A Novel Multitasking Framework for Feature Selection in Road Accident Severity Analysis

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*—In machine learning studies, feature selection presents a crucial step especially when handling complex and imbalanced datasets, such as those used in road traffic injury analysis. This study proposes a novel multitasking feature selection methodology that integrates the Grey Wolf Optimizer, knowledge transfer, and the CatBoost ensemble algorithm to enhance the performance and interpretability of road accident severity prediction. The main objective of this study is to identify critical features impacting the prediction of severe injury cases in road accidents. The proposed framework integrates several steps to handle the complexities related to feature selection. The fitness function of the Grey Wolf Optimizer model is designed to prioritize the classification accuracy of the severe injury class. To mitigate early convergence of the model, a knowledge transfer mechanism that generates new wolf instances based on a historical record of wolves used previously is integrated within a multitasking process. To evaluate the prediction performance of the generated feature subsets, the CatBoost algorithm is employed in the evaluation step to assess the effectiveness of the proposed approach. By Integrating these three step methodology which combine metaheuristic feature selection technique with knowledge transfer through a multitasking process, the proposed framework enhances generalization, reduces prediction models complexity and handles imbalanced distributions. It proposed a feature selection model that overcomes key limitations of traditional methods. Applied to real-world road crash data, the methodology significantly improves the identification of factors impacting the severity of injuries. Experimental results demonstrate enhanced model performance, reduced complexity, and deeper insights into the factors contributing to traffic injuries. These findings highlight the potential of advanced machine learning techniques in improving road safety analysis and supporting data-driven decision-making.**

***Keywords*—Feature selection; road accident; injury severity; Grey Wolf Optimizer; multitasking; knowledge transfer**

I. Introduction

Machine learning (ML) advancements have created interesting opportunities to solve complex problems in recent research studies. The large amount of collected data serves as an important source of information to train ML models. However, many datasets are subject to a common problem where certain classes, often the most critical, are significantly underrepresented. The analysis of such imbalanced datasets remains a persistent challenge.

This study aims to leverage advancements in machine learning techniques to identify factors associated with severe injuries in road traffic accidents. Through a detailed analysis of crash-related data, this work seeks to enhance the understanding of injury mechanisms and support the development of more effective safety policies and real-time intervention strategies.

In road safety studies, datasets often exhibit imbalanced data problems, and a large number of features are collected. Predicting severe injury resulting from road crashes often involves dealing with imbalanced data distributions. The underrepresentation of minority classes in such datasets, combined with the use of a large number of features, can impact the training and generalization capabilities of traditional machine learning models. It also impacts the complexity of ML models and could lead to overfitting. Guyon explains that domains with a large numbers of input features are susceptible to the curse of dimensionality and multivariate methods may lead to overfitting [1]. The reduction of the number of features can lead to more robust models by mitigating overfitting and enhancing generalization [2].

By isolating the most relevant features and reducing the dimensionality of datasets, feature selection improves the interpretability of prediction models and ensures better focus on minority class prediction. However, traditional feature selection methods often face challenges to balance the needs of minority classes in high-dimensional data, complex interactions between features can hinder models from identifying the critical variables that influence the accuracy of classification. For instance, in road crashes case studies, datasets consists of different feature domains including driver characteristics, crash dynamics, vehicle attributes, and environmental conditions. These various features can interact in non-linear ways, which make it difficult for conventional techniques to effectively identify the most relevant features [3].

These challenges are particularly pressing in the context of road traffic crashes, which remain a global issue, claiming 1.35 million lives and causing around 50 million injuries annually [4]. Such incidents are a leading cause of death, especially among individuals aged 15 to 29, and understanding the factors that influence injury severity is essential for developing effective safety interventions. However, the complex and multifaceted nature of road crash data makes it difficult to accurately identify the critical variables, further underscoring the need for advanced methodologies like the one proposed in this study.

This study proposes a novel feature selection methodology that leverages advanced machine learning models. The proposed framework includes the metaheuristic Grey Wolf Optimizer (GWO), knowledge transfer techniques throw a multitasking process, and the CatBoost ensemble algorithm as a predictor

model. To ensure the effectiveness of the model application in the case study of injury severity prediction in road accidents, a specific implementation of the fitness function of the Grey Wolf Optimizer algorithm is elaborated. By combining these techniques with a specific focus on the accuracy of severe injury predictions, the proposed framework aims to improve the identification of key factors influencing severe injury outcomes in road crashes, and overcome the limitations of traditional approaches.

This paper is organized as follows: The second section presents a literature review of feature selection and machine learning methodologies and techniques, with a focus on road accident feature analysis case studies. The third section outlines the proposed methodology for feature selection using a multitasking framework. The fourth section details the experiments conducted using the proposed feature selection framework. The fifth section presents the experimental results and their interpretations. Finally, the last section summarizes the work presented in this paper and highlights areas for future exploration.

II. Related Work

A. Feature Selection Techniques in Machine Learning

Feature selection (FS) step presents an important role in improving the performance of classification studies, especially when using complex and imbalanced datasets where, the presence of underrepresented classes can impact the performance of learning model. FS methods can be divided into two categories of approaches: data-centric and algorithm-centric. Data-centric techniques adjust data distribution to mitigate class imbalance effects through synthetic oversampling, instance weighting, or hybrid resampling strategies that integrate data augmentation with feature selection [5]. To minimize overfitting risks of these techniques, recent research has introduced adaptive synthetic sampling based on feature relevance and the assignment of instance-specific weights [6]. On the other hand, algorithm-centric methods introduce additional techniques to traditional feature selection paradigms (filter, wrapper, and embedded techniques) by incorporating cost-sensitive learning [7], alternative ranking criteria, or hybrid metaheuristics [8], to improve feature selection robustness in skewed distributions. Despite significant advancements in feature selection techniques and results, many challenges persist related to the identification of complex feature interactions, the reduction of computational time for model training and prediction in real-time applications, and early convergence which impacts the models ability to generalize learning in the presence of imbalanced class distributions. Emerging research introduce new feature selection techniques based on deep learning to dynamically weigh features [9]. Reinforcement learning-based feature selection models are used to iteratively refine feature subsets based on classification performance in imbalanced settings [10]. Another emerging technique called evolutionary computations aim to explore optimal feature subsets through population-based search strategies, such as Genetic Algorithms, Particle Swarm Optimization [11], and Grey Wolf Optimizer [12].

The points outlined below presents a detailed overview of cited feature selection methods and their relevance in selecting key factors influencing the performance of minority classes' prediction.

