi-Square Automatic Interaction Detector tree and the Linear Support Vector Machine models [15].

Wahab explores various machine-learning algorithms to predict motorcycle crash severity in Ghana, utilizing a medium-sized dataset consisting of 8,000 instances with 14 features. The Random Forest model exhibited greater accuracy than the other machine learning models tested. However, its individual accuracy for each injury severity class, especially for fatal injuries, was notably low [18]. A comparative analysis of Decision Trees, RF, and Naïve Bayes, performed on a reduced dataset using combined feature selection techniques (including two-way ANOVA, regression analysis, and chi-square), demonstrated that RF achieved the best performance [14]. Zhang recommends employing the Ordered Forest algorithm for analyzing injury severity in single bicycle crashes, as it showed better predictive performance compared to the traditional Random Forest model [10].

AdaBoost and Gradient Boosting models offer notable advantages for crash severity prediction. These advantages stem primarily from their capacity for global sensitivity analysis, which enables the assessment of both individual and combined effects of influencing factors [3]. Jiang applied these two boosting algorithms to predict crash injury severity and reported overall accuracy superior to that of other models. However, the prediction performance for minority classes, particularly fatal and severe injuries, remained low. Jiang recommends using the F1-score and monetized evaluation metrics to better assess model performance in imbalanced data settings.

Conducted on a large road crash dataset from France, the LightGBM model achieved the highest F1-score and accuracy for predicting non-severe injuries, which represent the majority class. In contrast, the Random Forest model obtained the highest precision for predicting severe injuries, the minority class [16].

Abdulazeez evaluated several machine-learning algorithms to predict child occupant crash injury severity in Abu Dhabi. Adaboost, Bagging REP, ZeroR, and XGBoost achieved the best overall accuracy both before and after data balancing, even without feature selection. However, the prediction results for the severe injury class, categorized as the minority class, remained low, except for XGBoost and CatBoost, which yielded better performance after data balancing and class consolidation [19].

Thus, the results vary significantly depending on the research methodology and the specific characteristics of the datasets used. Random Forest, Adaboost, and Gradient Boosting models demonstrated the highest overall accuracy on small to medium-sized datasets. However, the highest prediction accuracy for the severe injury class, identified as the minority class, was achieved using RF, stacking, and Support Vector Machine models, respectively, though the accuracy values remain relatively low. While reducing the number of classes can improve accuracy metrics, addressing data imbalance may lead to overfitting in certain algorithms. The performance of these models should be further evaluated on large datasets to assess their robustness in real-time operations.

III. Methodology

This study addresses gaps and limitations in the literature on severe injury prediction. These include data collection inconsistencies, low performance for the severe injury minority class, and challenges in selecting appropriate machine learning models.

To overcome these issues, a step-by-step methodology is proposed, as illustrated in Fig. 1.

The first step involves data preparation issues, from data collection and integration to data cleaning.

The second step addresses data engineering tasks, including feature transformations and data exploration, through four key dimensions: temporal and atmospheric conditions, road characteristics, vehicle characteristics, and user profiles.

The third step comprises a structured data processing phase in which the machine learning methodology is selected via a comparative analysis of four experiments. These experiments target two core challenges in multiclass injury severity prediction: selecting effective techniques to handle class imbalance and assessing how class grouping strategies influence predictive performance on minority classes.

Experiments N1–N2 retain the original four-class injury severity scale. Experiment N1 evaluates candidate models on the unmodified four-class dataset (nominal multiclass), while experiment N2 introduces an ordinal classification framework that explicitly encodes the ordered nature of the four severity levels.

Experiments N3–N4 adopt a binary formulation (severe vs non-severe). Experiment N3 trains the same models on the original imbalanced data to quantify the impact of multiclass reduction without correcting the imbalance. Experiment N4 applies SMOTE within each training fold of the cross-validation procedure. The data are balanced using oversampling performed only on the training subsets. The validation folds remain untouched, thereby preventing data leakage. Multiple classifiers are then evaluated on the resampled training data. Comparing these settings reveals the relative benefits of multiclass reduction and data balancing for severe injury detection and highlights which algorithms benefit most from SMOTE-like resampling.

