



OPEN

Identifying accident prone areas and factors influencing the severity of crashes using machine learning and spatial analyses

Yegane Khosravi¹, Farhad Hosseinali¹✉ & Mostafa Adresi²✉

Identifying accident prone areas and the contributing environmental factors can save lives and enhance infrastructure durability. This research presents an innovative use of two hierarchical clustering methods (Agglomerative Hierarchical and BIRCH clustering algorithms) to detect accident hotspots with high accident rates on the Yazd-Kerman road, in Iran. These approaches identified clusters of accidents, highlighting significant clusters and categorizing their overlap as two accident prone areas. The high percentage of overlapping results from both methods indicates a high level of consistency in the findings. Through observations, field visits, police report analysis, and interviews with locals, the primary causes of accidents in accident prone areas were identified. In one of the accident prone areas, the most significant reasons for accidents included the presence of a resting area, insufficient lighting at curves, and poor road signage, which created a dilemma zone for drivers. In another accident prone area, the main contributing factor was reduced visibility during dust storms. Then, K-Nearest Neighbors and Random Forests machine learning algorithms were employed to predict the severity of accidents, using various input attributes such as lighting, climate, alignment slope, and road geometry. The K-Nearest Neighbor surpassed the Random Forest technique, achieving an overall accuracy of 71% in contrast to 60%. This study effectively evaluated the outcomes of clustering, uncovering the underlying causes of accidents to inform future practical interventions. Moreover, by predicting the severity of accidents along the road, a framework was developed to propose strategies for risk reduction.

Keywords Intelligent transportation systems, Machine learning, Accident hotspots, Severity prediction, Clustering, Classification

Traffic accidents pose a significant threat that can cause human injuries and substantial financial and environmental losses¹. The World Health Organization (WHO) reports an annual death toll of 1.19 million due to traffic accidents, and 50 million injuries². Additionally, WHO reports that traffic accidents are the leading cause of death for children and young people aged 5–29³. Moreover, WHO has stated that 93% of all global deaths occur in low- and middle-income countries, which have about 60% of the world's vehicles. Since road traffic accidents account for nearly 3% to 5% of the nation's gross domestic product³, the investment in improving road safety would compensate for its cost by saving direct and indirect expenses of accidents, leading to enhanced sustainability in social, economic, and environmental aspects. According to the statistics released by the WHO, countries that have taken measures to raise road safety, have successfully reduced road accidents³. In this regard, many researchers have conducted extensive studies focusing on identifying accident prone areas, predicting the severity of accidents, and reducing the probability and severity of accidents. To understand accident patterns and accident prone areas, and to predict the severity of accidents, innovative techniques must analyse various parameters and provide an accurate result^{4,5}. Recently, advanced computer science techniques such as data mining are used to predict the severity of accidents⁵. In this method, data related to accidents, vehicle type, speed, weather conditions, and time are considered. Then, a model for predicting the severity of accidents can be generated using machine learning algorithms. By employing these machine learning algorithms, accident prone areas and the severity of accidents can be predicted more accurately, and measures can be taken to reduce

¹Department of Surveying Engineering, Faculty of Civil Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran. ²Department of Geotechnics and Water Engineering, Faculty of Civil Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran. ✉email: f.hosseinali@sru.ac.ir; m.adresi@sru.ac.ir

accident frequency⁶. For example, based on this prediction, measures can be taken to improve road infrastructure, limit the speed of vehicles, and raise awareness about safe driving practices in accident prone areas. Certainly, much attention has been recently paid to the problem of identifying accident hotspots and predicting the severity of accidents. For example, Ghaffari et al. introduced a method based on reliability to identify accident hotspots and compared the results with the results of Frequency and Empirical Bayesian methods, utilizing simulated data⁷. They found their introduced method to be better in prediction than the two others. Xu and Tao used the principal component clustering to identify accident hotspots on a road in China⁸. This technique can be employed to evaluate and quantify the safety levels of various roads by extracting principal components and conducting clustering for accident hotspots. Using the Classification and Regression Tree as a data mining method, Tavakoli Kashani et al. evaluated the factors affecting the injuries caused by road accidents due to fatigue and drowsiness within seven years in three provinces of Iran and identified several factors. Geometric factors were found to have an impact on them⁹. The accident data that was clustered was also classified and analyzed according to various parameters, including rural and urban areas, residential and non-residential areas, and the drivers' gender. Agrawal et al., Hajela et al., and Karami and Johansson identified accident hotspots using the DBSCAN clustering method. DBSCAN clustering requires two primary parameters, which are Eps and MinPts^{10–13}. Finding the optimal value of these parameters is a rigorous task. Consequently, researchers are always trying to find different developed models to improve DBSCAN clustering. For instance, Akbari and Unland found a way to determine the initial parameters of DBSCAN automatically¹⁴. The idea was based on a statistical technique for outlier detection, namely the empirical rule. However, the exact method used for determining the values of the initial parameters was not identified in the DBSCAN method. Numerous studies have clustered and identified hotspots using the K-means clustering method. Puspitasari et al. identified accident prone areas of highways in Indonesia using K-means clustering and finally analyzed each cluster separately to find the hotspots¹⁵. Anderson also clustered accidents using K-means clustering. This method is unreliable for identifying hotspots due to a lack of detection of outliers. It may put outliers in a cluster and cause clustering errors. Wan et al. employed the Spatial Agglomerative Hierarchical clustering method to cluster and identify hotspots of pick-up/drop-off passengers from a taxi¹⁶. This method provided good accuracy in this study and managed to identify hotspots accurately. Khosravi et al. determined hotspots of Dehbala road using K-means, K-medoids, and DBSCAN clustering. The findings indicated that climate, especially windy or rainy weather, and the road's geometrical aspects, such as slope and curvature, have a significant impact on accidents¹⁷. Accident prediction models enable various agencies to predict the severity of accidents that may occur in the future. Hence, the severity or the frequency of accidents can be reduced by accident prediction¹⁸. Beshah and Hill divided accidents into four classes: damage, injury, serious injury, and fatal, and then predicted the severity of accidents using KNN (K-Nearest Neighbor), DT (Decision Tree), and Naive Bayes algorithms¹⁹. It was concluded that the KNN algorithm had the best performance with an overall accuracy of 81%. Selvy et al. used KNN, RF, DT, and Logistic Regression (LR) for real-time crash prediction, determining the frequency and severity of crashes. Results proved that classification accuracy obtained from Random Forest (RF) is 96% surpassing that of other classification methods²⁰. Xing et al. evaluated the collision risk of unconstrained vehicle motions at the toll plaza diverging area using the LR model and five non-parametric models, including KNN, ANN (Artificial Neural Network), SVM (Support Vector Machines), DT, and RF²¹. They used these methods to examine the relationship between influencing factors and vehicle collisions. The result showed that KNN had the best prediction performance among the other methods. Iranitalab and Khattak also used KNN, MNL (Multi-Nomial Logit), and SVM to predict traffic crash severity²². The predictive rates demonstrated that KNN outperformed the other methods in terms of effectiveness. Ijaz et al. worked on a three-year accident data report of Pakistan to predict the severity of injury-causing accidents using classification methods, i.e., Decision jungle (DJ), RF, and DT. Results revealed that DJ outperformed the DT and RF by an overall accuracy of 83.7%. Spearman correlation analysis showed that factors such as lighting conditions and shiny weather conditions were more likely to worsen injury severity²³. Amiri et al. compared the Intelligent Genetic Algorithm with ANN by investigating fixed object crashes among elderly drivers²⁴. Compared to ANN, the Intelligent Genetic Algorithm was more capable of predicting high-severity crashes. ANN was more accurate in predicting low-severity crashes, but it failed to detect more severe ones. The light condition was identified as the most significant factor, followed by right and left shoulders. Numerous studies have been conducted to identify accident prone areas and predict the severity of accidents, but more research is needed due to differences in results. There is a research gap in advancing methodological approaches for accident analysis to obtain novel insights that can support road safety interventions and provide intelligent transportation systems. The occurrence and severity of accidents may be influenced by various factors, including vehicles' safety qualifications; drivers' drive attributes, and the ride quality of the roads. The police is now keeping a record of the properties of accidents, especially their locations. The current state-of-the-art lacks comprehensive analysis and insights into localized factors contributing to road accidents in Middle East countries. Research is still needed to explore the relationships between different environmental factors and apply advanced methods for the examination of road crash hotspots in Iran, improving infrastructure sustainability. To narrow these knowledge gaps, the current work aims to identify accident prone areas by considering environmental and road geometry properties and applying hierarchical clustering methods. Traffic accident data from a case study conducted by Yazd Traffic Agency in Iran, covering 5 years starting from 2014, was used. The present research provides the following key contributions: (1) identification of accident prone areas through the application of Agglomerative Hierarchical and BIRCH clustering methods, individually or in combination; (2) analysis of accidents in each hotspot to identify the determining parameters leading to accidents; (3) use of two different machine learning classification techniques, K-Nearest Neighbor (KNN) and Random Forest (RF), to predict the severity of accidents; and (4) use of data collected from the field to identify the causes of accidents in accident prone areas, by utilizing a machine learning methodology.

