applied sciences

MDPI

Article

Predicting Road Crash Severity Using Classifier Models and Crash Hotspots

Md. Kamrul Islam [1,*], Imran Reza [2,*], Uneb Gazder [3], Rocksana Akter [4,5], Md Arifuzzaman [1] and Muhammad Muhitur Rahman [1]

[1] Department of Civil and Environmental Engineering, College of Engineering, King Faisal University, Al-Ahsa 31982, Saudi Arabia
[2] Department of Civil, Architectural & Construction Management, University of Wyoming, Laramie, WY 82070, USA
[3] Department of Civil Engineering, University of Bahrain, Isa Town P.O. Box 32038, Bahrain
[4] Department of Civil Engineering, Dhaka University of Engineering and Technology, Gazipur 1707, Bangladesh
[5] School of Engineering, Monash University Malaysia, Subang Jaya 47500, Malaysia
* Correspondence: maislam@kfu.edu.sa (M.K.I.); ireza@uwyo.edu (I.R.)

Abstract: The rapid increase in traffic volume on urban roads, over time, has altered the global traffic scenario. Additionally, it has increased the number of road crashes, some of which are severe and fatal in nature. The identification of hazardous roadway sections using the spatial pattern analysis of crashes and recognition of the primary and contributing factors may assist in reducing the severity of road traffic crashes (R.T.C.s). For crash severity prediction, along with spatial patterns, various machine learning models are used, and the spatial relations of R.T.C.s with neighboring areas are evaluated. In this study, tree-based ensemble models (gradient boosting and random forest) and a logistic regression model are compared for the prediction of R.T.C. severity. Sample data of road crashes in Al-Ahsa, the eastern province of Saudi Arabia, were obtained from 2016 to 2018. Random forest (R.F.) identifies significant features strongly correlated with the severity of the R.T.C.s. The analysis findings showed that the cause of the crash and the type of collision are the most crucial elements affecting the severity of injuries in traffic crashes. Furthermore, the target-specific model interpretation results showed that distracted driving, speeding, and sudden lane changes significantly contributed to severe crashes. The random forest (R.F.) method surpassed other models in terms of injury severity, individual class accuracies, and collective prediction accuracy when using k-fold (k = 10) based on various performance metrics. In addition to taking into account the machine learning approach, this study also included spatial autocorrelation analysis based on G.I.S. for identifying crash hotspots, and Getis Ord G_i^* statistics were devised to locate cluster zones with high- and low-severity crashes. The results demonstrated that the research area's spatial dependence was very strong, and the spatial patterns were clustered with a distance threshold of 500 m. The analysis's approaches, which included Getis Ord G_i^*, the crash severity index, and the spatial autocorrelation of accident incidents according to Moran's I, were found to be a successful way of locating and rating crash hotspots and crash severity. The techniques used in this study could be applied to large-scale crash data analysis while providing a useful tool for policymakers looking to improve roadway safety.

Keywords: crash severity; crash types; ensembles; machine learning; hotspots

Citation: Islam, M.K.; Reza, I.; Gazder, U.; Akter, R.; Arifuzzaman, M.; Rahman, M.M. Predicting Road Crash Severity Using Classifier Models and Crash Hotspots. *Appl. Sci.* **2022**, *12*, 11354. https://doi.org/10.3390/app122211354

Academic Editors: Moongu Jeon, Kin-Choong Yow and Jeonghwan Gwak

Received: 10 October 2022
Accepted: 2 November 2022
Published: 9 November 2022

1. Introduction

One of the main causes of fatalities among people and property damage worldwide is traffic crashes [1,2]. Global Status Report on Road Safety 2015 gathered data from 180 nations, which claims that, every year, 1.25 million people are killed in traffic crashes [3]. Every day, 3000 people die due to traffic crashes around the world [4]. In another study related to Iran, the number of people who are injured in traffic crashes is reported to be

about ten times more than the number of deceased people, at around 240,000 cases per year. These high rates, combined with the significant share of passenger traffic, necessitate a thorough investigation into passenger safety [5]. For an intelligent transportation system to be deployed successfully and provide appropriate levels of medical care and transportation in a timely way, in the case of a traffic incident, an accurate and quick severity prediction algorithm is crucial. Different agencies may profit from the ability to estimate the severity of a reported incident with unknown severity or the severity of future crashes. Crash severity prediction is one of the most significant aspects of crash management; it helps rescuers assess the severity of traffic crashes and their potential impact and perform effective crash management methods. Injury severity is frequently regarded as a dependent (class/response) variable when studying traffic crash injury severity using historical crash records, whereas other crash components are referred to as independent (predictor) variables. Researchers use crash data to identify risk variables and execute effective interventions to improve road safety. Driver characteristics, roadway characteristics, vehicle characteristics, crash characteristics, and atmospheric elements are all factors that influence the severity and frequency of crashes [6,7]. Crash data are compiled into a massive database that includes various crash-related characteristics. In order to examine the crash database, many analytical methods have been applied in the literature.

Several studies have attempted to forecast the severity of crashes in recent years [8]. Many studies have attempted to analyze and identify the primary factors that influence the severity of road crashes [9]. The strategies most typically employed to undertake such analyses have been linear models, nonparametric models, and data mining approaches. These methods aid in determining the cause of a traffic crash. Statistical methods have been frequently employed to forecast the severity of traffic crashes. The logistic regression model, the ordered probit model, and the mixed logit model have all been proposed by different researchers and have been used to evaluate relevant data from traffic crashes and to examine the impact of various variables on the severity of traffic crashes [2,10–16]. These prediction models aim to predict the severity of traffic crashes. However, according to some researchers [7,17], the majority of regression models have established underlying correlations between dependent and independent variables and their own set of underlying assumptions (i.e., linear relations between the variables). If these assumptions are not met, the model may produce inaccurate estimates of the chance of serious injury. According to the reported literature, machine learning outperforms traditional statistical approaches in prediction, aided by the accessibility to a vast quantity of datasets [18]. Moreover, although crashes are random events that happen in space and time [19], they also exhibit spatial dependency and spatial autocorrelation, which should be considered during their evaluation [20]. Hotspot techniques should be used to manage road safety initiatives that envision minimizing traffic crashes with limited cost, as proven by [21]. Similarly, Australia's national hotspot program helped reduce fatal crashes by 30% [22]. In Belgium and Denmark, hotspot applications showed similar performance [23].

The present study is focused on three objectives. Firstly, developing an accurate model for predicting crash severity for the study area. Such a model will be helpful for medical and emergency services to provide rescue services in a timely manner and analyze crashes with unknown severity. Secondly, identifying parameters that affect and increase the severity of crashes, and thirdly, an approach that seeks to identify crash hotspots by emphasizing places where crashes are most likely to occur under specific conditions has also been determined in this study. For the prediction of road crash severity, tree-based ensemble models (random forest and gradient boosting) and a logistic regression model are examined in this work. In addition to taking into account machine learning approaches, this study also makes an effort to analyze the spatial consequences of crashes in an effort to point out and identify potential hotspots utilizing various variables. As a result, machine learning models can identify hotspots rather than just depending on the number and location of crashes by learning data patterns in variables. The current study is important because it shows the value of using statistical methods and spatial autocorrelation to

identify crash-intensive-prone zones and successfully applies these methods using data from crashes on rural and urban roads in the Saudi Arabian province of Al-Ahsa, over a period from 2016 to 2018. This effort will be beneficial for traffic management authorities to devise effective policies for reducing traffic crashes and mitigating their impacts.

The rest of this paper Is structured as follows. A summary of the most recent research in this field is given in Section 2. Data source and data processing are explained along with descriptive statistics in Section 3. Machine learning models taken into account, in this paper, are described in Section 4. Section 5 presents exploratory data analysis, while Section 6 discusses the study's findings. Section 7 concludes with a brief summary of the findings and suggestions for future research.

2. Literature Review

Traditional methods for predicting collisions and classifying their seriousness use statistical modeling [24]. Poisson, binomial, Poisson–lognormal, negative binomial, gamma, and zero-inflated regressions, negative multinomial models, generalized estimation equations, random effects models, and random parameters models have all been applied in this situation [25]. The Bayesian hierarchical binomial logit, Bayesian-ordered probit, log-linear model, extended-ordered logit, multinomial logit, multivariate probit, ordered probit, and ordered logit are among the models that have also been utilized [26] for the estimation of crash severity. Despite the advances made with these techniques, statistical modeling has limitations because each model includes its predetermined correlations between the dependent and independent variables a and assumptions [27]. Statistical modeling, according to [28,29], necessitates assumptions about data distribution. Such premises could be false and, as a result, violated. On the other hand, the drawbacks of this strategy have been extensively investigated, presenting an opportunity to employ fresh approaches. Machine learning methods, such as decision trees (D.T.s), nearest neighbor classification (K.N.N.), support vector machines (SVMs), evolutionary algorithms (E.A.s), deep learning models, and artificial neural networks (ANNs), have been employed to analyze data for a variety of road safety issues and devised as data analytic methods due to their capacity to handle enormous volumes of multidimensional data and overcome the shortcomings of statistical methodologies. Furthermore, the flexibility of machine learning modeling techniques, generalization and learning capabilities, and great predictive capacity led to its acceptance as an accurate, generic, and practical mathematical model in road safety.

ANN is a useful tool for solving problems in a variety of fields. It can be utilized for traffic incident detection [30], public transportation [27,31], and road planning [32]. ANN is a complex, non-linear, parallel processor with a natural tendency to store and retrieve experimental knowledge [33]. The possible non-linear relationships between injury severity levels and crash causes were modeled by Delen et al. [34] using a number of ANN models. Bayesian neural network (B.N.N.) and ANN models have been used to research road safety concerns for many years. Although the multilevel network architecture of ANN and B.N.N. models are similar, they differ in their capacity to predict outcome variables [35]. Moghaddam et al. [36] employed ANNs to predict and estimate crash severity in urban roadways, as well as to uncover major crash-related parameters. Variables, such as roadway width and head-on collisions, were found to have an impact on the outcomes.

The effectiveness of B.N.N. models for forecasting road crashes was evaluated by Xie, Lord, and Zhang [37] using data gathered on rural areas in Texas. Ona et al. [17] used a Bayesian network to identify the elements that influence injury severity, which was divided into two categories: slightly hurt and killed/severely injured. Simoncic [38] created a Bayesian network for injury severity analysis. The findings revealed that Bayesian networks could be used to model road crashes. It also highlighted some benefits of using a Bayesian network over a regression model, such as the ability to include more variables and a larger data set.

The statistical learning theory underpins the support vector machine (SVM) approach [39]. SVM models [40,41], a novel class of models based on structural risk reduction

and statistical learning theory [42], have lately been developed for road safety studies. Furthermore, the findings revealed that SVM models outperform B.P.N.N. models (or are at least equal) and do not overfit the data. Nearest neighbor classification (K.N.N.) is a basic yet groundbreaking machine learning algorithm. In a prediction problem, the K.N.N. classifies an observation based on the k-nearest observations [43].

For classification problems, the D.T. technique is highly beneficial. A training set of inputs and outputs (i.e., classes) is produced during the development of a tree. D.T.s are data mining techniques [44,45], which can be utilized to address classification difficulties. One of the key advantages of D.T.s is that their structure allows for the extraction of 'if-then' decision rules (D.R.s). Such DRs may reveal behaviors that are unique to a particular data collection. They are not only practical but also simple to comprehend from the standpoint of safety analysis [46]. Evolutionary algorithms (E.A.s) are based on natural selection processes, in which only the strongest individuals survive [47], and use stochastic search strategies. The two fundamental types of evolutionary algorithms are genetic algorithms and genetic programming, and they are particularly useful for optimizing problems typically related to other approaches.

