Traffic & Transportation

Impact of Data Balancing and Feature Engineering on Accident Severity Models

Fayez ALANAZI[1], Aminu SULEIMAN[2]

Original Scientific Paper
Submitted: 27 Aug 2024
Accepted: 13 Dec 2024

[1] fkalanazi@ju.edu.sa, Jouf University, College of Engineering, Civil Engineering Department
[2] asuleiman.civ@buk.edu.ng, Bayero University Kano, Faculty of Engineering, Department of Civil Engineering

Publisher:
Faculty of Transport and Traffic Sciences, University of Zagreb

ABSTRACT
This study investigates the impacts of feature engineering techniques, including Clustering, Target Encoding and Anomaly Detection, in conjunction with data balancing methods, on the efficacy of machine learning models for predicting road accident severity. Automated Machine Learning (AutoML), Distributed Random Forest (DRF), Boosted Regression Trees (BRT) and Deep Learning models were evaluated on datasets that were balanced using the SMOTE (Synthetic Minority Over-Sampling Technique) and ADASYN (Adaptive Synthetic Sampling) techniques. Evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Log Loss, Area under the Curve (AUC), and Area under the Precision-Recall Curve (AUCPR) are employed. Results reveal that the AutoML consistently outperforms other models, achieving an 85% accuracy in predicting fatal accidents and 94% accuracy in predicting injuries. Deep Learning excels in injury accident prediction, with a 95% accuracy, but faces challenges with fatalities, achieving a 60% accuracy. The study underscores the critical role of feature engineering techniques and data balancing methods in enhancing predictive accuracy for accident severity classification. Specifically, the incorporation of Clustering, Target Encoding and Anomaly Detection techniques alongside SMOTE and ADASYN balancing methods significantly improves the model performance. Further refinement and validation are crucial for optimising model performance in real-world traffic safety management applications.

KEYWORDS
accident severity; traffic safety; machine learning; feature engineering; data balancing; AutoML; ADASYN.

1. INTRODUCTION

Road traffic injuries (RTIs) pose a significant health issue globally, causing 1.35 million deaths annually and imposing a substantial burden on healthcare systems and economies worldwide with estimated costs of US$1.8 trillion from 2015 to 2030 [1–3]. Road traffic injuries (RTIs) are a major challenge for low and middle-income countries (LMICs), which now bear the brunt of this global issue. Despite having far fewer vehicles, these countries account for a staggering 90% of worldwide road deaths [3]. The Gulf Cooperation Council (GCC) countries, including Saudi Arabia, face road safety challenges, exhibiting higher accident rates compared to western nations like USA and the UK [4–9]. In response, the Saudi government has committed to substantial infrastructure investments, exemplified by the SR500 billion investment goal by 2030, aimed at bolstering economic growth through enhanced transportation networks [10]. The development of extensive highways and road infrastructure underscores Saudi Arabia's dedication to fostering economic prosperity and social connectivity.

The Eastern Province of Saudi Arabia, vital to the nation's transportation and logistics sector, boasts an extensive road network crucial for economic development [11, 12]. However, this development coincides with increased road accidents of varying severity levels, raising concerns about road safety and its economic and

social ramifications. The economic toll of road traffic injuries (RTIs) in Saudi Arabia is substantial, equivalent to 4.3% of the country's GDP, with treatment costs surpassing average earnings [13, 14]. Poor driving standards, characterised by reckless behaviours such as mobile phone usage, speeding and red light violations, contribute to the high accident rates [14–16]. Inadequate enforcement of traffic regulations exacerbates the situation, necessitating a shift from reactive to proactive road safety approaches. Traditional reactive approaches to road safety, focused on post-accident interventions, have limitations in addressing the escalating rate of road traffic accidents (RTAs) [17, 18]. Proactive prevention strategies are imperative to mitigate accidents and their repercussions effectively.

This study proposes a data-driven approach, leveraging machine learning techniques, to predict road accident severity pre-emptively. The primary objective of this study is to evaluate the impacts of data balancing and feature engineering techniques on the performance of machine learning models in predicting road accident severity. Specifically, the study aims to investigate the effectiveness of the Synthetic Minority Over-Sampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN) in addressing class imbalance within the accident dataset, determining which balancing technique yields superior results for different machine learning models. Additionally, it explores advanced feature engineering methods, such as clustering, target encoding, and anomaly detection, to enhance model accuracy and identify the most effective features for improving predictive performance. The study also compares the predictive performance of various machine learning models, including AutoML, Deep Learning, Distributed Random Forest (DRF) and Gradient Boosting Machine (GBM), using multiple evaluation metrics. The core idea is to analyse past accident data, incorporating data balancing and feature engineering, to build models that predict potential accident severity. These models can then guide targeted interventions and policy decisions for accident prevention and mitigating their severity. This would empower authorities to strategically allocate resources and prioritise safety measures in high-risk areas. Furthermore, by integrating real-time data with the models and continuously updating them, proactive measures can be taken to reduce both the likelihood and impact of accidents.

The subsequent sections of this paper delineate the methodology and findings of the study. Section II reviews pertinent literature on road safety, focusing on the Eastern Province of Saudi Arabia, and critiques traditional reactive approaches. Section III outlines the data collection and pre-processing methods, including sources and cleaning techniques. Section IV elucidates the data balancing techniques utilised to mitigate class imbalance in the dataset. Section V expounds on the feature engineering methods employed to augment the predictive prowess of machine learning models. Section VI introduces the machine learning models utilised for accident severity prediction, encompassing deep learning, distributed Random Forest, GBM and AutoML. Section VII evaluates the performance of these models and furnishes actionable insights and recommendations for road safety enhancement. Finally, Section VIII concludes the paper, discussing its implications for future research and road safety strategies in the Eastern Province and Saudi Arabia.

2. LITERATURE REVIEW

With a more powerful prediction tool, law enforcement and transportation authorities may take preventative action to increase road safety and lower the likelihood of serious accidents. As technology advances, the development of more complex machine learning algorithms and artificial intelligence solutions is anticipated to significantly improve transportation safety [19]. AI-powered vehicle inspection systems, intelligent traffic management systems, assisted driving technologies, anomaly and intrusion detection, and crash prediction models are just a few examples of how AI is being leveraged to enhance safety across different modes of transportation [20, 21]. Machine learning algorithms have become instrumental in enhancing transportation safety by enabling authorities to analyse vast amounts of traffic data to identify patterns and trends that may not be immediately apparent. Transportation authorities can now leverage machine learning algorithms, such as tree-based methods (e.g. decision trees and random forests), support vector machines (SVM) [22] and neural networks, to analyse large volumes of traffic safety data. These algorithms excel in classifying crash data, predicting crash severity and identifying risky driving behaviours based on historical and real-time data, enabling the detection of patterns and trends that may not be immediately apparent [23, 24]. These insights inform proactive and targeted interventions, such as implementing traffic signal adjustments, enhancing road signage and deploying targeted enforcement measures in high-risk areas. Research indicates that the placement of speed cameras based on thorough data analysis can lead to significant reductions in traffic collisions. For instance, Tilahun [25] highlights the effectiveness of automated speed cameras in decreasing injury crashes, emphasising the importance of using accident data to inform camera locations. Similarly, Kalambay and

Pulugurtha, [26] discuss how traffic speed patterns can be analysed to identify suitable sites for implementing variable speed limit signs, which may include speed cameras. Furthermore, Li, Zhang and Ren [27] provide evidence that the safety impacts of speed cameras are enhanced when their placement is guided by accident data, underscoring the necessity of a data-driven approach in traffic management. Furthermore, machine learning can facilitate the development of mobile applications that alert drivers to hazardous conditions or risky behaviours, ultimately creating a safer driving environment for all road users [28].

Road safety prediction is a critical area of research that has drawn substantial interests in recent years due to the potential of machine learning techniques to improve the accuracy of predictions and inform proactive interventions [29]. Several studies have shown significant advancements in using machine learning predictive modelling to address road safety concerns [30–35]. These studies have analysed various factors such as driver behaviour, road conditions and weather patterns, and have been able to develop robust predictive models that provide important insights into mitigating the risks of severe road accidents [36–38].

A machine learning model was developed recently by Christofa et al. [39] to predict high-risk crash locations based on road characteristics. The study identified design speed, pavement markings, signage and road condition as key factors in determining crash risk. These findings can inform targeted interventions to enhance road safety. In a comprehensive evaluation of the literature, Silva et al. [37] investigated three distinct methods for employing machine learning techniques to predict accidents, emphasising the application of neural networks as a promising crash prediction strategy and highlighting the advantages of machine learning models. Artificial neural networks offer an advantage by learning from data without requiring any assumptions. They can capture complex patterns in data more effectively than traditional statistical methods [40]. Similarly, a study by Zhang et al. [41] improved the effectiveness of speed camera placement using the generalised random forest. This method estimated heterogeneous treatment effects in traffic safety studies which provides authorities with a more comprehensive information.

Several researchers have used statistical techniques [42], reinforcement learning approaches [43], hybrid models [44] and deep learning models [45] to predict traffic accident severity. Although, statistical models are limited by their reliance on assumptions about data distribution and predefined relationships between variables [46], they offer significant advantages in terms of interpretability and the ability to reveal heterogeneity caused by unobserved factors [47, 48]. These strengths make statistical models valuable tools in understanding the underlying causes of crashes and can complement machine learning techniques in hybrid approaches for crash prediction [49–51].

In Saudi Arabia and surrounding regions, research studies on road safety prediction using machine learning have yielded promising results. For example, machine learning models for predicting accident causes and injury severity have been developed using various machine learning algorithms. These models have been applied to analyse crash data and identify key risk factors related to road accident [41, 52]. Aldhari et al. [53] explored accident severity prediction, demonstrating the effectiveness of machine learning for this purpose. However, most of these research studies have limitations which include limited scope, lack of real-time data, restricted exploration of deep learning techniques and insufficient explanation of black box models. For instance, studies often focus on specific regions or datasets, which may not be representative of broader trends [54]. Additionally, many models rely on historical data without integrating real-time information, which can hinder their predictive accuracy [55]. While some research has ventured into deep learning methods, the exploration remains limited compared to the potential of these techniques [53]. Finally, the complexity of machine learning models often leads to a lack of transparency, making it difficult for stakeholders to understand the decision-making processes involved [56].

Despite the potential of machine learning techniques, there are limitations to their application in road safety prediction. According to a study by Wang et al. [57], the choice of proper methodology determines the quality of research; for example, machine learning approaches need the right data analysis techniques in order to identify the causes of accidents in certain study regions or zones [58]. Additionally, the use of a single machine-learning algorithm may not be sufficient to achieve the intended outcomes, with multiple analytical techniques often combined to enhance the analysis of results. Other issues include data imbalance, feature engineering and model interpretability [57, 59, 60]. Recent studies have focused on improving road accident severity machine learning based models through data balancing and feature engineering (Fiorentini & Losa, 2020; Mohammad pour et al., 2023; Ogungbire & Pulugurtha, 2024; Sarkar et al., 2020). These research studies have shown encouraging outcomes in terms of prediction model optimization for precisely determining the risk variables linked to serious accidents.

The accuracy of machine learning based accident severity models is largely dependent on treatment of issue the imbalance of data [58, 65]. This means that there may be significantly more instances of minor accidents than severe or fatal ones, resulting in a skewed dataset [61, 66, 67]. For instance, it may be challenging for the machine learning model to predict the severity of future accidents in a dataset of road accidents when 90% of the incidents result in injuries and just 10% in fatalities [68]. This imbalance may cause projections to be skewed or erroneous, which would reduce the efficacy of preventive and focused efforts meant to increase road safety [61, 69]. However, by employing techniques like oversampling or undersampling to increase the representation of minority class instances, transportation departments can mitigate the impact of data imbalance, leading to more accurate predictions and targeted interventions for accident prevention [67, 70, 71].

Different approaches to data balancing in enhancing the efficacy of machine learning models exist, hence the need for selecting appropriate methods for crash severity prediction models. Some methods of data balancing include oversampling minority classes [72, 73] and undersampling majority classes [74–76] to create a more even distribution of data. These methods can aid in correcting dataset imbalances and enhance the ability of the machine learning algorithms to predict traffic accidents severity [77]. The strength of generating synthetic samples using techniques such as SMOTE [72] or ADASYN [78] is that it can correct dataset imbalances and enhance the ability of the machine learning algorithms to forecast the severity of traffic accidents. The drawback of these methods is that undersampling might cause significant information to be lost, while oversampling could result in overfitting [79]. Other balancing methods that can be considered for enhancing road safety prediction models include ensemble learning techniques like Random Forest [80], which can handle imbalanced datasets by aggregating the predictions of multiple classifiers. Additionally, cost-sensitive learning approaches [81] assign different costs to misclassifications based on class distribution, thereby placing more emphasis on minority class samples. Clustering-based methods such as Cluster-Based Over Sampling (CbOS) [82] can also be used to generate new samples by clustering similar datapoints together and oversampling from these clusters.