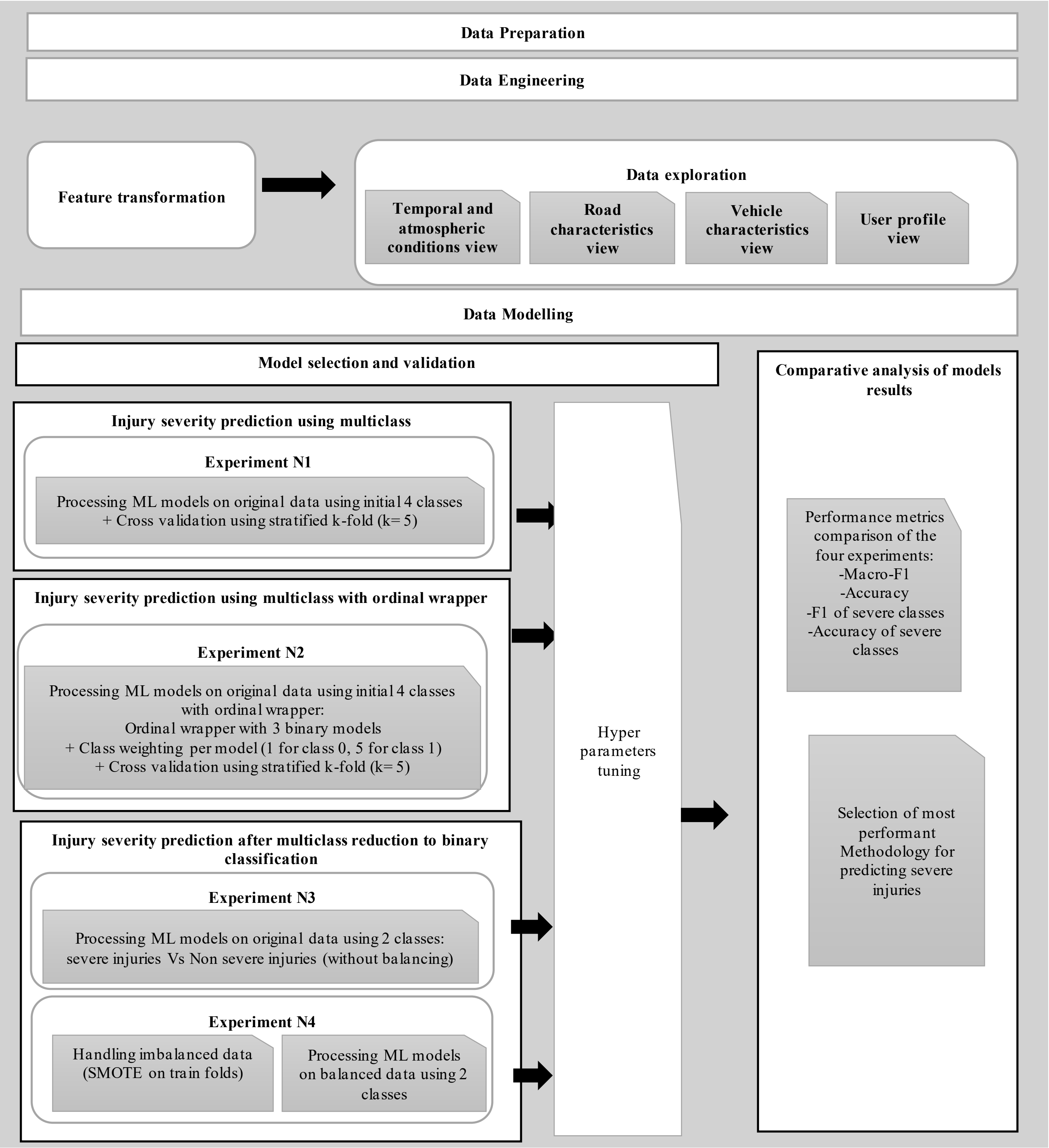
Data Preparation
Data Engineering
Feature transformation
Data exploration
Temporal and atmospheric conditions view
Road characteristics view
Vehicle characteristics view
User profile view
Data Modelling
Model selection and validation
Injury severity prediction using multiclass
Experiment N1
Processing ML models on original data using initial 4 classes + Cross validation using stratified k-fold (k= 5)
Injury severity prediction using multiclass with ordinal wrapper
Experiment N2
Processing ML models on original data using initial 4 classes with ordinal wrapper:
Ordinal wrapper with 3 binary models
+ Class weighting per model (1 for class 0, 5 for class 1)
+ Cross validation using stratified k-fold (k= 5)
Injury severity prediction after multiclass reduction to binary classification
Experiment N3
Processing ML models on original data using 2 classes: severe injuries Vs Non severe injuries (without balancing)
Experiment N4
Handling imbalanced data (SMOTE on train folds)
Processing ML models on balanced data using 2 classes
Hyper parameters tuning
Comparative analysis of models results
Performance metrics comparison of the four experiments:
-Macro-F1
-Accuracy
-F1 of severe classes
-Accuracy of severe classes
Selection of most performant Methodology for predicting severe injuries

Fig. 1. Framework of the proposed prediction methodology.

Jointly analyzing results across all four experiments enables the identification of the most effective combination of model, imbalance-handling strategy, and class representation scheme for accurately predicting severe injuries in highly imbalanced road safety data.

IV. Experimentation

A. Data Description

This research is conducted on real data sourced from the annual databases of road traffic accidents maintained by the French National Interdepartmental Observatory of Road Safety.

The annual databases extracted from the Bodily Accident Analysis Report [25] encompass all bodily traffic accidents occurring in metropolitan France, as well as in overseas departments and other overseas territories. This study conducted experiments using datasets from the years 2005 to 2020. The recorded data on road traffic accidents contains multiple layers of information, including features related to crash characteristics, location, involved vehicles, and road users.

To create the input dataset for this study, tables were joined based on the foreign keys specified in each data file, resulting in a comprehensive global dataset. Following the integration of 64 data files, corresponding to four files for each year, the final dataset comprised a total of 2,380,573 entries and 57 features, which served as the input data for this research.

B. Data Cleaning

In classification tasks, data cleaning focuses on enhancing classification performance by improving the quality of the training data. Common data cleaning techniques include outlier removal [26], missing value imputation [27], normalization and standardization [28], error detection and correction, handling duplicate values, noise reduction [29] and categorical data handling [30].

In this study, the initial step involved merging the collected datasets before applying data cleaning techniques. Error detection and correction were performed by casting column types (date, numeric, string, etc.) and correcting incorrect value formats for numerical columns as well as erroneous entries in categorical columns. Missing value imputation techniques were employed to address null values. A percentage count of missing values for each feature was generated and classified to identify unusable columns or records.

To avoid imputation of non-representative data for certain features, fixed intervals were established for handling missing data, as outlined in Fig. 2, which describes the data cleaning process and corresponding actions. Columns with over 89.9% missing values were removed, except in cases where these columns contained information specific to certain types of road users, such as pedestrians. In such cases, null values are not indicative of missing data but rather signify that this information is irrelevant to other types of road users. For instance, null values in the columns “locp” (pedestrian location) and “etatp” (pedestrian status, e.g. alone, accompanied) are not considered missing data for records of crashes involving only vehicles and no pedestrians.

Records with less than 5% missing data in any column were deleted. For columns with between 5% and 89.9% missing values, various imputation methods were applied: mean imputation was used for numerical data, mode imputation for categorical data, and, in some cases, imputation by deduction based on a data dictionary. This latter method was applied to categorical features where undefined values could be reasonably inferred based on similar meanings within the data.

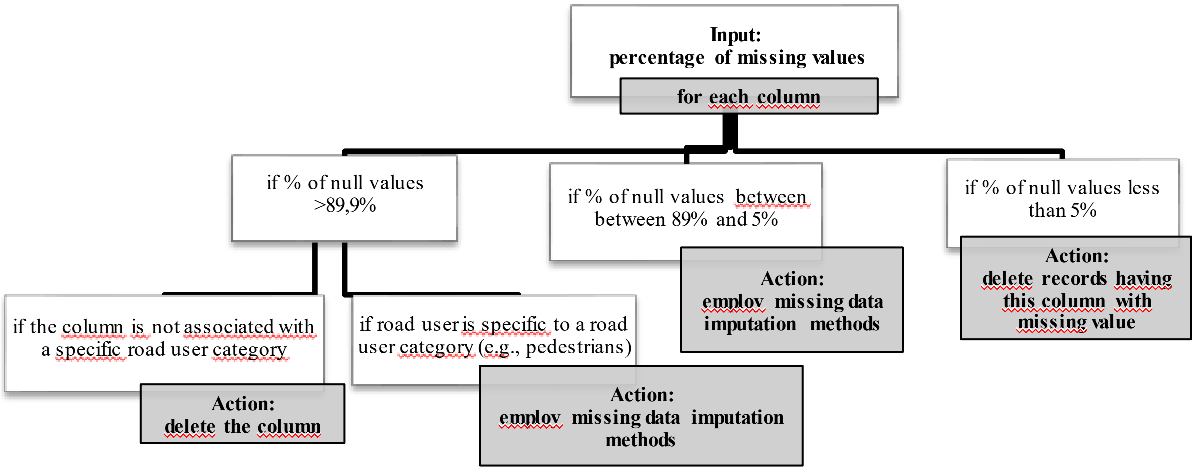
Input:
percentage of missing values
for each column
if % of null values >89,9%
if % of null values between between 89% and 5%
if % of null values less than 5%
Action: employ missing data imputation methods
Action: delete records having this column with missing value
if the column is not associated with a specific road user category
if road user is specific to a road user category (e.g., pedestrians)
Action: delete the column
Action: employ missing data imputation methods

Fig. 2. Data cleaning process and actions.

