A novel one-vs-rest consensus learning method for crash severity prediction

Hussain, Syed Fawad; Ashraf, Muhammad Mansoor

DOI:
10.1016/j.eswa.2023.120443

Document Version
Publisher's PDF, also known as Version of record

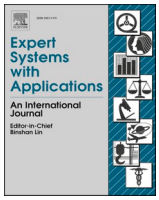
Citation for published version (Harvard):
Hussain, SF & Ashraf, MM 2023, 'A novel one-vs-rest consensus learning method for crash severity prediction', *Expert Systems with Applications*, vol. 228, 120443. https://doi.org/10.1016/j.eswa.2023.120443

Expert Systems With Applications 228 (2023) 120443

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



A novel one-vs-rest consensus learning method for crash severity prediction

Syed Fawad Hussain [a,c,*,1], Muhammad Mansoor Ashraf [b,c]

[a] School of Computer Science, University of Birmingham, UK
[b] L2TI, Institut Galilée, Université Sorbonne Paris Nord, Villetaneuse 93430, France
[c] Machine Learning and Data Science (MDS) Lab, Faculty of CS & Engg., G.I.K Institute, Topi 23640, Pakistan

ARTICLE INFO

Keywords:
Crash severity prediction
Ensemble learning
Mutual information
One-vs-Rest

ABSTRACT

Research in crash severity prediction is necessary to allow safety planners to take precautionary measures and enable first aiders to remain prepared for assisting the injured. Existing literature in the field of crash severity prediction is mostly focused on generating the attributes for predicting the severity. However, in reality, not all features are discriminating, and certain classes are challenging to detect even employing the entire feature set. Although to tackle these problems several techniques are developed in the Machine Learning (ML) literature. But their application to crash severity prediction and an optimal strategy for the best combination of features or classifiers for achieving high accuracy is a less studied area. To address these problems, this work first provides a comparison of widely used classifiers for predicting crash severity; and secondly, by combining class-wise majority voting with One-vs-Rest (OvR) approach, a novel classification framework named, OvR consensus learning (OvRCL) is proposed. The proposed method avail a feature selection technique, Mutual information (MI), to acquire the most relevant feature set regarding the output class (i.e. severity). Moreover, to differentiate each class in the multi-class data, OvRCL iteratively runs ML algorithms as binary classifiers in an ensemble framework to significantly ameliorate classification performance. In our experiments, we use four classifiers, namely, the *k*-Nearest Neighbors (*k*-NN), Support Vector Machines (SVM), Random Forest (RF), and Bagging classifier, to get the consensus. Analysis was done using a real crash dataset obtained from an open data source of Leeds city council. A four-year crash data (2015–2018) is used for training and the OvRCL is tested on the 2019 data. Moreover, to validate the performance of OvRCL, this study also utilizes two more datasets with high-class imbalance. In contrast to conventional ML algorithms, our experiments depict that the OvRCL is a potent method for forecasting crash severity levels on the data under test.

1. Introduction

Roadways play an important role in the development of a nation's economy and bring many cultural and social benefits to the community. The road infrastructure of a country is vital to its growth as it provides access to employment, social, health, and education services making it among the most significant of all public assets. However, the burden on roads also increases with the increase in the population and economic activities. This can lead to a surge in the number of road accidents which causes huge losses both in terms of life and property in the case of severe crashes (we use the terms *accident* and *crash* interchangeably and prefer the former when denoting the generic term and the later when talking about its severity). For instance, according to (Pakistan Statistical Year Book, 2020), around 10,000 road crashes were reported in 2019–20 involving roughly 13,000 vehicles, and approximately 5,500 lives were lost. Many more suffer injuries or are paralyzed for life. If we could predict how severe the collision would be and, consequently, the likely injuries, lives might be saved and better healthcare could be given.

With the increase in the number of road crashes in the past few decades, researchers have increasingly focused on developing various models for predicting the severity of injuries in road traffic crashes. Traditionally, these models depended on mathematical and statistical analysis, but more recently, Artificial Intelligence (AI), in particular Machine Learning (ML) algorithms, are being applied to anticipate various crash-causing factors and to predict the severity of the crashes. Conventional ML and Deep Learning (DL) algorithms, such as Support Vector Machines (SVM), Deep Neural Networks (DNN), etc. have resulted in relatively low performance when it comes to predicting the

* Corresponding author.
E-mail address: s.f.hussain@bham.ac.uk (S.F. Hussain).
[1] https://orcid.org/0000-0001-9122-6029

https://doi.org/10.1016/j.eswa.2023.120443
Received 16 July 2022; Received in revised form 13 March 2023; Accepted 7 May 2023
Available online 11 May 2023

severity of vehicular crashes (Iranitalab & Khattak, 2017; Rahim & Hassan, 2021; Z. Yang et al., 2022). This could be because some of the classes are not easily distinguishable, such as between severe and fatal injuries, when considering a multi-class problem as is usually the case. In literature, feature selection, One-vs-One (OvO), One-vs-Rest (OvR), and ensemble techniques, are some of the methods that are used in classifying difficult-to-separate classes where the accuracy scores are low (Dong et al., 2020; Hussain et al., 2020; Hussain, Khan, et al., 2022; Q. Li et al., 2020; Ramírez et al., 2018).

A range of ML models was employed in the literature on accident severity prediction to examine the severity of single and two-car crashes as well as collisions involving pedestrians. In (Abdel-Aty & Abdelwahab, 2004), the authors adopted ANN and Adaptive Resonance Theory to study driver injury using crash data from Central Florida. Another study (Delen et al., 2006) also leveraged ANN and determined that the relationship between the severity level and the factors determining the severity level is nonlinear. Later, more sophisticated models such as regression and multinomial Logit (MNL) have been employed (Çelik & Oktay, 2014; Z. Li et al., 2012). Moreover, some studies presented the use of feature selection and ensemble learning for crash severity prediction (Chandra et al., 2019). Random Forest (RF) and Mixed Logit Models were utilized by (Haleem et al., 2015) to analyze different factors affecting crash severity whereas (Kabeer, 2016) studied Decision Trees (DTs) and Ensemble techniques to achieve higher accuracy scores. Recently, (Iranitalab & Khattak, 2017) employ clustering in conjunction with classification to further improve accuracy, while (Katanalp & Eren, 2020) examine features that are well suited for severity prediction. However, these studies have relatively low performance in predicting the severity on an imbalanced dataset. Furthermore, to the best of our knowledge, there exists no literature where the combination of multiple techniques, particularly those suited for identifying hard-to-detect multiple classes has been studied for crash severity prediction.

Motivated by the above discussion, this paper presents a new approach for predicting the crash severity level by combining multiple ideas from feature selection, multi-class classification, and ensemble learning. These concepts are merged uniquely in the proposed model to elevate the classification performance on an unbalanced dataset. The proposed algorithm is used to predict the severity level of a crash that may happen under a particular set of conditions, for example, the road type, road condition, weather and light conditions, etc. The main objective of this research is to allow road safety planners to develop accurate countermeasures to prevent crashes in such conditions as well as to assist emergency responders to judge the potential medical assistance that might be needed for the injured. Our results show that the proposed approach is better able to classify the data with the added advantage of reducing the data size (hence, the time to predict severity).

1.1. Contributions

The specific contributions of this work are as follows:

- We utilize a feature selection algorithm, Mutual Information (MI), to extract the most discriminating features,
- To improve accuracy and reduce the bias from any single algorithm, an ensemble technique is employed,
- The novel One-vs-Rest Consensus Learning (OvRCL) technique is proposed for identifying hard-to-classify classes.

We begin by employing features taken from different existing literature related to crash severity prediction. Instead of using the complete feature set, we use MI to select only the most discriminating features. An ensemble of four classifiers, namely Support Vector Machines (SVM), *k*-Nearest Neighbors (*k*-NN), Random Forests (RF), and Bagging classifier (BG) is used to improve the accuracy and reduce the bias. These models were selected as they have been widely used in the literature on crash severity prediction and show good results on a variety of datasets (Haleem et al., 2015; Iranitalab & Khattak, 2017; Kabeer, 2016; Kannojiya et al., 2020; Yan et al., 2021; Z. Yang et al., 2022). However, OvRCL is a framework in which the chosen classifiers are not determinant of the results. OvRCL allows for the unique and iterative identification of classes while each classifier employs an OvR method for every individual class. Fig. 1 illustrates the abstract view of the main contribution of this study.

The rest of the paper is organized as follows: Section 2 presents the literature review and gives a summary of the related work. In Section 3, we briefly describe the various classifiers used in our ensemble approach and detail the proposed OvR consensus learning model. Section 4 provides the dataset description and evaluation techniques used in this study. The results are presented and analyzed in Section 5, and we conclude the discussion and provide future directions in Section 6.

2. Literature review

This literature review focuses on learning the crash severity model and predicting the injuries/fatality of casualties from road accidents. Significant literature on crash severity modeling uses statistical techniques where the severity is taken as a dependent model while other features form the independent variable. These features usually consist of the type of vehicle, the gender or age of the driver, the condition of the roadway, and/or the weather conditions. Some studies implemented statistical and probabilistic models for crash severity analysis, (Fan et al., 2016) for a binary class problem, (Ye & Lord, 2014) for multi-class prediction, and (Yasmin & Eluru, 2013) to address the correlation of the unobserved fraction of utility of crash severity levels. The focus of these models is more on the modeling and crash severity prediction is simply used as a validation technique (Iranitalab & Khattak, 2017). However, many recent works choose to classify the accident severity using ML approaches. The rest of this section also focuses on these machine learning models.

Amongst the earlier studies, the work of (Abdel-Aty & Abdelwahab, 2004) compares two different Artificial Neural Networks (ANN) paradigms – the Multilayer Perceptron (MLP), and the fuzzy Adaptive Resonance Theory (ART), to predict the severity level of a crash. The results were compared with the calibrated ordered probit model, and their work showed that ANN performs better in predicting the severity level of a crash. Similarly, (Delen et al., 2006) also made use of ANN to model the non-linear relationship between severity level and factors causing the crash. Their work showed the existence of a non-linear relationship between the intensity of the features and the crash severity. They also suggested the characteristics that are influential in making any difference in the severity level. Moreover, (Xie et al., 2009) compared the performance of traditional Ordered Probit and Bayesian inference Ordered Probit (BOP) models in driver's injury severity, and it was shown that the BOP model works better when the size of the dataset is small. However, when the dataset size increases, BOP's performance decreases. To model and categorize the data, k-NN and C-means clustering were respectively adopted (Lv et al., 2009). This work shows that using a clustering technique in combination with classifiers can help increase performance – an idea also used by later authors. The study analyzes traffic data and suggests crash-causing conditions. (Kim et al., 2010) presented the probabilistic analysis of the severity of pedestrians in pedestrian-vehicle crashes by using the Mixed Logit model. It was found that numerous factors like darkness, freeway, over-speed vehicles, or if the vehicles were trucks, double the probability of fatal injury to pedestrians. Following in the footsteps of (Lv et al., 2009), a comparison of SVM with order probit is shown by (Z. Li et al., 2012). Their findings also imply that SVM, a ML model, outperforms other methods in determining the severity level.