In addition to data balance, feature engineering is also essential for enhancing machine learning model performance [39, 60, 83]. Feature engineering involves extraction of pertinent features from raw data, such as temporal variables, road conditions, geometric attributes and other parameters that can be readily available in real time or historic, before the accident or immediately after an accident [84]. These engineered features are then utilised to train machine learning based accident severity prediction models. However, the selection and engineering of these features in road safety context can be challenging. To address this challenge, our paper proposes the use of advanced feature engineering techniques, including clustering algorithms [85], target encoding [86] and anomaly detection [87], to improve the capacity of our machine learning models for accident severity prediction. Ultimately, by combining road safety prediction models with data balancing and feature engineering techniques, we can create a more robust and accurate system for predicting and preventing accidents. By leveraging technology and data-driven approaches, we can move towards a future where road safety is maximised and accidents are minimised. This collective effort will not only save lives but also build a more connected and secure transportation infrastructure for everyone.

3. METHODOLOGY

3.1 Crash data for the modelling

The dataset for this study was acquired from the Saudi Arabian Oil Group, ARAMCO (Arabian-American Oil Company), from 2018 to 2022, covering the Eastern Province of Saudi Arabia (see *Figure 1*). It contains information on 9548 road accidents that occurred in Saudi Arabia during this period. The dataset includes various attributes related to each accident, such as the number of vehicles involved, the severity of the accident, the type of accident, the reason for the accident, the number of fatalities and injuries, the number of vehicles involved, the location of the accident, whether it occurred in or out of the city and the coordinates of the accident location. The severity of the accidents ranges from injury to fatal, with a total of 2527 fatalities and 7021 injuries recorded. The accidents were caused by various factors, including swerve accidents, overturn accidents, run-over accidents, stationary object accidents and utility pole accidents.

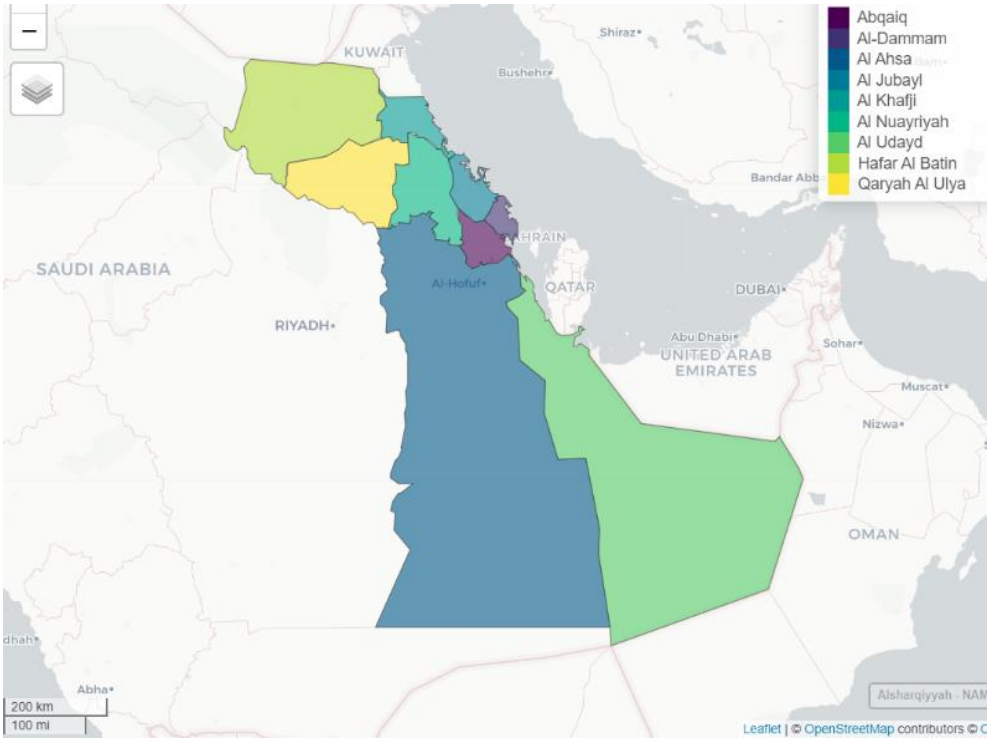

Figure 1 – Eastern Province of Saudi Arabia (Study area)

The study followed a structured research methodology as shown in *Figure 2*. The key stages involved in a data-driven project. It begins with data collection, where relevant data is gathered from ARAMCO. This is followed by data pre-processing, which involves cleaning and preparing the data for analysis. Next, the data balancing stage ensures that the dataset is representative and free from biases. Then, feature engineering is performed to create new variables that enhance the predictive power of the models. This is succeeded by feature selection, where the most relevant features are identified to improve model performance. The process then transitions to model selection and training, where different algorithms are evaluated to determine the best fit for the data and select the best model for the dataset. Finally, the methodology culminates in model evaluation, assessing the model's performance and generating insights and recommendations based on the findings.

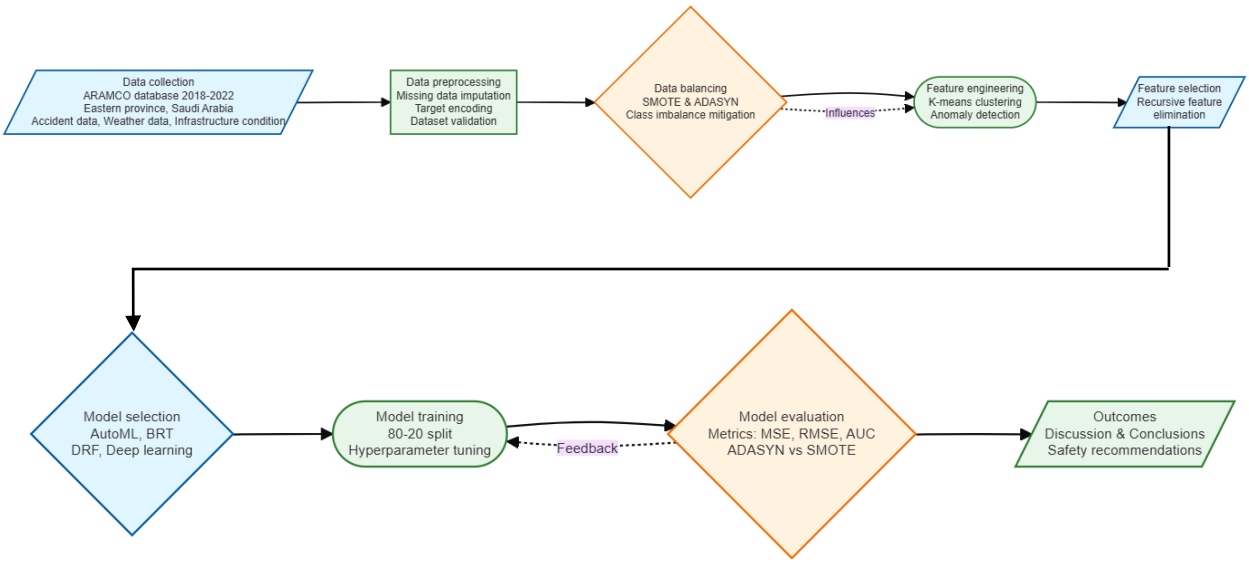

Figure 2 – Study framework

3.2 Data pre-processing

In the data pre-processing phase, various techniques were employed to ensure the quality and readiness of the dataset for machine learning analysis.

Data imputation

A comprehensive cleaning procedure that addressed missing values, discrepancies and outliers was carried out to guarantee the dataset's integrity and completeness. In particular, the missForest R package was used to restore 765 missing coordinates. This imputation technique was developed based on the random forest algorithm suggested by Stekhoven and Buhlmann [88]. One key advantage is its capacity to simultaneously handle multiple data types – numeric and categorical variables, which are commonly found in road safety datasets [88]. This flexibility sets it apart from methods like KNN-Impute that struggle with categorical variables. MissForest also excels in capturing the distribution of the data and can effectively capture complex interactions and non-linear relationships [89]. This is particularly important when modelling the intricate factors involved in road crashes. Comparative studies across various domains have shown that MissForest outperforms other common imputation methods, such as k-nearest neighbours (KNN) and multivariate imputation by chained equations (MICE), in terms of providing more accurate imputed values [90]. Additionally, MissForest provides integrated out-of-bag error estimates so that it is possible to evaluate the imputation quality without provision for test data. This is a useful feature since it enables the evaluation of imputation accuracy for every variable. MissForest is also computationally efficient and capable of handling high-dimensional data with many variables, which is a frequent feature of datasets related to road safety that include a large number of possible risk factors [88]. The availability of the missForest R package, which is freely available and easy to use, further enhances its accessibility for road safety researchers.

Using variables with missing values as the target variable and full observations as training data, the approach iteratively trains Random Forest models, predicting missing values for observations with incomplete data. This iterative procedure keeps going until certain convergence requirements are satisfied, including a cap on the number of iterations or a minimal shift in the imputed values in between rounds. The quality of the dataset was also preserved by identifying and addressing outliers and inconsistencies using data validation and visualization tools.

Target encoding for efficient categorical variable encoding based on accident severity

Categorical variables can be challenging to encode for machine learning models, as they often require one-hot encoding, which can lead to a high-dimensional feature space [91]. Target encoding is a versatile method for encoding categorical variables in machine learning modelling. It efficiently handles high cardinality variables, providing a more meaningful representation of the data [92]. It supports both continuous and binary target variables, making it suitable for various tasks. In order to provide a more comprehensive knowledge of the elements impacting the outcomes of traffic crashes, target encoding captures complicated interactions between categorical variables and the target variable. Model correctness and computing efficiency are improved over techniques such as ordinal encoding and one-hot encoding [93]. Its user-friendly R package, H2O, makes it a convenient choice for researchers [86].

In order to convert categorical characteristics in the crash dataset into a format that machine learning algorithms could use, they were encoded for this study. By capturing the link between category factors and the target variable, this encoding approach contributes to the prediction of accident severity. The mathematical expression for target encoding is shown in *Equation 1*:

$$Target\ encoding(ci) = \frac{\sum_{j=i}^{N} y_i I(xj=ci)}{\sum_{j=i}^{N} I(xj=ci)} \quad (1)$$

where target encoding(ci) represents the target-encoded value for category ci; y_j is the target variable (e.g. response variable) for the hj^{th} observation; x_j is the categorical variable for the h_j^{th} observation; (x_j=c_i) is an indicator function that equals 1 if x_j is equal to category c_i, and 0 otherwise; N is the total number of observations.

In this expression, the numerator calculates the sum of the target variable for all observations where the categorical variable *x* takes the value *ci*, while the denominator calculates the count of observations with $x=c_i$. Thus, the target-encoded value for category *ci* is the mean of the target variable for observations with $x=c_i$.

3.3 Data imbalance

Traffic accident data often exhibits class imbalance, meaning the distribution of accident severities (e.g. fatal, injury, minor) is uneven. Typically, there are far fewer severe accidents compared to minor ones. Machine learning models may have difficulties as a result of this imbalance since they often lean in favour of the majority class and do a poor job of anticipating the less common but possibly more serious, catastrophic incidents. In order to resolve class imbalance in the traffic accident dataset, this section describes the use of data balancing strategies using the UBL R package [94] to address class imbalance in the traffic accident dataset. In the UBL R package, the Synthetic Minority Over-Sampling Technique (SMOTE) [72] and the Adaptive Synthetic Sampling Method (ADASYN) [78] are used in the UBL R package to handle class imbalance by creating synthetic samples for the minority class. Comparing these techniques helps researchers determine the most effective approach for their specific dataset and problem domain. ADASYN, unlike SMOTE, focuses on densely populated regions, offering a more targeted strategy for addressing class imbalance [95]. This approach can potentially yield superior performance in scenarios where the minority class clusters in specific areas of the feature space as in the case of crash severity data. Empirical evidence shows that ADASYN can outperform SMOTE under certain experimental conditions, emphasising the importance of exploring multiple data balancing techniques when dealing with imbalanced datasets [96]. However, the efficacy of ADASYN may vary depending on the dataset's unique characteristics, such as class imbalance and data point distribution [78]. Consequently, it is necessary to test a variety of resampling strategies and evaluate each one's effectiveness on an individual basis in order to determine the best method for predicting accident severity using various machine learning algorithms.

SMOTE (Synthetic Minority Over-Sampling Technique)

In order to create synthetic samples for the minority class, SMOTE (Chawla et al. 2002) interpolates between samples that already belong to the minority class. Using a user-defined parameter k, SMOTE picks the k nearest neighbours from the minority class given a minority class sample Sample_i. The line segment connecting Sample_i and one of its closest neighbours, $\text{Sample}_{\text{neighbour}}$, is then used to construct a new synthetic sample. To do this, multiply $\text{Sample}_{\text{neighbour}} - \text{Sample}_i$ by a random number λ, which ranges from 0 to 1, and then add the result to Samplei. *Equation 2* can be used to represent the mathematical expression for SMOTE:

$$\text{Synthetic Sample} = \text{Sample\_i} + \lambda \times (\ 〚\text{Sample}〛\_(\ 〚\text{neighbour}〛\_\ ) - 〚\text{Sample}〛\_\text{i}\ ) \quad (2)$$

ADASYN (Adaptive Synthetic Sampling Method)

Based on the distribution of instances in the feature space, ADASYN modifies the density distribution of synthetic samples. It focuses more on creating artificial samples for hard-to-learn areas of the minority class [71]. In areas with more severe class imbalances, ADASYN gives minority samples a larger weight based on its computation of the density distribution of minority class samples. The synthetic samples are then generated in proportion to these densities, similar to SMOTE. The mathematical expression for ADASYN follows the same formula as in SMOTE, but the value of λ is adjusted based on the local density ratio as shown in *Equation 3*:

$$\lambda = \frac{\min_{\text{examples}} - \text{examples}_i}{\min_{\text{examples}}} \quad (3)$$

where $\min_{\text{examples}}$ is the number of examples in the minority class with the fewest instances; examples_i is the number of minority class examples in the neighborhood of $Sample_i$.