C. Data Engineering

1) Feature transformation: To prepare the data for exploratory analysis, feature transformations were conducted, including the formatting of existing features and the generation of new ones. Eight additional features were derived from the columns "date," "time," and "user date of birth" to explicitly represent embedded information: "time slice", "year", "month", "day", "day of week", "is holiday", "age", and "age slice". The final task involved encoding the newly created categorical features to complete the data preparation and transformation process.

2) Exploratory data analysis: To analyze crash severity, an exploration of the distribution of accidents based on the injury severity level feature is conducted. This feature is categorized into four classes: unhurt, slightly injured, hospitalized injuries, and fatalities. Fig. 3 illustrates the distribution of injury severity levels. It demonstrates that fewer than 2% of road users fall into the killed category and 20% into hospitalized due to injury. These two categories are identified as minority classes within the dataset.

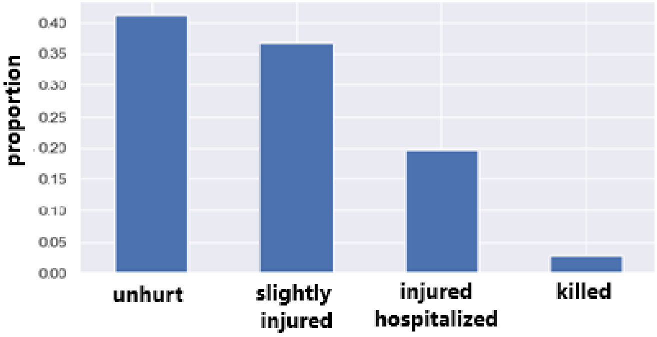

Fig. 3. Distribution of crashes by injury severity level.

3) Data modelling: The injury severity feature, originally comprising four levels (1=Unhurt, 2=Fatal, 3=Hospitalized, 4=Slight injury), was systematically remapped in Experiment N2. An ordinal scale was adopted to preserve the hierarchical relationships: 0 = Unhurt < 1 = Slight injury < 2 = Hospitalized < 3 = Fatal. In experiments N3 and N4, the four levels of injury severity are reduced to two classes: severe (fatal and hospitalized) and non-severe (unhurt and slight injury).

The methodological framework structures four complementary experiments with differentiated modeling strategies:

- Experiment N1: Multiclass prediction (4 ordinal classes) on original data.
- Experiment N2: Ordinal wrapper (3 binary classifiers) on original 4-class data with class weighting.
- Experiment N3: Binary classification on original imbalanced data.
- Experiment N4: Binary classification on SMOTE-balanced data (Non Severe vs Severe Injury).

This progression systematically evaluates binary vs multiclass formulations alongside data-level and algorithmic imbalance mitigation strategies.

4) Handling imbalanced data: Several machine learning studies address the issue of imbalanced data and attempt to mitigate its impact on the performance of models. Undersampling and oversampling methods, which aim to create equal amounts of data points for each class, are commonly employed to manage this imbalance. Subsequently, various developed models have been proposed. Galar categorizes techniques for handling imbalanced data into four types: algorithmic-level, data-level, cost-sensitive, and ensembles of classifiers [31].

In this study, class imbalance (Class “unhurt”: 41%, Class “fatal”: 2.7%) was addressed through complementary strategies across experiments, following Galar's established taxonomy:

- Experiments N1-N2: Algorithmic-level mitigation via stratified k-fold cross-validation (k=5) preserves original distribution while optimizing for rare classes.
- Experiment N2: Cost-sensitive ordinal wrapper implements class_weights= [1.0, 5.0] per binary classifier P(Y>k), assigning fivefold misclassification penalty to severe cases (>k threshold).
- Experiment N3: data-level evaluation using binary classification on original imbalanced data to establish the performance of multiclass reduction.
- Experiment N4: SMOTE oversampling applied to training folds generates synthetic minority class instances via k-nearest neighbor interpolation, proven effective for traffic crash severity prediction [32], [33]. It generates new synthetic instances of the minority class by interpolating values between the two nearest data points along a straight line [34].

This multi-strategy framework systematically compares data-level oversampling (SMOTE), cost-sensitive ordinal learning, and binary class reduction strategies for severe injury prediction.

5) ML modelling:

a) Model selection: In this work, the dataset is both large and highly imbalanced. In similar case studies, boosting algorithms, such as Gradient Boosting and the Random Forest algorithm, have demonstrated superior performance compared to other models. Therefore, four machine-learning algorithms are evaluated to analyze the severity level of injuries caused by road crashes: CatBoost, HistGradientBoosting Classifier, Random Forest, and Linear Support Vector Machine (SVM). The SVM model was selected for its well-documented effectiveness in classification tasks.

b) Hyperparameter tuning: Model hyperparameters were systematically optimized using stratified k-fold cross-validation (k=5, reduced to k=3 for 2000 iterations) with early stopping, following the methodological framework. Iteration counts were evaluated across {500, 800, 1000, 2000}, with Macro F1 serving as the primary optimization criterion across all experiments (N1-N4) and models. Optimal iterations were selected independently for each model-experiment combination, ensuring configuration-specific performance maximization while preventing overfitting. Cross-validation improvement curves demonstrating macro-F1 gains versus iteration count, along with final selected hyperparameters per model, are detailed in section V.

6) Evaluation metrics: To evaluate the performance of machine learning models, various metrics are employed by researchers. Accuracy, along with precision, recall, and F1-score, is among the most commonly used evaluation metrics.

In injury severity prediction studies, all of the metrics mentioned previously are employed. However, relying solely on the overall average prediction accuracy for highly imbalanced data can lead to biased interpretations of the results. Jiang suggests using additional metrics, such as the F1 score for each class and the monetized measurement (MM), which represents the weighted average of the F1 scores for each class [35].