The comparison of regression models and the Bayesian network by using the traffic police data from Jilin province in China is given by (Zong et al., 2013). The authors demonstrated that the application of Bayesian networks for severity level prediction outperformed regression

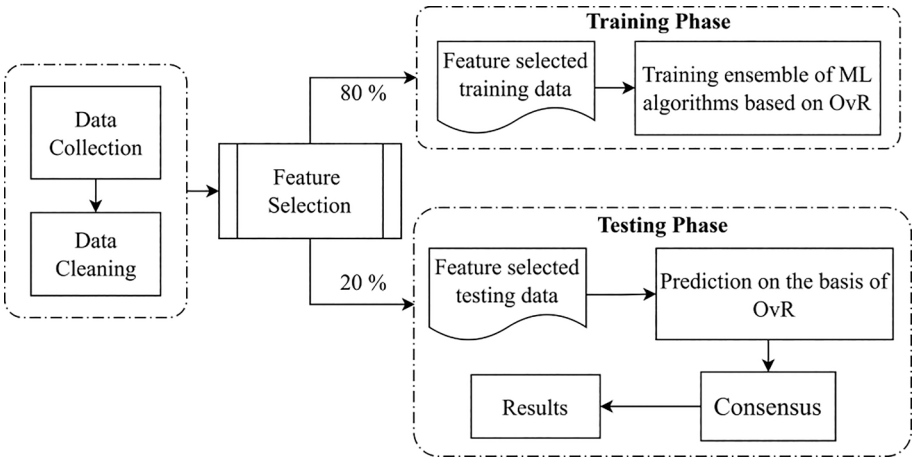


Fig. 1. Outline of the main contributions.

models. (Çelik & Oktay, 2014) presents another case study that examines the risk factors controlling the road traffic crash severity in the Erzurum and Kars provinces of Turkey. This study makes use of the Multi Nomial Logit (MNL) to classify three severity categories: fatal, injury, and no injury. In addition to this, factors affecting the severity of pedestrian accidents were discussed by (Haleem et al., 2015). Authors in this study utilized RF to rank the important parameters and a mixed logit model for the severity prediction. In the recent past, advanced techniques developed in ML have also been employed. For instance, (Kabeer, 2016) proposes the use of an ensemble technique using Naïve Bayes and DT classifiers. It is shown that ensemble techniques can increase the prediction accuracy to 78.03% while the accuracy of Naïve Bayes and DTs was 58.76% and 51.22%, respectively. A combination of clustering and classification is used by (Iranitalab & Khattak, 2017) to improve classification accuracy. (Chandra et al., 2019) compares RF and logistic regression and shows that, RF has high performance in predicting crash severity. (Kannojiya et al., 2020) presented the comparison of two models using different soft computing methods. (Katanalp & Eren, 2020) proposed a new classification technique (i.e., DT-based revised fuzzy logic (DT-RFL) and DT-based converted fuzzy logic (DT-CFL) using C4.5 DTs) to predict the cyclist crash severity. Results show that the DT-RFL has higher accuracy compared to conventional C4.5 DTs and DT-CFL algorithms. Other methods were also employed; for example, (Lee et al., 2018) explored structural equation modeling, and (Fountas et al., 2020) investigated a zero-inflated hierarchical ordered probit model.

More recently, tree-based approaches have also been employed by various researchers for crash severity prediction and to determine the factors affecting crashes. For instance, (Ijaz et al., 2021) presented a comparative study of DT, Decision jungle, and RF for severity prediction involving crashes of three-wheeled motorized rickshaws. According to the results from a stratified 10-fold cross-validation, the decision jungle surpassed the DT and RF with an overall accuracy of 83.7%. The authors also used Spearman correlation analysis to determine that lightning conditions, crashes involving young drivers (between 20 and 30 years), crashes on highways with high-speed limits, and shiny weather conditions result in more severe crashes. In addition to this, a comparative study of tree-based and non-parametric models for the prediction of severity in single-vehicle crashes was given by (Yan et al., 2021). This study uses five crash severity sub-datasets and found that in each dataset, urban freeways are a decisive factor that causes crashes, whereas rural freeways are more closely associated with more serious crashes. Furthermore, DT, RF, and Gradient Boosted Trees were utilized by (Amini et al., 2022) for the development of a hybrid framework incorporating explainable AI, predictive analytics, and heuristic optimization methods. The proposed methodology aims to examine and explain the risk variables for injury severity in automobile accidents. Apart from tree-based approaches, (Hou et al., 2022) investigates random parameters logit models for out-of-sample prediction, quantifying marginal effects, and analyzing temporal instability for crash severity.

Besides classical ML models, (Rahim & Hassan, 2021) applied Convolutional Neural Network (CNN) with customized loss function for the prediction of crash severity. In this study, the dataset is first transformed into images using t-SNE and the convex hull algorithm. Afterward, instead of a fully connected layer for classification, the authors use batch normalization, dropout, and ReLU activation functions in a two-layer neural network at the end of pre-trained EfficientNet-B7. Moreover, Deep Learning based approaches have also been implemented by (Ma et al., 2021; Sattar et al., 2022) for traffic crash severity prediction. (Ma et al., 2021) employed stacked sparse autoencoder (SSAE) for severity prediction in addition to the Catboost algorithm for analyzing the importance of contributing factors. Whereas, (Sattar et al., 2022) presented the comparative study of Vanilla-MLP along with MLP with embedded layers and TabNet for feature importance analysis. Authors in this study addressed the binary class problem by merging the severe and fatal classes, hence two classes are formed (i.e. severe and non-severe). Moreover, (Z. Yang et al., 2022) presented a case study of Chinese traffic accident data by proposing a Deep Neural Network (DNN) architecture for injury, death, and property loss prediction. Table 1 below provides a summary of the significant accident analysis-related literature in the past decade (Table 2).

In terms of SVM, both the OvO and OvR strategy has been widely used in the literature for multi-class classification. For instance, (Xu, 2011) extend the one-vs-rest SVM by introducing an approximate ranking loss as its empirical loss term with improved results. The one-vs-rest scheme is not confined to SVMs but rather it had been used with other ML models. Such as, (Ramírez et al., 2018) presented an ensemble of RF using one-vs-rest for the classification of Alzheimer's Disease and Mild Cognitive Impairment. The proposed model also shows higher precision compared to other state-of-the-art models. Moreover, the consensus learning technique has also been used by (Liu et al., 2021; Tang et al., 2018) for clustering problems, and shows good results. Additionally, (Hussain & Qaisar, 2022) have demonstrated that using OvR ensemble learning can significantly enhance the results. The authors employed the scheme on epileptic seizure classification and showed boosted classification accuracy, particularly for the multi-class classification problem. This forms the basis of our work to implement feature selection, and consensus learning for individual-class identification from difficult-to-detect classes in an imbalanced dataset for crash severity prediction.

Table 1
Summary of reviewed literature of the past 10 years in chronological order.

Reference	Algorithms	Dataset	Analysis
(Zong et al., 2013)	Bayesian networks, Binary logit model, Ordered probit model	Police reported traffic accident records, Jilin province, China.	Examined the crash severity based on the number of fatalities, the number of injuries, and property damage.
(Celik & Oktay, 2014)	Multinomial Logit (MNL).	Traffic accident data, Erzurum and Kars provinces, Turkey.	Investigates the risk factor affecting the traffic crash severity.
(Haleem et al., 2015)	Random Forest, Mixed Logit models.	Crash data from the Florida Department of Transportation (FDOT).	Compares different factors influencing pedestrian crash severity levels at signalized and un-signalized intersections.
(Kabeer, 2016)	Naïve Bayes, Decision trees, Ensemble technique.	Road traffic crash datasets from Leeds city council, UK.	Determines the crash severity level.
(Iranitalab & Khattak, 2017)	Random Forest, SVM, *k*-NN, Multinomial Logit, *k*-Means, latent class clustering.	Crash datasets from Nebraska, United States.	Predicts crash severity level and proposed cost-based approach for accidents.
(Lee et al., 2018)	Structural Equation Modeling (SEM).	Traffic accident data from Traffic Accident Analysis System (TAAS), Seoul, Korea.	Examine the influence of water level and rainfall on the crash severity.
(Chandra et al., 2019)	Random Forest, Logistic regression.	Data of traffic accidents and vehicles, UK.	Examine crash severity prediction with and without feature extraction.
(Kannojiya et al., 2020)	SVM, *k*-NN, RF, Logistic Regression, Naïve Bayes.	Road traffic accident datasets from Leeds city council, UK.	Analogize weather conditions that may cause accidents.
(Katanalp & Eren, 2020)	C4.5 algorithm, Decision Tree-based converted fuzzy logic (DT-CFL), Decision Tree-based revised fuzzy logic (DT-RFL).	Cycle-vehicle crashes, Adana City, Turkey.	Examine the effects of different factors responsible for injury severity in cyclists. DT-RFL has comparatively high accuracy.
(Fountas et al., 2020)	Zero-inflated hierarchical ordered probit model.	Single vehicle accidents, Scotland, UK.	Inspects the effect of various accident-causing features on severity level under different light and weather conditions.
(Ma et al., 2021)	Stacked sparse autoencoder (SSAE), Catboost, K-means clustering.	Data of traffic accident, UK.	Analytic framework to determine the importance of various factors, clustering co-related data, and binary class prediction (i.e. serious or non-serious).
(Rahim & Hassan, 2021)	t-SNE, Convex hull algorithm, EfficientNet-B7, SMOTE, CART, MARS	Louisiana Department of Transportation and Development (LDOTD).	Comparison of Deep learning model with SVM on the feature selected and balanced dataset.
(Yan et al., 2021)	DT, RF, Adaboost, Gradient Boosting Decision Tree, Extreme Gradient Boosting, Quadratic Discriminant Analysis, SVM, k-NN, Bernoulli Naive Bayes, MLP.	Dataset of the single vehicle crashed by University of North Carolina Highway Safety Research Center, US.	Comparison of ML algorithms for crash severity prediction and analysis of factors responsible for crashes.
(Ijaz et al., 2021)	DT, RF, Decision Jungle	Dataset of three-wheeled motorized rickshaw by provisional emergency response service, RESCUE 1122, Rawalpindi, Pakistan.	Analysis of severe crash-causing factors and comparison of ML models for predicting crash severity.
(Amini et al., 2022)	DT, RF Gradient Boosted Trees (GBT), Leave-one-covariate-out (LOCO), TreeExplainer, Variable neighborhood search (VNS).	Crash Report Sampling System (CRSS) by National Highway Traffic Safety Administration (NHTSA), US.	Analysis of the factors responsible for more severe accidents.
(Z. Yang et al., 2022)	DNN, SVM, RF, Logistic Regression	Chinese accident dataset by Key Laboratory for Urban Transportation Complex Systems, Theory, and Technology of Ministry of Education, China.	Investigation of the DNN model for prediction and analysis of factors affecting injury severity, death severity, and property loss severity.
(Hou et al., 2022)	Fixed parameter multinomial logit model, random parameters logic model, random parameters logic model with heterogeneity in means, random parameters logic model with heterogeneity in means and variances.	Police-reported crash dataset from Heilongjiang Province, China.	Predicts out-of-sample injury severities, along with the investigation of marginal effects and temporal instability of crash severity.
(Sattar et al., 2022)	Vanilla-MLP, MLP with embedded layers, TabNet	Data of traffic accident, UK.	Predicts the severity of crashes along with the determination of factors responsible for severe crashes.

3. Methodology

In this study, an ensemble of four classical classifiers is used to find a consensus of classification labels. Moreover, we also compute the discriminatory power of each feature in the crash prediction dataset. This enables the classifiers to use a subset of the feature that helps in the classification process with the added advantage of reducing the size of the data, thereby improving its efficiency. This section describes the feature selection strategy, the classifiers used in our approach, and the proposed approach of using OvRCL for achieving consensus-based class prediction.

3.1. Notations

Let $\mathbf{X}$ denote the data matrix which is divided into $\mathbf{X}_{train}$and $\mathbf{X}_{test}$ to represent the train and test data, respectively. N is the number of instances in $\mathbf{X}$ and F is the set of features. Let f_i be a single feature from F, and S be the set of selected features whereas $S^{'}$ be the set of non-selected features, (i.e., $S^{'} = F - S$). We denote $\mathbf{X}^{S}_{train}$ to represent the training dataset with only the selected features while the corresponding test dataset is represented by $\mathbf{X}^{S}_{test}$. C is the set of labels containing all the

Table 2
Pseudocode for the OvRCL algorithm.

ALGORITHM: OvR Consensus Learning (OvRCL)

Inputs: Classifiers CL_i $i = 1..P$, train and test data (i.e., X_{train}, X_{test}), class labels $\boldsymbol{y}$.
Outputs: Consensus output label, $\hat{y}$.
Initialization: Set of selected features $S = \{\phi\}$,
Set of non-selected features $S^{'} = \{\phi\}$,
Initialize $\hat{y}_i = 0 \; \forall \; i = 1..N_{test}$.

Begin

for each f_i in F
  for each c_j in C
    compute $I(f_i,c_j)$
  endfor
  compute $DMI(f_i)$ using Eq. (3)
endfor
select top $|S|$ features using Eq. (4)
select the dataset X^S_{train} and X^S_{test}
for each classifier $i = 1..P$
  train CL_i where $i \in$ {RF, SVM, k-NN, BG}
endfor
for each $i = 1..N_{test}$
  $\hat{y}_i^{RF} CL_{RF}(X^S_{train}, X^S_{test})$ $\hat{y}_i^{SVM} CL_{SVM}(X^S_{train}, X^S_{test})$ $\hat{y}_i^{k-NN} CL_{k-NN}(X^S_{train}, X^S_{test})$ $\hat{y}_i^{BG} CL_{BG}(X^S_{train}, X^S_{test})$ compute $\hat{y}_i$using Eq. (6)
end for
return$\hat{y}$

End

classes $\{c_1, \cdots, c_K\}$, where c_j denotes the single class and K is the total number of classes (not to be confused with k used for nearest neighbors in k-NN). We denote the i^{th} classifier as $CL_i(i = 1 \cdots P)$ where P is the number of classifiers in the ensemble, and the function $CL_i(\mathbf{X}^S_{train})$ returns the labels predicted by the classifier using the data $\mathbf{X}^S_{train}$. Let $\mathbf{y}$ be the class labels such that $y_i \in \{1 \cdots \text{K}\}$ belonging to the instance i.