The artificial samples are created repeatedly in both SMOTE and ADASYN until the appropriate degree of class balance is attained. In this study, class balancing was performed using the SMOTE and ADASYN technique implemented in the UBL R package, with the class percentage (C.perc) parameter set to maintain the injury class at its original level (1) while oversampling the fatal class by 2.5 times its original size. This

approach ensured an improved representation of the minority class without enforcing a strict 1:1 balance, allowing for flexibility based on the dataset's characteristics and the goals of the analysis.

3.4 Feature Engineering

Feature engineering is an essential stage in the machine learning process that turns unstructured input into meaningful and instructive features, which has a substantial influence on model performance [60]. This process could play a vital role in enhancing the performance of the machine learning models for accident severity prediction. Feature engineering bridges the gap between raw data and the model's internal representation, enabling improvements in model interpretability, reduction in model complexity [83] and enhancement of model performance by crafting features that effectively represent relationships between variables and the target variable (accident severity) [97]. K-means clustering, target encoding and anomaly detection are chosen for feature engineering in accident severity due to their unique strengths and complementarity. K-means clustering helps identify similar accident groups based on various features [98], target encoding encodes categorical variables [99] and anomaly detection identifies outliers or unusual patterns in data [100]. By combining these techniques, a larger variety of patterns and correlations within the data are captured, resulting in a more complete and accurate collection of characteristics for forecasting accident severity. This method enhances the models' capacity for generalisation while lowering the danger of overfitting. By utilising a range of techniques in feature engineering, predictive models become more accurate and efficient [101]. Feature selection and transformation creates more efficient and interpretable models that provide valuable insights for accident prevention strategies [102].

Feature K-means clustering for spatial insights and accident pattern identification

K-means clustering is a well-known unsupervised learning method that groups together similar data points [85]. K-means clustering may be used to locate spatial patterns and accident clusters based on their geographic coordinates in relation to road safety prediction [103, 104]. By clustering accidents based on their spatial distribution, we can gain insights into high-risk areas and identify potential factors contributing to accidents in those areas. The K-means clustering technique in the H2O R package divides data into k groups according to their similarity. The mathematical expression for K-means clustering algorithm is as follows:

Given a dataset with n observations and p features represented as $X = \{x_1, x_2, \ldots, x_n\}$, where x_i is a p-dimensional feature vector, the aim of K-means clustering is to split the data into k clusters $(C_1, C_2, \ldots, C_k)$ so that the within-cluster sum of squares (WCSS) is minimised.

The objective function is the sum of the squared Euclidean distances between each observation (x_i) and the centroid (μ_j) of the cluster (C_j) to which it is allocated. The method iteratively minimises this function. It is possible to express the objective function mathematically as shown in *Equation 4*:

$$\text{minimise: } \sum_{j=1}^{k} \sum_{x_i \in c_i} \left\| x_i - \mu_j \right\|^2 \tag{4}$$

where k is the number of clusters; μ_j is the centroid of cluster C_j; $||\cdot||$ represents the Euclidean norm.

The K-means algorithm proceeds through the following steps:

1) **Initialisation**: Randomly initialise k centroids $\mu 1, \mu 2, \ldots, \mu k$.

2) **Assignment step**: Assign each observation xi to the cluster with the nearest centroid. This can be expressed as in *Equation 5:*

$$argminj \; || \; xi - \mu j \; ||^2 \tag{5}$$

3) **Update step**: Update the centroids μj of each cluster as the mean of the feature vectors of the observations assigned to that cluster as in *Equation 6*.

$$\text{Mathematically: } \mu_j = \frac{1}{|C_j|} \sum x_i \in C_j x_i \tag{6}$$

Iteration: Continue steps 2 and 3 until convergence, which happens when there is no more substantial change in the centroids or when the allotted number of iterations is achieved. The K-means method ultimately yields a collection of k clusters, each of which is represented by a centroid. Each observation is allocated to a cluster according to how close it is to one of the centroids [105]. In the H2O R package, the K-means clustering algorithm is efficiently implemented to handle large-scale datasets and provides options for parallelisation and distributed computing to accelerate the clustering process.

Anomaly detection

Anomaly detection, essential for identifying unusual or abnormal observations in a dataset, is a valuable technique in road safety prediction, aiding in the identification of outliers and the creation of features for enhanced high-severity accident prediction [87]. Implemented through various algorithms in the H2O R package, such as Isolation Forest [106], anomaly detection isolates anomalies by randomly selecting features and split values, recursively creating a binary tree structure. Anomalies, expected to have shorter paths in these trees, are identified by higher anomaly scores calculated based on the average path length from the root node to the terminal node in multiple trees. Isolation Forest provides an efficient and scalable approach to anomaly detection, particularly in high-dimensional datasets [107]. In this study, the H2O R package facilitated anomaly detection for identifying outliers and creating features to improve high-severity accident prediction.

Isolation trees: Create t isolation trees, with each one being built as follows:

1) Choose a portion of the data at random.
2) Choose a feature at random from the subset.
3) Choose a split value at random from the lowest and maximum values of the chosen feature.
4) Divide the data recursively using the chosen feature and split value until every data point is isolated in a separate leaf node.

Anomaly Score Calculation: For each data point x_i, compute the average path length $h(x_i)$ from the root node to the terminal node across all isolation trees. The anomaly score $s(x_i)$ for x_i is then calculated using *Equation 7*:

$$s(x_i) = 2^{\frac{-E(h(x_i))}{C(n)}} \tag{7}$$

where $E(h(xi))$ is the average path length of xi across all trees; $c(n)$ is the average path length of a failed search in a binary tree of *n* data points, given by *Equation 8*:

$$2H(n-1) - \frac{2(n-1)}{n} \tag{8}$$

where $H(n)$ is the harmonic number.

Anomaly detection: Anomalies are identified as data points with higher anomaly scores, exceeding a certain threshold.

3.5 Feature selection

In order to increase model performance, lower computational complexity and improve interpretability, feature selection is an essential stage in the machine learning process [37]. In the context of road safety prediction, where datasets often contain a multitude of features, feature selection becomes imperative to identify the key factors influencing accident severity accurately. The technique of selecting features for this study involves first eliminating variables from the dataset that would be challenging to find after an incident. The accident type, accident reason, total number of fatalities and total number of injuries are among the criteria that have been eliminated. To train and monitor the performance of the several machine learning models used for the study, new features produced from the clustering, anomaly detection and target encoding were then introduced one at a time. Using the four chosen machine learning methods, many models were created. For each algorithm, models were trained on either SMOTE or ADASYN balanced data containing the observed selected variables (Accident-ID, Vehicle-Count, Accident-Severity, City, In-Or-Out-City, X-Coordinate, Y-Coordinate, Year) and any or all of the derived features. The final models were selected based on their performance in predicting Accident-Severity.

3.6 Machine learning prediction models

The selection of machine learning models – Deep Learning, Distributed Random Forest (DRF), Gradient Boosting Machine (GBM), and AutoML – was carefully considered based on their unique strengths and suitability for the task. Deep Learning, chosen for its ability to capture complex patterns, is well-suited for uncovering intricate relationships [108]. DRF, recognised for scalability and interpretability [109], provides

insights into feature importance and decision-making processes. GBM, a robust ensemble method, was selected for its capability to handle heterogeneous data and complex interactions ([110–112]. AutoML, streamlining the model selection process efficiently, proves valuable when resources or time are limited [113]. This ensemble aims to maximise predictive accuracy and interpretability, offering a robust framework for deriving actionable insights and making informed decisions from the road accident dataset [114].

3.7 Model selection criteria

The selection of the best performing model among the trained machine learning models was guided by several criteria to ensure optimal performance and generalisation. Cross-validation techniques were employed to evaluate models across multiple data subsets, mitigating overfitting and providing robust performance estimates [115]. Models were compared based on task-specific metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Area under the Curve (AUC) or log loss [116–118]. A preference for simpler models over complex ones was adopted to improve interpretability and avoid overfitting, particularly crucial with limited training data or computational resources. This systematic evaluation aimed to select a model balancing accuracy, interpretability and computational practicality, crucial for effective road safety prediction.

Deep learning

In this study, deep learning, specifically utilising deep neural networks (DNNs) [119] was employed to capture intricate relationships within accident data for severity prediction. Deep neural networks (DNNs) are a powerful tool for tasks like image recognition and natural language processing. They can be implemented using the H2o R package and are skilled at learning intricate patterns from large datasets. The architecture includes an input layer for receiving pre-processed accident data, hidden layers for learning non-linear relationships, and an output layer predicting accident severity [120]. While DNNs offer high predictive power, they require computational resources and careful tuning to prevent overfitting [121]. Strengths of deep learning lie in its automatic feature extraction, scalability to large datasets and ability to handle high-dimensional data, making it particularly suitable for this task [122]. The training process involves forward and backward propagation with hyperparameters like layer depth, neuron count and learning rate determining model performance [123, 124].

Mathematical foundations: The foundation of deep learning is the use of many hidden layer artificial neural networks (ANNs), which allow the model to extract intricate patterns and representations from the input. The mathematical expression for a basic feedforward neural network can be expressed as in *Equation 9*:

$$\hat{y} = f(W2 \cdot f(W1 \cdot x + b1) + b2) \tag{9}$$

where x is the input vector; W_1 and W_2 are weight matrices for the hidden layers; b_1 and b_2 are bias vectors; $f(\cdot)$ represents the activation function (e.g. ReLU, sigmoid).
It assesses their accuracy and discusses the impact of feature engineering and data balancing techniques.

Distributed random forest (DRF)

In this study, Distributed Random Forest (DRF) from the H2O package was utilised for predicting road accident severity, offering scalability for large datasets and high-dimensional feature spaces [125]. DRF, an ensemble method of multiple decision trees, is particularly adept at handling non-linear data and providing feature importance rankings [80].

The mathematical expression for a decision tree in DRF is shown in *Equation 10:*

$$y = T(x), \tag{10}$$

where *y* is the output of the tree, *T* is the tree function and *x* is the input vector.

DRF's training process involves parallel growth of trees on random subsets of features and data points, mitigating overfitting. Its strengths include robustness, scalability and the ability to handle missing values, making it suitable for classification and regression tasks on potentially large traffic accident datasets.

Gradient Boosting Machine (GBM)

Gradient Boosting Machine (GBM) from the H2O package was used in this work to predict the severity of traffic accidents by utilising its ability to sequentially create an ensemble of decision trees. In order to capture complicated non-linear correlations in the data, GBM, an ensemble learning technique, iteratively adds new models, such as decision trees, to rectify errors produced by prior ones [126]. The mathematical expression for GBM involves a weighted sum of individual trees as in *Equation 11*:

$$\hat{y} = \sum_{i=1}^{N} \propto_i h_i(x), \tag{11}$$

where $\propto_i$ is the weight assigned to the i^{th} tree and $h_i(x)$ is the prediction of the i^{th} tree.

GBM's strengths include its suitability for tasks with mixed feature types and complex interactions, robustness to overfitting and interpretability of feature importance. The training process for GBM entails sequentially adding decision trees to minimise a differentiable loss function, with hyperparameters including the learning rate, tree depth, number of trees and regularisation parameters [111].

AutoML (Automated Machine Learning)

By using machine learning algorithms such as decision trees, gradient boosting, random forests, support vector machines and neural networks, the AutoML algorithm automates the process of selecting and tuning models [127]. AutoML empowers data scientists by providing efficient model selection and tuning processes, allowing them to focus on interpreting results and making informed decisions [128]. As AutoML evolves, it will become an indispensable tool for businesses to leverage data, drive innovation and make data-driven decisions [129]. The AutoML training process involves exploring a predefined search space of ML algorithms and hyperparameters, utilising techniques such as cross-validation and hyperparameter optimisation. AutoML's strengths lie in its efficiency, scalability and adaptability to different datasets and tasks, making it suitable for users with limited ML expertise who seek to quickly build high-performing models [130]. Using established search methods and assessment criteria, the training process involves comparing the performance of many models through cross-validation and identifying the top-performing model. In this work, machine learning models for forecasting the severity of traffic accidents were automatically selected and fine-tuned by using the AutoML from the R H2O package.