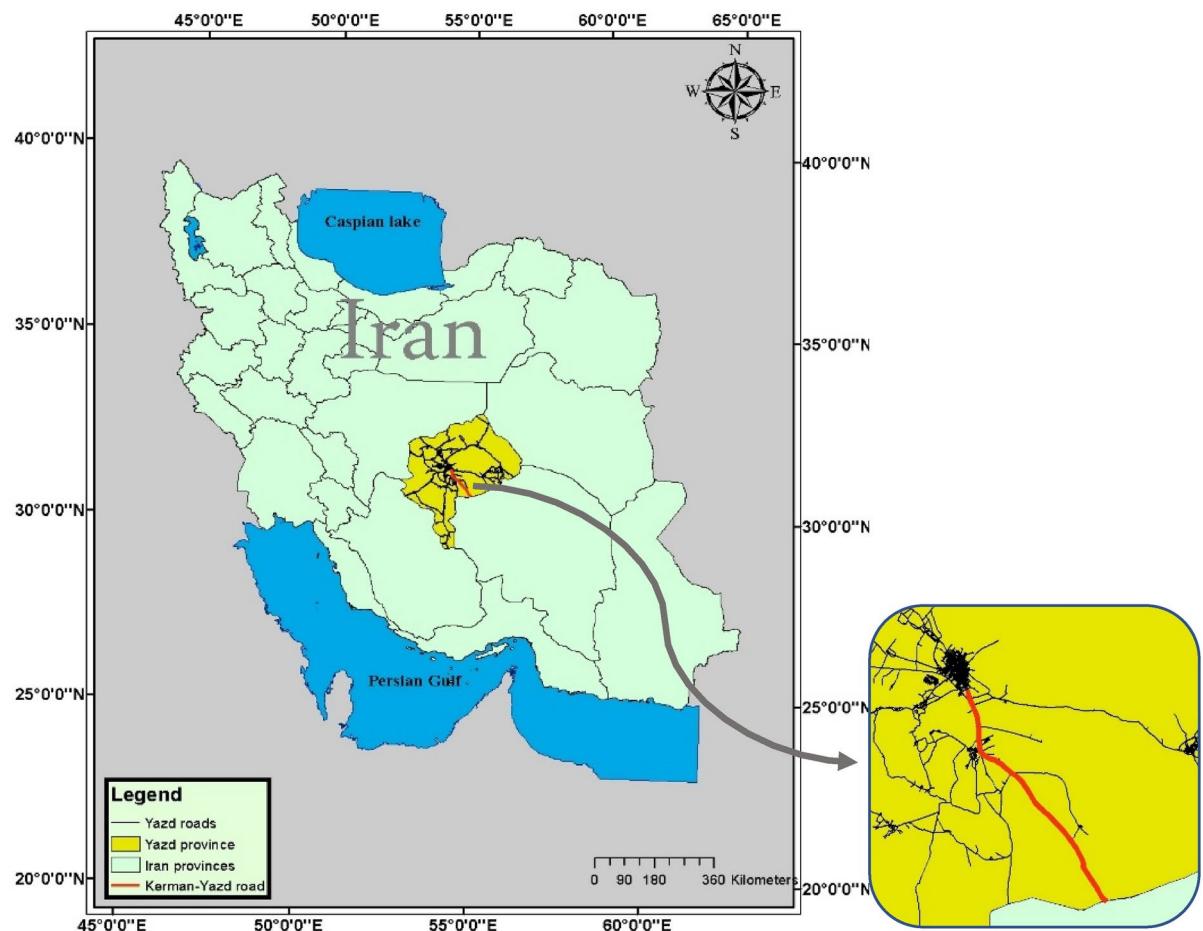


Fig. 1. Study area (created by ArcGIS 10.4 software <http://www.esri.com>).

Factor	Description
Climate	Clear/cloudy/rainy/snowy/windy/foggy/sandstorm/dusty
Lighting	Day/night/sunrise/sunset
Slope	Upward/downward/level
Alignment	Curve/straight

Table 1. Factors investigated in this research.