3.2. Feature selection strategy

3.2.1. Mutual information (MI)

Feature selection is important to remove redundancy among features of the data and keep only the discriminating features that help classifiers to differentiate between instances of different classes. Most feature selection methods associate a score with each feature that signifies how "discriminatory" the feature is in differentiating instances of different classes. One such popular and widely used method to compute the feature's score is its MI with the class. MI is the measure of the amount of information or relevance, on average, that one attribute contains about the other. A high MI score means that using that feature can help in discriminating instances of different classes. The MI score is computed between a feature and each of the classes.

Let $P_r$denote the probability than the mutual information $I(f_i; c_j)$between f_i and c_j is given by:

$$I(f_i; c_j) = P_r(f_i, c_j)\left[log_2 \frac{P_r(f_i, c_j)}{P_r(f_i)P_r(c_j)}\right] + P_r(\bar{f}_i, c_j)\left[log_2 \frac{P_r(\bar{f}_i, c_j)}{P_r(\bar{f}_i)P_r(c_j)}\right] \quad (1)$$

where $\bar{f}_i$ means that the feature f_i does not occur. The value of $I(f_i; c_j)$ is zero if f_i and c_j are independent. The above equation shows the relation between one feature f_i and one given class c_j. To compute the overall feature score, we can use the sum, average, or maximum values with all classes as demonstrated by (Hussain, et al., 2022; Yang & Pedersen, 1997). Here, we use the sum given by:

$$I_{avg}(f_i; C) = \sum_{j=1}^{n} I(f_i, c_j) \quad (2)$$

3.2.2. MI-based feature selection

For feature selection, (Hussain et al., 2020) exploited a derivative of the MI, called Discriminative Mutual Information (DMI). Features are scored in a way such that the ratio of the MI score with one class against the rest is maximized. Mathematically, DMI is computed as:

$$DMI(f_i) = \frac{N_i}{\sum_{j=1, j\neq i}^{n} N_j} \times \frac{P_r(f_i, c_j)\left[log_2 \frac{P_r(f_i, c_j)}{P_r(f_i)P_r(c_j)}\right] + P_r(\bar{f}_i, c_j)\left[log_2 \frac{P_r(\bar{f}_i, c_j)}{P_r(\bar{f}_i)P_r(c_j)}\right]}{\sum_{j=1, j\neq i}^{n} P_r(f_i, c_j)\left[log_2 \frac{P_r(f_i, c_j)}{P_r(f_i)P_r(c_j)}\right] + P_r(\bar{f}_i, c_j)\left[log_2 \frac{P_r(\bar{f}_i, c_j)}{P_r(\bar{f}_i)P_r(c_j)}\right]} \quad (3)$$

where N_i is the number of elements in class i. Afterward, the first feature is selected having the maximum DMI and the remaining features are selected if they maximize:

$$DMI(f_i; C) - \beta \sum_{f_s \in S} I_{avg}(f_s; f_i) \quad (4)$$

where β is a user-defined parameter to manage the relative importance of redundancy between a particular feature and already selected features. In Eq. (4), the first part estimates the mutual information between a particular feature and the class, whereas the second part gives the redundancy of i^{th} feature f_i and already selected set of features, S. Therefore, in addition to choosing the most discriminating features, we also ensure that any chosen features bring new information that was not catered to using the existing set of selected features.

3.3. Proposed One-vs-Rest consensus learning (OvRCL) methodology

Ensemble learning is a machine learning technique that merges the predictions of multiple classifiers to make a strong predictive model (Bauer & Kohavi, 1999). Several algorithms are run independently on the multi-class problem and the final result is acquired by voting or averaging the prediction from each algorithm. If the constituent algorithms struggle on the multi-class problem, for instance, if a class or set of classes is difficult to separate, consensus voting may also struggle.

This paper integrates binary classification with consensus voting to propose a new OvRCL framework. The OvRCL is a heuristic approach for

multi-class classification by repeatedly exploiting binary classification as it divides the multi-class data into multiple binary classification models. A binary classifier (CL) is then trained on each model and the final prediction is made based on the highest score. Instead of executing OvR for complete classification and then obtaining the majority vote, OvRCL performs consensus voting individually at the class level. The classifiers are better able to perform for difficult-to-detect classes since the objective is to distinguish only one class from the others in an OvR scenario. For a dataset with K classes, an algorithm is iterated for $j = 1 \cdots K$ times, such that $c_j = 1$ if $CL(\boldsymbol{x}_i) = j$, and 2 otherwise $\forall\, i = 1 \cdots N$, where $\boldsymbol{x}_i$ is an instance of $\boldsymbol{X}$ having N number of instances. Furthermore, during the testing phase, each model determines whether the sample belongs to class 1 or contrarily to $j = 2 \cdots K$. At the end of the j^{th} iteration, every instance has i predictions belonging to class j where i is the number of classifiers in the ensemble setting. The final class is predicted by taking a majority vote of j^{th}class from the prediction of all the classifiers. This is given by the following Eq. (6):

$$\widehat{y}_{x_{test}} = argmax CL_i\left(\boldsymbol{x}^S_{test}\right) \tag{6}$$

where $\widehat{y}_{x_{test}}$ denotes the predicted class and $\boldsymbol{x}^S_{test}$ denotes an instance of the test data with the selected feature set.

The process is repeated for $j = 1 \cdots K$ times until all instances are labeled by the OvR consensus-based classifier. Note that there may arise a case where an instance previously classified as belonging to class j may be predicted as belonging to class k, $1 \leq j,\ k \leq K$. The label for this instance will only be updated if the number of classifiers predicting the label to be k is greater than the number of classifiers predicting the label to be j. Let the total number of classifiers used in the ensemble learning be P and the number of classifiers predicting the label as j be denoted as p_j, then the label of an instances $\boldsymbol{x}_i$ may only be updated if $\frac{p_k}{P} > \frac{p_j}{P}$. The overall algorithm is shown in Fig. 2.

The MI score for feature selection is computed on the training dataset and the same selected feature set is then used for the test data. Thereafter, each classifier is trained on the training dataset using the one-vs-rest strategy and the OvRCL algorithm is used for the final prediction. The following shows the algorithm of the proposed OvRCL algorithm.

3.4. Prediction methods

A variety of classification techniques can be used to predict the label (Ashraf et al., 2022; Bano & Hussain, 2021; Hameed et al., 2020; Si et al., 2022; Zelenkov & Volodarskiy, 2021). These models are trained on the training data and predict the labels on the test data. This research implements consensus-based learning by employing k-NN, SVM, RF, and Bagging classifiers because they have been extensively employed in the literature on crash severity prediction on a variety of datasets and show comparatively good results (Haleem et al., 2015; Iranitalab & Khattak, 2017; Kabeer, 2016; Kannojiya et al., 2020; Yan et al., 2021; Yang et al., 2022). We briefly describe these classifiers below.

3.4.1. Random Forest (RF)

Random Forest (RF) is a technique that leverages DTs and was first proposed as a classifier by (Breiman, 2001). It comprises a multitude of tree-structured predictors in such a way that each tree depends on the value of independent and same random vectors, while they would cast an individual vote for the most popular class. In RF, each node is split using the best features, from a subset of indicators arbitrarily picked at that node. This method is robust and performs well in comparison with other commonly used classifiers (Qaisar & Hussain, 2020, 2021). Its implementation requires the identification of the number of trees to grow and the number of candidates randomly sampled at each stage. To train the RF model, in this study we fix the number of trees at 60. Usually, the number of trees in the forest is a user-defined parameter and can be set by starting with a lower value and gradually increasing it. This can be done using a training-validation split strategy.

3.4.2. Support Vector Machines (SVM)

Primarily SVM is a binary classification algorithm first proposed by (Cortes & Vapnik, 1995) and can be used for both regression and classification problems. SVM works by finding an optimal hyperplane to separate instances of different classes and is based on statistical learning theory and structural minimization. In cases where a linear separation between classes is not possible, a "kernel trick" is used to find distances between instances in a transformed higher dimensional space defined by the kernel function in which the data is linearly separable. Popular kernel methods include the Radial Basis Function (RBF), Perceptron (MLP), and polynomial functions (Hussain, 2019). The choice of kernel method is both problem and data-dependent and multiple kernels can be tried to select the one with the best performance using a validation set. In this paper, we use the third-order polynomial kernel function to achieve separability as the data is not linearly separable. This was selected after trying multiple kernel methods and varying the C

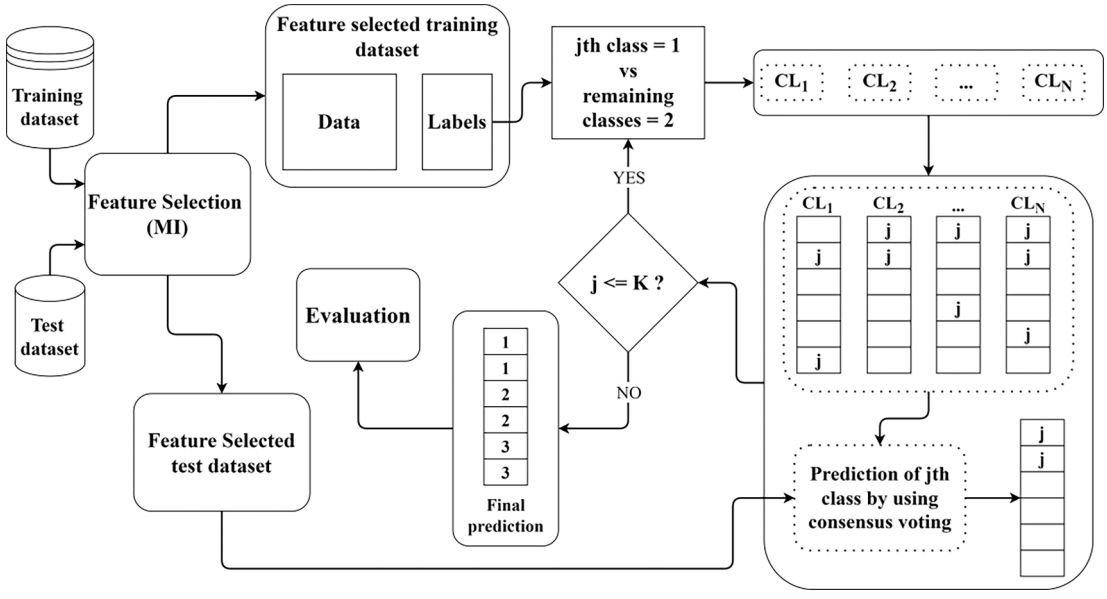


Fig. 2. The proposed OvRCL framework using N classifiers (CL_1, …, CL_N).

parameter using cross-validation.

3.4.3. k-Nearest Neighbors (k-NN)

The k-Nearest Neighbor is a simple classifier proposed by (Cover & Hart, 1967). It is a prediction method that predicts the unclassified test data by looking at the closest set of classified points. Two things must be kept in mind while implementing the k-NN: the value of k or the number of neighbors that should be considered to classify the test point, and the distance function to be used to measure the distance between the test point and the set of closest points in the training data. In this paper, we set the value of k to be 3 but similar results are achieved for $k = 5$ or 7. The City-block (also called Manhattan) distance is used as the distance function which calculates the absolute difference between two points or coordinates. If d_{ij} denotes the distance between observations i and j and x_{il} and x_{jl}represent the values of the l^{th} variable for observations i and j, respectively, then the city-block distance can be computed as:

$$d_{ij} = \sum_{l=1}^{|F|} |x_{il} - x_{jl}| \tag{5}$$

3.4.4. Bagging classifier (BG)

In (Breiman, 1996), bagging predictors as an algorithm for generating various predictors and using them to make an aggregated predictor was first introduced. It is an ensemble technique designed to improve the stability and accuracy of ML algorithms. It aggregates the averages over the variants while foreseeing a numerical outcome and the final class is predicted by a majority vote. The basic idea of BG is that it creates many "weak learners" and uses them to build a "strong learner". Generally, it makes many DTs that are weak learners and combines them to create a strong classifier. Each tree gives a vote to a class and the final prediction of the new class is gained by the class that has the most votes.