1) Filter-based methods: These methods use statistical measures to evaluate features independently of the model. Common techniques are:

- Pearson and Spearman correlation which assess the statistical relationship between features and the target variable [1].
- Chi-square test which evaluates the statistical dependence between categorical features, comparing the observed data with the expected values [13].
- Two-Way ANOVA which is a statistical test used to identify the significant impact of features between two data groups (input and target). The test helps determine whether to accept or reject the null hypothesis [13].

2) Wrapper-based methods: Wrapper-based methods use machine learning algorithms to evaluate different feature subsets through three main steps: Generation of feature subsets, training and evaluation of the chosen machine learning model for each subset, and the identification of the best subset that represents the relevant features impacting the target variable [14]. Common wrapper-based techniques include forward selection, backward elimination, stepwise selection, recursive feature elimination (RFE) [15] and genetic algorithms [16]. These methods are model-specific, which allows them to optimize feature selection based on the model's performance. However, they tend to be computationally expensive and are susceptible to overfitting [1].

3) Embedded methods: Under the third category of embedded methods, feature selection is seamlessly integrated into the machine learning algorithm itself. These methods not only identify relevant features but also actively suppress the influence of less informative ones, offering a highly efficient solution to feature selection. Key techniques include:

- L1 and L2 Regularization (Lasso and Ridge): These methods incorporate regularization terms into the loss function, shrinking the coefficients of less significant features and promoting sparse, interpretable models.
- Decision Trees and Random Forests: These algorithms inherently measure feature importance by analyzing how frequently a feature contributes to optimal node splits. In Random Forests, the Gini index is commonly employed to quantify this importance, ensuring robust feature evaluation [17].
- Ensemble Methods: Advanced techniques such as gradient boosting and CatBoost go further by quantifying feature contributions to the overall model performance. This enables precise ranking of features based on their predictive power [1].

4) Hybrid methods: To identify the most relevant and coherent features, many researchers combine in practice multiple feature selection techniques. This combination of feature selection techniques is referred to as hybrid methods.

Bhyuian used Chi-square, Two-way ANOVA, and regression analysis to identify nine key factors impacting road crash severity from a set of fourteen features [13]. In a similar context or road accident severity prediction, Alkheder employed Chi- square automatic interaction detector trees, Bayesian networks, and linear SVM to identify risk factors and improve classification performance, achieving a testing accuracy of 66% for correct predictions [18]. Kashifi used SHAP analysis and the Gated Recurrent Convolution Network model to identify complex relationships in road accident data [19].

The combination of feature selection techniques in hybrid methods is valuable in complex domains such as road crash injury prediction. It can enhance the robustness of feature selection and improve model performance.

Despite advancements of feature methods such as filter, wrapper, and embedded techniques, these approaches present several limitations related to computational inefficiency, sensitivity to noise, and the risk of suboptimal feature subsets due to local minima entrapment [1], [20].

To overcome these limitations, metaheuristic algorithms have emerged as powerful alternatives using efficient global search strategies. Among these algorithms, the Grey Wolf Optimizer, which is inspired by the social hierarchy and cooperative hunting behavior of grey wolves, demonstrated its ability to balance exploration and exploitation [21]. The concept of this algorithm aims to identify optimal feature subsets, it consists of dynamically updating candidate solutions based on a fitness function to assess the relevance of generated candidates. However, due to its random initialization of candidate solutions, the standard GWO model may face stagnation in later iterations and sensitivity to initial parameter settings. This necessitates further enhancements to improve its robustness and adaptability, such as hybridization with machine learning techniques [22], [23].

B. Road Accident Features Analysis

Feature selection is closely linked to the choice of data architecture model during the data collection phase. Data architecture determines the number of collected features, consistency, and detail of features. Well-chosen features can lead to accurate and meaningful insights, while poor feature selection may result in misleading conclusions.

In road crash studies, key features for accident analysis have been refined over the years by road safety experts. The European Road Assessment Program (EuroRAP) established standardized protocols to display the safety level of a road, offering a common framework for communication [24]. Regular updates are recommended to adjust the evolving nature of road and environmental factors, vehicle characteristics, and driver profiles. These features also vary depending on the national context and the specific road safety strategies in place, adapting to the unique challenges and priorities of each region. However, during the data engineering phase, data analysts often create additional features to highlight new aspects that are not adequately represented by the original, collected features. These engineered features provide a deeper understanding of the data, revealing hidden patterns and relationships.

To cover sector-specific aspects of this study, an analysis of data dictionaries of road crash injury studies has been conducted to identify the key characteristics of the data architecture model for road crash injury datasets and elaborate a specific feature engineering map for road crashes datasets (see Fig. 1).

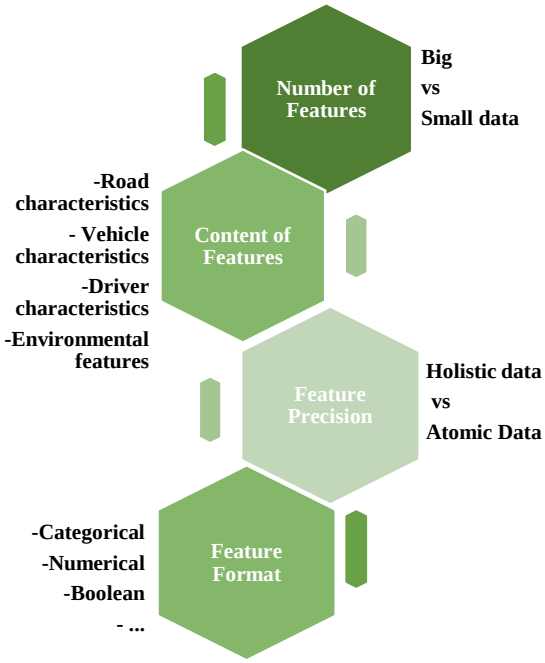

Fig. 1. Feature engineering map for road crashes datasets.

The presented road crash feature map highlights four main characteristics of the architecture of a road accident data inventory:

- Number of features: The number of features varies significantly across studies, ranging from as few as 8 to as many as 50 features. A key measure is introduced to differentiate between Big and Small Datasets, referring to the volume of data, whether large-scale or limited in scope.
- Content of features: The features are typically classified into four broad categories: road features, vehicle features, driver features, and environmental features.
- Feature value format: Features vary in their data format, including categorical, numerical, and Boolean types.
- Precision of feature description: The level of detail in feature descriptions, particularly for categorical values, differs significantly. Some studies provide specific and detailed descriptions (e.g., road surface type, weather conditions), while others are less detailed (e.g., broad classifications of road conditions). Holistic/Atomic Data: Indicating whether the dataset includes broad, comprehensive features (holistic) or more granular, individual characteristics (atomic).