In this study, Macro-F1 constitutes the primary evaluation metric across all four experiments, computing the unweighted average of per-class F1-scores to ensure equitable minority class contribution despite severe imbalance. This metric robustly penalizes poor severe injury detection masked by majority class dominance, representing the established standard for multiclass imbalanced traffic safety prediction [36].

A comprehensive set of evaluation metrics was applied consistently across all experiments. For each configuration (Experiments N1–N4), model performance was primarily assessed using macro-F1. This was complemented by per-class precision, recall, and F1-scores, as well as weighted F1 and overall accuracy. These metrics were used to capture both global and class-specific behavior under class imbalance.

For Experiment N2, which models the ordinal nature of injury severity, quadratic weighted kappa was additionally employed. It was implemented using the cohen_kappa_score function with quadratic weights. This metric quantifies agreement beyond chance and penalizes distant misclassifications more strongly than adjacent ones. Such a property is particularly appropriate for hierarchical severity scales.

This unified evaluation framework enables comparison across Experiments N1–N4. It supports the selection of the most appropriate methodology for severe injury prediction by jointly considering overall predictive performance and the accurate identification of clinically critical yet rare cases.

V. Results and Discussion

This study primarily aims to examine the challenges in predicting road crash injury severity, particularly for minority classes, and to identify the most effective machine learning approach for accurate prediction. To this end, hyperparameter tuning results are first analyzed for selected models across each experiment to determine optimal parameters. Model performance is then evaluated individually within the four experiments, followed by a comprehensive comparison that highlights strengths, weaknesses, and the superior methodology for injury severity prediction.

A. Hyperparameter Tuning Results

To ensure optimal model configuration across all four experiments and algorithms, hyperparameters were tuned using stratified cross-validation and early stopping, following the methodological framework outlined in Fig. 1.

1) CatBoost hyperparameter evaluation: For the CatBoost model in experiment N1 (initial 4-class configuration), hyperparameter tuning focused on varying the number of iterations (500-2000) while fixing learning_rate=0.05 and depth=6.

Fig. 4 presents the macro F1-score evolution across iteration counts. Optimal performance was achieved with iterations=2000, learning_rate=0.05, depth=6 (F1_macro = 0,648), representing the best configuration identified through systematic grid evaluation.

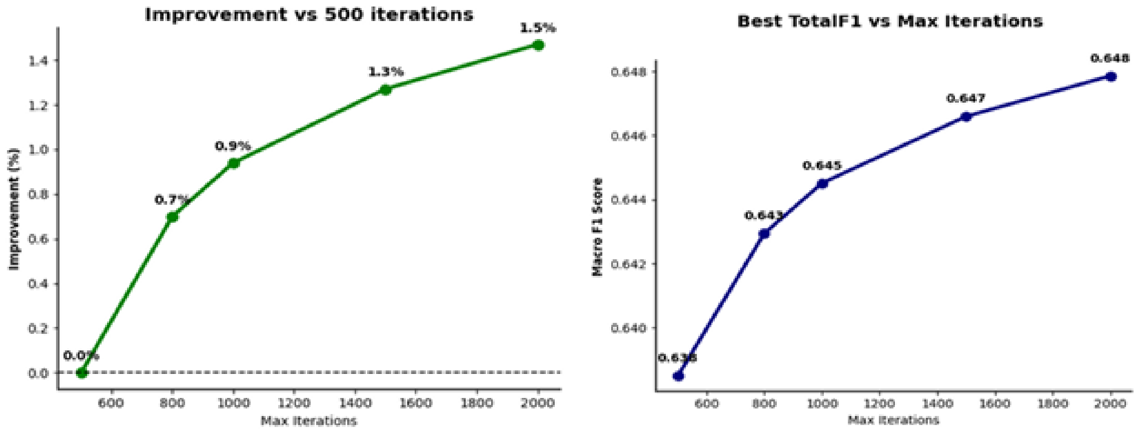

Fig. 4. Macro F1 evolution of experiment N1 across iterations.

For subsequent experiments, the same hyperparameter tuning procedure was applied to identify the optimal configuration.

2) Histgradient boosting classifier hyperparameter evaluation: For the HistGradientBoostingClassifier, hyperparameter tuning was conducted using RandomizedSearchCV with 12 iterations and 5-fold cross-validation, systematically exploring combinations of learning_rate (0.05-0.20), max_iter (100-300), max_leaf_nodes (15-31), min_samples_leaf (10-20), and max_depth (3-5 or None).

Fig. 5 presents the macro F1-score performance across hyperparameter ranges applied in experiment N1 using the initial four classes of injury severity. Optimal performance was achieved with min_samples_leaf=20, max_leaf_nodes=31, max_iter=100, max_depth=5, learning_rate=0.2 (F1_macro = 0.484, std=0.013), representing the best configuration identified through randomized search.