3.8 Machine learning model training process

During the training phase, models were built using the training set and evaluated on the validation set, with data split accordingly [131]. For the Deep Learning model, hyperparameters such as the number of hidden layers, neurons per layer, activation functions and optimisation techniques were adjusted [132]. Grid search optimisation was employed for hidden layers, neurons and L1 norms, and the best model was identified using the AUC from 5-fold cross-validation. For ensemble models like Distributed Random Forest (DRF) and Gradient Boosted Machines (GBM), hyperparameters such as the number of trees, maximum tree depth and other relevant parameters were selected to maximise performance [133]. In Boosted Regression Trees (BRT), iterative testing informed the selection of hyperparameters, including n-trees (1000), max_depth (5) and learn_rate (0.1), based on recommendations from the literature [134]. Similarly, Random Forest models employed n-trees (1000) and max_depth (5), balancing accuracy and computational efficiency. Additionally, the H2O AutoML framework streamlined the process by automatically tuning parameters, exploring diverse algorithms and selecting the top-ranked model based on performance metrics within a fixed runtime of 30 seconds. The H2O R package was instrumental in facilitating hyperparameter tuning, providing a user-friendly interface for iterative model optimisation. Multiple training iterations were conducted to refine hyperparameters and enhance model generalisation.

3.9 Model evaluation

Data splitting process

To ensure reliable model assessment when using the H2O R package, the process of dividing data into training and validation sets follows a standard protocol. To do this, the dataset must be divided into two subsets:

a training set and a validation set. The machine learning models are trained on the training set, and their performance on unseen data is assessed using the validation set. The H2O R package provides various functions to facilitate the data splitting process, such as h2o.splitFrame(), which allows users to split the dataset based on specified proportions and/or random sampling. This ensures that both the training and validation sets represent the underlying data distribution adequately, minimising bias in model evaluation. In this study the data was split into 80% training and 20% validation [135].

Performance metrics

When evaluating machine learning models trained using the H2O R package, several performance metrics can be utilised to assess their effectiveness in solving specific tasks. These metrics provide insights into the model's predictive accuracy, generalisation capability and ability to discriminate between classes [117,136]. Commonly used performance metrics include:

a. Mean Squared Error (MSE): Defined as the average of the squared differences between the predicted and actual values in regression tasks. Mathematically, MSE is expressed in *Equation 12*:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 \tag{12}$$

where y_i represents the actual target value, $\hat{y_i}$ represents the predicted value and n is the number of samples.

b. Root Mean Squared Error (RMSE): The square root of MSE, providing a measure of the standard deviation of the residuals. RMSE is calculated using *Equation 13*:

$$RMSE = \sqrt{MSE} \tag{13}$$

c. Log Loss: A measure of the accuracy of a classification model, calculated as the negative logarithm of the predicted probability of the true class. Lower log loss values indicate better performance.

Log Loss is estimated as: $\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log\left(\hat{y}_i\right) + (1 - y_i)\log\left(1 - \hat{y}_i\right)\right]$ (14)

where y_i represents the actual binary target (0 or 1) and $\hat{y}_i$ represents the predicted probability of the positive class.

d. Mean per Class Error: Calculates the average error rate across classes in a classification task, providing insights into class-specific performance.

$$\text{Mean Per Class Error} = \frac{1}{K}\sum_{j=1}^{K}\frac{PF+FN}{TP+FP+FN+YN} \tag{15}$$

where *K* is the number of classes, FP is the number of false positives, FN is the number of false negatives, TP is the number of true positives and TN is the number of true negatives.

e. Area under the Curve (AUC): Measures the model's ability to discriminate between positive and negative classes in binary classification tasks. AUC values close to 1 indicate excellent discrimination, while values close to 0.5 suggest random performance. AUC is calculated using the trapezoidal rule or integrating the ROC curve (see *Equation 16*):

$$AUC = \int_0^1 TPR(fpr)d(fpr) \tag{16}$$

where TPR is the true positive rate (sensitivity) and fpr is the false positive rate $(1 - specificity)$.

f. Area under the Precision-Recall Curve (AUCPR): Similar to AUC, but it focuses on the precision-recall trade-off, particularly useful for imbalanced datasets (See *Equation 17*).

$$\text{AUCPR} = \int_0^1 \text{Precision}(recall)d(recall) \tag{17}$$

where Precision is the positive predictive value and recall is the true positive rate (sensitivity).

g. Gini Coefficient: It is a metric derived from the Lorenz curve that measures the inequality in a dataset, commonly used as a performance metric for binary classification models. It is estimated by using *Equation 18*.

$$\text{Gini} = \frac{2 \times AUC - 1}{2} \tag{18}$$

h. These mathematical expressions provide a quantitative measure of the model's performance in regression and classification tasks, guiding the selection of the best performing model based on predefined criteria and business objectives.

4. RESULTS

This section evaluates the performance of the Deep Learning, DRF, GBM and AutoML models in predicting accident severity. It assesses their accuracy and discusses the impact of feature engineering and data balancing techniques.

4.1 Characteristics of accidents in the Eastern Province of the KSA

The accident data covering the period between 2018 and 2022 as shown in *Table 1* in the Eastern Province of Saudi Arabia show that Dammam has the highest frequency of accidents, while Dhahran has the least.

Table 1 – Summary of the accident data

Feature	Minimum	1st quartile	Median	Mean	3rd quartile	Maximum
Year	2018	2019	2020	2020	2021	2022
Accident ID	1	2388	4774	4774	7161	9548
Y-coordinate	19.26	25.56	26.39	26.41	26.89	29.39
X-coordinate	43.89	48.6	49.64	49.11	50.04	55.53
City (Most frequent)	-	-	-	Dammam (4995)	-	-
City (Least frequent)	-	-	-	Dhahran (536)	-	-
Location (In-city)	4995	-	-	-	-	-
Location (Out-city)	4553	-	-	-	-	-
Vehicle count	1	1	2	1.791	2	9
Number of victims	0	1	1	1.682	2	24
Accident severity (Fatal)	2527	-	-	-	-	-
Accident severity (Injury)	7021	-				

The median year aligns with most numerical features, indicating a stable trend in the accident severity. Accident locations are evenly distributed between in-city and out-of-city areas. The median vehicle count per accident is 2, with the majority resulting in injuries (n=7021) and fatalities (n=2527). The median year shows a stable trend in accident severity, with accidents evenly distributed between in-city and out-of-city areas. The majority of accidents result in injuries and fatalities, with a large portion involving multiple vehicles. The data suggests improvements in road safety measures and emergency response protocols. Further analysis of the specific causes of accidents could help implement preventative measures. Consistent efforts are needed to address underlying factors contributing to accidents, with targeted interventions and accurate prediction models crucial. This can reduce accident numbers and save lives. Continuous evaluation and improvement of road safety measures and emergency response protocols are essential for ensuring road user safety.

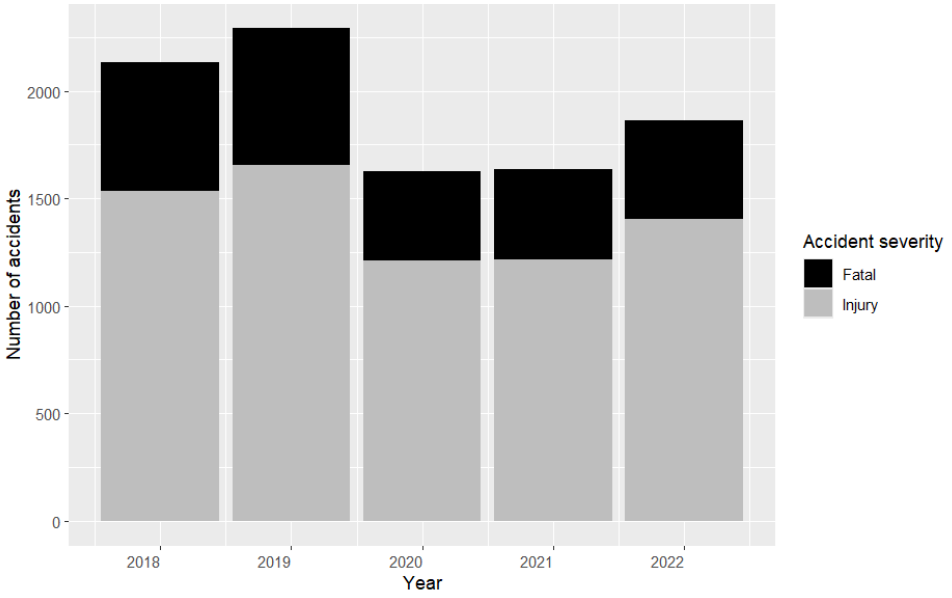

Figure 3 – Distribution of accident severity by year

Over a five-year period, injury accidents have consistently outnumbered fatal accidents, with 1150 injury accidents occurring in 2021 compared to 400 fatal accidents as shown in *Figure 3*. The pattern of accident severity remains consistent, with a peak in 2019 exceeding 2000 and slightly fewer accidents in 2020 and 2021. The importance of robust machine learning models in predicting and mitigating accident severity is highlighted. Historical data trends can help algorithms identify patterns and factors contributing to accident severity, enabling researchers and policymakers to develop targeted interventions and strategies to reduce injury accidents and save lives. Implementing machine learning models can lead to more accurate predictions and proactive measures to prevent severe accidents. Continuous updates and refining of these models with real-time data can help stakeholders stay ahead of emerging trends and adapt their strategies. This proactive approach can significantly reduce the impact of accidents on society.

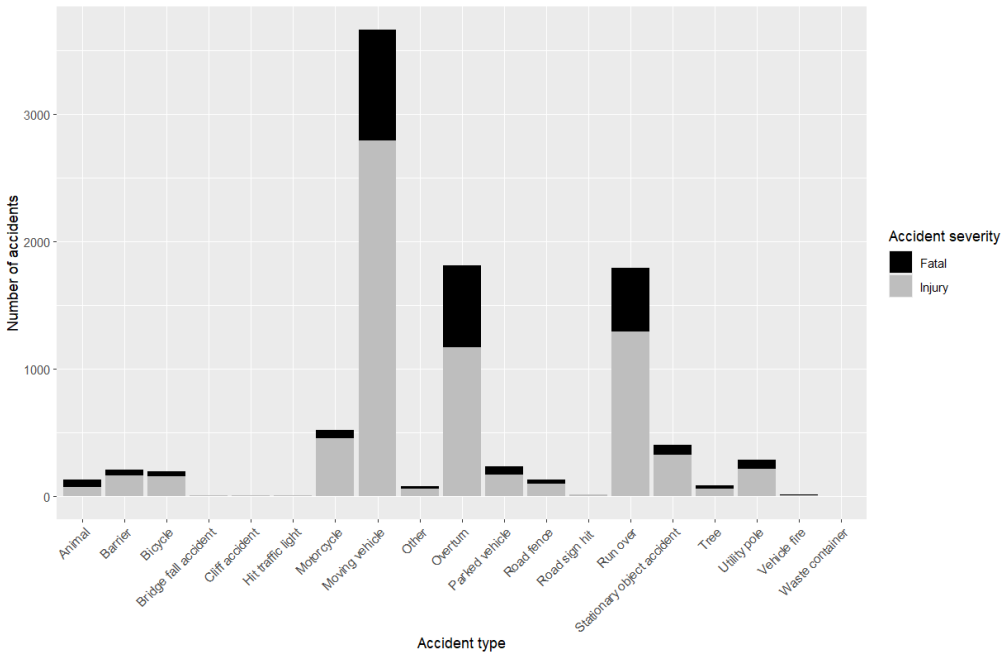

Figure 4 – Distribution of accident severity by accident type

The distribution of accident severity across different types of accidents shown in *Figure 4* offers valuable insights into their prevalence and impact. Injury accidents outnumber fatal accidents, with moving vehicle accidents being the most common. Overturn and runover accidents are the most common, with motorcycles

and stationary objects also contributing to fatalities and injuries. Overturn accidents have a higher fatality rate than runover accidents, indicating the need for targeted interventions. Moving vehicle accidents pose a greater risk to road safety than those involving parked vehicles. Implementing measures to reduce the number of moving vehicle accidents, overturn and runover accidents, and collisions involving motorcycles and stationary objects can help reduce fatalities and injuries. Targeted interventions addressing specific factors leading to overturn accidents can help decrease the proportion of fatalities associated with these incidents. Preventive measures, such as improved road design, regular maintenance and stricter enforcement of traffic laws, can contribute to creating a safer environment for drivers, pedestrians and cyclists. Promoting awareness campaigns and educating individuals on safe driving practices can further reduce accident likelihood. A proactive approach to road safety can lead to a significant reduction in road-related injuries and fatalities, creating a more secure and sustainable transportation system for all.