Data and methods

Study area

This study utilized data from accidents that occurred on the Yazd-Kerman road between 2014 and 2018. The data included 465 accidents during this period. Yazd-Kerman road is a quite crowded transit road and a part of an important route that connects the capital of Iran (Tehran) to the most important southern port, Bandar Abbas. The study focused on 107 km of Yazd-Kerman road in Yazd Province. The accident coordinates on this road were recorded using GPS receivers and were referenced using WGS1984 reference ellipsoid and UTM projection systems, as shown in Fig. 1. The data was provided by the Roads and Transportation Organization of Yazd Province and included various pieces of information, such as geographic coordinates, time and date of the accident, the cause of the accident, vehicle type, and alignment. The primary focus of this study is to examine how various geometric and environmental factors contribute to the frequency and severity of accidents. Table 1 outlines the factors that were considered in this study.

The factors taken into account by this study are introduced briefly:

Climate: this factor describes the weather conditions at the time and place of the accident, which include clear, cloudy, rainy, snowy, stormy, sandstorm, foggy, or dusty conditions. This data was extracted from Wunderground.com.

Lighting: this factor includes the state of road lighting in four modes; day, night, sunrise, and sunset. The accidents occur in one of these time intervals.

Slope: this factor includes the road slope in three modes of upward, downward, and zero slopes (level), which were extracted from the road's center-line levels from the DEM (Digital Elevation Model) of the area.

Alignment: this factor describes the properties of the road's layout (i.e. curve radius, which can be effective on accident positions).

The data used in this article includes various parameters such as the cause of the accident, type of vehicle, climate, slope, date and time of the accident, and alignment. The four most critical parameters in the study area were selected by an expert and incorporated into the model. The lighting is divided into four classes, namely day, night, sunrise, and sunset, based on the recorded time of the accident. In Lighting classes, the times of sunrise and sunset in different months and seasons are also considered. In the classification of slopes, slopes of 0 to $\pm 2\%$ were classified as level, slopes of more than 2% were classified as an upgrade, and slopes lower than -2% were classified as a downgrade. Selecting the intervals for the slopes was based on the average slope of the road. To classify the alignment, the curves of the road were identified and their radius and curvature were also obtained. The alignment was classified as the curve and the straight. As mentioned before, the data was provided by the Roads and Transportation Organization of Yazd Province, which has originally been recorded by the police. Because this data is produced exclusively by the police, it is very difficult, if not impossible, to validate it. One method to assess the accuracy of accident data (Only related to the number and not the location and time of accidents) is to compare the data of accidents with the records hospitals as demonstrated in studies by Curry et al.²⁵ and Soltani et al.¹⁸. These studies have shown that linked data can significantly enhance data quality for accident research. This comparison may help to elucidate potential discrepancies in the data; however, it is important to note that incidents without injuries will lack such information.

In order to identify accident hotspots, this study utilized the BIRCH clustering and Agglomerative Hierarchical algorithms, which were implemented using Python. The outputs were displayed using ArcGIS 10.4 software (<http://www.esri.com>). In addition, KNN and RF classification algorithms were utilized to predict the severity of accidents, which were implemented using RapidMiner software. The dataset was divided into 80% for calibration and 20% for validation. These methods were selected because they offer promising alternatives for obtaining comprehensive and accurate insights into the data, ensuring robust clustering results and reliable predictions of accident severity. Below are brief explanations of the mentioned methods.

Clustering

Clustering analysis is the process of dividing a heterogeneous population into several homogeneous subsets or clusters. In other words, it is a process by which a set of objects can fall into separate groups or clusters²⁶. In this study, Agglomerative and BIRCH hierarchical clustering algorithms have been used to cluster the accidents.

Agglomerative hierarchical

Agglomerative Hierarchical clustering is a bottom-up clustering method in which clusters have sub-clusters. This clustering merges observations or clusters with the least distance (most similarity) to form a new cluster. This process continues until only one cluster remains. The output of Agglomerative Hierarchical can be a dendrogram that can be cut according to the research requirements to reach the desired number of clusters²⁷. Figure 2 shows the process of this algorithm as a flowchart²⁸.

To perform the calculations related to this clustering method, we need two distance (similarity) criteria: (1) the distance between pairs of observations, and (2) the distance between the clusters. Euclidean distance (1) has been used in this research to measure the distance between observations. Also, a complete-linkage (2) has been used to measure the distance between the clusters²⁹. In Eq. (1), d is the Euclidean distance, (x_1, y_1) is the coordinate of the first point and (x_2, y_2) is the coordinate of the second point.

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (1)$$

BIRCH

The BIRCH clustering algorithm is one of the hierarchical clustering algorithms. This method is designed to work with big data. Moreover, BIRCH is user-friendly and can be easily implemented. This algorithm is based on the CF (clustering features) tree. In addition, this algorithm uses a tree-structured summary to create clusters. The BIRCH algorithm first scans all the data in memory and then compresses the data into small summaries. It then clusters small summaries using a hierarchical algorithm or other arbitrary algorithms. Finally, it refines the clusters to form less scattered and more homogeneous clusters. Considering that BIRCH does not directly cluster the datasets, it is often used with other clustering algorithms. The primary purposes of these algorithms are reducing the time of components and reducing the frequency of data scans, identifying dense areas, and mitigating noise³⁰. Figure 3 shows the process of this algorithm as a flowchart³¹.

The input parameters of this algorithm include:

- Threshold: it is the radius of the sub-cluster to get the new sample in it. The default value of the threshold is 0.5 and it should be as low as possible in the beginning.
- Branching factor: it is used to make the total number of sub-clusters in each node. If the new sample is entered after the mentioned value, then the sub-cluster splits further at that node. The default value is 50 branches.
- N_clusters: it is the number of clusters.

To calculate the goodness of the mentioned clustering algorithms, according to Eq. (2), the Silhouette index has been used. Its value ranges from -1 to 1.

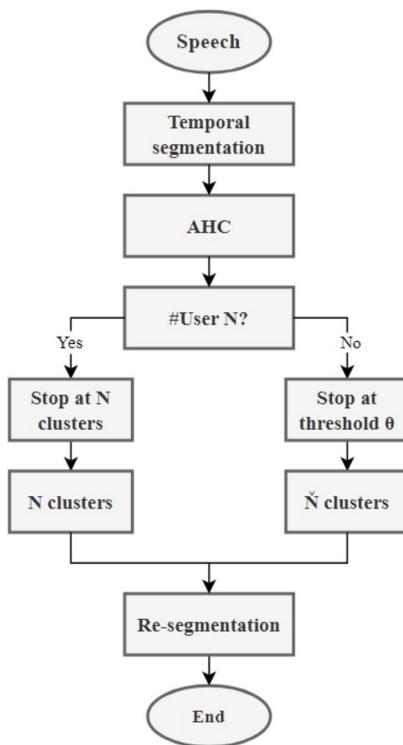


Fig. 2. Agglomerative Hierarchical algorithm flowchart²⁸.