Using vast data gathered on unsignalized crossings in Florida, multivariate adaptive regression splines (MARSs), a recently created machine learning technique, was used by Abdel-Aty and Haleem [48] to predict vehicle angle collisions. The data showed that MARS performed better than the N.B. models. After using R.F. to screen the covariates, the suggested MARS models produced promising results. According to the data, MARS appears to be an effective method for forecasting collisions at unsignalized intersections.

Deep learning is a novel approach that emerged from the advancement of machine learning. It examines many layers of non-linear information, both supervised and unsupervised, to interpret or categorize patterns [49]. A kind of machine learning called "deep learning" creates models that can extract traits from the most fundamental to the most apparent layers [50]. This capability eliminates the difficulty of raw data processing, which is a typical issue with many machine learning algorithms. Deep neural networks (D.N.N.s), convolutional neural networks (CNNs), and recurrent neural networks (R.N.N.s) are the most common examples of this approach, which are used in speech recognition, visual object recognition, and object detection [44]. In contrast to earlier machine learning techniques, which were constrained in their ability to read natural data in its raw form, deep learning develops computer models to extract inherent qualities from data at various levels. Deep learning applications in the field of transportation are limited, with only one study focusing on traffic flow prediction [51], despite the fact that deep learning methods have proven exceptional performance in a number of applications [52].

Moghaddam et al. [36] employed ANNs to predict and estimate crash severity in urban roadways, as well as to uncover major crash-related parameters. The results showed that the most important factors that increase crash severity on urban highways include variables such as highway width, head-on collisions, the type of vehicle at fault, disregarding lateral clearance, keeping track of distance, inability to control the vehicle, exceeding the permitted speed, and driver deviation to the left. Sharma et al. [51] used SVM and multilayer perception for predicting crash severity. They only used a small number of data samples to test their hypothesis. They looked at only two variables: speed and alcohol as major factors in car crashes. The SVM outperformed with a 94% accuracy rate. They stated that high-speed driving while inebriated was the cause of the crash. Tiwari et al. [53] used decision tree (D.T.), N.B., and SVM, as well as S.O.M. and k-modes, for clustering. They obtained superior results with the clustered dataset.

For traffic agencies, AlMamlook et al. [54] employed N.B., AdaBoost, R.F., and logistic regression (L.R.) to locate highways with high crash risk. They used A.U.C., R.O.C., recall, precision, and F-measure to evaluate their models. The R.F. outperformed with a 75% accuracy rate. In another study, Beshah and Hill (2010) [55] investigated crucial roadway-related characteristics that can influence the severity of road crashes. They created decision rules for road safety measures using D.T., NB, and K.N.N. They were primarily concerned

with drivers and pedestrians, with no regard for other factors such as time or speed. They did not report the accuracy parameters of their machine learning models in detecting crash severity risks.

A developing nonparametric tree-based model called the boosted regression tree model can capture the non-linear effects of both discrete and continuous variables without preprocessing the data. Using 5-year crash information for provincial highways in Ontario, Canada, Lee and Li (2015) [56] employed the boosted regression tree model to examine the severity of driver injury in addition to other nonparametric models. The boosted regression tree model's findings revealed a substantial correlation between the severity of driver injuries and vehicle ejection and head-on crashes. They found that for both single-vehicle and two-vehicle crashes, the boosted regression tree model predicted driver injury severity more accurately than the classification and regression tree model.

In order to determine the variables influencing the severity of pedestrian crashes in the Melbourne metropolitan region, where mid-block collisions accounted for 46% of all pedestrian collisions and mid-block collisions were the scene of 49% of pedestrian fatalities, Toran Pour et al. (2017) [57] developed three models employing various decision trees (D.T.s). They employed bagging and boosting approaches to enhance the D.T.s' accuracy, stability, and resilience. According to the study's findings, the boosting strategy increased each D.T. model's accuracy by 46%. Additionally, the results of raising D.T.s demonstrated that neighborhood social traits were just as significant in impacting the severity of pedestrian collisions as traffic and infrastructural factors.

A gradient boosting decision tree (G.B.D.T.) model was presented by Wu et al. (2019) [58] to examine the combined effects of crash-causing elements on four road crash indicators (i.e., injuries, deaths, number of crashes, and the financial loss). The economic, demographic, and road network conditions of Zhongshan, China, from 2000 to 2016, are studied using a total of 27 detailed influential elements. The findings demonstrate that the G.B.D.T. outperforms other conventional machine learning approaches in terms of prediction accuracy, handling multicollinearity across explanatory variables and, more crucially, ranking the influential factors on road crash prediction. The outcomes also demonstrate that there are parallels and variances among the most important determining factors for the crash indicator. Moreover, Wu et al. (2021) [59] applied a scaled stacking gradient boosting decision tree (SS-GBDT) for predicting bus passenger flow, which can also be used for road crash prediction. To analyze non-intersection accident severity data, Pande and Abdel-Aty (2009) [60] used association rule (or market basket analysis) learning directly for the first time. The analysis's findings provide straightforward criteria that show which crash features are related to one another. Using non-intersection crash data from the state of Florida for 2004, the application is illustrated. Later, De Ona et al. (2013a) [9] used association rules to draw out practical decision rules from the output of D.T. modeling. The percentage of samples that fit the rule among those that solely fit the left side, confidence, and lift was used to establish the importance of each generated rule. Support measures how often the rule occurs in the dataset (the statistical dependence of the rule). Other researchers (Montella, Aria; Montella, De O'a, Mauriello, Riccardi, and Silvestro, 2020) [61] extracted decision rules in the form of IF–THEN rules from CART results, where "IF" refers to the statuses of various independent variables and "THEN" is the status of the corresponding crash severity class variable. They contrasted the rules produced through association rule learning with the IF–THEN rules. The outcomes demonstrate the consistency of the two sets of regulations. Discovering relationships between independent factors and crash severity outcomes can be conducted by applying association rule learning, which is a straightforward process. The influence of independent factors on the seriousness of crashes in each rule cannot be measured, and its forecast accuracy is restricted in comparison to other ML techniques.

The breadth of traffic crash severity prediction is the subject of this research. The goal of traffic crash data mining research can be separated into two categories: (1) predicting the severity of traffic crashes and (2) identifying relevant elements determining the severity of crashes. The severity of road crashes is reliably predicted using ensemble learning

algorithms in this article. Gradient boosting machine and RF are two tree-based ensemble learning models employed in this investigation. Additionally, this study not only considers machine learning methods but also makes an effort to assess the spatial effects of crashes in order to highlight and pinpoint possible hotspots using various variables. Thus, the research contributes to the field of road safety by offering a model that would identify the important features for estimating, preventing, and forecasting crashes, particularly in the study region.

3. Data Collection and Data Processing

Al-Ahsa, the largest oasis in Saudi Arabia and the largest oil field in the world, is the largest governance of the eastern province of Saudi Arabia that stretches from Kuwait at 29°20′ N. to the southern tip of the Gulf of Bahrain at 25°10′ N (Figure 1). The case study is conducted in the Al-Ahsa region. This city was chosen for this study due to the city's high crash rate, which has been documented in recent literature [62]. Al-Ahsa recorded 31.9 percent of crashes in the Eastern Province on average between 2009 and 2016, the highest crash rate among other cities in the eastern region (Figure 2). Other cities reported fewer than 5% crashes. This unmistakably shows that cities in Al-Ahsa are significantly vulnerable to traffic crashes. It is noteworthy that UNESCO has recently listed the Al-Ahsa as a heritage site in Saudi Arabia. Thus, the city was recently recognized as a UNESCO-listed heritage site in Saudi Arabia, giving this city tremendous potential to become a popular international tourist destination. However, the high crash rate in this city may create adverse effects on such potentiality. Thus, it requires a proper response.

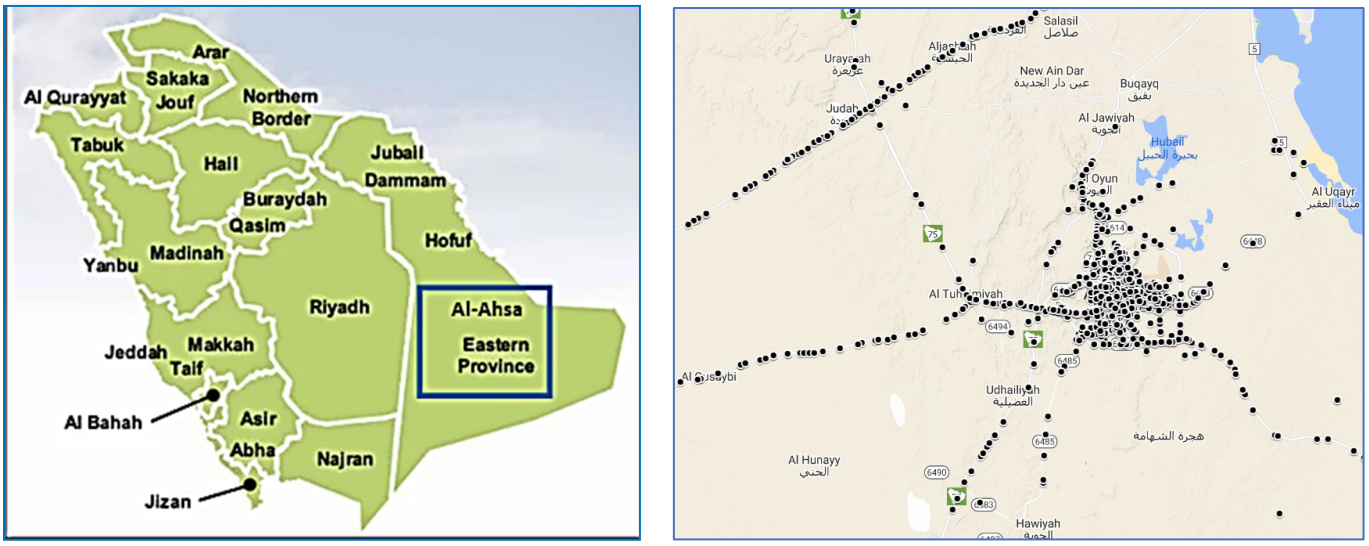

Figure 1. Study area and crash sites in Al-Ahsa between 2015 and 2018. (Adopted from [63]).

The crash data for Al-Ahsa was collected from the Traffic Control Authority and used in this study from October 2014 to May 2018. A total of 3994 crashes were reported in this city during that time. Vehicle collisions were the most frequent type of mishap, accounting for around 8 and 30%, respectively, of incidents that resulted in fatalities and injuries, according to an overall review of the data set (Figure 3). Another common crash type that resulted in 6.5 percent of death and 12.5 percent of injury incidents was vehicle overturning. Additionally, 12.5 percent of injury crashes and 2 percent of fatal crashes reflect the vulnerability of pedestrians. The probability of injury crashes is demonstrated by a number of incident types, such as collisions with parked cars, road railings, motorcyclists, and stationary objects, where fatal crash rates were very low.