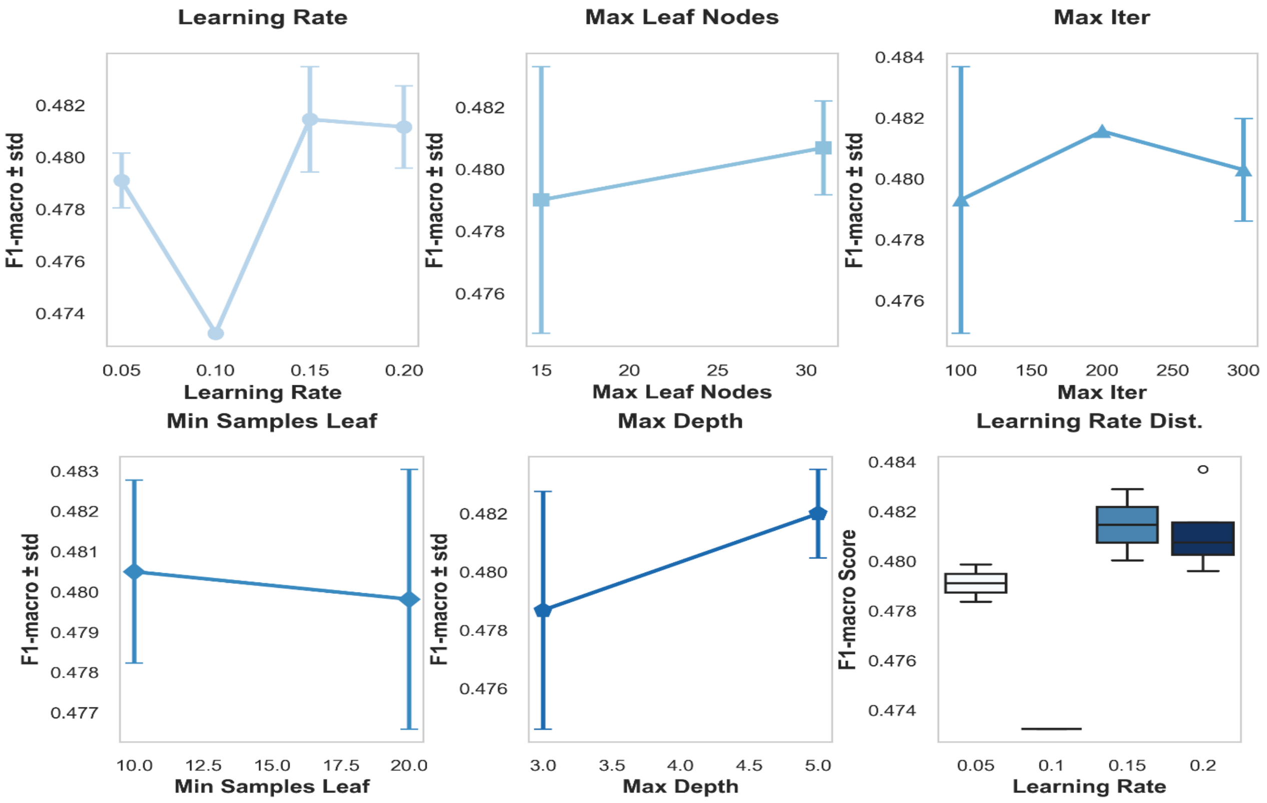

Fig. 5. Performance of hyperparameter tuning of the histgradient boosting classifier on Macro F1-score.

3) *Random forest hyperparameter evaluation:* A systematic grid search using cross_val_score (3-fold CV, F1_macro scoring) exhaustively evaluated eight configurations of RandomForestClassifier, focusing on n_estimators ∈ {100, 200} and max_depth ∈ {8, 12}, with fixed min_samples_split=5, min_samples_leaf=5, max_features=0.3 and using a balanced class_weight. The optimal configuration (n_estimators=100, max_depth=12) achieved the highest mean CV score, as confirmed by the hyperparameter validation curve presented in Fig. 6.

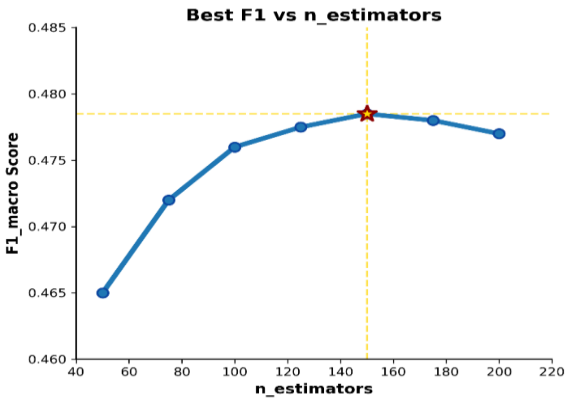

Fig. 6. Impact of RF hyper-parameter tuning of Macro-F1.

4) *SVM hyperparameter evaluation:* Hyperparameter tuning of the SVM classifier was conducted using RandomizedSearchCV with 10 iterations and 3-fold cross-validation, optimizing for the F1-macro score across four classes on a balanced dataset (class_weight='balanced'). The search explored a constrained parameter space comprising C values [0.1, 1, 10, 50, 100], RBF kernel with gamma values ['scale', 0.01, 0.1, 1.0], and fixed degree=3, completing 30 fits efficiently in a single-threaded configuration (n_jobs=1).

The optimal configuration (C=100, kernel='rbf', gamma=0.01, degree=3) achieved a cross-validated F1-macro score of 0.172 as presented in Fig. 7. This result indicates substantial challenges in multiclass classification, likely due to class imbalance and limited feature separability in the RBF-induced feature space.

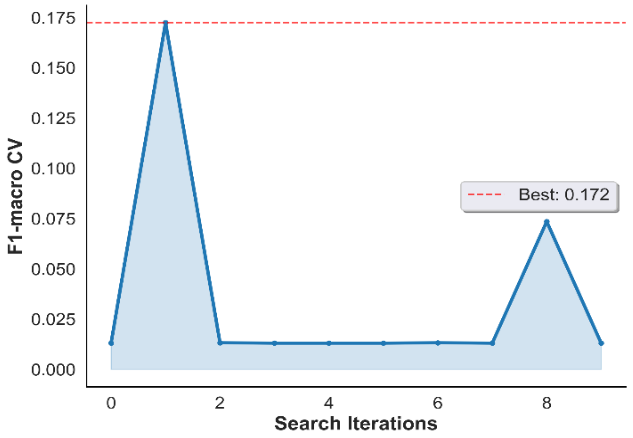

Fig. 7. SVM Hyperparameter tuning convergence.

B. Models Performance Analysis

1) Multiclass classification: The first two experiments within this methodological framework evaluate model performance using the initial four-class injury severity scheme. The second experiment incorporates an ordinal wrapper approach; together, these experiments compare model performance between native class weighting and ordinal wrapper configurations.

Table II presents performance metrics for CatBoost, HistGradientBoosting, Random Forest, and SVM models. The primary evaluation metrics, in order of priority, are macro F1-score, accuracy, and precision for severe classes (Hospitalized and Fatal).

TABLE II. MULTICLASS CLASSIFICATION PERFORMANCE OF EXPERIMENT N1 (ORIGINAL 4-CLASS) VS. EXPERIMENT N2 (ORDINAL WRAPPER)