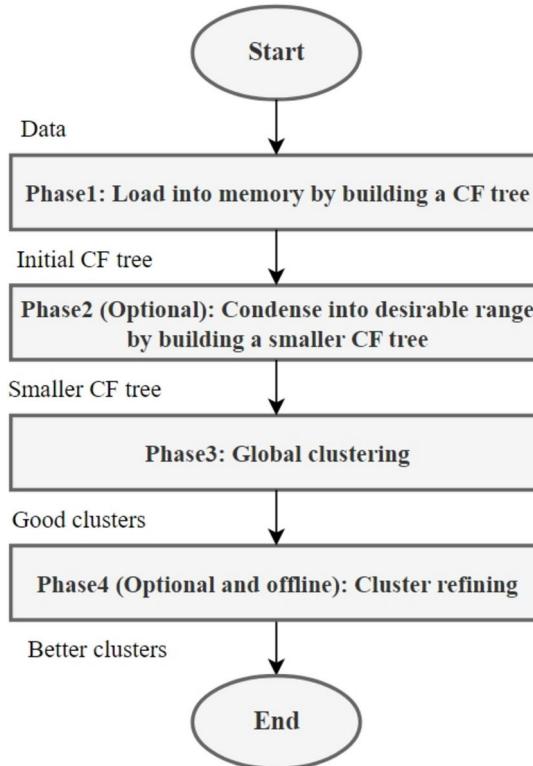


Fig. 3. The flowchart of BIRCH clustering³¹.

$$\text{Silhouette index} = \frac{(b - a)}{\max(a, b)} \quad (2)$$

where:

- a* is the average intra-cluster distance, i.e. the average distance between each point within a cluster;
- b* is the average inter-cluster distance, i.e. the average distance between all clusters.

Classification

Classification is a machine learning method that is used to learn how to assign a class label to input data. Because of the importance of extracting information from accident data, classification can be a helpful tool in road safety research. The accident severity prediction model enables various agencies to estimate the severity of a reported accident or the severity of an accident that may happen in a specific location in the future²². An accident is the result of the influence of many factors, so it is impossible to accurately predict which class of accident severity (damage, fatal, or injury) belongs to each accident. However, the severity of the accidents can be predicted to some extent. In this study, RF classification and KNN algorithms have been used to predict the severity of accidents.

K-nearest neighbor (KNN)

The KNN method is a supervised algorithm first developed and applied by Cover and Hart. This algorithm is based on the principle that "similar samples in a data set are often adjacent". The decision rule in this method is to assign an observation to the class with the highest number of votes among the K-nearest-neighbors³². KNN parameters include the number of close neighbors (K) and the distance function³³. The best choice of K depends on the data. In general, large values of K reduce the classification error but lower the resolution of the boundary between classes. One of the simplest ways to select the best K value is setting a range of neighbors and using validation criteria, such as overall accuracy³². Consistent with the majority of studies, Euclidean distance has been used in this research for the distance function parameter. The process of this classification algorithm is shown in Fig. 4.

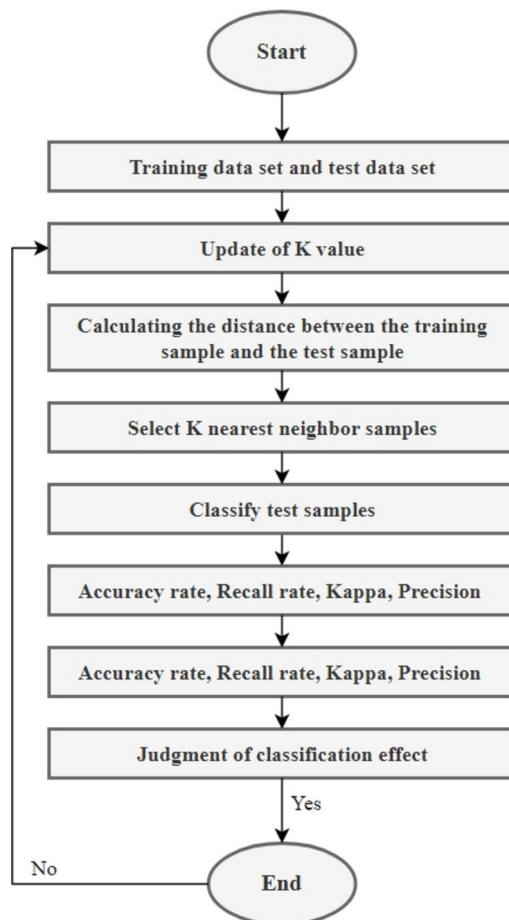


Fig. 4. KNN processing flowchart³⁴.

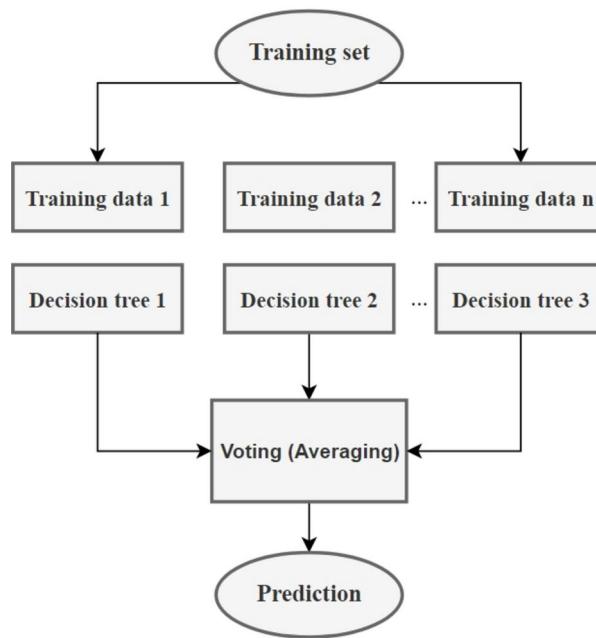


Fig. 5. RF processing flowchart³⁹.

		Predicted	
Actual	Negative	Negative	Positive
	Negative	TN	FP
	Positive	FN	TP

Table 2. Typical binary confusion matrix.

Random Forest (RF)

The Random Forest algorithm is one of the hybrid classification algorithms and one of the most widely used machine algorithms³⁵. The term “Random Forest” is derived from random decision forests, first coined by Ho in 1995 and completed by Amit and Geman³⁶. The RF algorithm is comprised of a group of decision trees; therefore, this method includes all the basic concepts of the decision tree. In other words, the RF algorithm combines several decision trees to make more accurate predictions³⁷. The RF algorithm aggregates the votes from different decision trees to decide the final class of the test object³⁸. The process of this classification algorithm is shown in Fig. 5³⁹.

Classification accuracy metrics

The confusion matrix is a commonly used measure for defining the validity of classification methods. It can be applied to binary and multi-class classification problems⁴⁰. The confusion matrix is a table that is used to define the performance of a classification algorithm. A confusion matrix depicts and summarizes the performance of a classification algorithm⁴¹. An example of a binary confusion matrix is shown in Table 2.