The proposed feature engineering map serves two main purposes:

- Pre-use Tool: It can be applied before the creation of a road accident dataset. In this phase, the map helps to identify the key features that need to be collected or derived, guiding the data acquisition process. This ensures that the dataset is built with relevant and meaningful features from the outset, facilitating a more efficient and effective analysis later on.
- Post-use Tool: Once a road accident dataset has been created and data is collected, the map can be utilized to classify the dataset based on the identified features. It helps in evaluating the quality of the features and the dataset as a whole, allowing for the selection of appropriate machine learning techniques. Based on the results from the feature engineering map, relevant algorithms and models can be chosen to enhance prediction accuracy, handle data imbalances, or optimize for specific outcomes.

III. Methodology

To perform the prediction capability of machine learning models using feature selection techniques, this paper proposed a multitasking feature selection framework using knowledge transfer and metaheuristic optimization algorithm. The proposed framework implements a feature selection process using the Grey Wolf Optimization, a metaheuristic optimization algorithm inspired by the hunting behavior of wolves. It aims to identify the most relevant features for a classification task. The proposed framework performs multiple tasks (feature selection iterations) where it optimizes feature subsets independently. For each task, the fitness function of GWO based on precision of severe injuries class evaluates the selected features using cross-validation and a CatBoost classifier.

To enhance the computational performance of the model, a knowledge transfer method is incorporated in the model by storing the historical wolves (feature subsets) evaluated in previous tasks and the best historical performance achieved by a feature subset which is represented by a wolf instance. Before each initialization of the wolf instance parameters, the model checks the historical wolves list, and generates new instance of wolf. This technique enhances the computational performance by avoiding redundant computations of used wolves.

One of the major limitation of wolf optimizer algorithm is the risk of stagnation in later iterations. To avoid this problem, the multitasking process is introduced in the proposed model. Combined to the knowledge transfer method described before, each task explores the historical list of wolves, generates new instances of wolves achieving a better performance than the stored best feature subset. This technique countermeasure an eventual fast convergence of the model.

Given the issue of imbalanced data and the strong representation of the non-severe accident class, the fitness function in the proposed model is designed to prioritize the precision of the severe injuries class. This configuration allows the model to focus its performance on improving the prediction of the minority class, which will result in feature subsets that primarily impact the severe injuries class.

Finally, the best feature subset is used to train a final model and evaluate its performance on a test set, focusing on precision for classifying the severe injuries class.

Fig. 2 presents the framework of the proposed multitasking feature selection model.

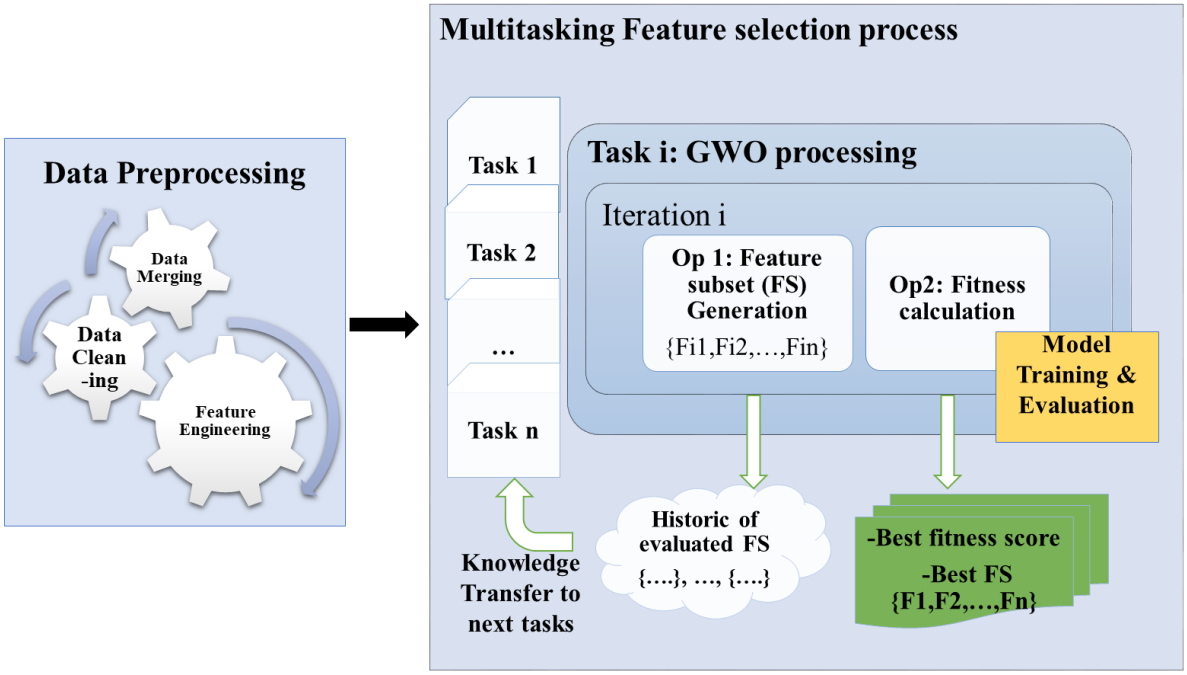

Fig. 2. Framework of the proposed multitasking feature selection model.

A. Grey Wolf Optimizer Processing

The proposed model employs Grey Wolf Optimization as the core of the multitasking feature selection framework. Inspired by the social hierarchy and hunting strategy of grey wolves, the search process of this metaheuristic algorithm is guided by four types of wolves called alpha, beta, delta, and omega. The alpha wolves represent the best solutions, while beta and delta are used to refine the search, and omega wolves explore new possibilities. The Grey Wolf Optimizer algorithm updates the positions of candidate solutions in a manner similar to the encircling, chasing, and attacking behaviors observed in wolf groups [21]. This allows the model to ensure an optimal selection of relevant features by balancing exploration (searching for new solutions) and exploitation (refining the best solutions).

GWO uses a fitness function to evaluate the score of selected wolves. The default fitness function is designed for general-purpose optimization tasks relying on minimizing an error function or maximizing an objective function without domain-specific adaptations. In this work, the proposed fitness function is tailored specifically to prioritize generated wolves involving the most performant classification accuracy of severe injury class. CatBoost classifier is used as the evaluation component in the classification step. In addition, cross-validation is employed to assess generalization ability and prevent overfitting. This specific adaptation of the fitness function of the Grey Wolf Optimizer algorithm ensures that the most informative features impacting severe injuries are retained, leading to a more accurate classification process and aligning the feature selection process with the specific objectives of this study.