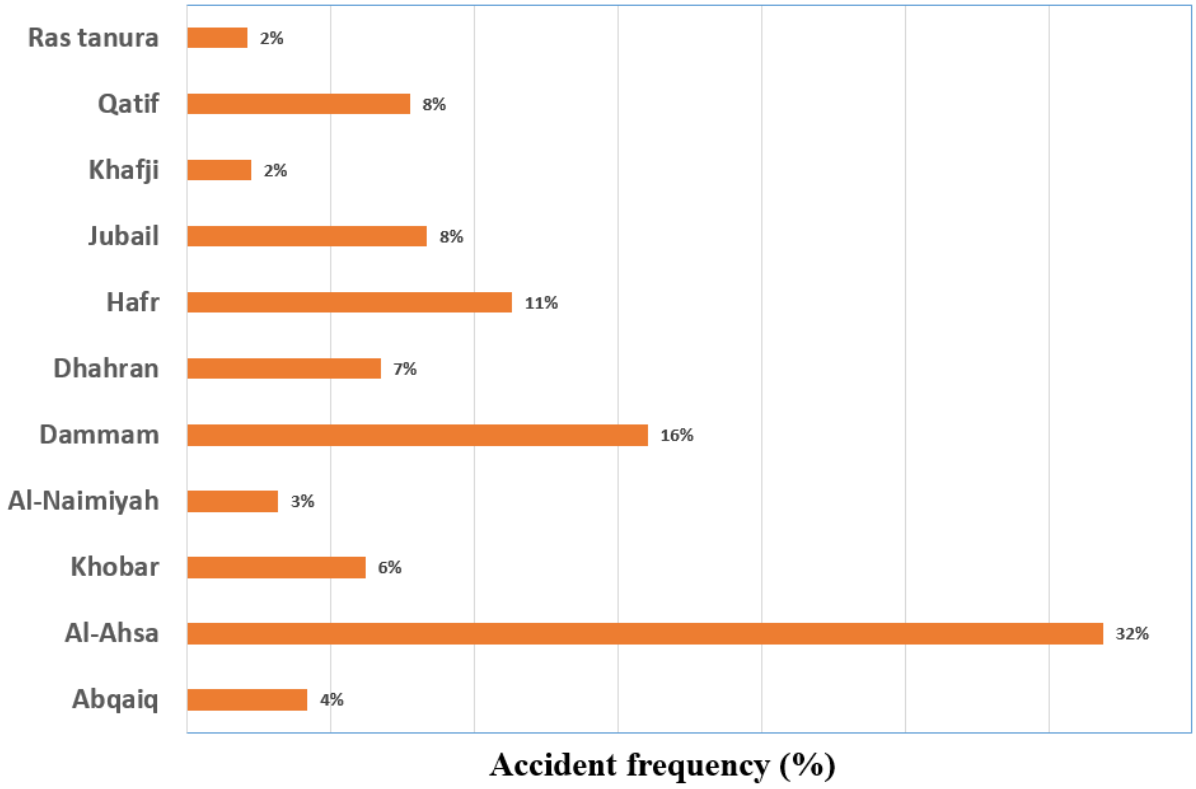

Figure 2. Frequency of road crashes in different cities in the Eastern region of Saudi Arabia (adopted from [62]).

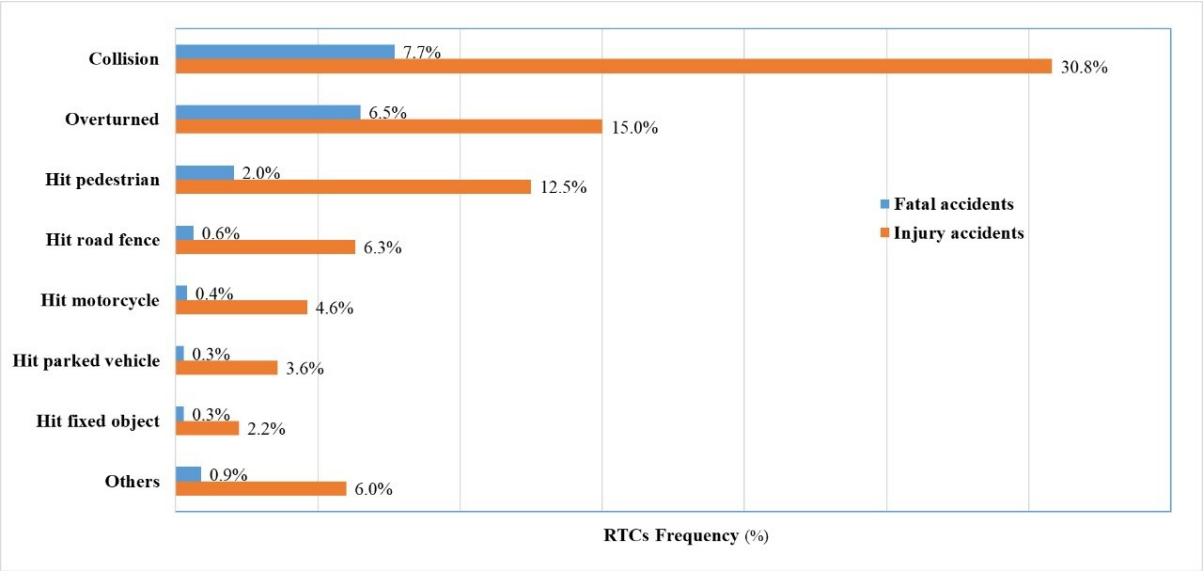

Figure 3. Crash frequency with crash types in the research area. (Adopted from [63]).

Data preprocessing and cleaning is an essential preliminary step for crash analysis using machine learning algorithms. The raw data had 8 different features with 8027 data points. The data were filtered in several steps. First, all the dates had to be converted from Hijri to Gregorian, and then three different columns were made for day, month, and year. Secondly, one of the features from the feature containing similar information was considered. One of the downsides of the data was that it had only two classes of injuries, namely, fatal and serious injuries. Because there was an imbalance in the target class, under-sampling technique was applied for the logistic regression model. The other two

classification models did not require treatment for data imbalance. The feature's number of deaths and injuries had some missing values. These data could not be imputed as they might have produced misleading inferences with the model accuracies. Missing data points were discarded from the final data set used for modeling.

Descriptive Statistics

A total of 4093 crashes from 2015 to 2018 made up the dataset used for the analysis. A total of 9031 people were hurt in the collisions. Nineteen percent were fatal crashes, and the remaining was severe injuries. The three leading causes of collisions—sudden turning, speeding, and failure to yield—accounted for roughly 47 percent, 18 percent, and 15 percent of collisions, respectively. Within the available dataset, the pattern of crashes over time demonstrates a decline in serious injury and fatal crashes, as illustrated in Figure 4. In their study, which similarly focused on Saudi Arabia, the authors of [64] ascribed the decline in serious crashes due to the installation of roadside cameras.

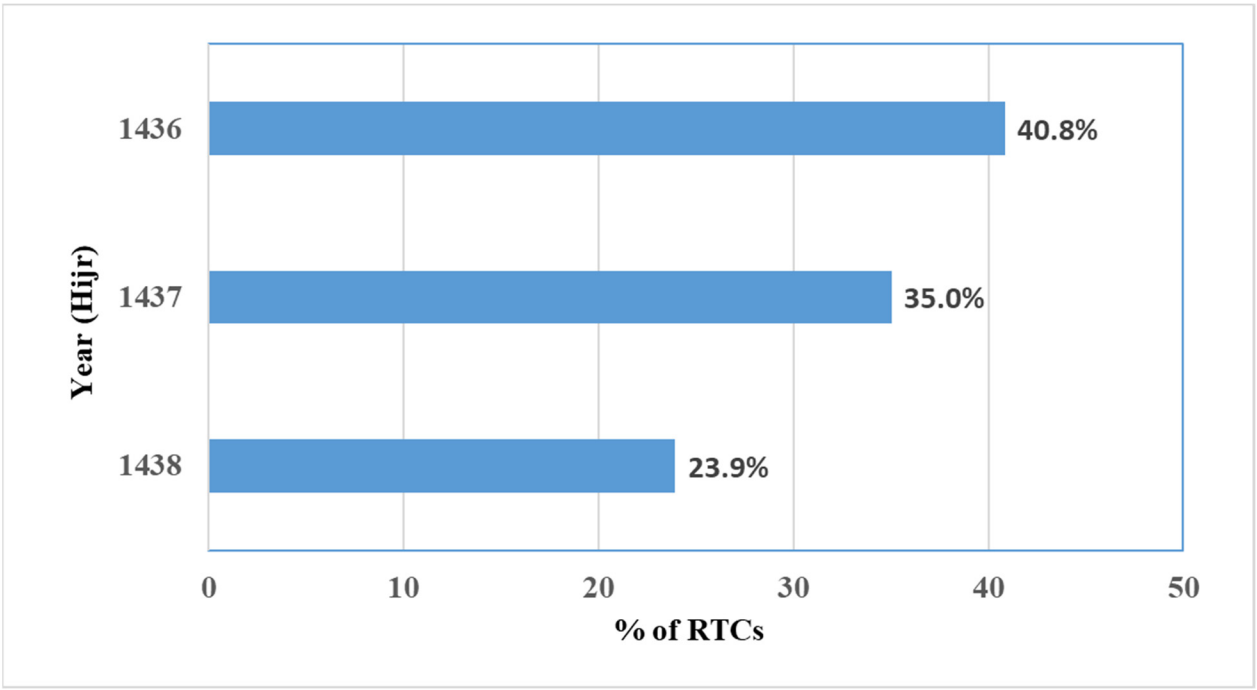

Figure 4. Chronological trend of serious injury and fatal crashes in the research area (adopted from [65]).

Fatality and serious injury R.T.C.s involving moving vehicles made up about 38 percent of the total number of crashes in the dataset that was made accessible. Numerous vehicle crashes (21 percent of R.T.C.s) and running over the road were two more common collisions (14 percent R.T.C.s). The authors of [62] cited the first two crash categories as being typical of those reported in the Eastern Province. This Saudi Arabian governorate includes the study region from which the data for this study was gathered.

The available dataset's descriptive statistics for the number of people who suffered fatalities and significant injuries in R.T.C.s are shown in Table 1. Because practically all R.T.C.s involve, on average, at least one fatality victim, the severity index is relatively high. Approximately 1.15 million people reside in the Al-Ahsa region [66]. Comparing Table 1's numbers to those from other research in the literature, they are high. According to the authors of [67], there were just 3.3 fatal crashes per year in Iowa city of U.S.A., as opposed to 352 in Al-Ahsa, in the available data that were made accessible.

Table 1. Statistics on serious injuries and fatalities in the study area (adopted from [65]).

Parameter	No. of Serious Injuries	No. of Deaths
Number of R.T.C.s	3245	756
Average per R.T.C. *	7	4
Total	22,787	3579
Range	0–46	0–13
Severity index = Deaths/No. of R.T.C.s	0.89	
The annual number of fatal R.T.C.s	352	
Number of annual fatalities	1193	
Deaths per 100,000 individuals	311	
R.T.C.s with fatalities and major injuries each year	1333	

* Number of serious injuries/number of RTCs with serious injuries; same was used for the average deaths per RTC.

Another study [68] found that the study area had 311 deaths per 100,000 persons, compared to 14.93 deaths per 100,000 in Columbia. It is interesting to note that the crash rates per year for the research area were not found to be more significant when compared to earlier studies. The yearly fatal crash rate in a study conducted by the authors of [67] was roughly 350, which is close to the number for the research area. Although the authors of [69] did not mention serious injury and fatal crash rates per year, the following studies further supported their argument. For fatal and significant injury R.T.C.s, Abdi et al. [70] recorded an annual rate of 1463 crashes for Addis Ababa, while [69] recorded an annual rate of 3631 for Torino (Italy). On the other side, there were 1333 fatal and serious injury R.T.C.s annually in the research area.