The confusion matrix counts the predicted and actual values by displaying them in a matrix. According to Table 2, the entries of the confusion matrix are defined as follows:

True positive value (TP): the total number of true results or predictions when the actual class is positive.

False positive value (FP): the total number of false results or predictions when the actual class is positive.

True negative value (TN): the total number of true results or predictions when the actual class is negative.

False negative value (FN): the total number of false results or predictions when the actual class is negative.

Some indices can be used to evaluate the accuracy and validity of the prediction results. The indices used in this study are defined below:

Overall accuracy (OA): is used to compare system performance. According to Eq. (3), OA determines the ratio of correctly predicted samples to all samples, which shows how correctly a classifier can predict the samples.

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

Precision: determines how many samples are classified correctly in each class based on the predicted labels. This criterion is calculated according to Eq. (4).

$$Precision = \frac{TP}{FN + TN} \quad (4)$$

Kappa: compares the existing classification algorithm with a random classification algorithm and explains to what extent the existing classification algorithm performed better than a random algorithm according to Eq. (5)⁴².

$$Kappa = \frac{2 \times (TP \times TN - FN \times FP)}{(TP + FP) \times (FP + TN) + (TP + FN) \times (FN + TN)} \quad (5)$$

The recall criterion is the ratio of the correctly predicted samples to the number of all real samples in the class, which is calculated according to Eq. (6)⁴².

$$Recall = \frac{TN}{FP + TN} \quad (6)$$

Results and discussions

Investigation of accident prone areas

For the Agglomerative Hierarchical algorithm, the number of clusters must be entered to complete the process; thus, the hierarchy of the algorithm continues until it reaches that number of clusters. The optimal number of clusters for Yazd-Kerman road was determined to be 18 using the silhouette validation index. Then, clusters with sharp differences in the number of accidents were selected based on natural breaks. Clusters 14 and 16 had the highest number of accidents with 114 and 87, respectively, in the Agglomerative Hierarchical algorithm. Figures 6, 7, 8, 9 shows the Silhouette validation index, Fig. 8 illustrates the clustering output, and Fig. 10 shows the number of crashes per cluster.

The BIRCH hierarchical algorithm requires the input of parameters such as threshold (T), branching factor (B), and the desired number of clusters (n). Through trial and error, a threshold value of T=0.02 was determined as the optimal starting point for this study. The default value of the branching factor, B=50, was employed. The optimal number of clusters was determined using the silhouette validation index, and 18 clusters were identified as being appropriate for the BIRCH hierarchical method. Clusters 3 and 4 had the highest number of accidents, with 54 and 64 accidents, respectively. The Silhouette validation index diagram is shown in Fig. 7, the clustering output is represented in Fig. 9, and the number of crashes per cluster is depicted in Fig. 11.

Clusters with a high number of accidents can be used to identify accident prone areas of Yazd-Kerman road. By joining clusters 14 and 16 of the Agglomerative Hierarchical algorithm and clusters 3 and 4 of BIRCH clustering, the accident prone areas were determined, which are shown in Fig. 12.

In accident prone area 1, shown in Fig. 12, the old road passes the front side of a resting area (point 3 in Fig. 13). Due to the increase in traffic volume over the years, a decision was made to construct a new road and convert the existing road into a one-way road. In the area of the resting area (Abolfazl Mosque), route A was

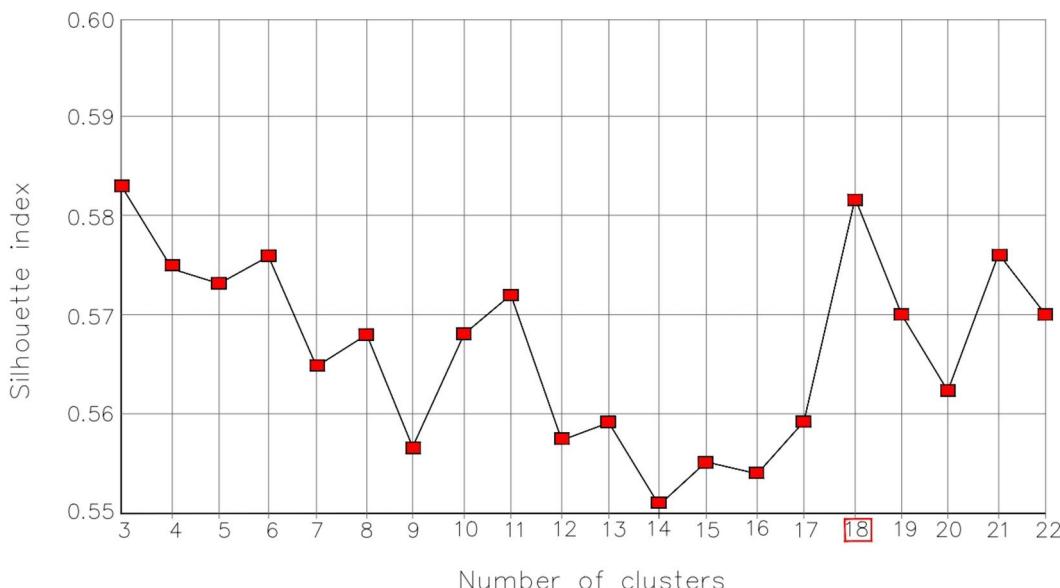


Fig. 6. Silhouette Index for Agglomerative Hierarchical Clustering (created by Excel software <http://www.microsoft.com>).