B. Multitasking Feature Selection

The second layer of the proposed framework consists of a multitasking process. This layer aims to enhance the robustness of the Grey Wolf Optimizer process and address its limitations related to the risk of stagnation in later iterations. In standard GWO, the search process may converge prematurely depending on the initially generated candidates. This may result to suboptimal feature subsets if diversity among candidate solutions is not ensured. The proposed framework integrates a process of multiple optimization tasks that run iteratively. Each task initializes the initial parameters and can then generate a new space of feature subsets. Inspired by multitasking evolutionary computation [25], this mechanism strengthens the exploration process operated by the Grey Wolf Optimizer. It enhances the overall feature selection process by ensuring that different feature subsets spaces are explored to identify a final subset that is both optimal and robust for classification.

C. Knowledge Transfer Processing

The third layer of the proposed framework aims to improve the efficiency of the multitasking layer by incorporating a knowledge transfer mechanism between the iterative tasks. The GWO algorithm randomly initialize the positions of candidate solutions (wolves), which represents potential solutions in a search space. The major limitation of using only the first layer of feature selection with GWO and multitasking is that tasks could be initialized with similar initial candidate solutions. This process may lead to a repetition of tasks that adds unnecessary computational time without providing additional value. The role of the knowledge transfer layer introduced in this framework is to transfer exploration information from previous tasks to subsequent ones, providing additional factors that refine the initialization of the Grey Wolf Optimizer parameters.

The knowledge-transfer layer records two main data types: the historical list of wolves explored in previous tasks, and a list of the best-performing solutions encountered earlier. When a new task begins, it first examines the data provided by the knowledge-transfer layer and then generates new instances of wolves, with the aim of improving classification performance based on the previously stored best subsets. By incorporating this knowledge, the optimization process benefits from the accumulated experiences of earlier tasks, leveraging them to find better feature subsets more efficiently.

Algorithm 1 describes the proposed framework including GWO processing, fitness function, multitasking feature selection and knowledge transfer mechanism.

Algorithm 1: Multitasking FS processing

```
Input:
   X, y: Original dataset.
   num_tasks: Number of optimization tasks.
   num_wolves: Number of wolves (binary feature
   selection vectors).
   max_iter: Maximum number of iterations.
Output:
   SF: Best Feature Subset;
   Model: Trained CatBoostClassifier.
Initialize
Compute
Split X and y into training (X_trainval, y_trainval)
   and test sets (X_test, y_test);
Define fitness_function(selected_features):
   | Extract selected columns from X_trainval
   |       based on the binary vector;
   | Train CatBoostClassifier using cross-validation
   | on the selected features;
   | Compute and return the mean precision score
   | for class of severe injuries;
Initialize
best_global_precision = 0 and previous_wolves = ∅;
For each task t = 1 to num_tasks do
   | Initialize wolves as random binary vectors
   |     (num_wolves × num_features);
   | Set local best fitness = 0;
   | For iteration i = 1 to max_iter do
   |     | For each wolf w = 1 to num_wolves do
   |     |     |   If wolf w exists in
   |     |     |   previous_wolves:
   |     |     |     Regenerate wolf w randomly;
   |     |     |   End
   |     |     |   Compute Fitness(w) =
   |     |     |   fitness_function(wolf w);
   |     |     |   Add wolf w to previous_wolves;
   |     |     |   End
   |     | End
   |     | Identify alpha, beta, delta as the top 3
   |     | wolves based on fitness;
   |     | For each wolf w = 1 to num_wolves do
   |     |     |   Update wolf w's position using
   |     |     |   GWO update formulas with
   |     |     |   alpha, beta, delta;
   |     | End
```

```
    |       | Update local best fitness if a better
    |       | solution is found;
    |   End
    |   If task's best fitness > best_global_precision,
          Update best_global_precision and SF;
    | End
End
Select features from X_trainval and X_test based on
SF;
Train final CatBoostClassifier on the selected
features of X_trainval;
Evaluate the model on X_test and compute the final
precision score for class 1;
Return SF and trained Model.
```

IV. Experimentation

A. Data Description

To verify the effectiveness of the proposed multitasking feature selection framework, an open data source of real data obtained from the annual road traffic accident databases managed by the French National Interdepartmental Observatory of Road Safety (ONISR) is used for the experiments. Each bodily injury accident -defined as an event occurring on a public road, involving at least one vehicle, and resulting in at least one victim requiring medical care- is recorded by law enforcement agencies that respond to the scene. This information is captured in a document called the Bodily Accident Analysis Report. The collection of these reports forms the national database of traffic-related bodily injuries, commonly referred to as the "BAAC file," overseen by ONISR.

The annual datasets extracted from the BAAC file include all bodily traffic accidents in mainland France. The research utilizes data from 2005 to 2020. The recorded accident data contains detailed information, covering aspects like crash characteristics, location, involved vehicles, and road users.

To create the input dataset for this study, tables were merged using foreign keys specified in each data file, resulting in a unified dataset. After combining 64 data files -four files for each year- the final dataset consisted of 2,380,573 entries and 57 features, which formed the basis for the analysis in this research.

B. Data Visualization

To comprehend the variation of features impacting the severity of injuries in road crashes, a univariate and multivariate statistical exploration of the dataset is conducted. The analysis was developed in accordance with a classification according to four views:

1) Temporal and atmospheric conditions view: Features involved in this exploration are year of crash, day of week, month, is holiday, in addition to atmospheric conditions and brightness. The temporal exploration shows a significant variation of killed and injured hospitalized road users when distribution is by month and day of week. An increase of the number of accidents is detected on summer and Fridays, road traffic at these periods should be investigated to ensure the real impact. Atmospheric conditions statistics show a slight amount of crashes with light rains (see Fig. 3).

2) Road characteristics view: This part of the analysis explores a bivariate statistical view of features related to road characteristics where crashes are produced. Statistics shows that seven features have a visible variation of number of crashes and severity injury: road localization, road category, type of intersection, mode of circulation, road profile, road plan shape and surface state. Crashes are more frequent at urban zones, outside of intersections and bidirectional roads. Departmental, municipal, national roads and highways are respectively road categories involving the highest number of crashes, especially hospitalized and killed ones (see Fig. 4). Most crashes occurred on flat roads and straight sections with normal surface state.