Model		**Injury severity level**	**Experiment N1: Original multiclass**			**Experiment N2: Ordinal wrapper multiclass**		
	Metric		***Precision***	***Recall***	***F1-score***	***Precision***	***Recall***	***F1-score***
Catboost		Unhurt	0.65	0.90	0.75	0.81	0.59	0.68
		Slight	0.67	0.47	0.55	0.49	0.38	0.43
		Hospitalized	0.50	0.44	0.47	0.35	0.75	0.47
		Fatal	0.54	0.03	0.06	0.29	0.18	0.22
	Accuracy				0.63			0.53
	Macro avg		0.59	0.46	0.46	0.48	0.47	0.45
	Weighted avg		0.63	0.63	0.61	0.59	0.53	0.54
Hist-gradient Boosting classifier		Unhurt	0.67	0.87	0.76	0.70	0.83	0.76
		Slight	0.70	0.34	0.46	0.64	0.26	0.37
		Hospitalized	0.40	0.43	0.41	0.30	0.32	0.31
		Fatal	0.16	0.52	0.24	0.11	0.65	0.19
	Accuracy				0.58			0.52
	Macro avg		0.48	0.54	0.47	0.44	0.52	0.41
	Weighted avg		0.62	0.58	0.57	0.59	0.52	0.52
Random Forest		Unhurt	0.71	0.77	0.74	0.46	0.91	0.61
		Slight	0.65	0.40	0.49	0.38	0.11	0.17
		Hospitalized	0.36	0.28	0.31	0.44	0.11	0.17
		Fatal	0.10	0.70	0.18	0.20	0.20	0.20
	Accuracy				0.54			0.44
	Macro avg		0.46	0.54	0.43	0.37	0.33	0.29
	Weighted avg		0.61	0.54	0.55	0.42	0.44	0.35
SVM		Unhurt	0.41	0.88	0.56	0.52	0.15	0.23
		Slight	0.04	0.07	0.05	0.38	0.42	0.40
		Hospitalized	0.16	0.05	0.08	0.21	0.48	0.29
		Fatal	0.00	0.00	0.00	0.03	0.05	0.04
	Accuracy				0.38			0.31
	Macro avg		0.15	0.25	0.17	0.28	0.27	0.24
	Weighted avg		0.20	0.38	0.25	0.40	0.31	0.30

2) Binary classification: Experiments N3 and N4 implement class reduction strategies to assess their impact on prediction performance. Experiment N3 collapses the four-class severity scheme into a binary classification framework (non-severe vs. severe). Experiment N4 introduces SMOTE oversampling as an additional balancing technique to evaluate its incremental benefit across the four models examined (CatBoost, HistGradientBoostingClassifier, Random Forest, and SVM).

Binary classification, as presented in Table III, reveals SMOTE's differential efficacy across algorithms: HistGradientBoosting shows a +13% improvement in accuracy (0.70 → 0.79) while the SVM achieves a substantial +169% increase in the severe-class F1-score (0.16→0.43). These results highlight the necessity of resampling for imbalance-sensitive architectures. In contrast, Random Forest and CatBoost demonstrate intrinsic robustness, maintaining consistently high performance (accuracy=0.83, ΔF1-severe < +2%).

Random Forest emerges as the overall top performer in binary classification (accuracy=0.83, F1-severe=0.50), exhibiting superior generalization across severity levels.

TABLE III. BINARY CLASSIFICATION PERFORMANCE OF EXPERIMENT N3 (CLASS REDUCTION) VS. EXPERIMENT N4 (WITH SMOTE OVERSAMPLING)

Model		Injury severity level	Experiment N3: Reduced to binary class			Experiment N4: Reduced to binary class + SMOTE balancing		
	Metric		*Precision*	*Recall*	*F1-score*	*Precision*	*Recall*	*F1-score*
Catboost		Non severe	0.85	0.95	0.90	0.86	0.94	0.90
		Severe	0.66	0.36	0.47	0.63	0.39	0.48
	Accuracy				0.83			0.83
	Macro avg		0.75	0.66	0.68	0.74	0.67	0.69
	Weighted avg		0.81	0.83	0.81	0.81	0.83	0.81
Hist-gradient Boosting classifier		Non severe	0.94	0.66	0.78	0.88	0.86	0.87
		Severe	0.39	0.84	0.54	0.50	0.55	0.53
	Accuracy				0.70			0.79
	Macro avg		0.67	0.75	0.66	0.69	0.71	0.70
	Weighted avg		0.83	0.70	0.73	0.80	0.79	0.80
Random Forest		Non severe	0.86	0.94	0.90	0.86	0.92	0.89
		Severe	0.63	0.41	0.50	0.60	0.44	0.51
	Accuracy				0.83			0.82
	Macro avg		0.74	0.67	0.70	0.73	0.68	0.70
	Weighted avg		0.81	0.83	0.81	0.81	0.82	0.81
SVM		Non severe	0.81	0.98	0.88	0.89	0.58	0.70
		Severe	0.50	0.10	0.16	0.31	0.71	0.43
	Accuracy				0.79			0.61
	Macro avg		0.65	0.54	0.52	0.60	0.65	0.57
	Weighted avg		0.7	0.79	0.73	0.77	0.61	0.65

C. Overall Performance Comparison

The proposed methodology systematically evaluates four machine learning models—CatBoost, HistGradientBoosting, Random Forest, and SVM—across multiclass injury severity prediction using an initial four-class scheme (Unhurt, Slight, Hospitalized, Fatal).

Experiment N1 applies native multiclass classification using stratified k-fold cross-validation. The results indicate moderate overall performance, primarily driven by tree-based ensemble methods. CatBoost achieves the highest accuracy (0.63) and macro-F1 score (0.46). However, all models show significant weaknesses in predicting rare severe classes, particularly the fatal category, with F1-scores below 0.25.

Experiment N2 introduces an ordinal wrapper with class weighting to account for injury severity's inherent ordering, yielding mixed outcomes. While recall for Hospitalized cases improves notably (e.g., CatBoost: 0.44 to 0.75), overall accuracy and Macro F1 decline slightly (CatBoost: 0.63 to 0.53; 0.46 to 0.45). The lower overall performance of the ordinal formulation may be partly influenced by class imbalance rather than reflecting an inherent limitation of the ordinal modeling framework. These results underscore the wrapper's limited efficacy in highly imbalanced multiclass settings and highlight the trade-offs of enforcing ordinal constraints without fully resolving class distribution skew.

Binary class reduction in experiment N3, collapsing the four classes into non-severe versus severe, markedly enhances predictive performance across all models, demonstrating the methodology's sensitivity to classification complexity. Accuracies increase to 0.79–0.83 for tree-based ensembles, compared to 0.38–0.63 in the multiclass setting. Severe-class F1-scores improve substantially (e.g., Random Forest reaches 0.50 from previous multiclass lows). This improvement results from a simplified decision boundary that reduces overfitting to the dominant unhurt and slight injury classes and increases focus on clinically relevant severe outcomes.

This shift validates the strategic class aggregation within the framework. The approach reduces modeling complexity while preserving prognostic utility. However, it sacrifices granular severity distinctions that are essential for detailed risk stratification.

Experiment N4 further addresses residual imbalance via SMOTE oversampling on training folds, producing algorithm-specific gains that affirm its value for imbalance-sensitive models. HistGradientBoosting accuracy rises 13% (0.70 to 0.79) with stable severe F1 (0.54 to 0.53), while SVM exhibits the most dramatic uplift in severe F1 (0.16 to 0.43, +169%), reflecting SMOTE's ability to augment synthetic severe minorities for linear classifiers. Conversely, intrinsically robust tree-based models such as Random Forest and CatBoost exhibit negligible performance variations (accuracies ~0.83 to 0.82;

severe F1 +2% max). These results indicate baseline resilience to class imbalance.