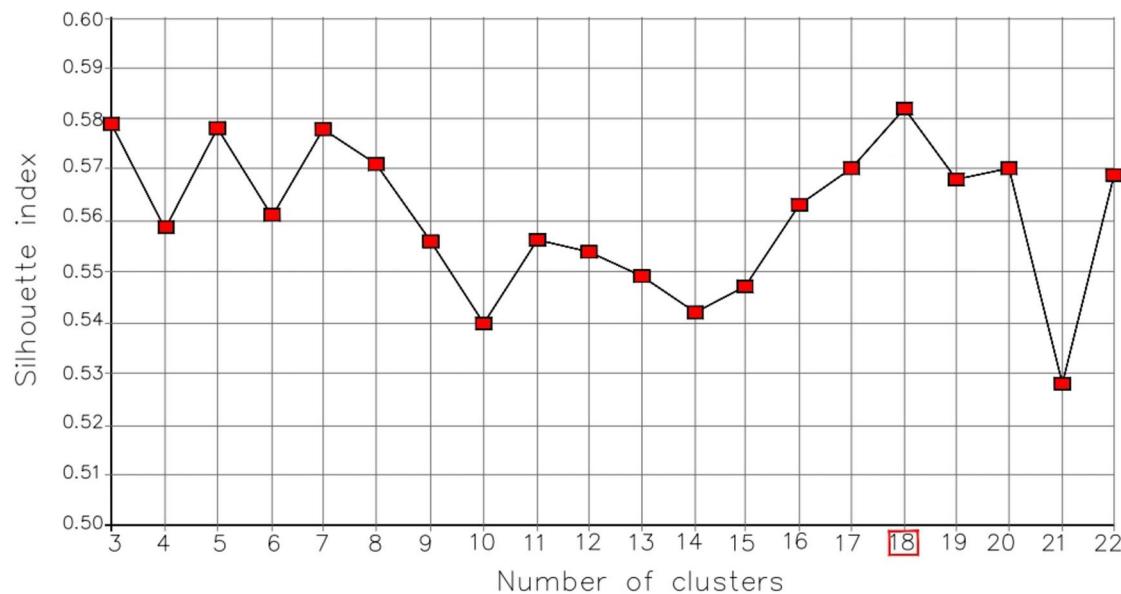


Fig. 7. Silhouette index for BIRCH clustering (created by Excel software <http://www.microsoft.com>).

connected to route B, and in the opposite direction, route C was connected to route D. By analyzing the accident data, as shown in Fig. 14, it was determined that most accidents in this accident prone area occurred during sunrise and sunset, as well as during night-time, when drivers may face reduced awareness levels and poor peripheral vision.

Field investigation revealed that in the direction from C to D, special topographical conditions characterized by a negative longitudinal slope, insufficient lighting at the curve, and inadequate road signage made it difficult for drivers to see the D path. As a result, drivers mistakenly perceived the extension of path C as A, which caused their vehicles to leave the road and overturn in this hazardous situation.

Furthermore, it was observed that the non-standard turning ramp at point 2 did not provide a sufficient length for the acceleration lane, which created turbulence in traffic flow and increased the risk of accidents. Additionally, the improper separation of paths between high-speed traffic and drivers intending to stop at Abolfazl mosque or leave the place and enter the road has resulted in severe accidents at point 3.

The overall findings indicate that the primary cause of accidents in the first accident prone area was inadequate geometric design for providing access to the main road for mosque-goers. It is clear that changing the mosque's land use is not feasible. Therefore, to mitigate the probability of accidents, it is essential to rectify the road's geometric design. The U-turns at the entrance and exit of this accident prone area should be redesigned, and the utilization of traffic signs should be optimized to enhance clarity and awareness for unfamiliar drivers with the road layout in that spot. Enhanced lighting can also significantly improve drivers' situational awareness. Additionally, a well-designed access-mobility level is crucial in managing the speed of vehicles leaving the main road to stop at the resting area and parking lots, and vice versa.

In the second accident prone area (depicted in Fig. 12), the causes of accidents were entirely different from those in the first area. Statistical data analysis (Fig. 15) indicates a strong correlation between the identified accident hotspots and the area's weather conditions. Based on field studies of the region and information obtained from social resources, it was discovered that the area experiences sudden rainfall during the summer. Due to various factors, including poor vegetation, proximity of the mountain to the road, and the slight slope of the land, a large area of the region is susceptible to low-speed flooding, which causes soil erosion and sedimentation of silt and clay from the mountainside to the roadside (as shown in Fig. 16).

Furthermore, the existence of a high-speed wind corridor in this part of the road results in the rapid formation of sandstorms, which decreases driving visibility due to increased wind speeds. Consequently, the majority of accidents in this area occur when a mass of soil passes over the Yazd-Kerman road during a sandstorm.

Moreover, the unsafe geometric design of the road in this area has significantly increased the risk of accidents. Inadequate consideration of the stopping sight distance in the design of the vertical curve and the presence of a horizontal curve immediately after the unsafe vertical curve have made this section of the road more hazardous, particularly at night or during a sandstorm when horizontal visibility is reduced. As a result, numerous vehicles have lost control and left the roadway.

The slope of alignments in the vertical curve can be reduced to improve the sight distance and enhance safety in the second accident prone area. Additionally, to further increase of the road safety in this area, it is advisable to widen the road by paving the unpaved shoulder, planting trees along the path, increasing the number of reflective signs, and using shy bars on the road borders. These measures are expected to diminish the probability of accidents in that hotspot.

Other factors, such as vehicle quality in Iran, intensify the accidents. The Pride model, manufactured by SAIPA Automaker Company, is prone to accidents 21% more than other cars in Iran. Research has shown that