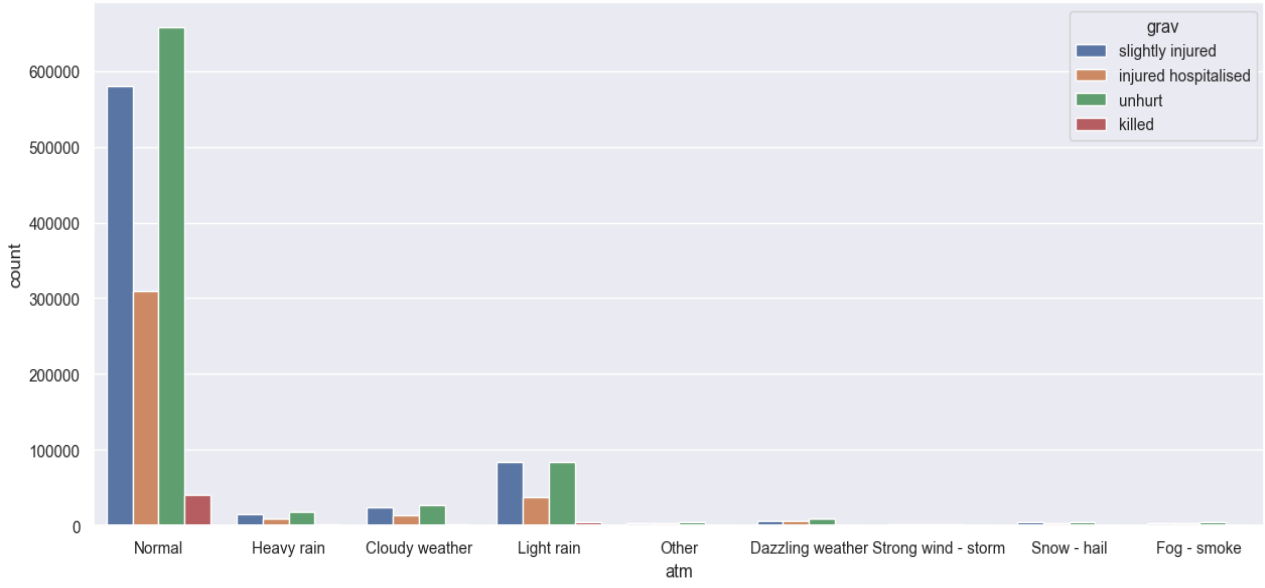

Fig. 3. Distribution by atmospheric conditions.

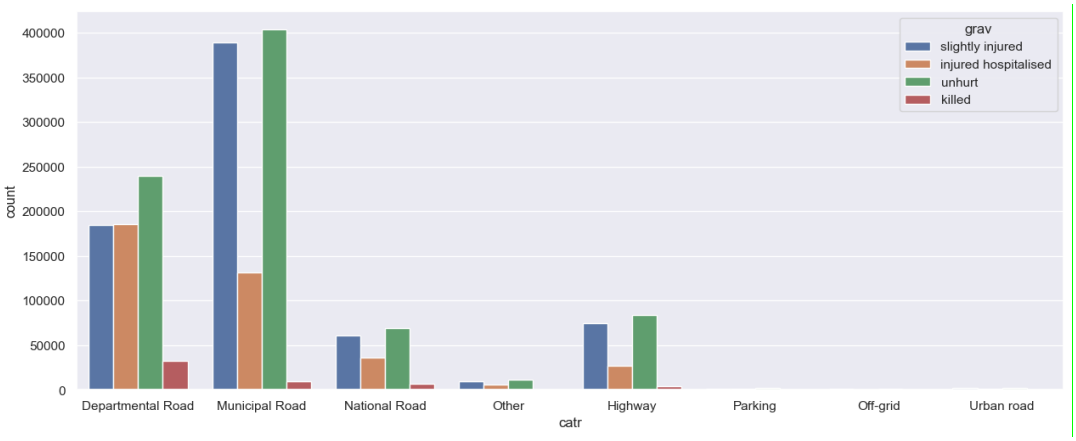

Fig. 4. Distribution by road category.

3) Vehicle characteristics view: Regarding vehicle features, features having the highest impact on number of crashes and severe (killed and hospitalized) injuries are category of vehicle, initial shock point, place of the road user into the vehicle, type of collision and main maneuver before the crash. Light vehicles alone have a domination of number of crashes and severe injuries. A significant impact is noted for frontal and side collisions of two vehicles and non-change direction maneuver before the crash. The place of the driver is the riskiest place in vehicles with a surrounding number of 300000 of hospitalized injuries and killed between 2005 and 2020.

This view aims to analyze features related to road user profile. An analysis of the distribution of crashes according to road user profile features (category, gender and age slice) and according to behavioral features (reason of travel at time of accident) is elaborated with a focus on localization on the road of pedestrian victims.

The statistical analysis shows that category of user displays a significant impact on injury severity: pedestrians and passengers face approximately same risk of being killed or hospitalized (see Fig. 5), but drivers are exposed to the highest risk. This statement matches with previous results related to the analysis of crashes' distribution by user place at vehicle.

- User profile view:

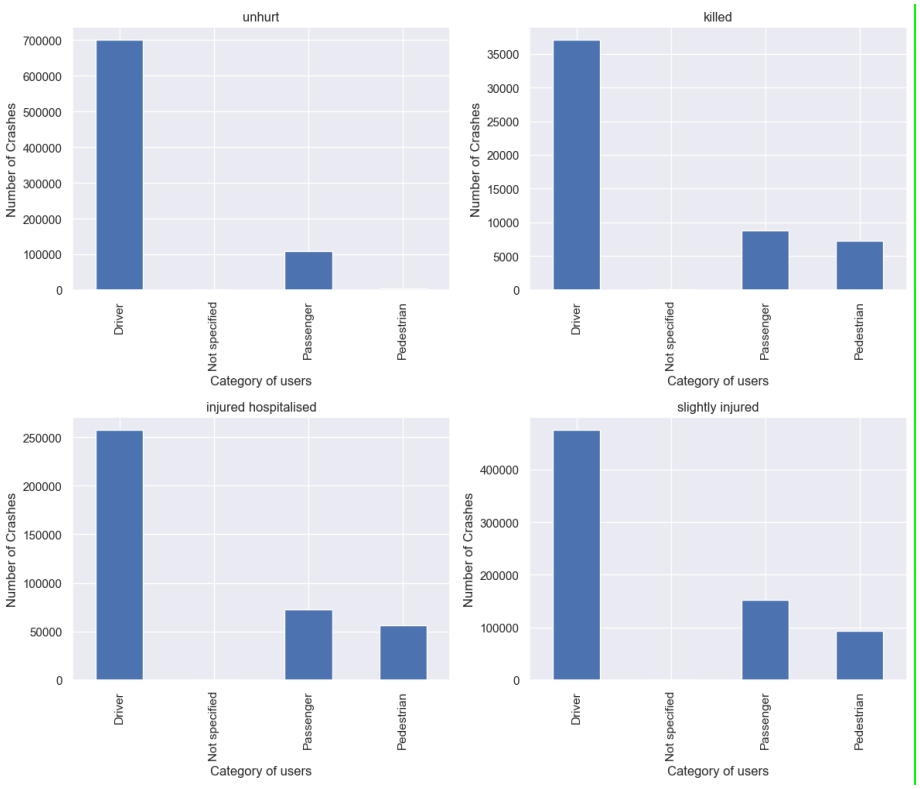

Fig. 5. Distribution by road user category.