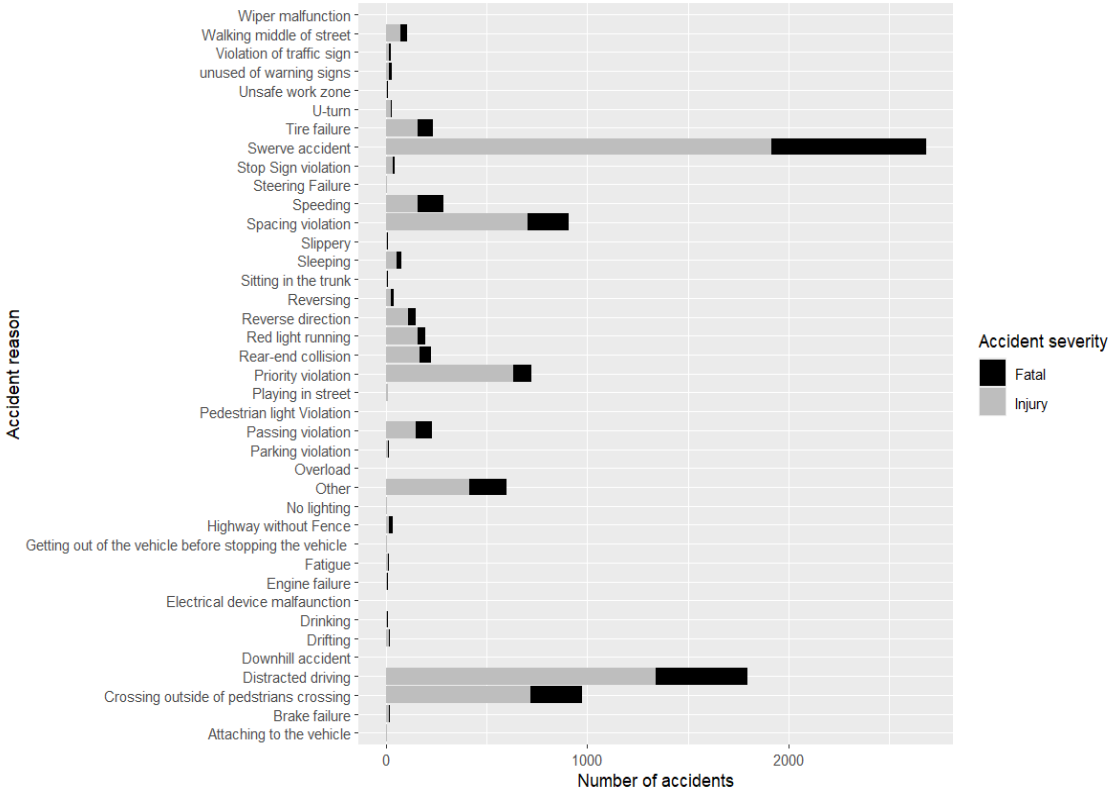

Figure 5 – Distribution of accident severity by the reason of accident

Figure 5 shows a correlation between accident reasons and severity in Saudi Arabia's Eastern Province in the period 2018–2022. Inattentive driving, speeding, unsafe lane changes, failing to yield and improper stopping/turning are the leading causes of accidents. Risky behaviours like speeding and inattention have a higher number of fatal accidents. This underscores the need for targeted interventions like stricter enforcement, public awareness campaigns and advanced driver-assistance systems. Responsible driving behaviour is crucial in reducing accident severity and ensuring road safety for all motorists. A collective effort from authorities and drivers can significantly reduce accidents and save lives. Promoting awareness and educating drivers on reckless behaviour can pave the way for a safer road environment. Enforcing strict penalties for traffic violations and consistently monitoring road conditions can further deter dangerous driving habits.

4.2 Evaluation of the performance of the machine learning models in predicting accident severity

The performance evaluation of the machine learning models, including AutoML, GBM, DRF and Deep Learning, was conducted across five datasets each for either SMOTE (Smt) or ADASYN (Ads) balanced data. Each dataset included observed variables and new features derived from techniques such as clustering, anomaly detection or target encoding. Evaluation metrics used to assess model performance include the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Logarithmic Loss (Log Loss), Mean per Class Error, Area under the Curve (AUC), Area under the Precision-Recall Curve (AUCPR) and Gini index. These metrics

offer a detailed understanding of how accurately the models predict outcomes and classify data, providing valuable insights into their effectiveness across various datasets and feature engineering methods.

Comparison of models based on modelling methods

The evaluation of model performance across different datasets reveals varying levels of efficacy across feature sets. This analysis delves into the performance of various machine learning models for road accident severity prediction, exploring different feature engineering techniques and data balancing methods, specifically comparing SMOTE and ADASYN balancing. The metrics used include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Log Loss, Area under the Curve (AUC), Area under the Precision-Recall Curve (AUCPR) and Gini as shown in *Table 2*.

Table 2 – Machine learning models evaluation (All dataset)

Metric	Model	ADASYN (Training)	ADASYN (Testing)	SMOTE (Training)	SMOTE (Testing)
MSE	**DeepLearning**	0.42	0.43	0.44	0.57
	GBM	0.08	0.13	0.05	0.22
	DRF	0.14	0.14	0.21	0.22
	AutoML	0.07	0.12	0.02	0.21
RMSE	**DeepLearning**	0.65	0.65	0.67	0.76
	GBM	0.29	0.36	0.22	0.47
	DRF	0.37	0.37	0.46	0.46
	AutoML	0.26	0.35	0.13	0.46
LogLoss	**DeepLearning**	1.48	1.52	1.77	2.33
	GBM	0.27	0.39	0.20	0.66
	DRF	0.40	0.41	0.61	0.62
	AutoML	0.23	0.37	0.11	0.65
Mean Per Class	**DeepLearning**	0.21	0.20	0.46	0.50
	GBM	0.10	0.18	0.03	0.48
	DRF	0.18	0.19	0.36	0.49
	AutoML	0.06	0.16	0.00	0.48
AUC	**DeepLearning**	0.84	0.82	0.58	0.60
	GBM	0.96	0.88	1.00	0.62
	DRF	0.88	0.86	0.74	0.66
	AutoML	0.99	0.88	1.00	0.63
AUCPR	**DeepLearning**	0.77	0.74	0.59	0.79
	GBM	0.95	0.81	1.00	0.81
	DRF	0.82	0.79	0.74	0.83
	AutoML	0.99	0.82	1.00	0.81
Gini	**DeepLearning**	0.67	0.65	0.16	0.20
	GBM	0.92	0.75	0.99	0.25
	DRF	0.75	0.72	0.47	0.32
	AutoML	0.98	0.77	1.00	0.27

AutoML consistently emerges as the top performer across all datasets, showcasing its adeptness in leveraging feature engineering techniques and machine learning models. However, the choice between SMOTE and ADASYN balancing appears to influence performance to some extent. Under SMOTE balancing, AutoML with all dataset excels, exhibiting lower MSE (0.02), RMSE (0.13) and LogLoss (0.11) compared to other models as shown in *Figure 6*. Conversely, under ADASYN balancing, AutoML maintains its superior performance metrics, indicating a slight edge over SMOTE balancing, particularly evident in the AutoML with all the variables (MSE: 0.07, RMSE: 0.26, LogLoss: 0.23).

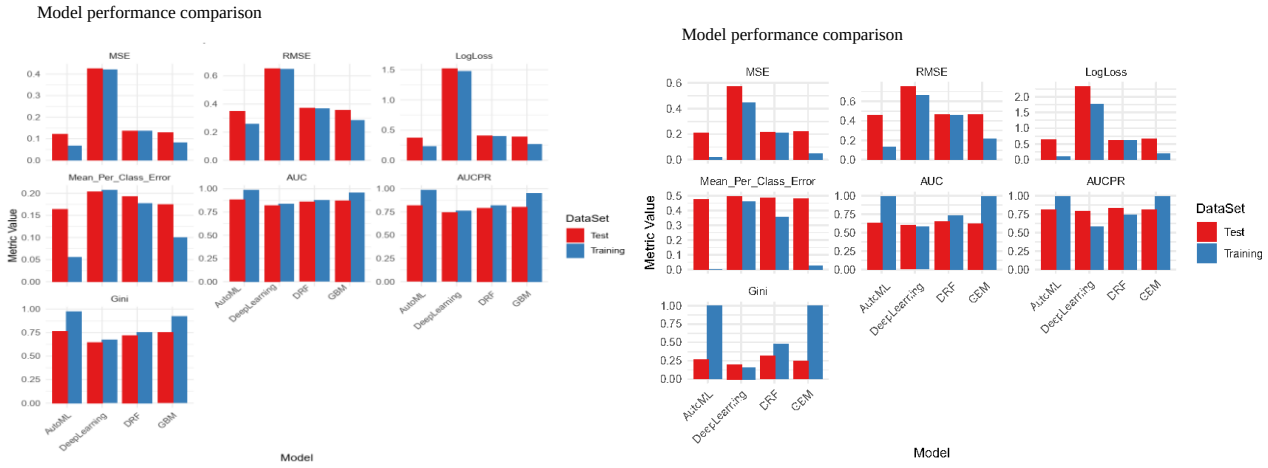

Figure 6 – Comparison of the performance of the machine learning models (Adsall – left; Smt all – right)

GBM demonstrates robust accuracy across datasets, with performance comparable between the SMOTE and ADASYN balancing methods. For instance, in the dataset with anomaly detection feature, GBM showcases competitive AUC (0.99) and Precision (0.98) under SMOTE balancing, similar to its performance under ADASYN balancing with the same dataset (AUC: 0.96, Precision: 0.89).

DRF shows less variation in performance across datasets and seems less influenced by specific feature engineering techniques or balancing methods. This consistency is observed in both SMOTE and ADASYN balanced datasets, suggesting DRF's resilience to the choice of the balancing method. For example, in the CL dataset with the clustering feature under SMOTE balancing, DRF has an AUC of 0.74 and a Precision of 0.6624, similar to its performance under ADASYN balancing with the same dataset (AUC: 0.89, Precision: 0.84).

Deep Learning appears better suited for capturing broader patterns, regardless of the balancing method employed. While it exhibits good AUC scores, its precision for individual predictions may be comparatively lower, as seen in both SMOTE and ADASYN balanced datasets. For instance, in the dataset with the target encoding feature under SMOTE balancing, Deep Learning has an AUC of 0.57 but a higher MSE (0.44) compared to the AutoML (AUC: 0.99, MSE: 0.04), aligning with the observations from the ADASYN results with similar datasets.

Overall, both SMOTE and ADASYN balancing techniques contribute to improved model performance, with advantages observed across all feature engineering techniques. AutoML demonstrates a slight preference for ADASYN balancing in this study, while GBM's performance appears less impacted by the choice of the balancing method. DRF exhibits consistent performance, and Deep Learning maintains its focus on broader patterns. Future research could further explore the nuances of the balancing method selection based on the specific model and feature engineering combinations.

Comparison based on accuracy in the Fatal and Injury Accident classifications

Figure 7 presents an evaluation of various machine learning models and balancing techniques used for predicting the severity of road traffic accidents, focusing on the correctly classified and misclassified cases for both fatal and injury outcomes. For fatal accidents, AutoML with SMOTE achieved the highest accuracy, correctly classifying 1208 cases (89.68%) and misclassifying 139 cases (10.32%). DRF with SMOTE also performed well, correctly classifying 1183 cases (87.82%) but with a slightly higher misclassification rate of 12.18%. BRT with SMOTE and AutoML with ADASYN showed good performance, correctly classifying

1152 (85.52%) and 864 (72.85%) cases, respectively, although with varying misclassification rates. On the other hand, Deep Learning with ADASYN and DRF with ADASYN exhibited lower accuracies, correctly classifying 714 (60.20%) and 793 (66.86%) cases, respectively. Deep Learning with SMOTE had the lowest accuracy for fatal predictions, correctly classifying only 69 cases (14.14%) and misclassifying 419 cases (85.86%).

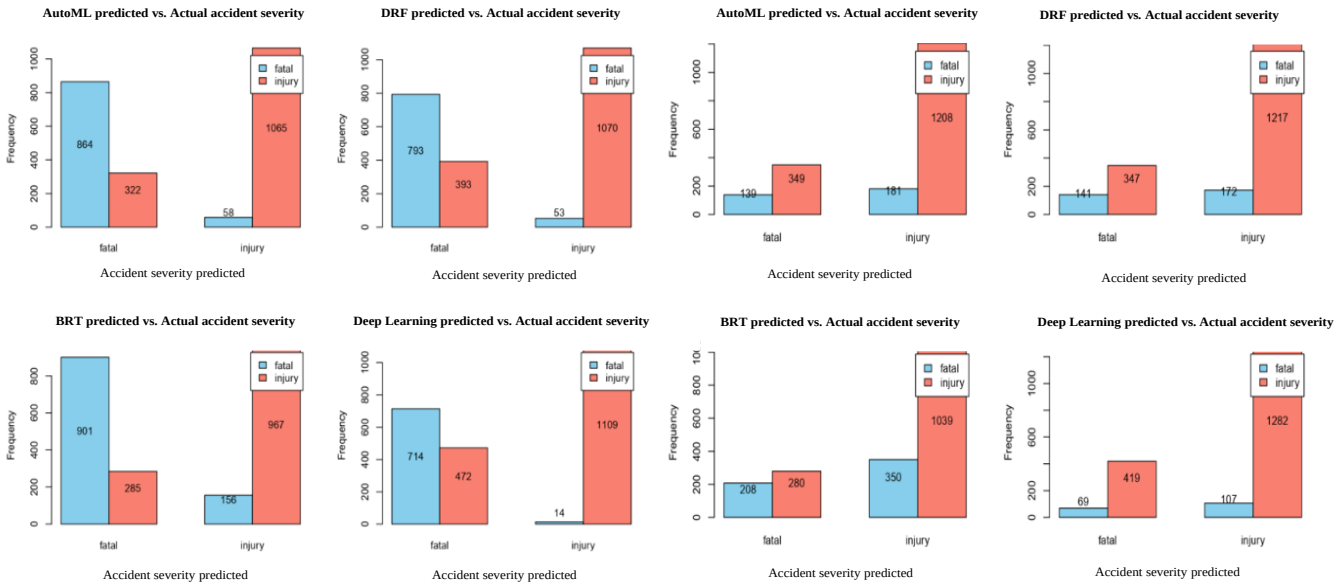

Figure 7 – Comparison of Fatal and Injury Accident predictions (ADASYN with all variables – left; SMOTE with all variables – right)

For injury accidents, Deep Learning with ADASYN performed exceptionally well, correctly classifying 1109 cases (98.75%) with only 14 misclassified cases (1.25%). DRF with ADASYN and AutoML with ADASYN also showed high accuracy, correctly classifying 1070 (95.28%) and 1065 (94.84%) cases, respectively. Deep Learning with SMOTE and BRT with ADASYN correctly classified 1282 (92.30%) and 967 (86.11%) cases, respectively. BRT with SMOTE showed moderate performance, correctly classifying 221 cases (41.70%) with a high misclassification rate of 58.30%. DRF with SMOTE and AutoML with SMOTE had lower accuracies, correctly classifying 204 (38.49%) and 181 (34.15%) cases, respectively.