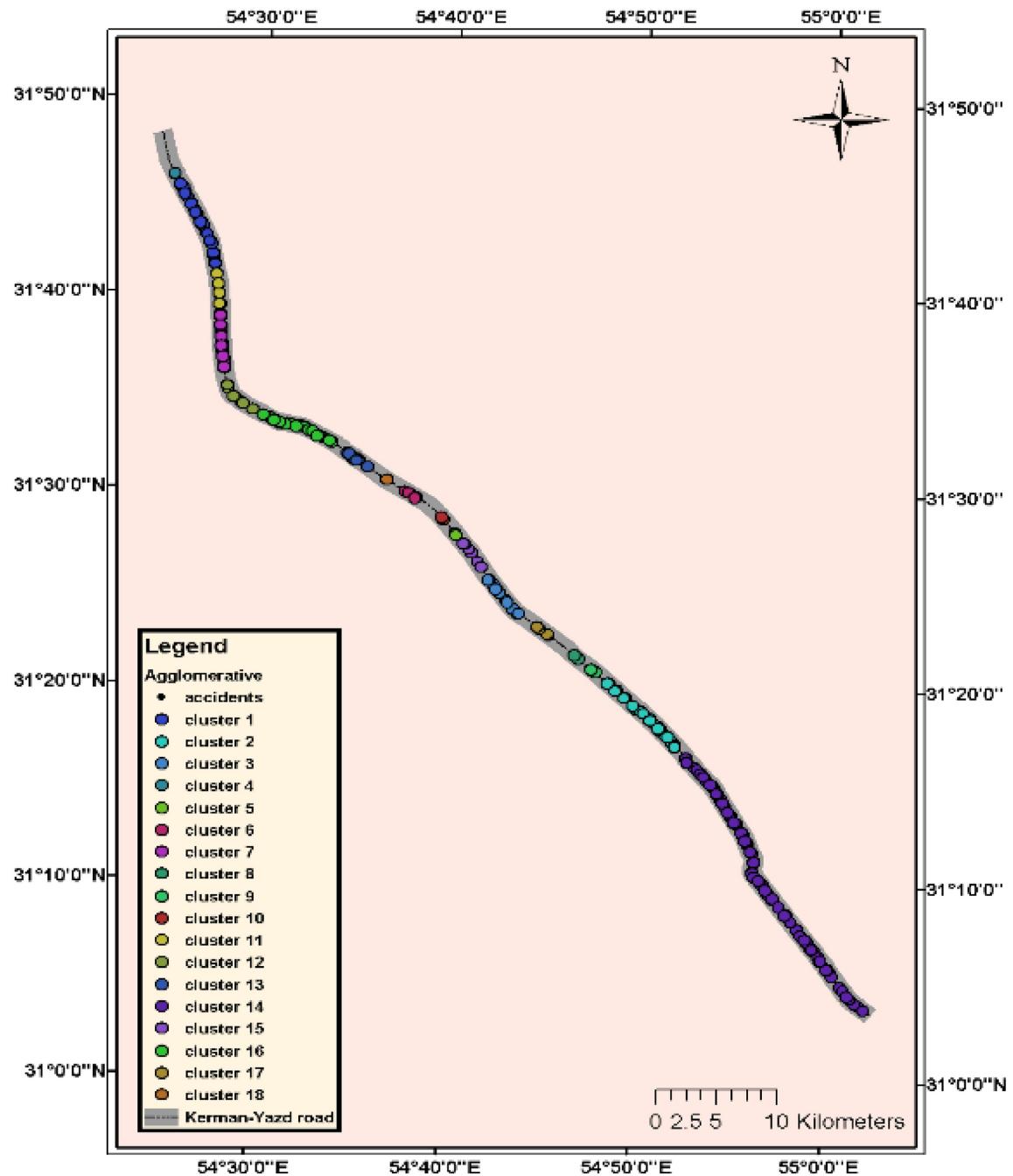


Fig. 8. Output of Agglomerative Hierarchical Clustering (created by ArcGIS 10.4 software <http://www.esri.com>).

accidents involving SAIPA's Pride car, the most affordable vehicle in Iran, are likely to result in higher road traffic injuries at the accident scene⁴³. However, considering that this case is particularly restricted to Iran, it has been excluded from the current research. Nonetheless, vehicle quality is an essential factor that should be considered in any study aimed at reducing the intensity of accidents.

Accident severity

In this study, the RF and KNN classification algorithms were employed to predict the severity of accidents along the road. To accomplish this, the data was divided into four classes, as shown in Table 3. Class 0 contains accident-free points and was selected based on the features listed in Table 1 along the path. Class 1 includes damage accidents, Class 2 comprises injury accidents, and Class 3 encompasses fatal accidents. By dividing the data into these classes, the classification algorithms were able to more accurately predict the severity of accidents based on their occurrence history.

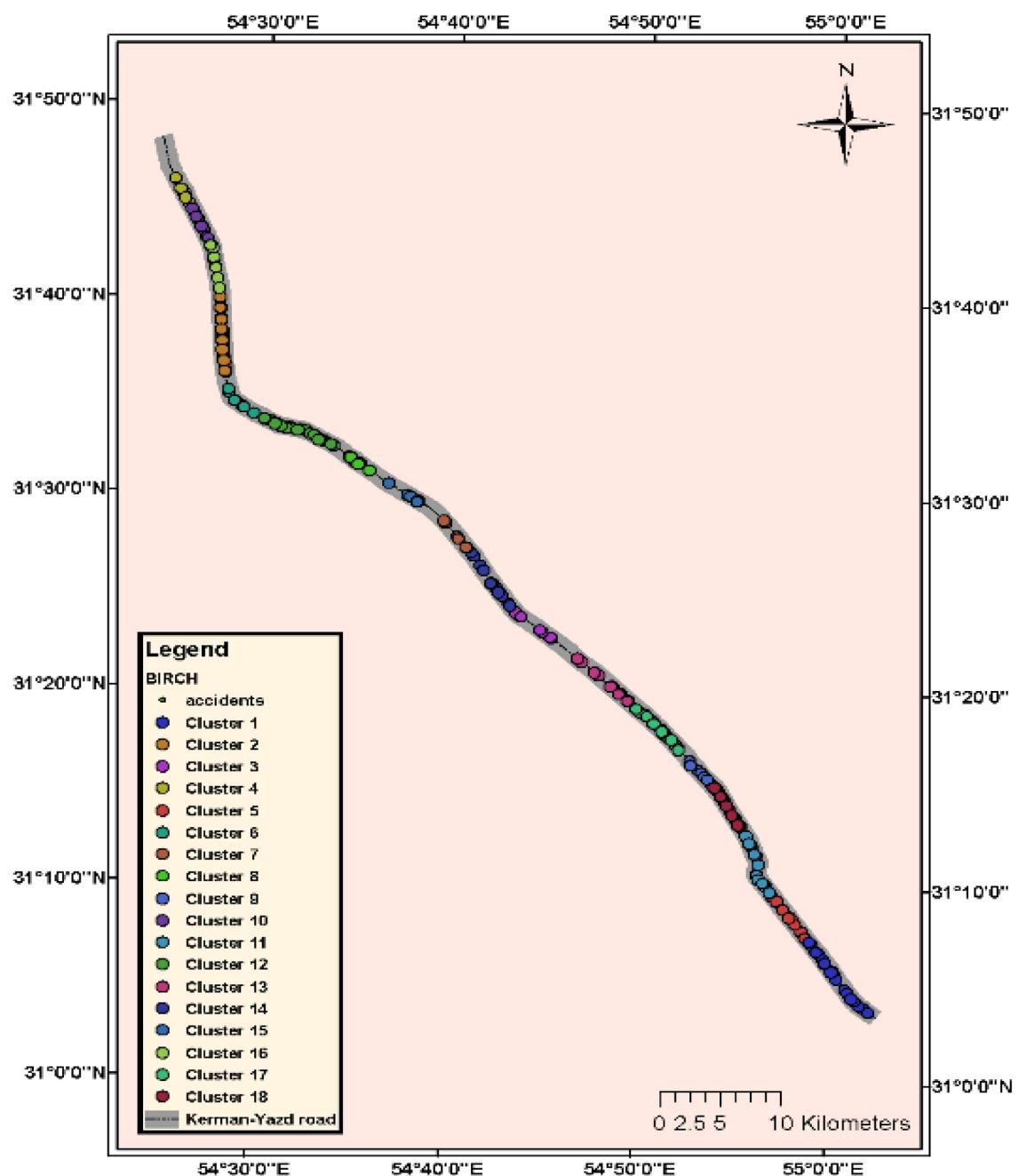


Fig. 9. Output of BIRCH clustering (created by ArcGIS 10.4 software <http://www.esri.com>).