This illustrates how R.T.C.s are greater in severity and involve more victims in the research area, although having lower occurrence rates than in other regions of the world. Further details about the crash dataset can be found in [63,65].

4. Machine Learning Models

Crash severity models involve a non-linear relationship between contributing factors, for which statistical models often become insufficient to explain the inner and intrinsic correlations at the same time; the statistical model usually has some predefined model assumptions, which may lead to an invalid model if violated [71]. Another issue with using a statistical model is determining which explanatory variables to include in the model using large-scale multivariate datasets. Machine learning (ML) models are increasingly used to overcome the limitations of statistical models for estimating the non-linear correlations between crash contributing factors and injury severity [60]. ML models have advantages over statistical models when dealing with outliers and noisy data. The proceeding subsections provide a few details related to the development algorithm of ML models used in this study, namely, random forest (R.F.), X.G. boosting and logistic regression.

4.1. Random Forest (R.F.)

R.F. is an ensemble learning algorithm that improves its prediction accuracy by stacking a considerable number of classifiers. The classifiers are the D.T., which are grouped, and their grouping is referred to as a forest, whose results are combined, thereby yielding better prediction results. R.F. algorithm is initiated with randomly selecting data points from the complete set, which is then used to construct the D.T.s. Lastly, collective voting is performed for results obtained from each D.T. The outcome class is assigned using the result with the highest votes [72].

Either the Gini index or entropy index can be used, as shown in Equations (1) and (2), to build nodes within D.T.s, representing the best splicing of data with a maximum distance between its branches [73].

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2 \quad (1)$$

$$Entropy = \sum_{i=1}^{C} -p_i * \log_2 p_i \quad (2)$$

where p_i is the relative frequency (or proportion) of the observed class and C represents the number of classes in the dataset. The final R.F. result can be expressed mathematically by Equation (3).

$$RF = argmax_{j\epsilon\{1,2,...C\}} \sum_{i=1}^{i} DT_{i,j} \quad (3)$$

where the argmax function represents the max value or majority vote, i is the D.T. number starting at one and ending with ith D.T., and j represents the number of classes available in the outcome feature.

Due to the incorporation of several D.T.s, the R.F. model can attain high precision values and avoids the overfitting problem. It can also be applied to datasets with gaps. However, the prediction process requires a longer time due to the length of the D.T.s, which also increases its complexity compared to other machine learning techniques, such as K.N.N. [74].

4.2. eXtreme Gradient Boosting (XGBoost)

XGBoost utilizes gradient descent optimization and any differentiable loss function for development. It is a scalable model as it enhances the computational limits of the machines. In this model, ensembles are built from the D.T.s, which are combined, one at a time, with adjustments made to reduce the prediction errors resulting from the prior models. XGBoost is reported to have high execution speed and better performance compared with R.F. models. However, it employs its own technique of constructing trees where two metrics are used to specify the best node splits: similarity score and Gain. The similarity score (S.C.) is expressed mathematically by Equation (4) [75].

$$Similarity\ Score\ (SC) = \frac{\left(\sum_{i=1}^{n} Residual_i\right)^2}{\sum_{i=1}^{n}\left(P_{pre_i} * \left(1 - P_{pre_i}\right)\right) + \lambda} \quad (4)$$

where *Residual* represents the difference between the actual value and the predicted value. P_{pre} represents the probability of an event calculated at a previous step. λ is a regularization parameter. After finding the similarity core for each leaf, the Gain is evaluated using Equation (5).

$$Gain = Left\ leaf_{SC} + Right\ leaf_{SC} - Root_{SC} \quad (5)$$

Based on the gain value, the node split with the highest value is then selected for the updated D.T. Additionally, XGBoost can be further refined by setting up the value of a hyperparameter called *Gamma*, which can be set manually or tuned using the grid search method. This hyperparameter can be used to prune the nodes with minimal *Gain* value, provided that they have $Gain - Gamma < 0$ [76].

4.3. Logistic Regression Model

Equation (6) illustrates how the concept of maximum likelihood is used to create logistic regression models. These models develop a linear functional form utility function, as seen in Equation (7). The coefficients of the utility function are estimated while maximizing

the log-likelihood function. The probability of an outcome is calculated by subjecting the utility function to a logistic function (see Equation (8)).

$$Modellikelihood = Max[\sqcap P_i(n)] \quad (6)$$

$$U_n = b + \sum_{j=1}^{I} Y_j X_j \quad (7)$$

$$P_i(n) = 1/(1 + e^{-Un}) \quad (8)$$

where $P_i(n)$ is the probability of ith sample in the dataset (of 'I' values) to have the outcome 'n', which is the actual outcome, b and Y are intercepts and coefficients for the utility function, respectively, and X is an array of independent variables having significant impact on the model [77]. Logistic regression models are very commonly used in the prediction of categorical responses, which includes predicting crash severity [78].

5. Exploratory Data Analysis

Spatial analysis of crash patterns is conducted in the study area. The maps in Figure 5 show the dataset's spatial distribution of fatal and other severe crashes. It could be observed from these figures that these crashes are primarily concentrated in the central part of the city. Other than that, there are a significant number of crashes on the highways that are used for intercity travel. In both cases, the locations of high crash frequencies correspond with locations of higher traffic volumes. Other researchers have also confirmed this trend in the past for other cities, such as [79]. The frequency of the crash reason is presented in Table 2. A total of 48 percent of the crashes were attributed to sudden turning. Speeding and not giving way were also among the top reasons of the crash. Tailgating and distractions of the drivers were also found to be noticeable reasons for a crash, with frequencies of 6.4% and 5.75%, respectively.

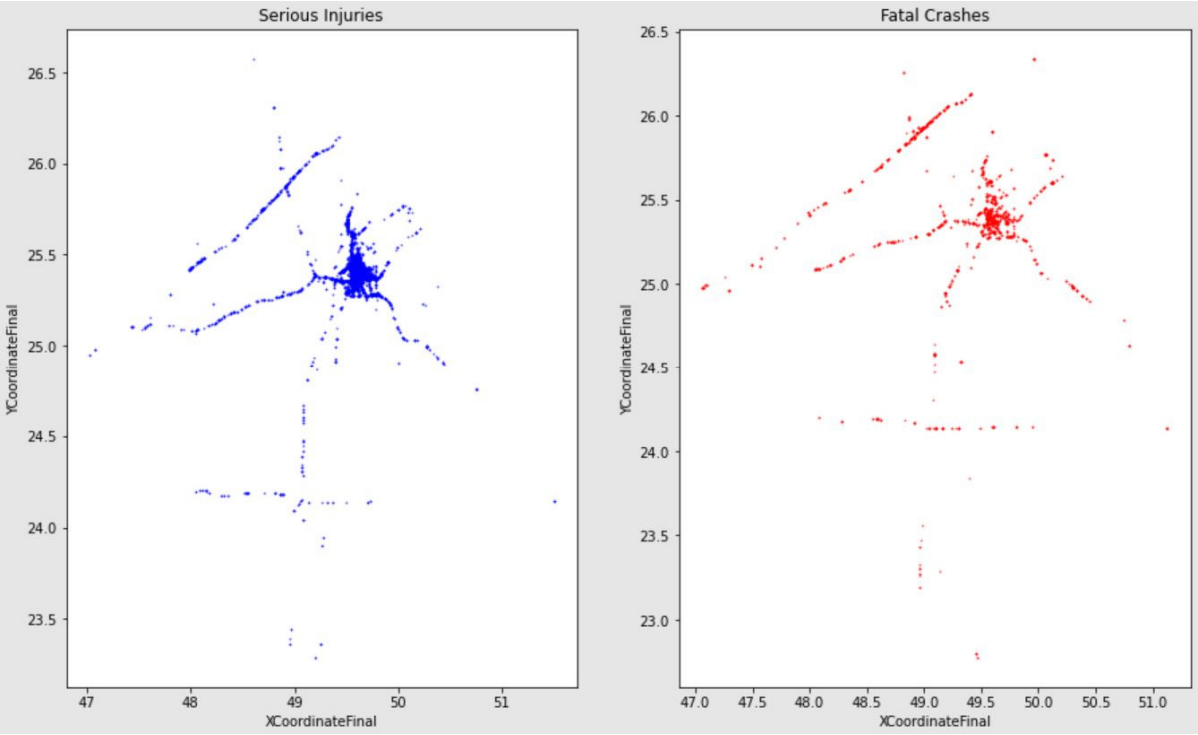

Figure 5. Map of fatal and serious injury crashes in Al-Ahsa.

Table 2. Frequency of crash reasons.

Crash Reasons	Frequency of Crash Reasons
Sudden lane changes	48.00%
Speeding	18.00%
Not giving way	15.00%
Insufficient safe distance	6.40%
Driver distraction	5.75%
Other	2.35%
Faulty tires	1.15%
Not using pedestrian crossing	1.00%
Illegal overtaking	0.55%
Red light violation	0.30%
Driving opposite to traffic	0.25%
Not stopping at STOP sign	0.20%
Drifting	0.10%
Falling asleep	0.05%
Getting out vehicle before stopping	0.05%
Hanging on the outside of vehicle	0.05%
Unsafe road works	0.05%
Exhaustion	0.02%
Violating pedestrian sign	0.02%
Downhill	0.02%
No warning signs	0.02%
Faulty breaks	0.02%
Faulty electrics	0.02%

5.1. K-Fold Cross-Validation

K-fold cross-validation is the most commonly applied process for model estimation. In this process, the available data are divided equally into 'K' parts. Each part is used for model validation (testing) in turn, while the remaining parts are used to train the model. In the end, the best model is the one that has better average performance for all different sets of validation subsets [80]. This process resolves overfitting issues and makes the model robust for application on new datasets [81].

5.2. Hyperparameter Optimization for Each Model

Characteristics of the model, which enable their customization to perform a certain task, are called hyperparameters [82]. There is no clear way to set them for a given set of data because these hyperparameters may interact in non-linear relations. Therefore, a search algorithm is required to set the optimum set of these parameters for any data set. This process is also known as hyperparameter tuning.

The optimization process includes the definition of a search space of n-dimensions, where 'n' is the number of parameters to be optimized and the size of each dimension is the possible range of values of alternatives that may be assigned to that hyperparameter. The outcome of the optimization process is a vector of hyperparameters which enables the best model performance.

Among the many algorithms available for optimization, random and grid search are considered the common and most robust procedures [83]. During the random search process, hyperparameter values are chosen randomly/arbitrarily, whereas in the grid

search, the space is divided into a grid network and values are evaluated in a systematic manner covering all cells of the grid. This study used the randomized search method to obtain the optimized hyperparameters that would produce the best accuracy for a specific model. With the random search, the user can set up a grid of hyperparameter values to select random combinations for training and evaluating the model. Using random search, we were able to explicitly control the number of parameter combinations. Table 3 shows the parameters, ranges, and optimized values.

Table 3. Optimized hyperparameters for ML classifier models.