Overall, AutoML with SMOTE and Deep Learning with ADASYN emerged as top performers for fatal and injury predictions, respectively, highlighting the importance of both the model and the balancing technique. ADASYN generally resulted in higher accuracy across different models for injury predictions compared to SMOTE. The variability in performance across models and techniques highlights the importance of model selection and data balancing methods tailored to specific outcomes (fatal vs. injury). Notably, Deep Learning with SMOTE for fatal accidents and DRF with SMOTE for injury accidents showed significantly lower performance, indicating potential issues with these combinations for accurate classification. The table underscores that the choice of machine learning model and data balancing technique significantly impacts the classification accuracy of accident severity predictions.

These findings underscore the significant impact of dataset characteristics and balancing techniques on the predictive performance of machine learning models. The observed variations in accuracy percentages highlight the necessity of selecting appropriate techniques tailored to the specific requirements of the prediction task. Such insights gleaned from comparative analyses enable informed decisions in model selection and data pre-processing, ultimately enhancing the efficacy of accident severity prediction systems.

Comparison with previous studies

The findings of this study on predicting accident severity using machine learning models can be compared and contrasted with the results reported in several related works. Consistent with the observations made by Aldhari et al. [53] and Akin et al. [52], this study emphasises the importance of feature engineering and data pre-processing in enhancing the predictive performance of machine learning models. The incorporation of

advanced techniques, such as clustering, anomaly detection and target encoding, to derive new features aligns with the strategies employed in these previous studies. The ability of the models to leverage these engineered features to improve accuracy in accident severity prediction underscores the value of a comprehensive approach to feature selection. In terms of model performance in accident severity prediction, the superior results achieved by AutoML in this study echo the findings of several studies [137–139], who also reported the effectiveness of automated machine learning techniques. The consistent outperformance of AutoML across various datasets and balancing methods further validates its robustness and adaptability, as observed in their work. The comparative analysis of different machine learning algorithms, including GBM, DRF and Deep Learning, corroborates the insights from studies by Jamal et al. [140] and Alrajhi and Kamel [56]. The superior performance of ensemble methods, such as GBM, in predicting accident severity is in line with the results reported in previous works [56]. However, the challenges faced by Deep Learning models in this study, particularly in terms of higher error rates, differ from the findings of Alrajhi and Kamel [56], who highlighted the potential of deep learning for accident risk prediction in the Saudi context. Regarding data balancing techniques, the slight advantage of ADASYN over SMOTE observed in this study aligns with the conclusions drawn by Mostafa [141] and Morris and Yang [67]. Their research also emphasised the importance of nuanced data balancing approaches in enhancing the performance of predictive models for accident severity. The need to strike a balance between accurately predicting different accident severity levels underscores the complexities involved in developing robust predictive models for road safety applications.

5. CONCLUSION

The study explores the predictive performance of machine learning models for accident severity categorization, revealing differences in model effectiveness across datasets and balancing techniques. AutoML emerged as the top-performing model, achieving high predictive accuracy in both fatal and injury accidents. Deep Learning showed promising results in predicting injury accidents, achieving 95% accuracy, but struggled with fatal accident predictions, achieving only 60% accuracy. Distributed Random Forest (DRF) and Gradient Boosting Machine (GBM) exhibited balanced performance across both categories, with DRF achieving an AUC of 0.88 and GBM reaching an AUC of 0.96 under ADASYN balancing. The study emphasises the importance of feature engineering and data balancing techniques in enhancing model performance. ADASYN generally yielded better results, with AutoML achieving a Root Mean Squared Error (RMSE) of 0.26 under ADASYN balancing.

Further research is required to improve the predictive accuracy and robustness of models for accident severity prediction, incorporating techniques like convolutional neural networks (CNNs) for spatial data and recurrent neural networks (RNNs) for temporal data. This study acknowledges limitations such as data quality dependence and low interpretability, with future work aimed at addressing these by using SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) for greater transparency. Additionally, machine learning models, supported by data balancing methods like SMOTE and ADASYN, can facilitate continuous data-driven decision-making and infrastructure improvements, such as enhanced lighting, clearer signage and traffic calming measures. Policy enhancements are needed to support the integration of predictive analytics in traffic management. Efforts should also focus on economic and social equity to prevent disproportionate impacts on disadvantaged communities, and predictive models should be used to optimise emergency response strategies. Implementing these data-driven recommendations will enhance the ability of transportation authorities to prevent road accidents and reduce their severity, contributing to safer roads and saving lives.

CONFLICT OF INTEREST

We declare that we have no financial, personal or professional interests that may have influenced the research or its interpretation presented in this manuscript. There are no conflicts of interest to disclose.

ACKNOWLEDGEMENTS

This work was funded by the Deanship of Graduate Studies and Scientific Research at Jouf University under grant No. (DGSSR-2023-02-02039).

REFERENCES

[1] Alsofayan YM, et al. Do crashes happen more frequently at sunset in Ramadan than the rest of the year? *Journal of Taibah University Medical Sciences*. 2022;17(6):1031–1038. DOI: 10.1016/j.jtumed.2022.06.002.

[2] Chen S, et al. The global macroeconomic burden of road injuries: Estimates and projections for 166 countries. *Lancet Planet Health*. 2019;3(9):e390 –e398. DOI: 10.1016/S2542-5196(19)30170-6.

[3] World Bank. The high toll of traffic injuries: Unacceptable and preventable. 2017. DOI: 10.1596/29129.

[4] Al-Madani HMN. Fatal crashes in GCC countries: Comparative analysis with EU countries for three decades. In: *Proceedings of SAFE 2013*. Rome, Italy; 2013. p. 471–482. DOI: 10.2495/SAFE130421.

[5] Awadalla DM, de Albuquerque FDB. Fatal road crashes in the emirate of Abu Dhabi: Contributing factors and data-driven safety recommendations. *Transportation Research Procedia*. 2021;52:260–267. DOI: 10.1016/j.trpro.2021.01.030.

[6] Bener A, et al. The impact of four-wheel drive on risky driver behaviours and road traffic accidents. *Transportation Research Part F: Traffic Psychology and Behaviour*. 2008;11(5):324–333. DOI: 10.1016/j.trf.2008.02.001.

[7] de Albuquerque FDB, Awadalla DM. Characterization of road crashes in the emirate of Abu Dhabi. *Transportation Research Procedia*. 2020;48:1095–1110. DOI: 10.1016/j.trpro.2020.08.136.

[8] Fadel H, et al. Vision zero: The journey to safer roads in the Middle East. *Federation Internationale de l'Automobile*. https://www.fia.com/news/gcc-countries-could-significantly-reduce-annual-road-traffic-fatalities-and-boost-economic [Accessed 3rd Aug. 2024].

[9] Rohrer WM. Road traffic accidents as public health challenge in the Gulf Cooperation Council (GCC) region. In: *Public Health - Open Journal*. 2016. p. e6–e7. DOI: 10.17140/PHOJ-1-e004.

[10] Saudi Vision 2030. National transformation program. *Vision 2030 Kingdom of Saudi Arabia*. http://www.vision2030.gov.sa/en/vision-2030/vrp/national-transformation-program/ [Accessed 3rd Aug. 2024].

[11] Alotaibi O, Potoglou D. Introducing public transport and relevant strategies in Riyadh City, Saudi Arabia: A stakeholders' perspective. *Urban, Planning and Transport Research*. 2018;6(1):35–53. DOI: 10.1080/21650020.2018.1463867.

[12] Moser S, Swain M, Alkhabbaz MH. King Abdullah economic city: Engineering Saudi Arabia's post-oil future. *Cities*. 2015;45:71–80. DOI: 10.1016/j.cities.2015.03.001.

[13] Alghnam S, et al. Healthcare costs of road injuries in Saudi Arabia: A quantile regression analysis. *Accident Analysis & Prevention*. 2021;159:106266. DOI: 10.1016/j.aap.2021.106266.

[14] Jamal A, Rahman MT, Al-Ahmadi HM, Mansoor U. The dilemma of road safety in the eastern province of Saudi Arabia: Consequences and prevention strategies. *International Journal of Environmental Research and Public Health*. 2020;17(1):Article 157. DOI: 10.3390/ijerph17010157.

[15] Wahaq AB, Bawazir A. Female drivers' attitudes and behavior regarding traffic regulations in Riyadh, Saudi Arabia. *Research Square*. 2021. DOI: 10.21203/rs.3.rs-179510/v1. https://www.researchsquare.com/article/rs-179510/v1 [Accessed 15th Feb. 2021].

[16] World Health Organization. Reducing road crash deaths in the Kingdom of Saudi Arabia. 2023. https://www.who.int/news/item/20-06-2023-reducing-road-crash-deaths-in-the-Kingdom-of-Saudi-Arabia [Accessed 27th Sep. 2023].

[17] Safarpour H, et al. The common road safety approaches: A scoping review and thematic analysis. *Chinese Journal of Traumatology*. 2020;23(2):113–121. DOI: 10.1016/j.cjtee.2020.02.005.

[18] Smith T. Fundamentals of the safe system approach. *Vision Zero Network*. 2024. https://visionzeronetwork.org/fundamentals-of-the-safe-system-approach/ [Accessed 16th Apr. 2024].

[19] Biddala SCR, Ibikunle O, Duffy VG. Systematic review on safety of artificial intelligence and transportation. In: Duffy VG, Krömker H, Streitz NA, Konomi S, editors. *HCI International 2023 – Late Breaking Papers*. Cham: Springer Nature Switzerland; 2023. p. 248–263. DOI: 10.1007/978-3-031-48047-8_16.

[20] Alqahtani H, Kumar G. Machine learning for enhancing transportation security: A comprehensive analysis of electric and flying vehicle systems. *Engineering Applications of Artificial Intelligence*. 2024;129:107667. DOI: 10.1016/j.engappai.2023.107667.

[21] Tselentis DI, et al. The usefulness of artificial intelligence for safety assessment of different transport modes. *Accident Analysis & Prevention*. 2023;186:107034. DOI: 10.1016/j.aap.2023.107034.

[22] Li Z, Liu P, Wang W, Xu C. Using support vector machine models for crash injury severity analysis. *Accident Analysis & Prevention*. 2012;45:478–486. DOI: 10.1016/j.aap.2011.08.016.

[23] Alsrehin NO, Klaib AF, Magableh A. Intelligent transportation and control systems using data mining and machine learning techniques: A comprehensive study. *IEEE Access*. 2019;7:49830–49857. DOI: 10.1109/ACCESS.2019.2909114.

[24] Neilson A, Indratmo, Daniel B, Tjandra S. Systematic review of the literature on big data in the transportation domain: Concepts and applications. *Big Data Research*. 2019;17:35–44. DOI: 10.1016/j.bdr.2019.03.001.

[25] Tilahun N. Safety impact of automated speed camera enforcement: Empirical findings based on Chicago's speed cameras. *Transportation Research Record: Journal of the Transportation Research Board*. 2023;2677(1):1490–1498. DOI: 10.1177/03611981221104808.

[26] Kalambay P, Pulugurtha SS. Data-driven exploration of traffic speed patterns to identify potential road links for variable speed limit sign implementation. *Urban, Planning and Transport Research*. 2024;12(1):2319711. DOI: 10.1080/21650020.2024.2319711.

[27] Li H, Zhang Y, Ren G. A causal analysis of time-varying speed camera safety effects based on the propensity score method. *Journal of Safety Research*. 2020;75:119–127. DOI: 10.1016/j.jsr.2020.08.007.

[28] Aghayari H, et al. Mobile applications for road traffic health and safety in the mirror of the Haddon's matrix. *BMC Medical Informatics and Decision Making*. 2021;21(1):230. DOI: 10.1186/s12911-021-01578-8.

[29] Zhang Z, Xu N, Liu J, Jones S. Exploring spatial heterogeneity in factors associated with injury severity in speeding-related crashes: An integrated machine learning and spatial modeling approach. *Accident Analysis & Prevention*. 2024;206:107697. DOI: 10.1016/j.aap.2024.107697.