4. Experimental setup

4.1. Dataset description

The experimental setup used in this paper is divided into two parts:

1. The first part is a detailed case study used to evaluate the important features that impact crash severity prediction and well as a qualitative analysis for policy decision making. That dataset used for this study was obtained from the open data source of Leeds city council[2]. We refer to this as the *Leeds* accidents dataset. It comprises information on accidents from different years across Leeds. In our study, the crash data from 2015−2018 is used as training data while the data from 2019 is used as the test dataset. In total, the dataset contains roughly 11,300 instances divided as follows – 9400 instances are used in the training data, whereas the test data contains approximately 1900 instances. The data records multiple features for each crash including the Reference Number, Grid Ref: Easting, Grid Ref: Northing, Number of Vehicles, crash Date, Time of the crash, 1st Road Class, Road Surface, Lighting Conditions, Weather Conditions, Type of Vehicle, Casualty Class, Casualty Severity, Sex of Casualty (referred to as Gender of casualty) and Age of Casualty.
2. The second part of the analysis is done to substantiate the performance of the OvRCL framework proposed in this study and its suitability for such as analysis. To do this, we leverage two more publicly available datasets with variations for crash severity prediction. The first is a cycling casualty in Leeds[3] data from 2009 to 2015 (referred to as *bicycle* crashes). This is a relatively small dataset with approximately two thousand instances and 16 features in total. The second relates to road casualty accidents from 2010 to 2021 in the city of Manchester[4] (referred to as *Manchester* accidents), having over 45 thousand records with 27 attributes. The two datasets are chosen to observe the performance on a small and medium sized dataset.

Preprocessing is the important step to remove features that may not be helpful in determining the information of interest. Therefore, all the datasets are cleaned by removing the spatial and temporal features as they are not relevant for crash severity prediction.

Summary of the Leeds accidents, which is used for a detailed case study, is given below. We remove features such as Reference Number, Grid Ref: Easting, Grid Ref: Northing, crash Date, Time, and Age of Casualty. The remaining data contains 8 features in Leeds accidents, 9 features in bicycle crashes, and 15 features in Manchester accidents. The classification problem tackled in this study is a multi-class problem with three classes. Regarding the casualty severity from all the datasets used, 1 denotes a fatal crash; 2 represents if the crash caused any serious injury to the driver, passenger(s), or any bystander; and 3 implies slight injuries to any of these subjects. Table 3 shows the summary of the attributes in the training and test dataset with exploratory data analysis. In terms of temporal distribution, the monthly and yearly distribution of the crashes with their severity count is given in Fig. 3. It can be seen that the data is highly imbalanced between the three classes.

4.2. Evaluation metrics

Many metrics can be used for measuring the performance of classifiers. One frequently used metric is the accuracy metric which measures the ratio of the correctly classified instances to the total number of instances. The benefit of using the accuracy measure is that it is readily interpretable. The accuracy score is computed as:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} \delta(y_i, \hat{y}_i) \tag{7}$$

where $\delta(y_i, \hat{y}_i)$ is 1 if both the actual and predicted labels belong to the same class and 0 otherwise.

A second evaluation metric used in this study is the Normalized Mutual Information (NMI) metric which is based on the informetric relationship between two variables. It depends on the entropy (H) between the actual labels y and the predicted labels, $\hat{y}$. Mathematically, this is given by:

$$NMI(y, \hat{y}) = \frac{H(y) + H(\hat{y})}{H(y, \hat{y})} \tag{8}$$

NMI is usually a preferred score when comparing multiple algorithms since it takes size and biased assignment into consideration. For instance, if 95 instances belong to class 1 and 5 belong to class 2, but a classifier assigns a single label to all elements, it will still result in an accuracy value of 95%. In such a case, the NMI score will be quite low even though accuracy is rather high indicating that the classifier did a poor job in the classification task.

In addition, this study also employs specificity and sensitivity to evaluate the performance of ML models. Specificity is the ability of a model to detect the True Negative (TN) of each output class whereas sensitivity is the detection of the True Positive (TP) by the classifier. Consequently, specificity is called the true Negative Rate (TNR), and sensitivity is referred to as True Positive Rate (TPR). If FP is the number of samples that were negative but predicted as positive, FN denotes the

[2] Leeds City Council dataset (https://www.data.gov.uk/dataset/6efe5505-941f-45bf-b576-4c1e09b579a1/road-traffic-accidents).

[3] The cycling casualty dataset in Leeds (https://datamillnorth.org/dataset/cycling-casualties-in-leeds).

[4] The Manchester road casualty accidents dataset (https://www.data.gov.uk/dataset/25170a92-0736–4090-baea-bf6add82d118/gm-road-casualty-accidents-full-stats19-data).

Table 3
Exploratory descriptive analysis of the Leeds council dataset.

Features	Description	Summary			
		Train dataset (2015–2018)		Test (2019)	
		Occurrence (*n*)	Frequency (%)	Occurrence (*n*)	Frequency (%)
1st Road Class	Motorway	736	0.0782	180	0.0943
	A(M)	210	0.0223	72	0.0377
	A	3133	0.3329	567	0.2973
	B	422	0.0448	80	0.0419
	C	19	0.0020	3	0.0015
	Unclassified	4891	0.5197	1005	0.5270
Road Surface	Dry	6989	0.7426	1359	0.7126
	Wet / Damp	2276	0.2418	521	0.2732
	Snow	53	0.0056	3	0.0015
	Frost / Ice	81	0.0086	17	0.0089
	Flood (surface water over 3 cm deep)	12	0.0012	5	0.0026
Light Conditions	Daylight: streetlights present	6498	0.6904	1389	0.7283
	Darkness: streetlights present and lit	2177	0.2313	436	0.2286
	Darkness: streetlights present but unlit	26	0.0027	9	0.0047
	Darkness: no street lighting	201	0.0213	43	0.0225
	Darkness: street lighting unknown	509	0.0540	30	0.0157
Weather Conditions	Fine without high winds	7949	0.8446	1633	0.8563
	Raining without high winds	826	0.0877	215	0.1127
	Snowing without high winds	41	0.0043	3	0.0015
	Fine with high winds	109	0.0115	19	0.0099
	Raining with high winds	135	0.0143	25	0.0131
	Snowing with high winds	20	0.0021	1	0.0005
	Fog or mist – if the hazard	29	0.0030	1	0.0005
	Other	50	0.0053	8	0.0041
	Unknown	252	0.02677	2	0.0010
Casualty Class	Driver or rider	5615	0.5966	1152	0.6040
	Vehicle or pillion passenger	2388	0.2537	428	0.2244
	Pedestrian	1408	0.1496	327	0.1714
Gender of Casualty	Male	5404	0.5742	1149	0.6025
	Female	4007	0.4257	758	0.3974
Type of Vehicle	Pedal cycle	1257	0.1335	236	0.1237
	M/cycle 50 cc and under	80	0.0085	16	0.0083
	Motorcycle over 50 cc and up to 125 cc	332	0.0352	77	0.0403
	Motorcycle over 125 cc and up to 500 cc	98	0.0104	17	0.0089
	Motorcycle over 500 cc	212	0.0225	39	0.0204
	Taxi/Private hire car	372	0.0395	60	0.0314
	Car	6082	0.6462	1295	0.6790
	Minibus (8 – 16 passenger seats)	30	0.0031	4	0.0020
	Bus or coach (17 or more passenger seats)	528	0.0561	65	0.0340
	Ridden horse	1	0.0001	2	0.0010
	Agricultural vehicle (including diggers etc.)	3	0.0003	1	0.0005
	Tram / Light rail	3	0.0003	0	0
	Goods vehicle 3.5 tons mgw and under	274	0.0291	73	0.0382
	Goods vehicle over 3.5 tons and under 7.5 tons mgw	46	0.0048	6	0.0031
	Goods vehicle 7.5 tons mgw and over	40	0.0042	8	0.0041
	Mobility Scooter	15	0.0015	3	0.0015
	Other Vehicle	33	0.0035	2	0.0010
	Motorcycle - Unknown CC	4	0.0004	3	0.0015
Casualty Severity	Fatal	66	0.0070	–	–
	Serious	1264	0.1343	–	–
	Slight	8081	0.8586	–	–

number of samples that were positive but forecasted as negative, TP denotes the number of samples that were positive and predicted as negative and TN represents the number of instances that were negative and are also predicted as negative and *j* represents the respective labels. Then sensitivity and specificity for *jth* class is given by:

$$Sensitivity_j = \frac{TP_j}{TP_j + FN_j} \tag{9}$$

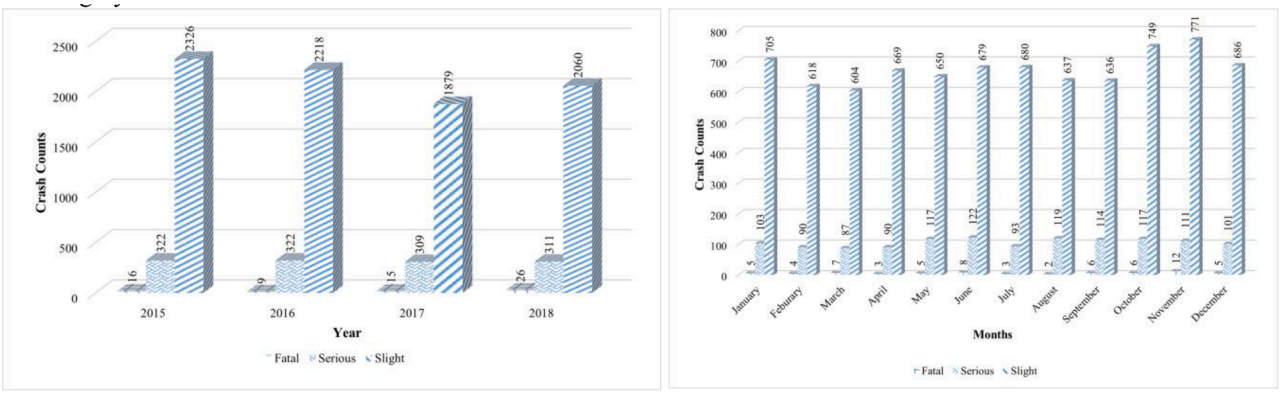
$$Specificity_j = \frac{TN_j}{TN_j + FP_j} \tag{10}$$

5. Results and discussion

The experimental setup for this study is designed to validate the intuition behind the proposed algorithm and compare its efficacy to existing state-of-the-art recent algorithms. To do this, we tested the proposed algorithm in a stepwise fashion by incorporating only a single step while keeping other factors constant. The results are compared in terms of accuracy since it has a straightforward interpretation and is one of the desirable properties to see how well an algorithm can classify compared to the ground truth. The NMI score is also used when analyzing the results of the algorithm in comparison with the individual classifiers used in the ensemble method.

5.1. Selecting the features

The first contribution of this study is to use an automated feature selection technique to select only the distinguishing features rather than



(a) (b)

Fig. 3. Distribution of crash severity by month and by year. (a) Counts of crashes by year. (b) Count of crashes by month.

the entire feature set. The DMI has the added effect of selecting features that only help to distinguish instances of one class from all other features as opposed to the standard MI technique which may select features that have a higher average or total MI score. To see which features are the most helpful, and to determine how many features should be selected, we compute the DMI score of each feature. The test dataset must not be used in any computation where prior knowledge of the class labels is necessary, hence we only compute these scores from the training dataset. The result is shown in Fig. 4.

Not all features in the dataset have equal importance in the classification process. Usually, the question of how many features to select is a balance between the size of the data (to reduce) and the information content to be retained. In our case, the type of vehicle, the casualty class (i.e., whether a pedestrian, driver, or passenger), the number of vehicles involved, the gender of the casualty, and the road class show the highest scores. Together, these five (05) features retain over 91% of the information content. Therefore, we select these five features when using feature selection. The rest of the features are left out in both the training and test dataset. The distribution of the selected features concerning the crash severity is given in Fig. 5, where *crash count* denotes the *occurrence* values in Table 3.

As observed in Fig. 4, the *type of vehicle* had the highest information score. Looking at the distribution, we see that in several cases simply knowing the type of vehicle is enough to predict the crash severity or eliminate certain types of severity. For instance, if a vehicle is less than 50 cc, a minibus, a coach, a horse, an agricultural vehicle, a tram, a goods vehicle with 3.5 to 7 tons of weight, a mobility scooter, or a motorcycle, there are fewer chances of a fatal crash. Similar information from other selected features also helps in the overall prediction of crash severity.