Distribution by age slice and gender presents a significant variation. Men and users having 15 to 34 years old and 45 to 65 presents the highest category of killed and hospitalized victims as shown in Fig. 6.

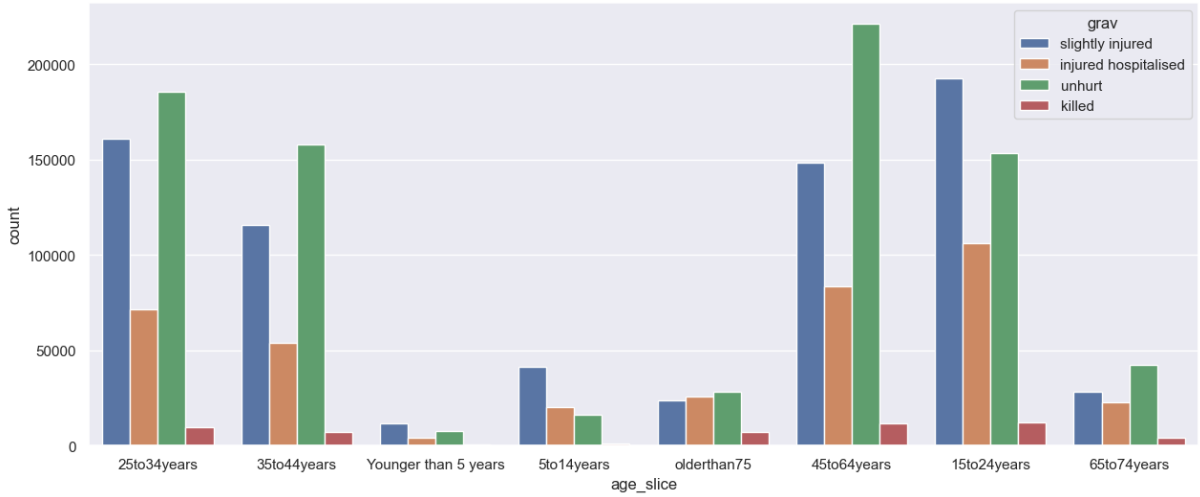

Fig. 6. Distribution by injury level and age slice.

The behavioral analysis presents a static peak of the value “leisure walk” for the fourth injury severity levels. Localization of pedestrians have also a direct impact on injury severity which is higher at areas far than 50m from the pedestrian crossing.

The bivariate statistical analysis highlights several features impacting the injury severity variation. Three groups of features from different views present converged results with a convergent impact. The 1st group of features (place, category of user) presents a significant impact on the risk of injury for drivers. The 2nd group (reason of walk, month, day of week, surface state) presents a neutral impact in leisure trips. The 3rd group (maneuver before the accident, mode of circulation, initial choc point, road plan shape, road profile, type of intersection) presents a higher impact on the risk of severe injuries for frontal crashes. Category of vehicle and road category present a significant singular impact on injury severity.

C. Data Preprocessing

To prepare the studied dataset, the first steps of data cleaning and feature engineering are elaborated. 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".

1) Multitasking feature selection process: The experimental process in this study aims to optimize feature selection for road traffic accident classification using the proposed multitasking feature selection framework (MFS) based on the Grey Wolf Optimizer. The objective is to identify the most relevant features from the dataset that contribute to accurately predicting severe accidents.

The dataset is initially split into training-validation (75%) and test (25%) subsets using stratified sampling to maintain class balance. The feature selection process is then executed over multiple tasks, where each task consists of several candidate solutions (wolves) exploring the feature space. Each wolf represents a binary vector indicating selected features.

The fitness function used in the MFS framework evaluates the precision of a CatBoost Classifier using 5-fold cross-validation. It measures the model's ability to correctly classify severe accidents (class 1) based on the selected features. The equation for the proposed fitness function is expressed as:

$$F(S) = \frac{1}{k} \sum_{i=1}^{k} Pi(S) \quad (1)$$

where,

- F(S) is the fitness value for a given subset of selected features S.
- k is the number of cross-validation folds (here, k=5).
- Pi(S) is the precision score for class 1 in the i-th cross-validation fold, computed as:

$$Pi(S) = \frac{TPi}{TPi+FPi} \quad (2)$$

where,

- TPi (True Positives) is the number of correctly predicted severe accidents in fold i.
- FPi (False Positives) is the number of non-severe accidents incorrectly classified as severe in fold i.

The objective is to maximize F(S), ensuring that the selected feature subset leads to the highest precision in classifying severe accidents.

The precision score for class 1 (severe accidents) is used as the performance metric. Throughout multiple iterations, the best-performing wolves (α, β, and δ) guide the position updates of the other wolves using adaptive coefficients. This iterative search process refines the feature selection, aiming to maximize classification precision.

Ultimately, this strategy helps the algorithm converge more effectively toward the best feature set, improving the classification model's precision in identifying severe accidents.

After completing all tasks, the best feature subset is selected based on the highest recorded precision. The final CatBoost model is then trained on the training-validation set using the selected features and evaluated on the independent test set. The test performance is measured using the precision score for class 1 to assess the model's ability to correctly identify severe accidents.

V. Results and Discussion

This section presents the experimental results for the multitasking feature selection framework using the road accident dataset. Three aspects are evaluated in this study: the model's performance; the effect of computational time on training and prediction; and the impact on model complexity, including an analysis of factors influencing the prediction of injury severity in road accidents.

A. Performance of the MFS Model

1) Convergence to the best feature subset: To evaluate the performance of the techniques used in the MFS model for identifying impacting factors and its capability to overcome the limitations of GWO through multitasking and knowledge transfer, an analysis of the generated wolves in each task and iteration during the data processing step is conducted. Table I presents an excerpt from the log file of the generated feature subsets during processing.