Within the framework's comparative validation, the binary Random Forest configuration emerges as the most suitable methodology for severe injury prediction.

Table IV quantifies performance uplifts across experiments for each model, using accuracy and severe F1-score (Hospitalized+Fatal multiclass; Severe binary) as key metrics.

In summary, the Random Forest model in the binary configuration (Experiment N3) emerges as the superior approach. It achieves the highest accuracy (0.83), a macro-F1 score of 0.70, a severe-class F1-score of 0.50, and a non-severe F1-score of 0.90. The model demonstrates exceptional stability across configurations, as shown by minimal performance degradation after SMOTE (-1.2% accuracy). It also records the largest improvement from multiclass to binary settings (+53.7% accuracy, +61.1% severe F1-score).

TABLE IV. PERFORMANCE IMPROVEMENT PERCENTAGES ACROSS EXPERIMENTAL CONFIGURATIONS

Model	Metric	Multiclass Exp. 1 → Binary Exp. 3 (%)	Binary Exp. 3 → SMOTE Exp. 4 (%)
CatBoost	Accuracy	+31.7% (0.63 → 0.83)	0.0% (0.83 → 0.83)
	Severe F1	+2.1% (0.27 → 0.47)*	+2.1% (0.47 → 0.48)
HistGradientBoosting	Accuracy	+20.7% (0.58 → 0.70)	+12.9% (0.70 → 0.79)
	Severe F1	+31.7% (0.33 → 0.54)*	-1.9% (0.54 → 0.53)
Random Forest	Accuracy	+53.7% (0.54 → 0.83)	-1.2% (0.83 → 0.82)
	Severe F1	+61.1% (0.31 → 0.50)*	+2.0% (0.50 → 0.51)
SVM	Accuracy	+107.9% (0.38 → 0.79)	-22.8% (0.79 → 0.61)
	Severe F1	+162.5% (0.08 → 0.16)*	+168.8% (0.16 → 0.43)

* Multiclass severe F1 approximated as avg(Hosp+Fatal).

This robustness stems from its intrinsic bagging mechanism, which effectively handles imbalance and feature interactions without relying on synthetic oversampling, outperforming even CatBoost's multiclass strengths and SVM's SMOTE-dependent gains.

VI. CONCLUSION

This study systematically evaluated machine learning models for predicting road crash injury severity within a structured methodological framework encompassing multiclass (original 4-class and ordinal wrapper), binary class reduction, and SMOTE oversampling configurations.

Multiclass experiments highlighted the dominance of tree-based ensemble methods, with CatBoost achieving the highest accuracy (0.63) and macro-F1 score (0.46). However, persistent difficulties in predicting rare severe classes emphasized the impact of class imbalance. Ordinal wrappers produced mixed improvements in recall, but often at the expense of overall performance metrics.

Binary class reduction substantially improved model performance across algorithms. Accuracies increased by 20–108%, and severe-class F1-scores improved by up to 162%. This strategy simplified decision boundaries while emphasizing clinically relevant severe outcomes. SMOTE further enhanced performance, particularly for linear classifiers such as SVM, with severe F1-score gains of up to 169%.

Random Forest with binary classification emerged as the optimal methodology, exhibiting unmatched stability across binary setups, validating the framework's comparative validation stage and affirming class aggregation's efficacy for imbalanced severity prediction.

These findings advance road safety analytics by demonstrating how strategic classification reduction outperforms complex ordinal modeling, enabling reliable severe injury flagging for proactive interventions in intelligent transportation systems.

However, limitations persist: the binary framing of the selected Random Forest model sacrifices multiclass granularity, potentially hindering fine-grained severity triage, while cross-validation results require external validation on diverse datasets to confirm generalizability beyond the current cohort. Future enhancements should incorporate explainability techniques (e.g., SHAP) to elucidate severe prediction drivers and bridge performance with clinical interpretability. Additionally, research should explore deep learning ensembles and multi-regional crash datasets.

Ultimately, deploying binary Random Forest models within real-time ITS frameworks holds promise for mitigating severe crash outcomes through data-driven risk prioritization

REFERENCES

[1] World Bank, "THE HIGH TOLL OF TRAFFIC INJURIES: Unacceptable and Preventable," 2017.

[2] World Road Association (PIARC), "ROAD SAFETY MANAGEMENT," in ROAD SAFETY MANUAL A GUIDE FOR PRACTITIONERS, in Version 4. , 2024. [Online]. Available: https://roadsafety.piarc.org/sites/safety/files/public/pdf/piarc_road_safety_management_2024_05_27_v4.pdf

[3] L. Jiang, Y. Xie, X. Wen, and T. Ren, "Modeling highly imbalanced crash severity data by ensemble methods and global sensitivity analysis," Journal of Transportation Safety & Security, vol. 14, no. 4, pp. 562–584, Apr. 2022, doi: 10.1080/19439962.2020.1796863.

[4] C. Carrodano, "Data-driven risk analysis of nonlinear factor interactions in road safety using Bayesian networks," Sci Rep, vol. 14, no. 1, p. 18948, Aug. 2024, doi: 10.1038/s41598-024-69740-6.

[5] W. L. Carlson, "Crash injury prediction model," Accident Analysis & Prevention, vol. 11, no. 2, pp. 137–153, Jun. 1979, doi: 10.1016/0001-4575(79)90022-8.

[6] S. C. Joshua and N. J. Garber, "Estimating truck accident rate and involvements using linear and Poisson regression models," Transportation

Planning and Technology, vol. 15, no. 1, pp. 41–58, Jun. 1990, doi: 10.1080/03081069008717439.

[7] A. S. Al-Ghamdi, "Using logistic regression to estimate the influence of accident factors on accident severity," Accident Analysis & Prevention, vol. 34, no. 6, pp. 729–741, Nov. 2002, doi: 10.1016/S0001-4575(01)00073-2.

[8] K. M. Kockelman and Y.-J. Kweon, "Driver injury severity: an application of ordered probit models," Accident Analysis & Prevention, vol. 34, no. 3, pp. 313–321, May 2002, doi: 10.1016/S0001-4575(01)00028-8.

[9] Md. E. Shaik, Md. M. Islam, and Q. S. Hossain, "A review on neural network techniques for the prediction of road traffic accident severity," Asian Transport Studies, vol. 7, p. 100040, 2021, doi: 10.1016/j.eastsj.2021.100040.