[30] Alkheder S, et al. Severity prediction of traffic accident using an artificial neural network. *Journal of Forecasting*. 2017;36(1):100–108. DOI: 10.1002/for.2425.

[31] Ijaz M, Lan L, Zahid M, Jamal A. A comparative study of machine learning classifiers for injury severity prediction of crashes involving three-wheeled motorized rickshaw. *Accident Analysis & Prevention*. 2021;154:106094. DOI: 10.1016/j.aap.2021.106094.

[32] Maghelal P, et al. Severity of vehicle-to-vehicle accidents in the UAE: An exploratory analysis using machine learning algorithms. *Heliyon*. 2023;9(10):e20694 . DOI: 10.1016/j.heliyon.2023.e20694.

[33] Mohamed SA, Kishta M, Al-Harthi HA. Investigating factors affecting the occurrence and severity of rear-end crashes. *Transportation Research Procedia*. 2017;25:2098–2107. DOI: 10.1016/j.trpro.2017.05.403.

[34] Panda C, Mishra AK, Dash AK, Nawab H. Predicting and explaining severity of road accident using artificial intelligence techniques, SHAP and feature analysis. *International Journal of Crashworthiness*. 2023;28(2):186–201. DOI: 10.1080/13588265.2022.2074643.

[35] Taamneh S, Taamneh M. Evaluation of the performance of random forests technique in predicting the severity of road traffic accidents. In: Stanton N, editor. *Advances in Human Aspects of Transportation*. Cham: Springer International Publishing; 2019. p. 840–847. DOI: 10.1007/978-3-319-93885-1_78.

[36] Júnior JF, et al. Driver behavior profiling: An investigation with different smartphone sensors and machine learning. *PLOS ONE*. 2017;12(4):e0174959 . DOI: 10.1371/journal.pone.0174959.

[37] Silva PB, Andrade M, Ferreira S. Machine learning applied to road safety modeling: A systematic literature review. *Journal of Traffic and Transportation Engineering (English Edition)*. 2020;7(6):775–790. DOI: 10.1016/j.jtte.2020.07.004.

[38] Zhang Z, et al. Machine learning based real-time prediction of freeway crash risk using crowdsourced probe vehicle data. *Journal of Intelligent Transportation Systems*. 2024;28(1):84–102. DOI: 10.1080/15472450.2022.2106564.

[39] Sarigiannis D, et al. Feature engineering and decision trees for predicting high crash-risk locations using roadway indicators. *Transportation Research Record*. 2024. DOI: 10.1177/03611981231217497.

[40] Qamar R, Zardari BA. Artificial neural networks: An overview. *Mesopotamian Journal of Computer Science*. 2023;2023:130–139. DOI: 10.58496/MJCSC/2023/015..

[41] Zhang Y, Li H, Ren G. Estimating heterogeneous treatment effects in road safety analysis using generalized random forests. *Accident Analysis & Prevention*. 2022;165:106507. DOI: 10.1016/j.aap.2021.106507.

[42] Schlögl M, et al. A comparison of statistical learning methods for deriving determining factors of accident occurrence from an imbalanced high resolution dataset. *Accident Analysis & Prevention*. 2019;127:134–149. DOI: 10.1016/j.aap.2019.02.008.

[43] Muzahid AJM, et al. Deep reinforcement learning-based driving strategy for avoidance of chain collisions and its safety efficiency analysis in autonomous vehicles. *IEEE Access*. 2022;10:43303–43319. DOI: 10.1109/ACCESS.2022.3167812.

[44] Sun Z, et al. A hybrid approach of random forest and random parameters logit model of injury severity modeling of vulnerable road users involved crashes. *Accident Analysis & Prevention*. 2023;192:107235. DOI: 10.1016/j.aap.2023.107235.

[45] Yang Z, Zhang W, Feng J. Predicting multiple types of traffic accident severity with explanations: A multi-task deep learning framework. *Safety Science*. 2022;146:105522. DOI: 10.1016/j.ssci.2021.105522.

[46] Xie Y. Values and limitations of statistical models. *Research in Social Stratification and Mobility*. 2011;29(3):343–349. DOI: 10.1016/j.rssm.2011.04.001.

[47] Mannering FL, Shankar V, Bhat CR. Unobserved heterogeneity and the statistical analysis of highway accident data. *Analytic Methods in Accident Research*. 2016;11:1–16. DOI: 10.1016/j.amar.2016.04.001.

[48] Wang C, Shao Y, Ye F, Zhu T. Injury severity analysis of e-bike riders in China based on the in-vehicle recording video crash data: A random parameter ordered logit model. *International Journal of Injury Control and Safety Promotion*. 2024:1–11. DOI: 10.1080/17457300.2024.2385102.

[49] Savolainen PT, et al. The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. *Accident Analysis & Prevention*. 2011;43(5):1666–1676. DOI: 10.1016/j.aap.2011.03.025.

[50] Ye F, et al. Investigating the severity of expressway crash based on the random parameter logit model accounting for unobserved heterogeneity. *Advances in Mechanical Engineering*. 2021;13(12):16878140211067278. DOI: 10.1177/16878140211067278.

[51] Zhang S, et al. Hybrid feature selection-based machine learning classification system for the prediction of injury severity in single and multiple-vehicle accidents. *PLOS ONE*. 2022;17(2):e0262941 . DOI: 10.1371/journal.pone.0262941.

[52] Akin D, et al. Identifying causes of traffic crashes associated with driver behavior using supervised machine learning methods: Case of highway 15 in Saudi Arabia. *Sustainability*. 2022;14(24):16654. DOI: 10.3390/su142416654.

[53] Aldhari I, et al. Severity prediction of highway crashes in Saudi Arabia using machine learning techniques. *Applied Sciences*. 2023;13(1):233. DOI: 10.3390/app13010233.

[54] Bachir H, Almannaa M. Crash severity predictive models using machine learning algorithms: A case study of Riyadh, Saudi Arabia. In: *Proceedings of the 13th Annual International Conference on Industrial Engineering and Operations Management*. Manila, Philippines: IEOM Society; 2023. DOI: 10.46254/AN13.20230244.

[55] Aboulola OI. Improving traffic accident severity prediction using MobileNet transfer learning model and SHAP XAI technique. *PLOS ONE*. 2024;19(4):e0300640 . DOI: 10.1371/journal.pone.0300640.

[56] Alrajhi M, Kamel M. A deep-learning model for predicting and visualizing the risk of road traffic accidents in Saudi Arabia: A tutorial approach. *International Journal of Advanced Computer Science and Applications*. 2019;10. DOI: 10.14569/IJACSA.2019.0101166.

[57] Wang H, et al. An interpretable deep embedding model for few and imbalanced biomedical data. *IEEE Journal of Biomedical and Health Informatics*. 2022:1–8. DOI: 10.1109/JBHI.2022.3223798.

[58] Wen X, et al. Applications of machine learning methods in traffic crash severity modelling: Current status and future directions. *Transport Reviews*. 2021;41(6):855–879. DOI: 10.1080/01441647.2021.1954108.

[59] Gao Y, Zhu Y, Zhao Y. Dealing with imbalanced data for interpretable defect prediction. *Information and Software Technology*. 2022;151:107016. DOI: 10.1016/j.infsof.2022.107016.

[60] Zheng A, Casari A. Feature engineering for machine learning: Principles and techniques for data scientists. O'Reilly Media; 2018.

[61] Fiorentini N, Losa M. Handling imbalanced data in road crash severity prediction by machine learning algorithms. *Infrastructures*. 2020;5(7):61. DOI: 10.3390/infrastructures5070061.

[62] Mohammadpour SI, et al. Classification of truck-involved crash severity: Dealing with missing, imbalanced, and high dimensional safety data. *PLOS ONE*. 2023;18(3):e0281901 . DOI: 10.1371/journal.pone.0281901.

[63] Ogungbire A, Pulugurtha SS. Effectiveness of data imbalance treatment in weather-related crash severity analysis. *Transportation Research Record*. 2024. DOI: 10.1177/03611981241239962.

[64] Sarkar S, et al. Predicting and analyzing injury severity: A machine learning-based approach using class-imbalanced proactive and reactive data. *Safety Science*. 2020;125:104616. DOI: 10.1016/j.ssci.2020.104616.

[65] Ali Y, Hussain F, Haque MM. Advances, challenges, and future research needs in machine learning-based crash prediction models: A systematic review. *Accident Analysis & Prevention*. 2024;194:107378. DOI: 10.1016/j.aap.2023.107378.

[66] Li G, et al. ReMAHA–CatBoost: Addressing imbalanced data in traffic accident prediction tasks. *Applied Sciences*. 2023;13(24):13123. DOI: 10.3390/app132413123.

[67] Morris C, Yang JJ. Effectiveness of resampling methods in coping with imbalanced crash data: Crash type analysis and predictive modeling. *Accident Analysis & Prevention*. 2021;159:106240. DOI: 10.1016/j.aap.2021.106240.

[68] Mohamed MG, et al. A clustering regression approach: A comprehensive injury severity analysis of pedestrian–vehicle crashes in New York, US and Montreal, Canada. *Safety Science*. 2013;54:27–37. DOI: 10.1016/j.ssci.2012.11.001.

[69] Wang K, Xue Q, Lu JJ. Risky driver recognition with class imbalance data and automated machine learning framework. *International Journal of Environmental Research and Public Health*. 2021;18(14):7534. DOI: 10.3390/ijerph18147534.

[70] Fernandez A, et al. SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*. 2018;61:863–905. DOI: 10.1613/jair.1.11192.

[71] Tang B, He H. KernelADASYN: Kernel based adaptive synthetic data generation for imbalanced learning. In: *2015 IEEE Congress on Evolutionary Computation (CEC)*. 2015. p. 664–671. DOI: 10.1109/CEC.2015.7256954.

[72] Chawla NV, et al. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*. 2002;16:321–357. DOI: 10.1613/jair.953.

[73] Sharma S, Gosain A, Jain S. A review of the oversampling techniques in class imbalance problem. In: Khanna A, et al., editors. *International Conference on Innovative Computing and Communications*. Singapore: Springer; 2022. p. 459–472. DOI: 10.1007/978-981-16-2594-7_38.

[74] Devi D, Biswas SK, Purkayastha B. A review on solution to class imbalance problem: Undersampling approaches. In: *2020 International Conference on Computational Performance Evaluation (ComPE)*. 2020. p. 626–631. DOI: 10.1109/ComPE49325.2020.9200087.

[75] Hasanin T, et al. Investigating random undersampling and feature selection on bioinformatics big data. In: *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)*. 2019. p. 346–356. DOI: 10.1109/BigDataService.2019.00063.

[76] Hasanin T, Khoshgoftaar T. The effects of random undersampling with simulated class imbalance for big data. In: *2018 IEEE International Conference on Information Reuse and Integration (IRI)*. 2018. p. 70–79. DOI: 10.1109/IRI.2018.00018.

[77] Mohammed R, Rawashdeh J, Abdullah M. Machine learning with oversampling and undersampling techniques: Overview study and experimental results. In: *2020 11th International Conference on Information and Communication Systems (ICICS)*. 2020. p. 243–248. DOI: 10.1109/ICICS49469.2020.239556.

[78] He H, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. 2008. p. 1322–1328. DOI: 10.1109/IJCNN.2008.4633969.

[79] Jian C, Gao J, Ao Y. A new sampling method for classifying imbalanced data based on support vector machine ensemble. *Neurocomputing*. 2016;193:115–122. DOI: 10.1016/j.neucom.2016.02.006.

[80] Louppe G. Understanding random forests: From theory to practice. *arXiv*. 2015. arXiv:1407.7502. DOI: 10.48550/arXiv.1407.7502.

[81] Fernández A, et al. Cost-sensitive learning. In: Fernández A, et al., editors. *Learning from Imbalanced Data Sets*. Cham: Springer International Publishing; 2018. p. 63–78. DOI: 10.1007/978-3-319-98074-4_4.

[82] Pereira J, Saraiva F. A comparative analysis of unbalanced data handling techniques for machine learning algorithms to electricity theft detection. In: *2020 IEEE Congress on Evolutionary Computation (CEC)*. 2020. p. 1–8. DOI: 10.1109/CEC48606.2020.9185822.

[83] Heaton J. An empirical analysis of feature engineering for predictive modeling. In: *SoutheastCon 2016*. 2016. p. 1–6. DOI: 10.1109/SECON.2016.7506650.

[84] Buian MFI, et al. Advanced analytics for predicting traffic collision severity assessment. *World Journal of Advanced Research and Reviews*. 2024;21(2):2007–2018. DOI: 10.30574/wjarr.2024.21.2.0704.

[85] Ikotun AM, et al. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*. 2023;622:178–210. DOI: 10.1016/j.ins.2022.11.139.

[86] H2O.ai. Target encoding — H2O 3.46.0.1 documentation. 2024. https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/target-encoding.html [Accessed 3rd May 2024].

[87] Thudumu S, et al. A comprehensive survey of anomaly detection techniques for high dimensional big data. *Journal of Big Data*. 2020;7(1):42. DOI: 10.1186/s40537-020-00320-x.