5.2. *Effect of feature selection*

We use the full version of the dataset $\boldsymbol{X}_{train}$ and $\boldsymbol{X}_{test}$ that employ the complete feature set and the reduced version of the dataset $\boldsymbol{X}^{S}_{train}$ and $\boldsymbol{X}^{S}_{test}$ with only the selected features. Each of the classifiers discussed in Section 3 is used to classify the two datasets along with the proposed OvRCL algorithm. Additionally, commonly used ML models like AdaBoost and ANN have also been compared with the OvRCL framework. As the data is linearly non-separable so the ANN with three hidden layers was implemented. This number was selected by trying a few other models with different layers and the model with the better result is selected. Furthermore, an Adaboost with 30 learners is used to predict the crash severity. The accuracy of the ML models is shown in Fig. 6 whereas the results for NMI are compared in Table 4.

From Fig. 6, we can see some interesting observations. Firstly, almost all the classical classification techniques result in very similar performance. In particular, Bagging is an ensemble classification technique, but its accuracy value is no better than using the standard k-NN or SVM. Therefore, using an ensemble technique in itself does not necessarily

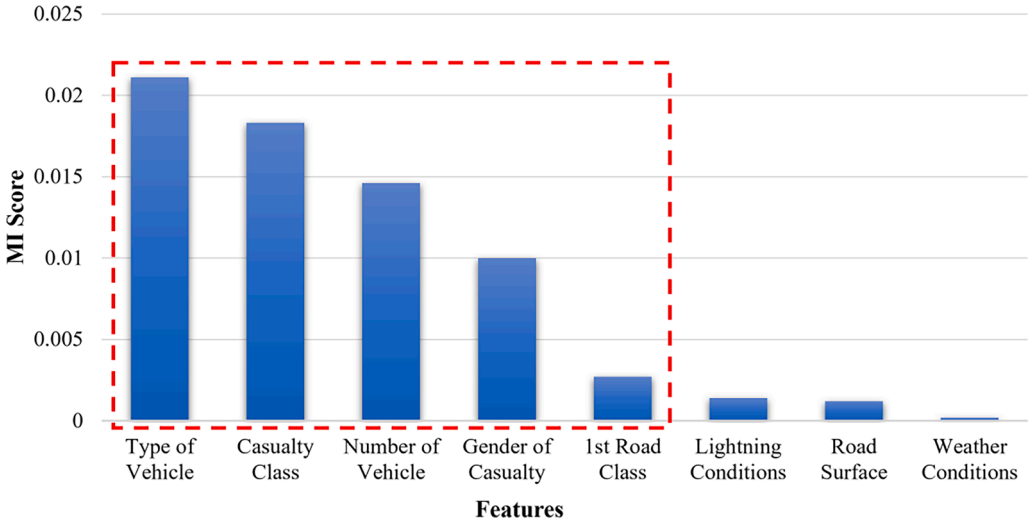


Fig. 4. Influence of each feature (using MI score) on crash severity in Leeds accidents.

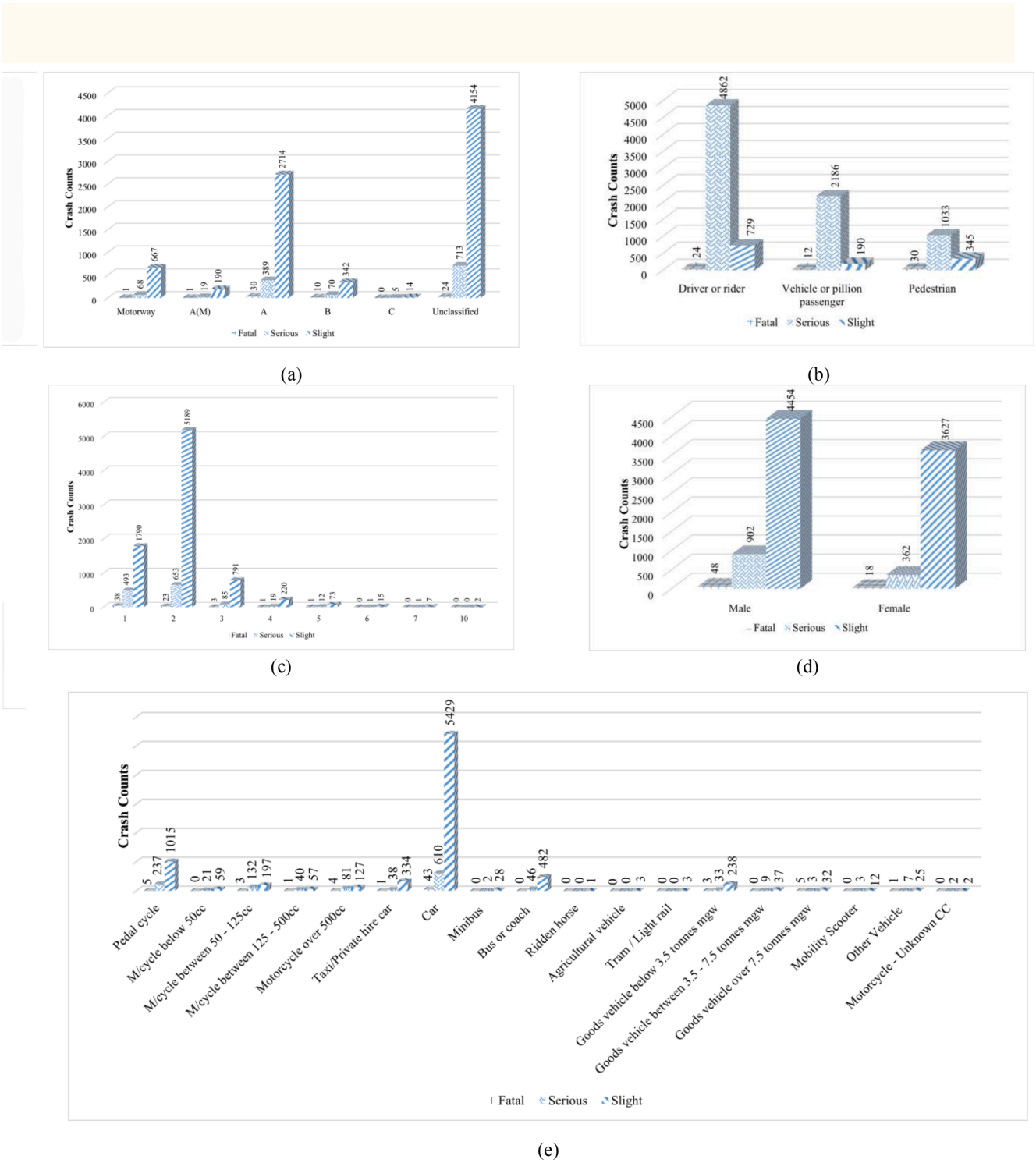


Fig. 5. Distribution of the selected features according to crash severity in Leeds accidents. (a) Distribution of severity across 1st road class. (b) Distribution of severity across casualty class. (c) Distribution of severity across the number of vehicles. (d) Distribution of severity across the gender of casualty. (e) Distribution of severity across the type of vehicle.

boost the performance in all cases. Secondly, it is also clear that feature selection helps in improving the accuracy score though in most cases the increase is only marginal. However, for the proposed approach, the jump in accuracy value is significant when using feature selection. In our view, this is because the features are selected according to their ability to distinguish a particular class and its effect is most useful when using the OvRCL approach.

5.3. Effect of OvR approach on accuracy

The OvR approach is traditionally used in such algorithms as SVM since it is much faster than the OvO approach which requires making binary combinations of all classes. SVM, however, is a single classifier so if the method is biased against any class or unable to handle imbalanced data, the classes may be confused which leads to lower accuracy. Table 5 shows the confusion matrix when using SVM with the OvR strategy for

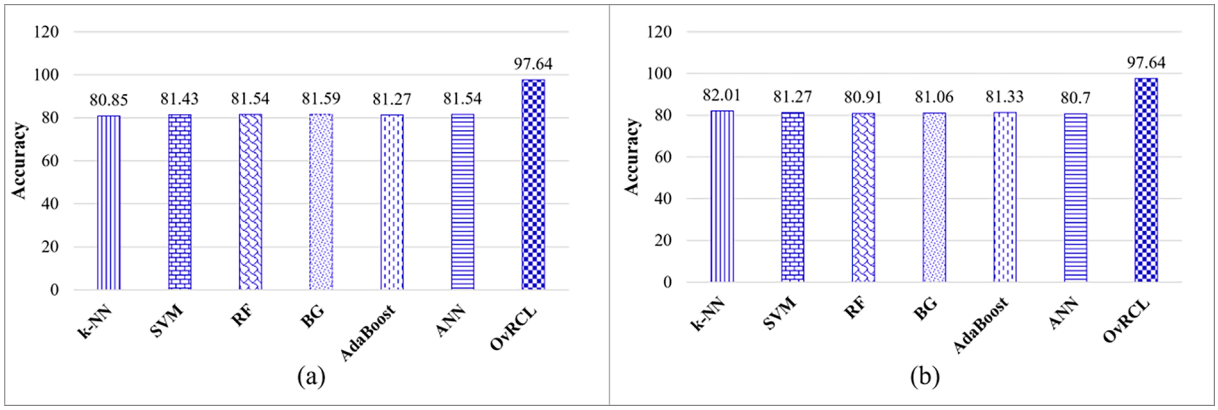


Fig. 6. Comparison of different classifiers and proposed model in Leeds accidents (a) Selected features (b) Complete features.

Table 4
Comparison of different classifiers and proposed method using NMI scores.

Models	Selected Features	Complete Features
k-NN	0.024	0.060
SVM	0.032	0.015
BG	0.043	0.040
RF	0.040	0.010
AdaBoost	0.011	0.015
ANN	0.026	0.016
OvRCL	**0.931**	**0.852**

Table 5
Confusion matrix for crash severity prediction using SVM with OvR strategy in Leeds accidents.

Actual Values	Predicted Values			
		Fatal	*Serious*	*Slight*
	Fatal	0	0	22
	Serious	0	0	334
	Slight	0	4	1547

predicting the crash severity. We see that all *fatal* and *serious* crashes were classified as *slight* injuries which is by far the class with the largest number of instances. This is the reason we had a moderate accuracy value when using SVM (Fig. 6), but a very low NMI score (Table 4) because the classes are imbalanced.

On the other hand, ensemble methods perform better since they are less biased due to a single classifier. However, most ensemble classifiers predict the final class from the overall prediction of the individual classifier rather than one class at a time as in the OvRCL strategy. Bagging is an ensemble method that uses multiple DTs to predict the labels. Table 6 shows the confusion matrix when using the Bagging algorithm. Compared to the SVM results, Bagging does a better job in predicting *fatal* and *serious* injuries, but it too confuses a significant amount of *fatal* and *serious* injuries with *slight* injuries, resulting in an overall low NMI score.

When using an ensemble method with the OvR strategy and aggregating after each class prediction rather than after the entire label prediction, the overall confusion between classes is significantly reduced, particularly where the instances were low (i.e., smallest class in unbalanced data). This is shown in the confusion matrix in Table 7, whereas Table 8 and Table 9 present the comparison using sensitivity and specificity between the proposed model and other machine learning algorithms respectively. It can be seen that the proposed OvRCL also has a higher true positive rate (sensitivity) and true negative rate (specificity) as compared to the conventional machine learning model.

Table 6
Sample confusion matrix of Bagging ensemble classifier in Leeds accidents.

Actual Values	Predicted Values			
		Fatal	*Serious*	*Slight*
	Fatal	1	2	19
	Serious	0	14	320
	Slight	2	10	1539

5.4. Validation of OvRCL algorithm

To acquire the most relevant features for crash severity prediction MI-based feature selection (Eq. (4) is applied to all the datasets. Fig. 7 and Fig. 8 illustrate the MI score of features and the selected subset in the dataset of cycling casualties in Leeds, and roads accident in Manchester and the UK, respectively. Feature subset is selected as they are more relevant and contain 90% of the information regarding the accident severity (as specified in section 5.1).

Furthermore, OvRCL is better able to detect the hard-to-detect classes irrespective of the size of the data as illustrated in Fig. 9 and Fig. 10. The proposed scheme also surpassed the conventional ML models in terms of NMI score as shown in Table 10 and Table 11, for sensitivity and specificity in Table 12 and Table 13 for the bicycle crashes and Manchester accidents data, respectively.

5.5. Comparison with other state-of-the-art methods

Finally, we compare the proposed method with several other state-of-the-art algorithms used in the literature on the Leeds and other datasets. The results are shown in Table 14 using the accuracy measure and different datasets. Since we do not have the code for the other methods, we simply cite the results as mentioned and explain the difference in the datasets used.