[10] Y. Zhang, H. Li, and G. Ren, "Analyzing the injury severity in single-bicycle crashes: An application of the ordered forest with some practical guidance," Accident Analysis & Prevention, vol. 189, p. 107126, Sep. 2023, doi: 10.1016/j.aap.2023.107126.

[11] M. Rezapour, A. Mehrara Molan, and K. Ksaibati, "Analyzing injury severity of motorcycle at-fault crashes using machine learning techniques, decision tree and logistic regression models," International Journal of Transportation Science and Technology, vol. 9, no. 2, pp. 89–99, Jun. 2020, doi: 10.1016/j.ijtst.2019.10.002.

[12] Z. Ma, G. Mei, and S. Cuomo, "An analytic framework using deep learning for prediction of traffic accident injury severity based on contributing factors," Accident Analysis & Prevention, vol. 160, p. 106322, Sep. 2021, doi: 10.1016/j.aap.2021.106322.

[13] L. Pérez-Sala, M. Curado, L. Tortosa, and J. F. Vicent, "Deep learning model of convolutional neural networks powered by a genetic algorithm for prevention of traffic accidents severity," Chaos, Solitons & Fractals, vol. 169, p. 113245, Apr. 2023, doi: 10.1016/j.chaos.2023.113245.

[14] H. Bhuiyan et al., "Crash severity analysis and risk factors identification based on an alternate data source: a case study of developing country," Sci Rep, vol. 12, no. 1, p. 21243, Dec. 2022, doi: 10.1038/s41598-022-25361-5.

[15] S. AlKheder, F. AlRukaibi, and A. Aiash, "Risk analysis of traffic accidents' severities: An application of three data mining models," ISA Transactions, vol. 106, pp. 213–220, Nov. 2020, doi: 10.1016/j.isatra.2020.06.018.

[16] K. Sefrioui Boujemaa, I. Berrada, K. Fardousse, O. Naggar, and F. Bourzeix, "Toward Road Safety Recommender Systems: Formal Concepts and Technical Basics," IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 6, pp. 5211–5230, Jun. 2022, doi: 10.1109/TITS.2021.3052771.

[17] L.-Y. Chang and J.-T. Chien, "Analysis of driver injury severity in truck-involved accidents using a non-parametric classification tree model," Safety Science, vol. 51, no. 1, pp. 17–22, Jan. 2013, doi: 10.1016/j.ssci.2012.06.017.

[18] L. Wahab and H. Jiang, "A comparative study on machine learning based algorithms for prediction of motorcycle crash severity," PLOS ONE, vol. 14, no. 4, p. e0214966, Apr. 2019, doi: 10.1371/journal.pone.0214966.

[19] M. U. Abdulazeez, W. Khan, and K. A. Abdullah, "Predicting child occupant crash injury severity in the United Arab Emirates using machine learning models for imbalanced dataset," IATSS Research, vol. 47, no. 2, pp. 134–159, Jul. 2023, doi: 10.1016/j.iatssr.2023.05.003.

[20] A. Fernández, V. López, M. Galar, M. J. del Jesus, and F. Herrera, "Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches," Knowledge-Based Systems, vol. 42, pp. 97–110, Apr. 2013, doi: 10.1016/j.knosys.2013.01.018.

[21] S. Rezvani and X. Wang, "A broad review on class imbalance learning techniques," Applied Soft Computing, vol. 143, p. 110415, Aug. 2023, doi: 10.1016/j.asoc.2023.110415.

[22] S. Shaffiee Haghshenas, G. Guido, S. Shaffiee Haghshenas, and V. Astarita, "Predicting the level of road crash severity: A comparative analysis of logit model and machine learning models," Transportation Engineering, vol. 20, p. 100323, Jun. 2025, doi: 10.1016/j.treng.2025.100323.

[23] K. Aziz, F. Chen, M. Ahmad, M. S. Khan, M. M. Sabri Sabri, and H. Almujibah, "An interpretable dynamic ensemble selection multiclass imbalance approach with ensemble imbalance learning for predicting road crash injury severity," Sci Rep, vol. 15, no. 1, p. 24666, Jul. 2025, doi: 10.1038/s41598-025-08935-x.

[24] S. Zhu, K. Wang, and C. Li, "Crash Injury Severity Prediction Using an Ordinal Classification Machine Learning Approach," Int J Environ Res Public Health, vol. 18, no. 21, p. 11564, Nov. 2021, doi: 10.3390/ijerph182111564.

[25] "Open Data | Observatoire national interministériel de la sécurité routière." Accessed: Nov. 11, 2024. [Online]. Available: https://www.onisr.securite-routiere.gouv.fr/outils-statistiques/open-data

[26] J. Freeman, "Outliers in Statistical Data (3rd edition)," Journal of the Operational Research Society, vol. 46, Aug. 1995, doi: 10.1057/jors.1995.142.

[27] G. E. A. P. A. Batista and M. C. Monard, "An analysis of four missing data treatment methods for supervised learning," Applied Artificial Intelligence, vol. 17, no. 5–6, pp. 519–533, May 2003, doi: 10.1080/713827181.

[28] Dalwinder Singh, Birmohan Singh, "Investigating the impact of data normalization on classification performance," Applied Soft Computing Journal 97 (2020) 105524, 2020, doi: 10.1016/j.asoc.2019.105524.

[29] X. Zhu and X. Wu, "Class Noise vs. Attribute Noise: A Quantitative Study," 2004, doi: 10.1007/s10462-004-0751-8.

[30] C. Seger, "An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing," 2018.

[31] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, vol. 42, no. 4, pp. 463–484, 2012, doi: 10.1109/TSMCC.2011.2161285.

[32] A. B. Parsa, H. Taghipour, S. Derrible, and A. (Kouros) Mohammadian, "Real-time accident detection: Coping with imbalanced data," Accident Analysis & Prevention, vol. 129, pp. 202–210, Aug. 2019, doi: 10.1016/j.aap.2019.05.014.

[33] A. B. Parsa, A. Movahedi, H. Taghipour, S. Derrible, and A. (Kouros) Mohammadian, "Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis," Accident Analysis & Prevention, vol. 136, p. 105405, Mar. 2020, doi: 10.1016/j.aap.2019.105405.

[34] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," jair, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[35] T. Harmon, G. Bahar, and F. Gross, "Crash Costs for Highway Safety Analysis," 2018.

[36] M. Kačan, M. Oršić, S. Šegvić, and M. Ševrović, "Multi-Task Learning for iRAP Attribute Classification and Road Safety Assessment," in 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), Sep. 2020, pp. 1–6. doi: 10.1109/ITSC45102.2020.9294305.