ML Model	Parameters	Range	Best Values
Logistic Regression	penalty	{l1, l2, elasticnet, none}	L2
	C	Default 1.0	1.38
	Solver	{'newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'}	Lbfgs
	multi_class	{'auto', 'ovr', 'multinomial'}	ovr
Random Forest (R.F.)	n_estimators	{0, 75, 100, 200, 300}	75
	max_features	{'auto', 'sqrt', 'log2'}	auto
	max_depth	{5, 10, 15, 20, 25}	5
	criterion	{'gini', 'entropy'}	'gini'
XGBoost	n_estimators	{50, 75, 100, 200, 300}	100
	learning_rate	{0.3, 0.01, 0.01, 0.05}	0.3
	max_depth	{5, 10, 15, 20, 25}	10
	gamma	{0.5, 1, 1.5, 2}	1
	booster	{'gbtree', 'gblinear'}	'gbtree'

5.3. Model Evaluation

The most popular performance indicators were used in this study to evaluate the effectiveness of the various strategies. These include the F1 score, precision, recall, and confusion matrix. The confusion matrix for classification issues consists of the four scenarios depicted in Table 4: true (T.P.) positive rate, true negative (T.N.) rate, false positive (F.P.), and false negatives (F.N.) rate. The accuracy of a sample's classification or prediction is measured as a percentage of all samples. Equation (10) is used to calculate the metric sensitivity. Similar calculations can be made for other performance metrics such as accuracy, F-measure, and specificity using Equations (9)–(13).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$Recall/Sensitivity = \frac{TP}{TP + FN} \tag{10}$$

$$Specificity = \frac{TN}{TN + FP} \tag{11}$$

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$F - Score = \frac{Precision * Recall}{Precision + Recall} \tag{13}$$

Table 4. Confusion matrix for evaluating model's performance.

	Prediction Condition	
Actual Conditions	**Positive**	**Negative**
Positive	True Positives (T.P.s)	False Negatives (F.N.s)
Negative	False Positives (F.P.s)	True Negatives (T.N.s)

6. Results and Discussion

A total of 8027 automobile collisions occurred throughout the study period, of which 21.7% resulted in fatalities and 78.3% in serious injuries. To create models that predict the degree of injury severity in collisions, eight predictor variables (attributes) were combined with the class variable of injury severity. The work also seeks to assess the importance of qualities (relative variable importance) in causing a crash.

The confusion matrix produced by each classification approach on the test data is summarized in Table 5. The confusion matrix for each class displays how instances belonging to that class are classified. The diagonal of the contingency table contains all of the correctly classified samples. Consequently, it is possible to visually check the matrix for errors.

Table 5. Confusion metrics for the classifier modes.

Classifier	Actual Class	Predicted Class	
		Fatal	Serious Injuries
Random Forest	Fatal	268	80
	Serious Injuries	13	1245
XGBoost	Fatal	254	94
	Serious Injuries	24	1234
Logistic Regression	Fatal	17	331
	Serious Injuries	12	1246

The confusion matrix was used to determine the various values of the classifiers' accuracy measure, which were then listed in Tables 4 and 5 to demonstrate how well the classification algorithms performed in relation to the crash in Al-Ahsa. The prediction accuracies for each class for all three classification algorithms are listed in Table 6. It expressly displays the sensitivity, specificity, accuracy, and recall results for all three machine learning approaches that were acquired using 10-fold cross-validation. The classifiers' performance metrics are provided in Table 6. The models were developed in a Python environment with Scikit-learn library. Random forest classifier achieves an accuracy of 94.00%, with a precision of 0.95 and 0.94 for fatal and serious injuries, respectively. The accuracy is 93.00%, with a precision of 0.91 and 0.93 for fatal and serious injuries when XGBoost was used, whereas using the logistic regression, the accuracy is 79.00% with a precision of 0.59 and 0.79 for fatal and serious injuries, respectively.

Table 6. Performance metrics for the classifier modes.

Classifier	Severity Class	Sensitivity	Specificity	Precision	Recall	F1-Score
Random Forest	Fatal	0.770	0.989	0.95	0.77	0.85
	Serious Injury			0.94	0.99	0.96
XG Boost	Fatal	0.730	0.981	0.91	0.73	0.81
	Serious Injury			0.93	0.98	0.95
Logistic Regression	Fatal	0.050	0.990	0.59	0.05	0.09
	Serious Injury			0.79	0.99	0.88

Comparing the accuracy parameters of these models reveals that the R.F. model had the best prediction accuracy in terms of recall, precision, and F1 score. The random forest model outperformed the other models by a small margin (F1 score—R.F. model: fatal 0.85 and serious injury 0.96; XGBoost: fatal 0.81 and serious injury 0.85). Technically, it can be inferred that the important feature of R.F. was better than the XGBoost; however, the difference in the feature list was not very significant. With slightly superior performance metrics than logistic regression, which came in last in the performance metrics ranking, XGBoost was the next-best classification algorithm. This clearly advocates the use of machine learning techniques in place of statistical techniques for predicting crash severity.

According to the gain ratio evaluator derived using random forest and XGBoost models, Figures 6 and 7 depict the relative priority ranking of each feature. Higher feature relevance is implied by an evaluator with a high value. The R.F. model suggests that explanatory variables that heavily contribute to the crash injury severity outcome are: faulty tires, hitting a moving vehicle, not giving way, running over, and sudden turning. Whereas the XGBoost model also produced similar important features with faulty tires, not giving way, failing to stop at a stop sign, and sudden turning among the top-ranked features responsible for the severe crash injury. In case of both machine learning models, age had relatively lower impact on the prediction of crash severity. However, with slight differences in the importance of other parameters. Because R.F. model showed better performance, in terms of predicting severity, hence, it would be reasonable to assume that the feature importance shown by this model is more reliable.

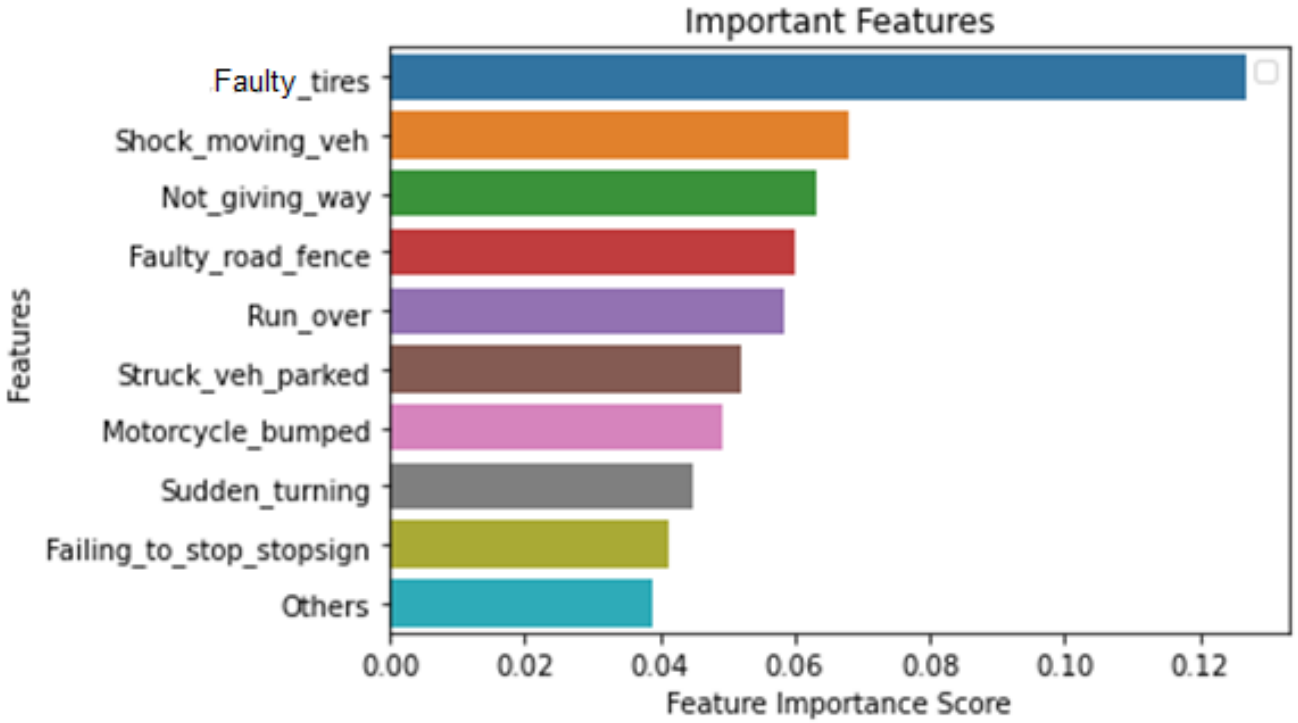

Figure 6. Variable importance using random forest classifier.

Finally, the classification model results for crash injury severity in Al-Ahsa were mostly in line with other regions' research on crash injury severity. Regarding the causes of the crash, our study from the same area indicated that the type of collision, the state of the road, the illumination, and speeding have the greatest effects [79].

Spatial Autocorrelation and Hotspot Analysis

The spatial distribution patterns of traffic accidents in the Al-Ahsa Region were identified using spatial autocorrelation (Global Moran's I). If the characteristics are spatially clustered, dispersed, or randomly distributed, it is shown by the values of Moran's I [84]. Equation (14) was used to calculate the Global Moran's I, z-score, and p-value using the spatial autocorrelation

$$I = \frac{N \sum_i \sum_j W_{i,j} (X_i - \overline{X}) (X_j - \overline{X})}{(\sum_i \sum_j W_{i,j}) \sum_i (X_i - \overline{X}) (X_j - \overline{X})^2} \quad (14)$$

where X_i is an attribute value of the target feature at location i, N is the total number of features, $W_{i,j}$ is the spatial weight between features at locations i and j, and the neighboring

feature at locations i and j has an attribute value of X_j [84]. The results of Moran's I statistic can be verified using z scores, where a confidence level is established. The temperature of cell I is similar to the crash characteristic of the nearby cells if cell I has a significant positive value (i.e., a positive number). I denote a substantial clustering range if it is a large positive value. The cell surface temperature of cell I is significantly different from the surrounding cells if the value of I is negative and significant, on the other hand, which denotes a negative spatial correlation. We can locate hotspots based on the dispersion of these sites. It assesses if the pattern expressed is random, clustered, or scattered using an associated attribute and the offered collection of features. Figure 8 displays the spatial autocorrelation graph using Moran's I statistics of the z-score and p-value for all traffic accidents between 2016 and 2018 in the Al-Ahsa region.

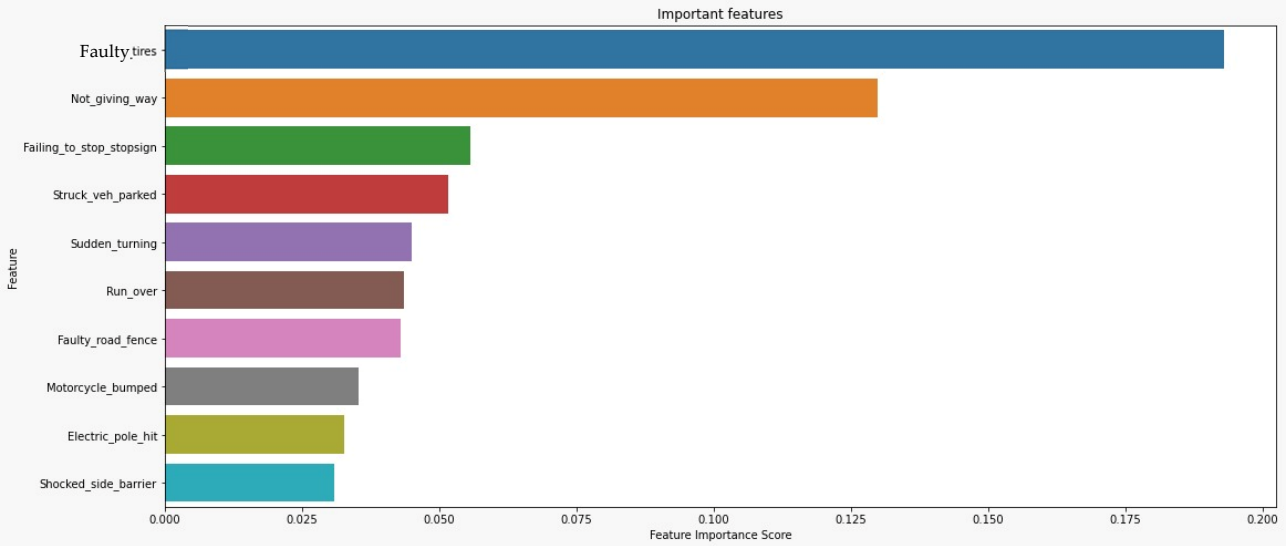

Figure 7. Variable importance using XGBoost Classifier.