[88] Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012;28(1):112–118. DOI: 10.1093/bioinformatics/btr597.

[89] Hong S, Lynn HS. Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. *BMC Medical Research Methodology*. 2020;20(1):199. DOI: 10.1186/s12874-020-01080-1.

[90] Li LI, Goshawk DP. Comparison of random forest and multiple imputation for imputing missing data: A case study of the education panel survey of the City of China.2015. https://www.albany.edu/chinanet/events/ucrn2016/papers/18_Comparison%20of%20Random%20Forest%20and%20Multiple%20Imputation%20for%20Imputing%20Missing%20Data.pdf [Accessed 10th May 2024].

[91] Micci-Barreca D. A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *SIGKDD Explorations Newsletter*. 2001;3(1):27–32. DOI: 10.1145/507533.507538.

[92] Prokhorenkova L, et al. CatBoost: Unbiased boosting with categorical features. In: *Advances in Neural Information Processing Systems*. 2018. https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html [Accessed 10th May 2024].

[93] Pargent F, et al. Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. *Computational Statistics*. 2022;37(5):2671–2692. DOI: 10.1007/s00180-022-01207-6.

[94] Branco P, Ribeiro RP, Torgo L. UBL: An R package for Utility-based Learning. *arXiv*. 2016. arXiv:1604.08079. http://arxiv.org/abs/1604.08079 [Accessed 3rd May 2024].

[95] Alex SA, Nayahi JJV, Kaddoura S. Deep convolutional neural networks with genetic algorithm-based synthetic minority over-sampling technique for improved imbalanced data classification. *Applied Soft Computing*. 2024;156:111491. DOI: 10.1016/j.asoc.2024.111491.

[96] Malhotra R, Kamal S. An empirical study to investigate oversampling methods for improving software defect prediction using imbalanced data. *Neurocomputing*. 2019;343:120–140. DOI: 10.1016/j.neucom.2018.04.090.

[97] Botelho AF, Baker RS, Heffernan NT. Machine-learned or expert-engineered features? Exploring feature engineering methods in detectors of student behavior and affect. *Twelfth International Conference on Educational Data Mining*. 2019. https://par.nsf.gov/biblio/10108548 [Accessed 10th May 2024].

[98] Yassin SS, Pooja. Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach. *SN Applied Sciences*. 2020;2(9):1576. DOI: 10.1007/s42452-020-3125-1.

[99] Bridgelall R, Tolliver DD. Railroad accident analysis by machine learning and natural language processing. *Journal of Rail Transport Planning & Management*. 2024;29:100429. DOI: 10.1016/j.jrtpm.2023.100429.

[100] Suh Y, Song B. Narrative texts-based anomaly detection using accident report documents: The case of chemical process safety. *Journal of Loss Prevention in the Process Industries*. 2019;57:47–54. DOI: 10.1016/j

[101] Katya E. Exploring feature engineering strategies for improving predictive models in data science. *Research Journal of Computer Systems and Engineering*. 2023;4(2). DOI: 10.52710/rjcse.88.

[102] Shi X, et al. A feature learning approach based on XGBoost for driving assessment and risk prediction. *Accident Analysis & Prevention*. 2019;129:170–179. DOI: 10.1016/j.aap.2019.05.005.

[103] Anderson TK. Kernel density estimation and K-means clustering to profile road accident hotspots. *Accident Analysis & Prevention*. 2009;41(3):359–364. DOI: 10.1016/j.aap.2008.12.014.

[104] Kazmi SSA, Ahmed M, Mumtaz R, Anwar Z. Spatiotemporal clustering and analysis of road accident hotspots by exploiting GIS technology and kernel density estimation. *The Computer Journal*. 2022;65(2):155–176. DOI: 10.1093/comjnl/bxz158.

[105] James G, et al. An introduction to statistical learning. Vol. 112. New York: Springer; 2021.

[106] Sterkenburg M. Theoretical and practical aspects of isolation forest. Master's thesis. Utrecht University; 2022. https://studenttheses.uu.nl/handle/20.500.12932/42666 [Accessed 4th May 2024].

[107] Laskar MTR, et al. Extending isolation forest for anomaly detection in big data via K-means. *ACM Transactions on Cyber-Physical Systems*. 2021;5(4):41:1–41:26. DOI: 10.1145/3460976.

[108] LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436–444. DOI: 10.1038/nature14539.

[109] Cevid D, et al. Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. *Journal of Machine Learning Research*. 2022;23(333):1–79. http://jmlr.org/papers/v23/21-0585.html [Accessed 10th May 2024].

[110] Friedman JH. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*. 2001;29(5):1189–1232. DOI: 10.1214/aos/1013203451.

[111] Konstantinov AV, Utkin LV. Interpretable machine learning with an ensemble of gradient boosting machines. *Knowledge-Based Systems*. 2021;222:106993. DOI: 10.1016/j.knosys.2021.106993.

[112] Oyedele A, et al. Deep learning and boosted trees for injuries prediction in power infrastructure projects. *Applied Soft Computing*. 2021;110:107587. DOI: 10.1016/j.asoc.2021.107587.

[113] Azevedo K, et al. A multivocal literature review on the benefits and limitations of automated machine learning tools. *arXiv*. 2024. arXiv:2401.11366. DOI: 10.48550/arXiv.2401.11366.

[114] Baykal T, et al. Accident severity prediction in big data using auto-machine learning. *Scientia Iranica*. 2023. DOI: 10.24200/sci.2023.60144.6626.

[115] Yates LA, et al. Cross validation for model selection: A review with examples from ecology. *Ecological Monographs*. 2023;93(1):e1557 . DOI: 10.1002/ecm.1557.

[116] Mbelwa J, et al. The effect of hyperparameter optimization on the estimation of performance metrics in network traffic prediction using the gradient boosting machine model. 2023. DOI: 10.48084/etasr.5548.

[117] Naser MZ, Alavi A. Insights into performance fitness and error metrics for machine learning. 2020. DOI: 10.1007/s44150-021-00015-8.

[118] Vujovic ŽĐ. Classification model evaluation metrics. *International Journal of Advanced Computer Science and Applications*. 2021;12(6). DOI: 10.14569/IJACSA.2021.0120670.

[119] Smys S, Chen JIZ, Shakya S. Survey on neural network architectures with deep learning. *Journal of Soft Computing Paradigm*. 2020;2(3):186–194. https://www.academia.edu/download/70861228/06.pdf [Accessed 4th May 2024].

[120] Li Z, He Q, Li J. A survey of deep learning-driven architecture for predictive maintenance. *Engineering Applications of Artificial Intelligence*. 2024;133:108285. DOI: 10.1016/j.engappai.2024.108285.

[121] Hussain H, et al. Design possibilities and challenges of DNN models: A review on the perspective of end devices. *Artificial Intelligence Review*. 2022;55(7):5109–5167. DOI: 10.1007/s10462-022-10138-z.

[122] Najafabadi MM, et al. Deep learning applications and challenges in big data analytics. *Journal of Big Data*. 2015;2(1):1. DOI: 10.1186/s40537-014-0007-7.

[123] Mehmood F, Ahmad S, Whangbo TK. An efficient optimization technique for training deep neural networks. *Mathematics*. 2023;11(6):1360. DOI: 10.3390/math11061360.

[124] Shrestha A, Mahmood A. Review of deep learning algorithms and architectures. *IEEE Access*. 2019;7:53040–53065. DOI: 10.1109/ACCESS.2019.2912200.

[125] Al-Allak A, Bertelli G, Lewis P. Random forests: The new generation of machine learning algorithms to predict survival in breast cancer. *International Journal of Surgery*. 2013;11(8):607. DOI: 10.1016/j.ijsu.2013.06.112.

[126] Natekin A, Knoll A. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*. 2013;7:21. DOI: 10.3389/fnbot.2013.00021.

[127] Chen Y-W, Song Q, Hu X. Techniques for automated machine learning. *SIGKDD Explorations Newsletter*. 2021;22(2):35–50. DOI: 10.1145/3447556.3447567.

[128] Kozak A, et al. Deciphering AutoML ensembles: Cattleia's assistance in decision-making. *arXiv*. 2024. arXiv:2403.12664. DOI: 10.48550/arXiv.2403.12664.

[129] Balaji A, Allen A. Benchmarking automatic machine learning frameworks. *arXiv*. 2018. arXiv:1808.06492. DOI: 10.48550/arXiv.1808.06492.

[130] Shen Z, et al. Automated machine learning: From principles to practices. *arXiv*. 2024. arXiv:1810.13306. DOI: 10.48550/arXiv.1810.13306.

[131] Vabalas A, et al. Machine learning algorithm validation with a limited sample size. *PLOS ONE*. 2019;14(11):e0224365 . DOI: 10.1371/journal.pone.0224365.

[132] Arora A, et al. Deep learning with H2O. H2O.ai; 2015.

[133] Fallatah O, et al. Factors controlling groundwater radioactivity in arid environments: An automated machine learning approach. *Science of The Total Environment*. 2022;830:154707. DOI: 10.1016/j.scitotenv.2022.154707.

[134] Yang R-M, et al. Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. *Ecological Indicators*. 2016;60:870–878. DOI: 10.1016/j.ecolind.2015.08.036.

[135] Lee CK, et al. Development and validation of a deep neural network model for prediction of postoperative in-hospital mortality. *Anesthesiology*. 2018;129(4):649–662. DOI: 10.1097/ALN.0000000000002186.

[136] Boehmke B, Greenwell BM. Hands-on machine learning with R. New York: Chapman and Hall/CRC; 2019. DOI: 10.1201/9780367816377.

[137] Aldhari I, et al. Severity prediction of highway crashes in Saudi Arabia using machine learning techniques. *Applied Sciences*. 2022;13(1):233. DOI: 10.3390/app13010233.

[138] Angarita-Zapata JS, et al. A case study of AutoML for supervised crash severity prediction. In: *Proceedings of the 19th World Congress of the International Fuzzy Systems Association (IFSA), 12th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT), and 11th International Summer School on Aggregation Operators (AGOP)*. Atlantis Press; 2021. p. 187–194. DOI: 10.2991/asum.k.210827.026.

[139] Angarita-Zapata JS, Maestre-Gongora G, Calderín JF. A bibliometric analysis and benchmark of machine learning and AutoML in crash severity prediction: The case study of three Colombian cities. *Sensors*. 2021;21(24):8401. DOI: 10.3390/s21248401.

[140] Toğan V, et al. Customized AutoML: An automated machine learning system for predicting severity of construction accidents. *Buildings*. 2022;12(11):1933. DOI: 10.3390/buildings12111933.

[141] Mostafa SM, Salem SA, Habashy SM. Predictive model for accident severity. *IAENG International Journal of Computer Science*. 2022;49(1). https://www.iaeng.org/IJCS/issues_v49/issue_1/IJCS_49_1_13.pdf [Accessed 6th May 2024].

تأثير موازنة البيانات وهندسة الميزات على نماذج شدة الحوادث

فايز العنزي، أمينو سليمان

الملخص

تُحقق هذه الدراسة في تأثير تقنيات هندسة الميزات، بما في ذلك التجميع، والترميز المستهدف، واكتشاف الشذوذ، إلى جانب أساليب موازنة البيانات، على كفاءة نماذج تعلم الآلة في التنبؤ بشدة حوادث الطرق. تم تقييم النماذج باستخدام التعلم الآلي الآلي، والغابات العشوائية الموزعة، ونماذج الأشجار التكيفية المعززة للتراجع، ونماذج التعلم العميق. استخدمت مجموعات البيانات الموزونة باستخدام تقنيتي الإفراط الاصطناعي للأقليات والتوازن الاصطناعي التكيفي. تضمنت معايير التقييم متوسط مربع الخطأ، والجذر التربيعي لمتوسط مربع الخطأ، والخسارة اللوغاريتمية، والمساحة تحت المنحنى، والمساحة تحت منحنى الاستدعاء الدقيق.

أظهرت النتائج أن التعلم الآلي الآلي يتفوق باستمرار على النماذج الأخرى، حيث حقق دقة بلغت 85% في التنبؤ بالحوادث المميتة و94% في التنبؤ بالإصابات. تميز التعلم العميق في التنبؤ بحوادث الإصابات بدقة بلغت 95%، ولكنه واجه تحديات في التنبؤ بالحوادث المميتة، حيث حقق دقة بلغت 60%. تؤكد الدراسة على الدور الحاسم لتقنيات هندسة الميزات وأساليب موازنة البيانات في تحسين دقة التنبؤ بتصنيف شدة الحوادث. على وجه الخصوص، أدى دمج تقنيات التجميع، والترميز المستهدف، واكتشاف الشذوذ مع تقنيتي الإفراط الاصطناعي للأقليات والتوازن الاصطناعي التكيفي إلى تحسين أداء النماذج بشكل ملحوظ. تظل الحاجة إلى مزيد من التعديل والتحقق أمراً ضرورياً لتحسين أداء النماذج في تطبيقات إدارة السلامة المرورية في العالم الحقيقي.

الكلمات المفتاحية

شدة الحوادث؛ السلامة المرورية؛ تعلم الآلة؛ هندسة الميزات؛ موازنة البيانات؛ التعلم الآلي الآلي؛ التوازن الاصطناعي التكيفي.