For the *k*-NN, SVM, Bagging, and Random Forest, the datasets are the same as the OvRCL algorithm. In which *k*-NN represents the baseline algorithm, SVM is a binary classifier and utilizes the OvR strategy for multi-class problems, while BG and RF represent ensemble techniques. Hence, these methods provide a good comparison to see the effect of using existing techniques of OvR and Ensemble in the existing literature.

Table 7
Sample confusion matrix using the proposed OvRCL algorithm in Leeds accidents.

	Predicted Values			
Actual Values		*Fatal*	*Serious*	*Slight*
	Fatal	20	2	0
	Serious	9	295	30
	Slight	1	10	1540

Table 8
Sensitivity values on the different models in feature-selected Leeds accidents.

Classes	*k*-NN	SVM	RF	BG	AdaBoost	ANN	OvRCL
Fatal	0	0	0	0.3333	0.0606	0.0606	0.9090
Serious	0.1796	0	0.0149	0.0419	0.0008	0.0467	0.8832
Slight	0.9368	0.9974	0.9909	0.9922	0.9989	0.9844	0.9929

Table 9
Specificity values on the different models in feature-selected Leeds accidents.

Classes	*k*-NN	SVM	RF	BG	AdaBoost	ANN	OvRCL
Fatal	0.9989	1	0.9994	0.9989	0.9999	0.9994	0.9946
Serious	0.9357	0.9974	0.9904	0.9923	0.9990	0.9845	0.9923
Slight	0.1825	0	0.0196	0.0477	0.0038	0.0519	0.9157

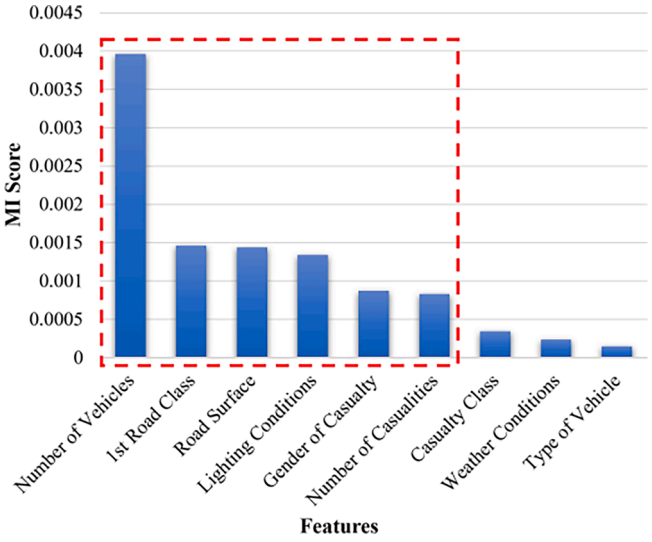


Fig. 7. Mutual Information Score of all the features and selected subset in bi-cycle crashes.

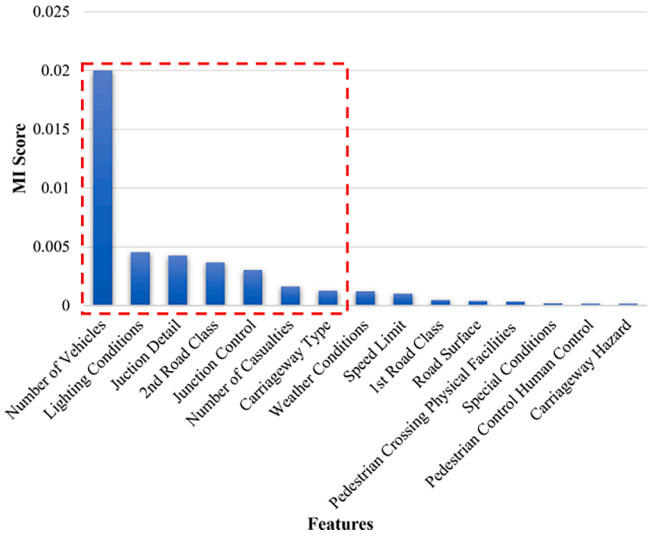


Fig. 8. Mutual Information Score of all the features and selected subset in Manchester accidents.

For (Kabeer, 2016) and (Kannojiya et al., 2020) the dataset is the same but differs in the years used. Whereas (Chandra et al., 2019; Ma et al., 2021), and (Sattar et al., 2022) make use of a different but comparable dataset for UK road traffic accidents. The results are evident from Table 8 in the previous section. On the same dataset, the existing machine learning classifiers of *k*-NN, SVM, Bagging and RF have comparable results of around 81–82% accuracy. From (Kannojiya et al., 2020) and (Chandra et al., 2019), Logistic regression and RF are more suitable for predicting crash severity. This may be because the crash data is usually highly imbalanced with distinctive features showing correlations with multiple classes under different conditions.

In the literature, many authors have tried to overcome this in two ways: firstly, by combining severity levels and transforming the problem into a binary classification problem and predicting only severe and non-severe crashes irrespective of whether the severe crashes were fatal or not (Sattar et al., 2022). Secondly, some techniques employ machine learning techniques to cluster the data using spatial dimensions (Iranitalab & Khattak, 2017; Ma et al., 2021). This can be done using a clustering method, such as the *k*-means or other clustering techniques (Hussain, 2011), to group accidents geographically. Their results show that it is easier to predict the severity of the crash after grouping them and predicting the severity of each group separately. However, in all the cases and using all the methods, the crash severity prediction remains below 90% on all the datasets.

6. Discussion

In this paper, we present a new machine learning approach for the classification of crash severity prediction by using a combination of feature selection and One-vs-Rest Consensus Learning (OvRCL). The previous sections show the performance of the proposed strategy on the Leeds accident dataset in comparison to established state-of-the-art techniques. In this section, we further analyze the features concerning the crash severity along with the advantages and limitations of the proposed model, and policy implications from this study.

6.1. Effect of classifiers on OvRCL

In discussed previously and shown in Fig. 2, any choice of classifiers could be used for the consensus learning. In the previous section, we chose SVM, K-NN, RF, and BG as classifiers because they have been extensively used in the literature for crash severity prediction on a variety of datasets and show comparatively good results (Haleem et al., 2015; Iranitalab & Khattak, 2017; Kabeer, 2016; Kannojiya et al., 2020; Yan et al., 2021; Yang et al., 2022).

Here we show that OvRCL improves on the classification accuracy for all the accident datasets compared to the individual classifiers and traditional ensemble approaches. To show this, we replace SVM and RF with Artificial Neural Network (ANN) and AdaBoost (AB) in the proposed OvRCL framework. It can be seen from Table 15 that both versions of OvRCL outperforms the individual models, including traditional ensemble methods like AdaBoost or Bagging. The two OvRCL differ in accuracy depending on the classifiers chosen. The choice of the

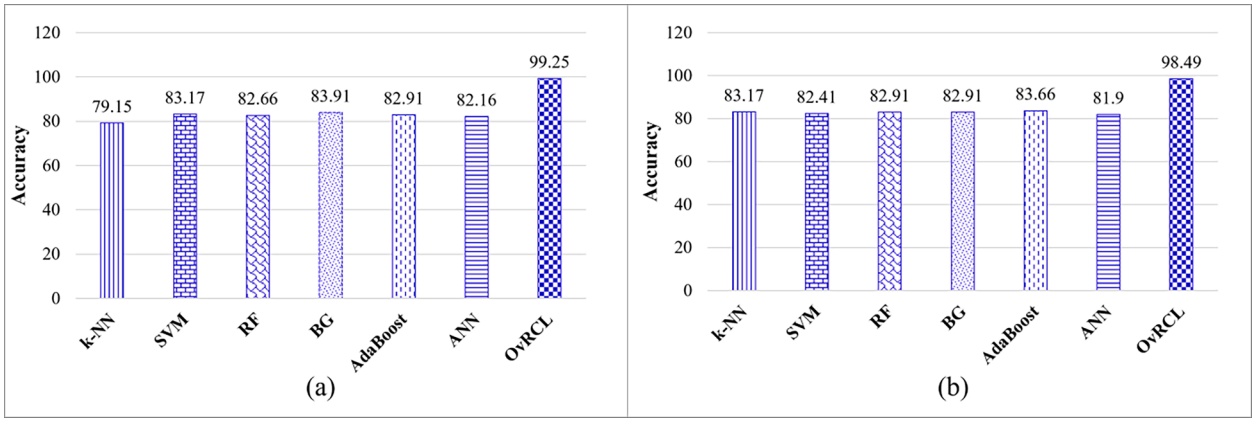


Fig. 9. Comparison of different classifiers and proposed model in bicycle crashes using (a) selected features (b) complete features.

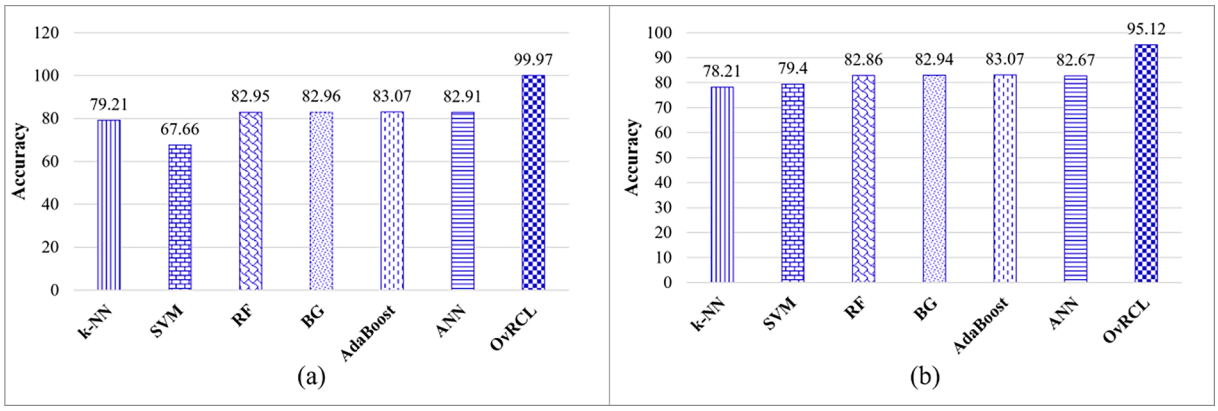


Fig. 10. Comparison of different classifiers and proposed model in Manchester accidents using (a) selected features (b) complete features.

Table 10
Comparison of different classifiers and proposed method in bicycle crashes using NMI scores.

Models	Selected Features	Complete Features
k-NN	0.0161	0.0268
SVM	0.0224	0.0089
BG	0.0046	0.0120
RF	0.0072	0.0046
AdaBoost	0.0163	0.0527
ANN	0.0028	0.0058
OvRCL	0.9288	0.8690

Table 11
Comparison of different classifiers and proposed method in Manchester accidents using NMI scores.

Models	Selected Features	Complete Features
k-NN	0.0019	0.0058
SVM	0.0002	0.0011
BG	0.0007	0.0057
RF	0.0006	0.0042
AdaBoost	0	0
ANN	0.0002	0.0020
OvRCL	0.9948	0.7122

classifiers were done based on previous literature where these methods have been shown to perform well when considering a single classifier. Hence, we see that OvRCL is a better approach than simply using a single classifier for all accident datasets used in this study.

6.2. Feature analysis

In this section, we analyze the selected features by using Shapley values (SHAP). (Lundberg & Lee, 2017) proposed to use the Shapley value to interpret the output of the model. Fig. 11 illustrates a SHAP summary plot that represents the range and distribution of the selected features to injury severity level. SHAP plot is an agnostic framework that orders features based on their influence to detect crashes *since it* qualitatively represents the relation of attributes to the output class (i.e., crash severity). The scatter here represents the different Shapley values of the features, and each point is colored by the value of the feature from low (blue) to high (red). The density of points demonstrates the distribution in the dataset. For instance, the type of vehicle has a greater negative value which means that changing the type of vehicle would have a greater influence on crashes with slight injuries. On the other hand, 1st road class has the highest influence on the seriousness of the crash. These findings are reasonable because the type of road increases the probability of serious crashes.

The SHAP feature dependency plot is primarily utilized to indicate the distribution and variation of Shapley values with features. Shap dependence plots are used to quantify how some variables might affect the output of the ML model. To better interpret the data, we analyze the SHAP feature dependency plot by employing the top 3 features with relatively great importance. Fig. 12(a) shows the effect of two contributing factors, the number of vehicles and Vehicle type. It can be seen that the number of crashes involving cars are higher as compared to other vehicle types. Fig. 12(b) illustrates the effect of two more contributing factors – the casualty class, and type of vehicle. It can be seen that the SHAP values of the driver/rider are positive. This shows that drivers or riders are more likely to suffer severe injuries. In other words, when

Table 12
Sensitivity and Specificity values on the different models in feature-selected bicycle crashes.