Figure 8 shows that the p-value was less than 0.001, the z-score for all crashes was 6.169, and Moran's I was 0.067. The positive Moran's I, high z-scores, and tiny p-values indicate that traffic crashes were spatially clustered, and there is less than a 1% chance that this clustered pattern arose by chance. As a result, this exhibited a statistically significant status, namely, densely clustered.

The Getis-Ord G_i^* function is widely used in this work to examine hotspots for automobile crashes. A statistic called Getis-Ord G_i^* points out high data point density where statistically significant point clusters are located in the neighborhood of a given point.

A hotspot can be identified if other features with high values are everywhere around a feature with a high-value density [83,84]. It is possible that features with a high-value density are not statistically significant hotspots. Equation (15) is used to construct Getis-Ord G_i^*, and Figure 9 displays a map of the spatial distribution of crash hotspots between 2016 and 2018. The road segments with statistically significant traffic crash hotspots are identified on this map.

$$G_i^* = \frac{\sum_{j=0}^{n} w_{i,j}x_j - \overline{X}\sum_{j=0}^{n} w_{i,j}}{\sqrt{\frac{\left[n\sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} w_{i,j}\right)^2\right]}{n-1}}} \quad (15)$$

where $w_{i,j}$ is the spatial weight matrix between objects i and j, x_j is the attribute value for item j, and n is the number of features.

A hotspot is indicated by a positive z-score, whereas a negative z-score shows a cold area. Z-score values that are more positive indicate a more intensive high-value clustering (i.e., stronger hotspots). Conversely, a lower z-score denotes a greater concentration of low

values (i.e., cold spots). Accordingly, points near the center of the city present stronger hotspots, while those away from the center and on small roads show cold spots.

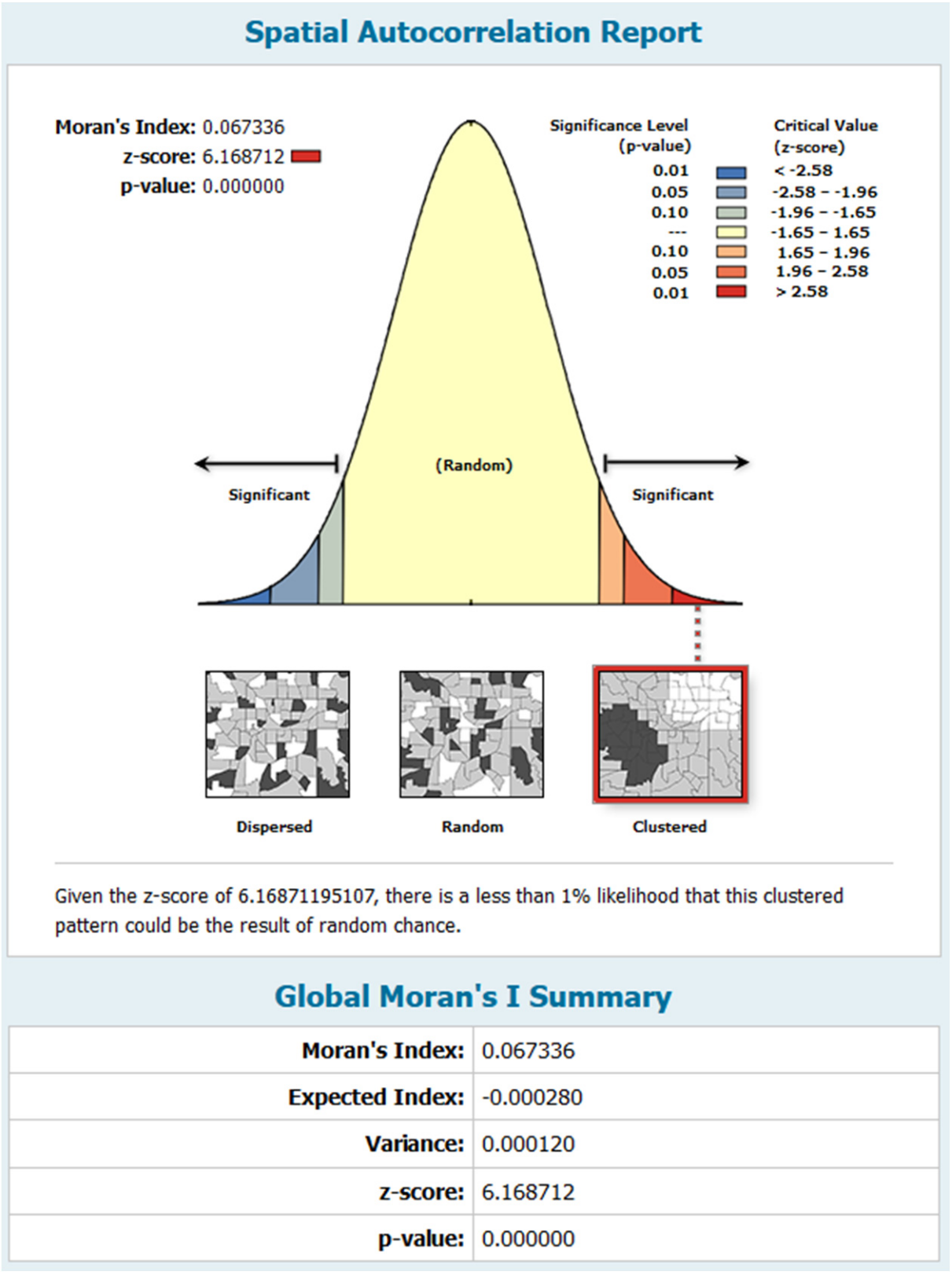

Figure 8. Traffic collisions in the province of Al-Ahsa: spatial autocorrelation report with significance graph of Moran's I, z-score, and *p*-value.

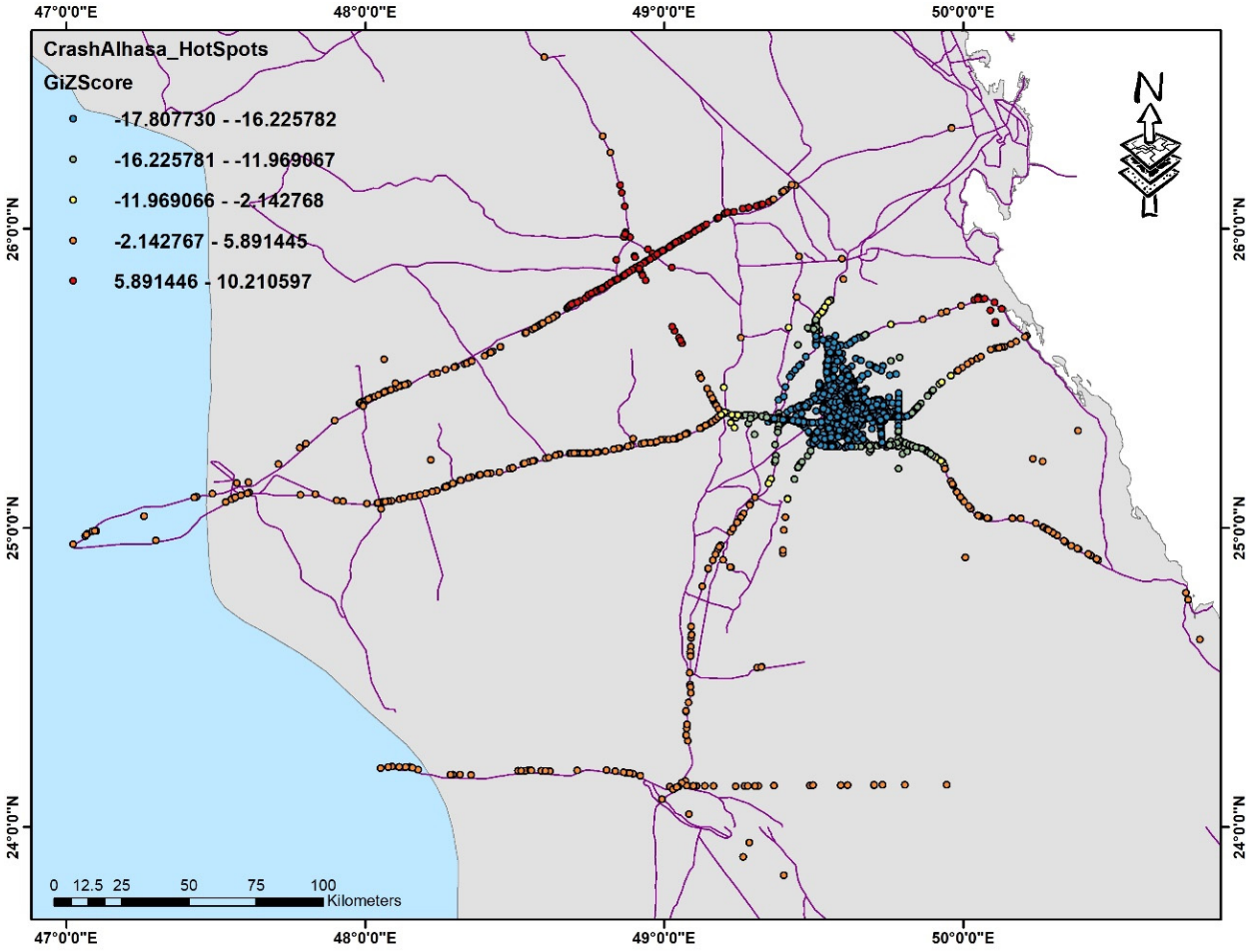

Figure 9. Composite spatial geographic distribution map of traffic crash hotspots based on the G_i^* score at the year level from 2016 to 2018.

7. Conclusions and Future Study

This study aimed to develop and compare models for predicting the severity of road crashes in Al-Ahsa city of Saudi Arabia, based on the crash data from 2016 to 2018. Two machine learning techniques, namely, R.F. and XGBoost, and the logistic regression model, as a statistical technique, were used. Among the developed models, R.F. had the best performance in terms of accuracy, precision, recall, and F1 score. The overall model accuracy in the severity predictions of RF, XGBoost, and logistic regression had accuracies of 94%, 93%, and 79%, respectively. The feature importance score from R.F. model suggests that faulty tires, not giving way, sudden turning, and running over were among the most important causes of severe crashes.

Over the past 20 years, methodologies for identifying crash hotspots have advanced, and they now play a critical role in implementing successful traffic safety management programs. In this study, the tendency of a crash pattern to cluster in space was determined using global autocorrelation analysis with Moran's I. The results demonstrate that high spatial dependence across the research region and the spatial patterns were clustered with a distance threshold of 500 m. The precise aggregated locations of traffic crashes across the research area were then determined using the local spatial autocorrelation Getis-Ord G_i^* statistic. The Getis-Ord G_i^* statistic successfully identified high-value clusters with a specified distance threshold. A region was not regarded as a hotspot if a high-value point was not surrounded by other high-value points. As a result, segments with high-value features could be predicted and given priority. The study will advise decision makers on the best places to invest or put safety measures in place.