TABLE I. Excerpt from the Log File of Generated Feature Subsets

Iteration	Wolf Number	Feature Subset	Fitness value
5	1	[0 2 4 5 8 9 10 11 12 13 14 15 16 17 18 22 24 25 28 30 31]	0,6430
5	2	[1 2 4 5 8 9 12 13 14 16 17 18 24 25 28 29 30 31]	0,6356
5	3	[0 2 4 5 8 11 13 14 15 16 17 18 25 28 29]	0,6201
5	4	[0 2 4 5 8 9 11 12 13 14 16 17 18 24 25 28 29 30 31]	0,6388
5	5	[2 5 9 10 12 13 14 16 24 25 30 31]	0,0000
5	6	[0 2 3 4 5 8 9 12 13 14 16 17 18 21 22 24 25 30]	0,0000
5	7	[0 2 4 5 8 9 10 12 13 14 16 17 21 22 24 25 29 30 31]	0,0000

The analysis of the log file presented in Table I reveals that GWO generates many new wolves (feature subsets) that had already been used in previous tasks. The additional layer of knowledge transfer introduced in the MFS framework effectively addresses this limitation. By leveraging the historical set of previously generated wolves, the model minimizes the reuse of feature sets. The fitness function of previously used wolves is automatically set to 0, as shown in Table I, encouraging the generation of new feature sets. This, in turn, enhances the chances of identifying the best feature subsets.

The proposed MFS model represents a significant improvement over classic GWO in generating impactful feature subsets while preventing rapid convergence to suboptimal solutions.

2) Prediction of severe injuries: The impact of the MFS framework on the prediction accuracy of injury severity levels is evaluated using classification metrics derived from the injury severity level predictions. Table II presents the prediction metrics obtained using the CatBoost classifier with the feature subset generated by the MFS framework and the metrics obtained using the CatBoost classifier on the entire dataset before applying feature selection.

TABLE II. Results of Injury Severity Level Prediction

Model	Prediction using MFS framework output			Prediction using Catboost without Feature selection		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Class 0	**0.87**	0.96	0.91	0.85	0.94	0.89
Class 1	0.65	0.35	0.45	**0.68**	0.41	0.51
Accuracy			**0.84**			0.83
Macro avg	0.76	0.65	0.68	0.76	0.68	0.70
Weighted avg	0.83	0.84	0.82	0.81	0.83	0.81

The results shows that the overall accuracy of the model is slightly higher when using the MFS framework output. While the precision for class 1 (severe injury) is slightly lower, the precision for the non-severe injury class is improved. The general analysis shows that the MFS framework maintains the prediction performance.

B. Computational Time of MFS Model

To evaluate the computational time gain of the proposed model, a comparison is made between the fitting and prediction times of the CatBoost model using features generated by the MFS framework and the CatBoost model using the initial features before the implementation of the MFS, as presented in Table III.

TABLE III. COMPUTATIONAL TIME COMPARISON

Computational time (seconds)	MFS framework	Initial Catboost	Percentage of gain
Fitting	**190.607**	209.35	**-9%**
Prediction	**0.212**	0.389	**-44,7%**

The computational time comparison shows that the MFS framework enhances efficiency over the initial CatBoost model, particularly in prediction time. The fitting time decreases from 209.35 seconds to 190.607 seconds, achieving a 9% reduction, indicating a slight improvement in training efficiency. More notably, the prediction time drops from 0.389 seconds to 0.212 seconds, resulting in a 44.7% reduction. This demonstrates that the MFS framework significantly improves computational efficiency by reducing both training and prediction times. The most remarkable gain is in prediction time, where the MFS framework nearly halves the required time, making it much more suitable for real-time or large-scale predictions.

C. Complexity of the Model

The results presented before shows that the proposed MFS framework help to identify a reduced feature subset that maintain the prediction accuracy of the injury severity level.

This feature selection process reduced the model's complexity from 35 features to 10, representing a 75% reduction in complexity.

D. Analysis of Impacting Factors

The highest precision is achieved using the selected subset of features, which includes: ['day', 'int', 'catr', 'circ', 'plan', 'surf', 'infra', 'situ', 'catv', 'obs', 'manv', 'catu', 'trajet', 'age_slice'].

This series of features generated using the MFS framework identify the factors that significantly influence the prediction of injury severity in road crashes. These features are systematically categorized into four principal groups, each representing a distinct dimension of the accident context and contributing to the overall predictive model.

1) Temporal and atmospheric conditions: By selecting only two key features -day of occurrence and surface conditions- to represent atmospheric conditions instead of incorporating a broader range of related variables, the overall model complexity is significantly reduced. This streamlined approach captures the essential environmental influences while mitigating redundancy and overfitting risks.

2) Road features: Features in this group relate to the geometric and infrastructural characteristics of the roadway. To reduce model complexity while retaining critical information, related features resulting from the MFS framework are: intersection typology, road classification, circulation modes, horizontal alignment (road layout), the state of road infrastructure, and the specific situational context of the accident. This curated selection effectively characterizes the essential physical environment in which the crash occurs, thereby playing a critical role in determining accident severity while mitigating redundancy.

3) Vehicle Features: The vehicle category is included in the selected feature subset. This inclusion may be attributed to the high incidence of road crashes involving light vehicles, which are classified as category 7, as illustrated in Fig. 7

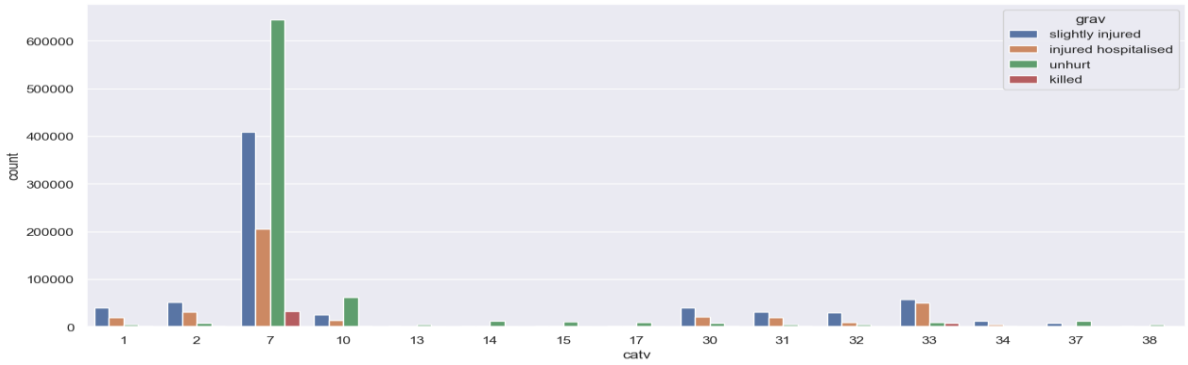

Fig. 7. Distribution of road crashes by category of vehicles and level of injury severity.

4) User profile: The user profile group is incorporated into the selected feature subset by including key demographic and behavioral indicator, specifically, age stratification and the primary reason for travel at the time of the accident. This streamlined selection effectively captures essential aspects of the human element in road safety, reflecting behavioral patterns and decision-making processes that are empirically linked to variations in injury severity, all while reducing overall model complexity.