	Classes	*k*-NN	SVM	RF	BG	AdaBoost	ANN	OvRCL
Sensitivity	*Fatal*	0	0	0	0	0	0	0
	Serious	0.0896	0.0448	0	0	0.0079	0.0119	1
	Slight	0.9364	0.9939	1	1	0.9925	0.9835	0.9939
Specificity	*Fatal*	0.9824	1	1	1	0.9981	0.9987	1
	Serious	0.9577	0.9940	1	1	0.9925	0.9828	0.9909
	Slight	0.0882	0.0441	0	0	0.0195	0.0234	1

Table 13
Sensitivity and Specificity values on the different models in feature-selected Manchester accidents.

	Classes	*k*-NN	SVM	RF	BG	AdaBoost	ANN	OvRCL
Sensitivity	*Fatal*	0.0082	0.0328	0	0	0	0	0.9918
	Serious	0.0590	0.1616	0.0035	0.0028	0.0003	0.0078	1
	Slight	0.9424	0.7837	0.9976	0.9992	0.9999	0.9957	0.9997
Specificity	*Fatal*	0.9913	0.9497	0.9998	0.9998	1	0.9999	1
	Serious	0.9491	0.8332	0.9978	0.9982	0.9999	0.9955	0.9996
	Slight	0.0764	0.2149	0.0039	0.0039	0.0003	0.0086	1

Table 14
Comparison of previous works with the OvRCL algorithm.

Method	dataset	Accuracy	Comments
k-Nearest Neighbors (Cover & Hart, 1967)	Leeds dataset	82.01%	The best result using *k*-NN (k = 3,5,7,9) with and without feature selection.
SVM (Cortes & Vapnik, 1995)	Leeds dataset	81.43%	The best result using SVM (RBF, Poly, Gaussian kernels), OvR, and OvO, with and without feature selection.
BG (Breiman, 1996)	Leeds dataset	81.54%	The best result using Bagging (10, 15, …, 50) with and without feature selection.
RF (Breiman, 2001)	Leeds dataset	81.59%	The best result using RF (10, 15, …, 50) with and without feature selection.
(Kabeer, 2016)	Leeds dataset	78.03%	Implemented naïve Bayes, decision trees, and ensemble technique.
(Kannojiya et al., 2020)	Leeds dataset	87.88%	Compared various ML models in which logistic regression has the highest accuracy.
(Chandra et al., 2019)	UK dataset	86–86.5%	RF shows better performance with and without feature selection in comparison with the logistic regression.
(Ma et al., 2021)	UK dataset	75–80%	Computed on 4 spatial classes with only injury classes, i.e., serious and non-serious, using Auto-encoders and Deep Neural Networks.
(Sattar et al., 2022)	UK dataset	73–75%	Compared various DL models in which Vanilla-MLP has the fastest training time whereas MLP with embedded layers has the highest accuracy while predicting binary classes (i.e. serious and non-serious).
OvRCL (proposed)	**Leeds dataset**	**97.64%**	**Applying discriminative feature selection along with class-wise OvR with ensemble classifier.**

Table 15
Comparison of OvRCL using different classifiers with ensemble and individual models.

Models	**Feature selected Leeds dataset**		**Feature selected bicycle crashes**		**Feature selected Manchester accidents**	
	Acc.	**NMI**	**Acc.**	**NMI**	**Acc.**	**NMI**
ANN	81.54	0.026	82.16	0.0028	82.91	0.0002
AdaBoost	81.27	0.011	82.91	0.0163	83.07	0
KNN	80.85	0.024	79.15	0.0161	79.21	0.0019
Bagging	81.59	0.043	83.91	0.0046	82.96	0.0007
OvRCL$_{(ANN,BG,AB,KNN)}$	**97.54**	**0.8453**	**94.47**	**0.6702**	**93.87**	**0.6636**
OvRCL$_{(SVM,KNN,RF,BG)}$	**97.64**	**0.931**	**99.25**	**0.9288**	**99.97**	**0.9948**

crashes occur, there would be a high severity of the driver as compared to the other casualty classes. However, negative SHAP values are seen when the casualty class is pedestrian which means that the injuries to the pedestrian are less severe.

6.3. Advantages and limitations of the proposed model

The key advantage of the proposed model is its ability to identify crash severity in a multi-class problem even in the presence of highly imbalanced data. While several machine learning algorithms have been used in the past, they are not able to differentiate between hard-to-classify classes, such as between *slight* and *serious* severity and between *fatal* and *serious* severity. The proposed model employs a feature selection technique that reduces the feature set, only retaining the discriminatory features. This has the added advantage of giving feature importance ranking which can be used to select pertinent features for the crash severity modeling. The proposed framework predicted indistinguishable classes in the imbalanced dataset by implementing class-wise OvR along with the ensemble of binary learning models for classification.

The proposed method also comes with some overhead in that multiple models need to be trained on each class prediction. Therefore, the model will be similar to the OvR model and the overall complexity will be similar to the algorithm requiring the highest complexity among the ensemble models. The good thing is that the models can be executed in parallel as each prediction is independent of the previous class (section 3.4), thereby, requiring little overhead time as the consensus building is a simple voting scheme.

6.4. Policy implications

The proposed algorithm developed in this paper has been evaluated on a real dataset. Firstly, using the proposed method provides better performance, in terms of evaluation metrics, than existing models in the literature. Secondly, the analysis of the features used, and their rankings can also provide helpful insight to safety planners and policymakers.

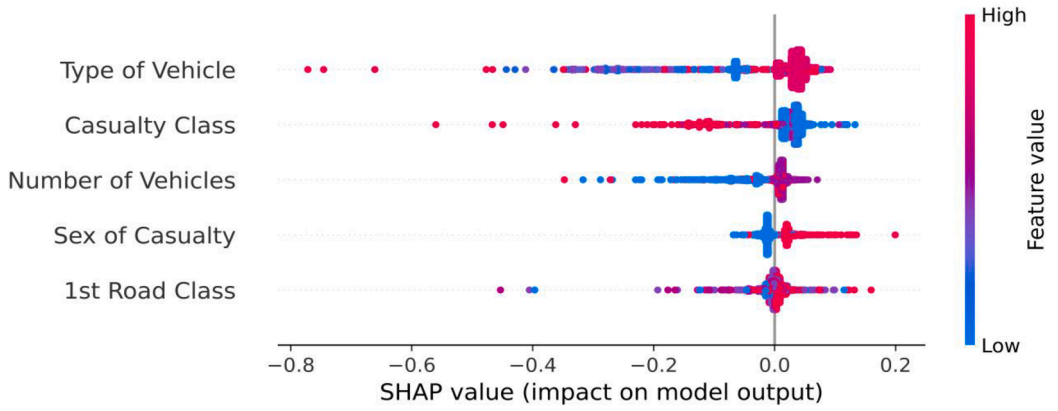
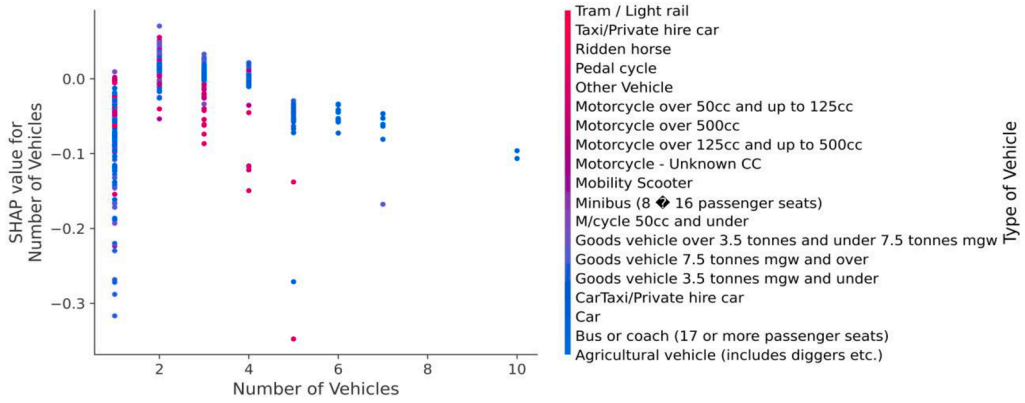


Fig. 11. SHAP summary plot.



(a)

(b)

Fig. 12. SHAP Dependence plot.

SHAPley values are also utilized for the analysis of selected features regarding the crash severity.

Based on the above analysis, the severity of a casualty in an accident involving a single vehicle would be minimal, and as the number of vehicles involved increased, the severity level also increased. This is in line with earlier research (Ma et al., 2021) that found that the number of vehicles increases the likelihood of accidents and the severity of injuries. Similarly, the chances of a pedestrian getting hit by a car are more when compared to the other casualty class. This is consistent with the previous study that pedestrians are hit more by automobiles than by any other vehicle (Ding et al., 2018).

Although the work in this paper follows previously reported experiential and statistical analysis, it also goes further to provide new insights. For instance, it might be conceivable to convey such data to emergency services and medical facilities using feature ranking and a reduced set as a result of feature selection. This will make it easier for hospitals to assess the crash's severity swiftly and set up the appropriate medical care for the injured person. Such information can easily be gathered and transferred using sensors embedded in vehicles or smart apps on mobile phones (for pedestrians). Already, such apps are being developed by insurance companies for accident detection and response, and by manufacturers of medical devices for fall detection.

The results from the model can also be helpful for policymakers. Safety planners may use this study to assess the severity of a crash given its circumstantial parameters. As an example, since the model predicts that a high number of pedestrians are involved in crashes (Fig. 12b), one can suggest making adequate facilities for pedestrians. To avoid an accident involving pedestrians, planners can recommend installing safety guardrails and pedestrian crossovers near driveways with high private car ownership. Moreover, it can also be seen (Fig. 12a) that the crashes of Buses or coaches with 17 or more seats and the crashes of cars are more frequent, among which the crashes of single vehicles are more common. Overall, cars have the highest frequency of crashes regardless of the severity type. Therefore, private car owners should be the subject of focus for reducing severe and fatal accidents, particularly when it comes to pedestrian safety. However, more information is needed to assess the causes that lead to this considerable number of single-vehicle crashes.

7. Conclusion

This research investigates the performance of different classification methods for crash severity prediction. A new method for crash severity prediction is proposed which uses a combination of selecting discriminatory features, class-wise OvR, and ensemble-based consensus learning. In the past, traditional machine learning algorithms have been employed for predicting crash severity but with limited success. Contrarily, our proposed method can produce better results for crash severity prediction. We attribute this result to selecting only those features that help in discriminating among the different types of severity rather than those having a simple correlation with the crash severity. Moreover, to help distinguish between the close and imbalanced classes, the ensemble technique embedded with OvR classification helps to predict severity levels with fewer instances. The binary classification in the OvR strategy helps in predicting instances of a single class particularly when the features have been selected to help identify one class from the others. The overall result is the higher performance with low complexity since we do not employ any deep networks or any extra steps of data clustering.

This study also analyzes the features that are mostly correlated to the severity of the crash. This can also help in further focused research to help avoid crashes that may lead to severe injuries or fatalities. In the future, we plan to extend this work and test the proposed method on other detailed datasets to analyze the cause of crashes, for instance by associating the selected features to the time and location of the accident so that the number of crashes resulting in severe injuries can be reduced.

8. Declaration

This work was done partly while both authors were with the Ghulam Ishaq Khan Institute of Engineering Science and Technology. The first author has since joined the School of Computer Science at the University of Birmingham while the second author is currently at the Institut Galilée, Université Sorbonne Paris Nord. Substantial work and revision were done while the authors were at their current positions.

CRediT authorship contribution statement

Syed Fawad Hussain: Conceptualization, Formal analysis, Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Muhammad Mansoor Ashraf:** Data curation, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Links to dataset are given in the manuscript

References

Abdel-Aty, M. A., & Abdelwahab, H. T. (2004). Predicting injury severity levels in traffic crashes: A modeling comparison. *Journal of Transportation Engineering, 130*(2), 204–210. https://doi.org/10.1061/(ASCE)0733-947X(2004)130:2(204)

Amini, M., Bagheri, A., & Delen, D. (2022). Discovering injury severity risk factors in automobile crashes: A hybrid explainable AI framework for decision support. *Reliability Engineering and System Safety, 226*(June), 108720. https://doi.org/10.1016/j.ress.2022.108720

Ashraf, M. M., Waqas, M., Abbas, G., Baker, T., Abbas, Z. H., & Alasmary, H. (2022). FedDP: A Privacy-Protecting Theft Detection Scheme in Smart Grids Using Federated Learning. *Energies 2022, Vol. 15, Page 6241*, *15*(17), 6241. https://doi.org/10.3390/EN15176241.