This study had some limitations with implications for future research and can be extended in the future from various perspectives. First, for the crash severity, there were only two available classes (fatal and serious injury); P.D.O. was absent due to data scarcity, which otherwise consists of the biggest share of the crash severity class. Moreover, the analysis period used is from 2016 to 2018, which can be considered a short period for this study. This is because of the unavailability of an extended data set in the study area. Future studies can utilize detailed and comprehensive datasets containing information on other explanatory variables, such as road geometric features, weather conditions, driving behaviour, road conditions, etc. One of the ML models' criticisms is their inability to interpret the model. Machine learning interpretation techniques such as SHapley Additive exPlanations (S.H.A.P.) may be utilized to better visualize the feature sensitivity on the model output for a better model interpretation. Moreover, it is recommended for future studies to apply a larger dataset, including different cities of Saudi Arabia, along with more severe classes. In addition, we also recommend performing further analysis to investigate the temporal trends of crashes. In this regard, the use of time series analysis is recommended. Provided more data are available, deep learning techniques are also promising in crash severity studies.

Author Contributions: Conceptualization, M.K.I., I.R., R.A., U.G., M.M.R. and M.A.; methodology and software, I.R. and M.K.I.; validation, and formal analysis, I.R., U.G. and M.K.I.; resources and data curation, M.K.I.; writing—original draft preparation, review and editing, M.K.I., U.G., M.A., I.R., M.M.R. and R.A.; project administration, M.K.I.; funding acquisition, M.K.I. All authors have read and agreed to the published version of the manuscript.

Funding: This work was financially supported by Deanship of Scientific Research in the King Faisal University, Saudi Arabia [Grant 1668].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding authors Md. Kamrul Islam (maislam@kfu.edu.sa) and Imran Reza (ireza@uwyo.edu) upon reasonable request.

Conflicts of Interest: The authors would like to declare that there are no conflict of interest.

References

1. Zhang, G.; Yau, K.K.W.; Chen, G. Risk Factors Associated with Traffic Violations and Accident Severity in China. *Accid. Anal. Prev.* **2013**, *59*, 18–25. [CrossRef] [PubMed]
2. Klauer, S.G.; Guo, F.; Simons-Morton, B.G.; Ouimet, M.C.; Lee, S.E.; Dingus, T.A. Distracted Driving and Risk of Road Crashes among Novice and Experienced Drivers. *N. Engl. J. Med.* **2014**, *370*, 54–59. [CrossRef] [PubMed]
3. World Health Organization. *Global Status Report on Road Safety 2015*; World Health Organization: Geneva, Switzerland, 2015.
4. Lee, J.; Yoon, T.; Kwon, S.; Lee, J. Model Evaluation for Forecasting Traffic Accident Severity in Rainy Seasons Using Machine Learning Algorithms: Seoul City Study. *Appl. Sci.* **2020**, *10*, 129. [CrossRef]
5. Kashani, A.T.; Mohaymany, A.S. Analysis of the Traffic Injury Severity on Two-Lane, Two-Way Rural Roads Based on Classification Tree Models. *Saf. Sci.* **2011**, *49*, 1314–1320. [CrossRef]
6. Kopelias, P.; Papadimitriou, F.; Papandreou, K.; Prevedouros, P. Urban Freeway Crash Analysis Geometric, Operational, and Weather Effects on Crash Number and Severity. *Transp. Res. Rec.* **2007**, *2015*, 123–131. [CrossRef]
7. Chang, L.Y.; Wang, H.W. Analysis of Traffic Injury Severity: An Application of Nonparametric Classification Tree Techniques. *Accid. Anal. Prev.* **2006**, *38*, 1019–1027. [CrossRef] [PubMed]
8. Alikhani, M.; Nedaie, A.; Ahmadvand, A. Presentation of Clustering-Classification Heuristic Method for Improvement Accuracy in Classification of Severity of Road Accidents in Iran. *Saf. Sci.* **2013**, *60*, 142–150. [CrossRef]
9. De Oña, J.; López, G.; Abellán, J. Extracting Decision Rules from Police Accident Reports through Decision Trees. *Accid. Anal. Prev.* **2013**, *50*, 1151–1160. [CrossRef] [PubMed]
10. Wanjau, S.K.; Muketha, G.M. Improving Student Enrollment Prediction Using Ensemble Classifiers. *Int. J. Comput. Appl. Technol. Res.* **2018**, *07*, 122–128. [CrossRef]
11. McClafferty, J.; Hankey, J.M. *100-Car Reanalysis: Summary of Primary and Secondary Driver Characteristics*; Virginia Tech Transportation Institute: Blacksburg, VA, USA, 2010.

12. Tian, R.; Li, L.; Chen, M.; Chen, Y.; Witt, G.J. Studying the Effects of Driver Distraction and Traffic Density on the Probability of Crash and Near-Crash Events in Naturalistic Driving Environment. *IEEE Trans. Intell. Transp. Syst.* **2013**, *14*, 1547–1555. [CrossRef]
13. Klauer, S.G.; Guo, F.; Sudweeks, J.D.; Dingus, T.A. *An Analysis of Driver Inattention Using a Case-Crossover Approach on 100-Car Data: Final Report*; US Department of Transportation National Highway Traffic Safety Administration; Virginia Tech Transportation Institute: Blacksburg, VA, USA, 2010; 148p.
14. Klauer, S.G.; Dingus, T.A.; Neale, V.L.; Sudweeks, J.D.; Ramsey, D.J. *The Impact of Driver Inattention on Near Crash/Crash Risk: An Analysis Using the 100-Car Naturalistic Driving Study Data Analysis*; Virginia Tech Transportation Institute: Blacksburg, VA, USA, 2006; p. 226.
15. Guo, F. Individual Driver Risk Analysis Using Naturalistic. Road Safety Simulation. Ph.D. Thesis, Virginia Tech Transportation Institute, Virginia Tech, Blacksburg, VA, USA, 2011.
16. Berdoulat, E.; Vavassori, D.; Sastre, M.T.M. Driving Anger, Emotional and Instrumental Aggressiveness, and Impulsiveness in the Prediction of Aggressive and Transgressive Driving. *Accid. Anal. Prev.* **2013**, *50*, 758–767. [CrossRef] [PubMed]
17. De Oña, J.; Mujalli, R.O.; Calvo, F.J. Analysis of Traffic Accident Injury Severity on Spanish Rural Highways Using Bayesian Networks. *Accid. Anal. Prev.* **2011**, *43*, 402–411. [CrossRef] [PubMed]
18. Sarkar, S.; Vinay, S.; Raj, R.; Maiti, J.; Mitra, P. Application of Optimized Machine Learning Techniques for Prediction of Occupational Accidents. *Comput. Oper. Res.* **2019**, *106*, 210–224. [CrossRef]
19. Becky, P.Y.l.; Anderson, T.K. 21 September 2015, Road Safety as a Public Health Issue from: Spatial Analysis Methods of Road Traffic Collisions CRC Press. Available online: https://www.routledgehandbooks.com/doi/10.1201/b18937-4 (accessed on 1 November 2022).
20. Yao, S.; Loo, B.P.Y.; Yang, B.Z. Traffic Collisions in Space: Four Decades of Advancement in Applied, G.I.S. *Ann. GIS* **2016**, *22*, 1–14. [CrossRef]
21. Xu, Q.; Tao, G. Traffic Accident Hotspots Identification Based on Clustering Ensemble Model. In Proceedings of the 5th International Conference on Cyber Security and Cloud Computing/4th International Conference on Edge Computing and Scalable Cloud (CSCloud/EdgeCom), Shanghai, China, 22–24 June 2018; IEEE: New York, NY, USA, 2018; pp. 1–4. [CrossRef]
22. Bureau of Infrastructure, Transport and Regional Economics (BITRE), 2012, Evaluation of the National Black Spot Program Volume 1 BITRE Report 126, Canberra ACT. Available online: https://www.bitre.gov.au/publications/2012/report_126 (accessed on 1 November 2022).
23. Geurts, K.; Wets, G. Black Spot Analysis Methods: Literature Review. Onderz. *Kennis Verkeersonveiligheid* **2003**, *1*, 32.
24. Kidando, E.; Moses, R.; Ozguven, E.E.; Sando, T. Incorporating Travel Time Reliability in Predicting the Likelihood of Severe Crashes on Arterial Highways Using Non-Parametric Random-Effect Regression. *J. Traffic Transp. Eng.* **2019**, *6*, 470–481. [CrossRef]
25. Lord, D.; Mannering, F. The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives. *Transp. Res. Part A Policy Pract.* **2010**, *44*, 291–305. [CrossRef]
26. Savolainen, P.T.; Mannering, F.L.; Lord, D.; Quddus, M.A. The Statistical Analysis of Highway Crash-Injury Severities: A Review and Assessment of Methodological Alternatives. *Accid. Anal. Prev.* **2011**, *43*, 1666–1676. [CrossRef] [PubMed]
27. Zeng, Q.; Huang, H.; Pei, X.; Wong, S.C.; Gao, M. Rule Extraction from an Optimized Neural Network for Traffic Crash Frequency Modeling. *Accid. Anal. Prev.* **2016**, *97*, 87–95. [CrossRef] [PubMed]
28. Li, H.; Graham, D.J.; Majumdar, A. The Effects of Congestion Charging on Road Traffic Casualties: A Causal Analysis Using Difference-in-Difference Estimation. *Accid. Anal. Prev.* **2012**, *49*, 366–377. [CrossRef] [PubMed]
29. Chang, L.Y. Analysis of Freeway Accident Frequencies: Negative Binomial Regression versus Artificial Neural Network. *Saf. Sci.* **2005**, *43*, 541–557. [CrossRef]
30. Doğan, E.; Akgüngör, A.P. Forecasting Highway Casualties under the Effect of Railway Development Policy in Turkey Using Artificial Neural Networks. *Neural Comput. Appl.* **2013**, *22*, 869–877. [CrossRef]
31. Budalakoti, S.; Srivastava, A.N.; Otey, M.E. Anomaly Detection and Diagnosis Algorithms for Discrete Symbol Sequences with Applications to Airline Safety. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **2009**, *39*, 101–113. [CrossRef]
32. Dia, H.; Rose, G. Development and Evaluation of Neural Network Freeway Incident Detection Models Using Field Data. *Transp. Res. Part C Emerg. Technol.* **1997**, *5*, 313–331. [CrossRef]
33. Haykin, S. *Neural Networks and Learning Machines*, 3rd ed.; Prentice Hall: New York, NY, USA, 2009.
34. Delen, D.; Sharda, R.; Bessonov, M. Identifying Significant Predictors of Injury Severity in Traffic Accidents Using a Series of Artificial Neural Networks. *Accid. Anal. Prev.* **2006**, *38*, 434–444. [CrossRef] [PubMed]
35. Dong, C.; Shao, C.; Li, J.; Xiong, Z. An Improved Deep Learning Model for Traffic Crash Prediction. *J. Adv. Transp.* **2018**, *2018*, 3869106. [CrossRef]
36. Moghaddam, F.R.; Afandizadeh, S.; Ziyadi, M. Prediction of Accident Severity Using Artificial Neural Networks. *Int. J. Civ. Eng.* **2011**, *9*, 41–49.
37. Xie, Y.; Lord, D.; Zhang, Y. Predicting Motor Vehicle Collisions Using Bayesian Neural Network Models: An Empirical Analysis. *Accid. Anal. Prev.* **2007**, *39*, 922–933. [CrossRef] [PubMed]
38. Simoncic, M. A Bayesian Network Model of Two-Car Accidents. *J. Transp. Stat.* **2004**, *7*, 13–25.
39. Smola, A.J.; Scholkopf, B. A Tutorial on Support Vector Regression. *Stat. Comput.* **2004**, *14*, 199–222. [CrossRef]