The comprehensive integration of these selectively chosen feature categories underscores the multifactorial and complex interplay among environmental, infrastructural, vehicular, and human factors that collectively determine injury severity in road crashes. The focus on the most representative features from each group resulting from the multitasking feature selection framework, reduces model complexity while preserving critical information. Consequently, it enhances the understanding of factors impacting road safety and facilitates the development of more robust and interpretable predictive models for injury severity assessment.

VI. CONCLUSION

This paper presents a multitasking feature selection framework for predicting severe injuries caused by road crashes. This novel approach to data analysis and feature selection

combines three layers of learning to identify features impacting severe injuries in road crashes. It combines the strengths of the Grey Wolf Optimizer, the advantages of multitasking, the knowledge-transfer mechanism and the Catboost classifier to effectively reduce the complexity of large datasets and improve the classification performance of predictive models.

The metaheuristic algorithm GWO serves as a robust optimization tool to identify relevant features impacting the classification of injury severity. The multitasking process ensures a wide exploration of potential feature subsets. On the other hand, the knowledge transfer mechanism ensures the efficiency of the multitasking process leading to improved generalization and faster convergence.

Experimental results validate the efficacy of the proposed framework, it demonstrates its superiority over conventional methods in terms of feature selection efficiency, complexity reduction, predictive performance, and significant reduction in computational time. These findings suggest that the framework can be held for improving the performance of machine learning models in road safety data analysis and even across a variety of other domains. In future work, the MFS framework could be integrated into a safety countermeasure system, offering the possibility to adjust factors influencing severe injuries in real time. Due to its adaptability, the framework could be further refined and applied to more complex datasets across various domains, especially in real-world applications dealing with large-scale data. However, the overall performance of the proposed framework depends on the initial performance of the model used to compute the fitness function within the GWO algorithm. This dependency may limit the framework's adaptability and generalizability.

ACKNOWLEDGMENT

The experimental work presented in this paper was developed using the HPC-MARWAN computing cluster provided by the National Center for Scientific and Technical Research (CNRST) in Rabat, Morocco.

REFERENCES

[1] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection".

[2] H. Liu and H. Motoda, Eds., Computational Methods of Feature Selection. New York: Chapman and Hall/CRC, 2007.

[3] B. Wali, A. J. Khattak, and T. Karnowski, "The relationship between driving volatility in time to collision and crash-injury severity in a naturalistic driving environment," Analytic Methods in Accident Research, vol. 28, p. 100136, Dec. 2020.

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

[5] K. Kalaiselvi and S. B. V. J. Sara, "A Hybrid Filter Wrapper Embedded-Based Feature Selection for Selecting Important Attributes and Prediction of Chronic Kidney Disease," in International Conference on Computing, Communication, Electrical and Biomedical Systems, A. Ramu, C. Chee Onn, and M. G. Sumithra, Eds., Cham: Springer International Publishing, 2022, pp. 137–153.

[6] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Jun. 2008, pp. 1322–1328.

[7] Y. Li, C. Ma, Y. Tao, Z. Hu, Z. Su, and M. Liu, "A Robust Cost-Sensitive Feature Selection Via Self-Paced Learning Regularization," Neural Process Lett, vol. 54, no. 4, pp. 2571–2588, Aug. 2022.

[8] O. M. Alyasiri, Y.-N. Cheah, A. K. Abasi, and O. M. Al-Janabi, "Wrapper and Hybrid Feature Selection Methods Using Metaheuristic Algorithms for English Text Classification: A Systematic Review," IEEE Access, vol. 10, pp. 39833–39852, 2022.

[9] H. Tian, S.-C. Chen, and M.-L. Shyu, "Evolutionary Programming Based Deep Learning Feature Selection and Network Construction for Visual Data Classification," Inf Syst Front, vol. 22, no. 5, pp. 1053–1066, Oct. 2020.

[10] 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.

[11] H. B. Nguyen, B. Xue, P. Andreae, and M. Zhang, "Particle Swarm Optimisation with genetic operators for feature selection," in 2017 IEEE Congress on Evolutionary Computation (CEC), Jun. 2017, pp. 286–293.

[12] F. G. Mohammadi, M. H. Amini, and H. R. Arabnia, "Evolutionary Computation, Optimization, and Learning Algorithms for Data Science," in Optimization, Learning, and Control for Interdependent Complex Networks, M. H. Amini, Ed., Cham: Springer International Publishing, 2020, pp. 37–65.

[13] 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.

[14] D. Patel, A. Saxena, and J. Wang, "A Machine Learning-Based Wrapper Method for Feature Selection," IJDWM, vol. 20, no. 1, pp. 1–33, Jan. 2024.

[15] 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.

[16] E. M. Maseno and Z. Wang, "Hybrid wrapper feature selection method based on genetic algorithm and extreme learning machine for intrusion detection," Journal of Big Data, vol. 11, no. 1, p. 24, Feb. 2024.

[17] G. Pillajo-Quijia, B. Arenas-Ramírez, C. González-Fernández, and F. Aparicio-Izquierdo, "Influential Factors on Injury Severity for Drivers of Light Trucks and Vans with Machine Learning Methods," Sustainability, vol. 12, no. 4, Art. no. 4, Jan. 2020.

[18] 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.

[19] M. T. Kashifi, "Robust spatiotemporal crash risk prediction with gated recurrent convolution network and interpretable insights from SHapley additive explanations," Engineering Applications of Artificial Intelligence, vol. 127, p. 107379, Jan. 2024.

[20] Z. Sadeghian, E. Akbari, and H. Nematzadeh, "A hybrid feature selection method based on information theory and binary butterfly optimization algorithm," Engineering Applications of Artificial Intelligence, vol. 97, p. 104079, Jan. 2021.

[21] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey Wolf Optimizer," Advances in Engineering Software, vol. 69, pp. 46–61, Mar. 2014.

[22] M. H. Nadimi-Shahraki, S. Taghian, and S. Mirjalili, "An improved grey wolf optimizer for solving engineering problems," Expert Systems with Applications, vol. 166, p. 113917, Mar. 2021.

[23] I. Dagal, A.-W. Ibrahim, A. Harrison, W. F. Mbasso, A. O. Hourani, and I. Zaitsev, "Hierarchical multi step Gray Wolf optimization algorithm for energy systems optimization," Sci Rep, vol. 15, no. 1, p. 8973, Mar. 2025.

[24] B. Green, "iRAP Star Rating and Investment Plan Manual Version 1.0," iRAP.

[25] A. Gupta, Y.-S. Ong, L. Feng, and K. C. Tan, "Multiobjective Multifactorial Optimization in Evolutionary Multitasking," IEEE Transactions on Cybernetics, vol. 47, no. 7, pp. 1652–1665, Jul. 2017.