Bano, S., & Hussain, S. F. (2021). Prediction of Covid-19 and post Covid-19 patients with reduced feature extraction using Machine Learning Techniques. In *Proceedings - 2021 International Conference on Frontiers of Information Technology, FIT 2021*, 37–42. https://doi.org/10.1109/FIT53504.2021.00017.

Bauer, E., & Kohavi, R. (1999). Empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning, 36*(1), 105–139. https://doi.org/10.1023/a:1007515423169

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140. https://doi.org/10.1007/bf00058655

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Çelik, A. K., & Oktay, E. (2014). A multinomial logit analysis of risk factors influencing road traffic injury severities in the Erzurum and Kars Provinces of Turkey. *Accident Analysis and Prevention, 72*, 66–77. https://doi.org/10.1016/j.aap.2014.06.010

Chandra, S., Kaur, P., Sharma, H., Varshney, V., & Sharma, M. (2019). In-Database Analysis of Road Safety and Prediction of Accident Severity. *International Conference on Information Management and Machine Intelligence (ICIMMI), 277–283*. https://doi.org/10.1007/978-981-15-4936-6_31

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*, 273–297.

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

Delen, D., Sharda, R., & Bessonov, M. (2006). Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. *Accident Analysis and Prevention, 38*(3), 434–444. https://doi.org/10.1016/j.aap.2005.06.024

Ding, C., Chen, P., & Jiao, J. (2018). Non-linear effects of the built environment on automobile-involved pedestrian crash frequency: A machine learning approach. *Accident Analysis and Prevention, 112*(August 2017), 116–126. https://doi.org/10.1016/j.aap.2017.12.026

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science, 14*(2), 241–258. https://doi.org/10.1007/S11704-019-8208-Z/METRICS

Fan, W. (David), Gong, L., Washing, E. M., Yu, M., & Haile, E. (2016). Identifying and Quantifying Factors Affecting Vehicle Crash Severity at Highway-Rail Grade Crossings: Models and Their Comparison. *Transportation Research Board 95th Annual Meeting*.

Fountas, G., Fonzone, A., Gharavi, N., & Rye, T. (2020). The joint effect of weather and lighting conditions on injury severities of single-vehicle accidents. *Analytic Methods in Accident Research, 27*, 100124. https://doi.org/10.1016/j.amar.2020.100124

Haleem, K., Alluri, P., & Gan, A. (2015). Analyzing pedestrian crash injury severity at signalized and non-signalized locations. *Accident Analysis and Prevention, 81*, 14–23. https://doi.org/10.1016/j.aap.2015.04.025

Hameed, N., Shabut, A. M., Ghosh, M. K., & Hossain, M. A. (2020). Multi-class multi-level classification algorithm for skin lesions classification using machine learning techniques. *Expert Systems with Applications, 141*, 112961. https://doi.org/10.1016/J.ESWA.2019.112961

Hou, Q., Huo, X., Leng, J., & Mannering, F. (2022). A note on out-of-sample prediction, marginal effects computations, and temporal testing with random parameters crash-injury severity models. *Analytic Methods in Accident Research, 33*, 100191. https://doi.org/10.1016/j.amar.2021.100191

Hussain, S. F. (2019). A novel robust kernel for classifying high-dimensional data using Support Vector Machines. *Expert Systems with Applications, 131*, 116–131. https://doi.org/10.1016/J.ESWA.2019.04.037

Hussain, S. F. (2011). Bi-clustering gene expression data using co-similarity. *International Conference on Advanced Data Mining and Applications, 7120 LNAI(PART 1)*, 190–200. https://doi.org/10.1007/978-3-642-25853-4_15/COVER

Hussain, S. F., Babar, H.-Z.-U.-D., Khalil, A., Jillani, R. M., Hanif, M., & Khurshid, K. (2020). A fast non-redundant feature selection technique for text data. *IEEE Access, 8*, 181763–181781. https://doi.org/10.1109/access.2020.3028469

Hussain, S. F., Khan, K., & Jillani, R. (2022). Weighted multi-view co-clustering (WMVCC) for sparse data. *Applied Intelligence, 52*(1), 398–416. https://doi.org/10.1007/S10489-021-02405-3/TABLES/3

Hussain, S. F., & Qaisar, S. M. (2022). Epileptic seizure classification using level-crossing EEG sampling and ensemble of sub-problems classifier. *Expert Systems with Applications, 191*, 116356. https://doi.org/10.1016/J.ESWA.2021.116356

Hussain, S. F., Shahzadi, F., & Munir, B. (2022). Constrained class-wise feature selection (CCFS). *International Journal of Machine Learning and Cybernetics, 13*(10), 3211–3224. https://doi.org/10.1007/S13042-022-01589-5/METRICS

Ijaz, M., lan, L., Zahid, M., & Jamal, A. (2021). A comparative study of machine learning classifiers for injury severity prediction of crashes involving three-wheeled motorized rickshaw. *Accident Analysis and Prevention*, *154*(March), 106094. https://doi.org/10.1016/j.aap.2021.106094.

Iranitalab, A., & Khattak, A. (2017). Comparison of four statistical and machine learning methods for crash severity prediction. *Accident Analysis and Prevention, 108* (February), 27–36. https://doi.org/10.1016/j.aap.2017.08.008

Kabeer, S. I. (2016). *Analysis of Road accident in Leeds Syed*. National College of Ireland.

Kannojiya, A. K., Maurya, R., & Rajitha, B. (2020). Survey on soft computing methods for accident condition and severity predictions. *Advances in Intelligent Systems and Computing, 1039*, 584–591. https://doi.org/10.1007/978-3-030-30465-2_65

Katanalp, B. Y., & Eren, E. (2020). The novel approaches to classify cyclist accident injury-severity: Hybrid fuzzy decision mechanisms. *Accident Analysis and Prevention, 144*(March), 105590. https://doi.org/10.1016/j.aap.2020.105590

Kim, J. K., Ulfarssom, G. F., Shankar, V. N., & Mannering, F. L. (2010). A note on modeling pedestrian-injury severity in motor-vehicle crashes with the mixed logit model. *Accident Analysis and Prevention, 42*(6), 1751–1758. https://doi.org/10.1016/j.aap.2010.04.016

Lee, J., Chae, J., Yoon, T., & Yang, H. (2018). Traffic accident severity analysis with rain-related factors using structural equation modeling–A case study of Seoul City. *Accident Analysis & Prevention, 112*, 1–10.

Li, Q., Song, Y., Zhang, J., & Sheng, V. S. (2020). Multiclass imbalanced learning with one-versus-one decomposition and spectral clustering. *Expert Systems with Applications, 147*, 113152. https://doi.org/10.1016/J.ESWA.2019.113152

Li, Z., Liu, P., Wang, W., & Xu, C. (2012). Using support vector machine models for crash injury severity analysis. *Accident Analysis and Prevention, 45*, 478–486. https://doi.org/10.1016/j.aap.2011.08.016

Liu, J., Teng, S., Fei, L., Zhang, W., Fang, X., Zhang, Z., & Wu, N. (2021). A novel consensus learning approach to incomplete multi-view clustering. *Pattern Recognition, 115*. https://doi.org/10.1016/j.patcog.2021.107890

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *2017-December*, 4766–4775. https://arxiv.org/abs/1705.07874v2.

Lv, Y., Tang, S., & Zhao, H. (2009). Real-time highway traffic accident prediction based on the k-nearest neighbor method. In *2009 International Conference on Measuring Technology and Mechatronics Automation, ICMTMA 2009*, *3*, 547–550. https://doi.org/10.1109/ICMTMA.2009.657.

Ma, Z., Mei, G., & Cuomo, S. (2021). An analytic framework using deep learning for prediction of traffic accident injury severity based on contributing factors. *Accident Analysis and Prevention, 160*(July), 106322. https://doi.org/10.1016/j.aap.2021.106322

Pakistan Statistical Year Book 2020 | Pakistan Bureau of Statistics. (n.d.). Retrieved December 16, 2022, from https://www.pbs.gov.pk/publication/pakistan-statistical-year-book-2020.

Qaisar, S. M., & Hussain, S. F. (2020). Arrhythmia Diagnosis by Using Level-Crossing ECG Sampling and Sub-Bands Features Extraction for Mobile Healthcare. *Sensors 2020, Vol. 20, Page 2252*, *20*(8), 2252. https://doi.org/10.3390/S20082252.

Qaisar, S. M., & Hussain, S. F. (2021). Effective epileptic seizure detection by using level-crossing EEG sampling sub-bands statistical features selection and machine learning for mobile healthcare. *Computer Methods and Programs in Biomedicine, 203*. https://doi.org/10.1016/J.CMPB.2021.106034

Rahim, M. A., & Hassan, H. M. (2021). A deep learning based traffic crash severity prediction framework. *Accident Analysis and Prevention, 154*(March), 106090. https://doi.org/10.1016/j.aap.2021.106090

Ramírez, J., Górriz, J. M., Ortiz, A., Martínez-Murcia, F. J., Segovia, F., Salas-Gonzalez, D., … Puntonet, C. G. (2018). Ensemble of random forests One vs. Rest classifiers for MCI and AD prediction using ANOVA cortical and subcortical feature selection and partial least squares. *Journal of Neuroscience Methods, 302*, 47–57. https://doi.org/10.1016/j.jneumeth.2017.12.005

Sattar, K., Chikh Oughali, F., Assi, K., Ratrout, N., Jamal, A., & Masiur Rahman, S. (2022). Transparent deep machine learning framework for predicting traffic crash severity. *Neural Computing and Applications, 2*. https://doi.org/10.1007/s00521-022-07769-2

Si, T., Bagchi, J., & Miranda, P. B. C. (2022). Artificial Neural Network training using metaheuristics for medical data classification: An experimental study. *Expert Systems with Applications, 193*, 116423. https://doi.org/10.1016/J.ESWA.2021.116423

Tang, C., Chen, J., Liu, X., Li, M., Wang, P., Wang, M., & Lu, P. (2018). Consensus learning guided multi-view unsupervised feature selection. *Knowledge-Based Systems*, *160*(December 2017), 49–60. https://doi.org/10.1016/j.knosys.2018.06.016.

Xie, Y., Zhang, Y., & Liang, F. (2009). Crash injury severity analysis using Bayesian ordered probit models. *Journal of Transportation Engineering, 135*(1), 18–25. https://doi.org/10.1061/(ASCE)0733-947X(2009)135:1(18)

Xu, J. (2011). An extended one-versus-rest support vector machine for multi-label classification. *Neurocomputing, 74*(17), 3114–3124. https://doi.org/10.1016/j.neucom.2011.04.024

Yan, X., He, J., Zhang, C., Liu, Z., Qiao, B., & Zhang, H. (2021). Single-vehicle crash severity outcome prediction and determinant extraction using tree-based and other non-parametric models. *Accident Analysis and Prevention, 153*(January), 106034. https://doi.org/10.1016/j.aap.2021.106034

Yang, Y., & Pedersen, J. O. (1997). A Comparative Study on Feature Selection in Text Categorization. *Int. Conf. Mach. Learn. (ICML)*, 412–420.

Yang, Z., Zhang, W., & Feng, J. (2022). Predicting multiple types of traffic accident severity with explanations: A multi-task deep learning framework. *Safety Science*, *146*(September 2021), 105522. https://doi.org/10.1016/j.ssci.2021.105522.

Yasmin, S., & Eluru, N. (2013). Evaluating alternate discrete outcome frameworks for modeling crash injury severity. *Accident Analysis and Prevention, 59*, 506–521. https://doi.org/10.1016/j.aap.2013.06.040

Ye, F., & Lord, D. (2014). Comparing three commonly used crash severity models on sample size requirements: Multinomial logit, ordered probit and mixed logit models. *Analytic Methods in Accident Research, 1*, 72–85. https://doi.org/10.1016/J.AMAR.2013.03.001

Zelenkov, Y., & Volodarskiy, N. (2021). Bankruptcy prediction on the base of the unbalanced data using multi-objective selection of classifiers. *Expert Systems with Applications, 185*, 115559. https://doi.org/10.1016/J.ESWA.2021.115559

Zong, F., Xu, H., & Zhang, H. (2013). Prediction for traffic accident severity: Comparing the bayesian network and regression models1. *Mathematical Problems in Engineering, 2013*. https://doi.org/10.1155/2013/475194