40. Li, X.; Lord, D.; Zhang, Y.; Xie, Y. Predicting Motor Vehicle Crashes Using Support Vector Machine Models. *Accid. Anal. Prev.* **2008**, *40*, 1611–1618. [CrossRef] [PubMed]
41. Zhang, Y.; Xie, Y. Forecasting of Short-Term Freeway Volume with v-Support Vector Machines. *Transp. Res. Rec.* **2007**, *2024*, 92–99. [CrossRef]
42. Kecman, V. Support Vector Machines–An Introduction. In *Support Vector Machines: Theory and Applications*; Heidelberg, S.B., Ed.; Springer: Berlin/Heidelberg, Germany, 2005.
43. Silva, P.B.; Andrade, M.; Ferreira, S. Machine Learning Applied to Road Safety Modeling: A Systematic Literature Review. *J. Traffic Transp. Eng.* **2020**, *7*, 775–790. [CrossRef]
44. Abellán, J.; López, G.; De Oña, J. Analysis of Traffic Accident Severity Using Decision Rules via Decision Trees. *Expert Syst. Appl.* **2013**, *40*, 6047–6054. [CrossRef]
45. Gharehchopogh, F.S.; Dizaji, Z.A.; Aghighi, Z. Evaluation of Particle Swarm Optimization Algorithm in Prediction of the Car Accidents on the Roads: A Case Study. *IATSS Res.* **2013**, *3*, 1–12. [CrossRef]
46. Morcillo, L.G.; Poyo, F.J.C.; Maldonado, G.L. Using Decision Trees for Comparing Different Consistency Models. *Procedia-Soc. Behav. Sci.* **2014**, *160*, 332–341. [CrossRef]
47. Holland, J.H. *Adaptation in Natural and Artificial Systems*; University of Michigan Press: Cambridge, MA, USA, 1975.
48. Abdel-Aty, M.; Haleem, K. Analyzing Angle Crashes at Unsignalized Intersections Using Machine Learning Techniques. *Accid. Anal. Prev.* **2011**, *43*, 461–470. [CrossRef]
49. Deng, L.; Yu, D. Deep Learning Methods and Applications. *Found. Trends Signal Process.* **2014**, *7*, 197–387.
50. Lecun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
51. Sharma, B.; Katiyar, V.K.; Kumar, K. Traffic Accident Prediction Model Using Support Vector Machines with Gaussian Kernel. *Adv. Intell. Syst. Comput.* **2016**, *437*, 1–10. [CrossRef]
52. Huang, W.; Song, G.; Hong, H.; Xie, K. Deep Architecture for Traffic Flow Prediction: Deep Belief Networks with Multitask Learning. *IEEE Trans. Intell. Transp. Syst.* **2014**, *15*, 2191–2201. [CrossRef]
53. Tiwari, P.; Kumar, S.; Kalitin, D. Road-User Specific Analysis of Traffic Accident Using Data Mining Techniques. In Proceedings of the International Conference on Computational Intelligence, Communications, and Business Analytics, Kolkata, India, 24–25 March 2017; Mandal, J., Dutta, P., Mukhopadhyay, S., Eds.; Springer: Singapore, 2017; pp. 398–410.
54. Almamlook, R.E.; Kwayu, K.M.; Alkasisbeh, M.R.; Frefer, A.A. Comparison of Machine Learning Algorithms for Predicting Traffic Accident Severity. In Proceedings of the Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), Amman, Jordan, 9–11 April 2019; IEEE: New York, NY, USA, 2019; pp. 272–276. [CrossRef]
55. Beshah, T.; Hill, S. Mining Road Traffic Accident Data to Improve Safety: Role of Road-Related Factors on Accident Severity in Ethiopia. *AAAI Spring Symp. Tech. Rep.* **2010**, *SS-10-01*, 14–19.
56. Lee, C.; Li, X.C. Predicting Driver Injury Severity in Single-Vehicle and Two-Vehicle Crashes with Boosted Regression Trees. *Transp. Res. Rec. J. Transp. Res. Board* **2015**, *2514*, 138–148.
57. Toran Pour, A.; Moridpour, S.; Tay, R.; Rajabifard, A. Modelling Pedestrian Crash Severity at Mid-Blocks. *Transp. A Transp. Sci.* **2017**, *13*, 273–297. [CrossRef]
58. Wu, W.; Jiang, S.; Liu, R.; Jin, W.; Ma, C. Economic Development, Demographic Characteristics, Road Network and Traffic Accidents in Zhongshan, China: Gradient Boosting Decision Tree Model. *Transp. A Transp. Sci.* **2020**, *16*, 359–387. [CrossRef]
59. Wu, W.; Xia, Y.; Jin, W. Predicting Bus Passenger Flow and Prioritizing Influential Factors Using Multi-Source Data: Scaled Stacking Gradient Boosting Decision Trees. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 2510–2523. [CrossRef]
60. Abdel-Aty, M.A.; Abdelwahab, H.T. Predicting Injury Severity Levels in Traffic Crashes: A Modeling Comparison. *J. Transp. Eng.* **2004**, *130*, 204–210. [CrossRef]
61. Montella, A.; de Oña, R.; Mauriello, F.; Rella Riccardi, M.; Silvestro, G. A Data Mining Approach to Investigate Patterns of Powered Two-Wheeler Crashes in Spain. *Accid. Anal. Prev.* **2020**, *134*, 105251. [CrossRef]
62. Jamal, A.; Rahman, M.T.; Al-Ahmadi, H.M.; Mansoor, U. The Dilemma of Road Safety in the Eastern Province of Saudi Arabia: Consequences and Prevention Strategies. *Int. J. Environ. Res. Public Health* **2020**, *17*, 157. [CrossRef]
63. Rahman, M.M.; Islam, K.; Al-Shayeb, A.; Arifuzzaman, M. Towards Sustainable Road Safety in Saudi Arabia: Exploring Traffic Accident Causes Associated with Driving Behavior Using a Bayesian Belief Network. *Sustainability* **2022**, *14*, 6315. [CrossRef]
64. Almoshaogeh, M.; Abdulrehman, R.; Haider, H.; Alharbi, F.; Jamal, A.; Alarifi, S.; Shafiquzzaman, M.D. Traffic Accident Risk Assessment Framework for Qassim, Saudi Arabia: Evaluating the Impact of Speed Cameras. *Appl. Sci.* **2021**, *11*, 6682. [CrossRef]
65. Islam, M.K.; Gazder, U.; Akter, R.; Arifuzzaman, M. Involvement of Road Users from the Productive Age Group in Traffic Crashes in Saudi Arabia: An Investigative Study Using Statistical and Machine Learning Techniques. *Appl. Sci.* **2022**, *12*, 6368. [CrossRef]
66. City-Fact. Available online: https://www.city-facts.com/al-ahsa (accessed on 6 November 2022).
67. Liu, C.; Sharma, A. Exploring Spatio-Temporal Effects in Traffic Crash Trend Analysis. *Anal. Methods Accid. Res.* **2017**, *16*, 104–116. [CrossRef]
68. Arévalo-Támara, A.; Orozco-Fontalvo, M.; Cantillo, V. Factors Influencing Crash Frequency on Colombian Rural Roads. *Promet-Traffic&Traffico* **2020**, *32*, 449–460. [CrossRef]
69. Bassani, M.; Rossetti, L.; Catani, L. Spatial Analysis of Road Crashes Involving Vulnerable Road Users in Support of Road Safety Management Strategies. *Transp. Res. Procedia* **2020**, *45*, 394–401.

70. Abdi, T.A.; Hailu, B.H.; Adal, T.A.; Van Gelder, P.H.; Hagenzieker, M.P.; Carbon, C.-C. Road Crashes in Addis Ababa, Ethiopia: Empirical Findings between the Years 2010 and 2014. *Afr. Res. Rev.* **2017**, *11*, 1–13. [CrossRef]
71. Mannering, F.L.; Bhat, C.R. Analytic Methods in Accident Research: Methodological Frontier and Future Directions. *Anal. Methods Accid. Res.* **2014**, *1*, 1–22. [CrossRef]
72. Belgiu, M.; Drăgu, L. Random Forest in Remote Sensing: A Review of Applications and Future Directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31. [CrossRef]
73. Flach, P. *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*; Cambridge University Press: Cambridge, UK, 2012.
74. AlThuwaynee, O.F.; Kim, S.W.; Najemaden, M.A.; Aydda, A.; Balogun, A.L.; Fayyadh, M.M.; Park, H.J. Demystifying Uncertainty in PM10 Susceptibility Mapping Using Variable Drop-off in Extreme-Gradient Boosting (X.G.B.) and Random Forest (R.F.) Algorithms. *Environ. Sci. Pollut. Res.* **2021**, *28*, 43544–43566. [CrossRef]
75. Boschetti, A.; Massaron, L. *Python Data Science Essentials: A Practitioner's Guide Covering Essential Data Science Principles, Tools, and Techniques*; Packt Publishing Ltd.: Birmingham, UK, 2018.
76. Budholiya, K.; Shrivastava, S.K.; Sharma, V. An Optimized XGBoost Based Diagnostic System for Effective Prediction of Heart Disease. *J. King Saud Univ.-Comput. Inf. Sci.* **2020**, *34*, 4514–4523. [CrossRef]
77. Fiorentini, N.; Losa, M. Handling Imbalanced Data in Road Crash Severity Prediction by Machine Learning Algorithms. *Infrastructures* **2020**, *5*, 61. [CrossRef]
78. Jeong, H.; Jang, Y.; Bowman, P.J.; Masoud, N. Classification of Motor Vehicle Crash Injury Severity: A Hybrid Approach for Imbalanced Data. *Accid. Anal. Prev.* **2018**, *120*, 250–261. [CrossRef] [PubMed]
79. Jamal, A.; Zahid, M.; Tauhidur Rahman, M.; Al-Ahmadi, H.M.; Almoshaogeh, M.; Farooq, D.; Ahmad, M. Injury Severity Prediction of Traffic Crashes with Ensemble Machine Learning Techniques: A Comparative Study. *Int. J. Inj. Contr. Saf. Promot.* **2021**, *28*, 408–427. [CrossRef]
80. Murphy, K.P. *Machine Learning: A Probabilistic Perspective*; M.I.T. Press: Cambridge, MA, USA, 2012.
81. Géron, A. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*; O'Reilly Media: Newton, MA, USA, 2019.
82. Ali, M.; Polat, N. Boşanma Verilerinin Coğrafi Bilgi Sistemleri Destekli Mekânsal İstatistiksel Yöntemler Ile İrdelenmesi. Investigation of Divorce Data with Spatial Statistical Methods Aided Geographic Information Systems Investigation of Divorce Data with Spatial Statistic. *Harran Univ. J. Eng.* **2018**, *3*, 112–118.
83. Ulak, M.B.; Ozguven, E.E.; Vanli, O.A.; Horner, M.W. Exploring Alternative Spatial Weights to Detect Crash Hotspots. *Comput. Environ. Urban Syst.* **2019**, *78*, 101398. [CrossRef]
84. Hazaymeh, K.; Almagbile, A.; Alomari, A.H. Spatiotemporal Analysis of Traffic Accidents Hotspots Based on Geospatial Techniques. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 260. [CrossRef]