To predict the severity of accidents using the KNN algorithm, the parameter and distance function should be determined. In this study, the Euclidean distance function was utilized, as it has been shown to perform well in previous research. During the training phase, different values for the K parameter were tested, ranging from 1 to 12. Ultimately, a value of K=2 was selected for the Yazd-Kerman road, as it led to the best overall accuracy, which was consistent with previous studies, such as⁴² and⁴⁴. As demonstrated in Fig. 17, this model achieved an accuracy of over 71% by examining the two nearest neighbors. The confusion matrix is presented in Table 4. Table 5 show the validation criteria for this method using the test data collection.

According to Fig. 17, increasing the K value decreases overall accuracy. The reason is that the features that are spatially close to each other possess close attributes because they all have similar slopes, climates, alignment, and lighting. In contrast, those accidents that are far from the intended feature have different attributes to that feature because they are in different locations with different attributes. In this study, as illustrated in Fig. 17, analyzing two nearest neighbors can yield an accuracy exceeding 71%. By using more essential attributes, a higher level of accuracy can be achieved.

The RF algorithm randomly selects several predictive variables, which are a subset of the total variables. To predict the severity of accidents, this algorithm requires two critical parameters: the number of trees (n-tree),

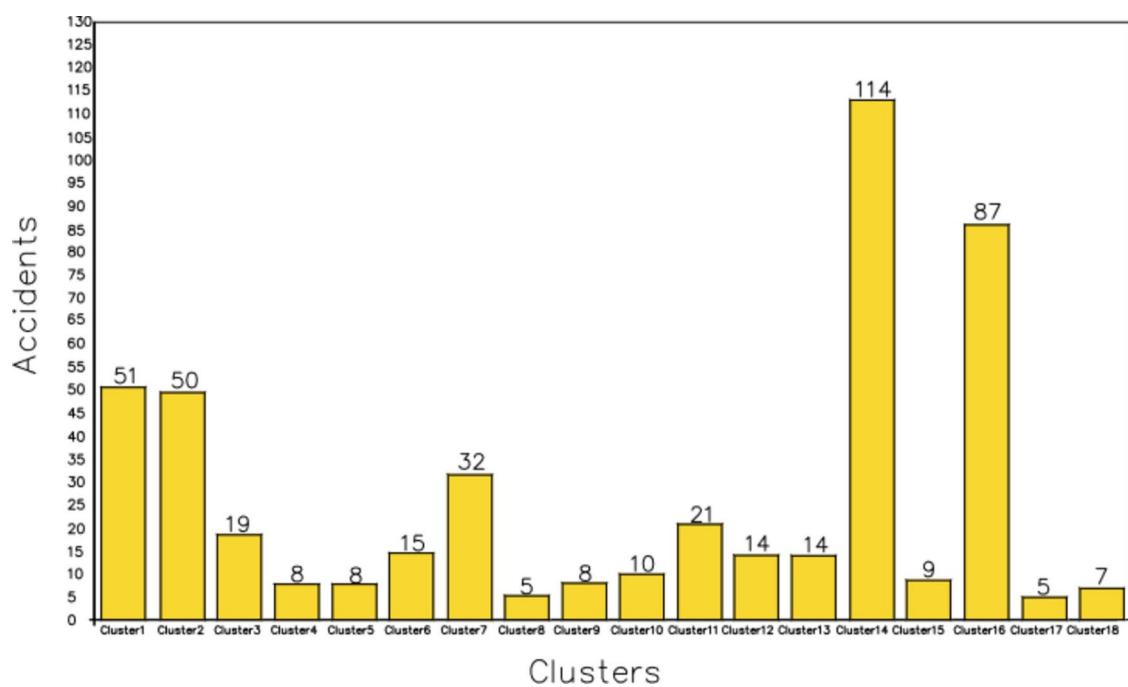


Fig. 10. Accident per cluster in Agglomerative Hierarchical Clustering (created by Rapidminer software docs. rapidminer.com).

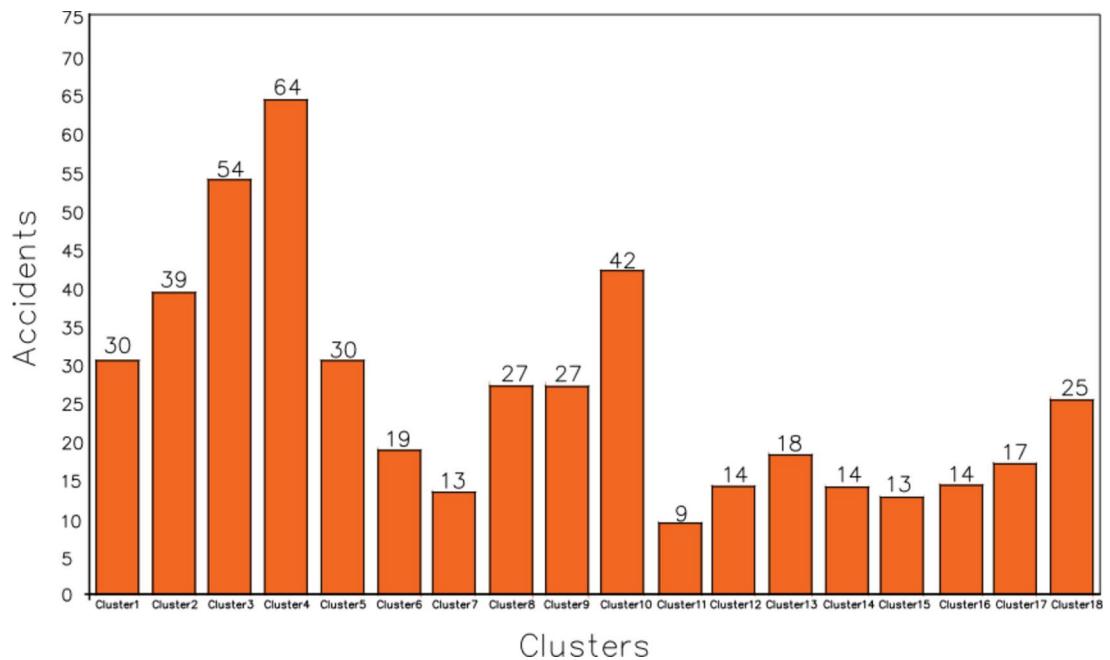


Fig. 11. Accident per cluster in BIRCH clustering (created by Rapidminer software docs.rapidminer.com).

and the number of random samplings (MTRY). During the training process, the first parameter was determined, as displayed in Fig. 18. The second parameter was initiated based on the first parameter during the algorithm's implementation. Figure 18 indicates that the construction of 15 decision trees enables the Forest algorithm to attain an overall accuracy of 60%. The confusion matrix is presented in Table 6. Table 7 displays the validation criteria for this method using the test dataset.

After evaluating the outcomes of the RF and KNN algorithms, it was discovered that the KNN algorithm provides higher accuracy than the RF algorithm. Consequently, accidents that occur close to each other share nearly identical characteristics. By analyzing the attributes of two neighboring accidents, the severity of an

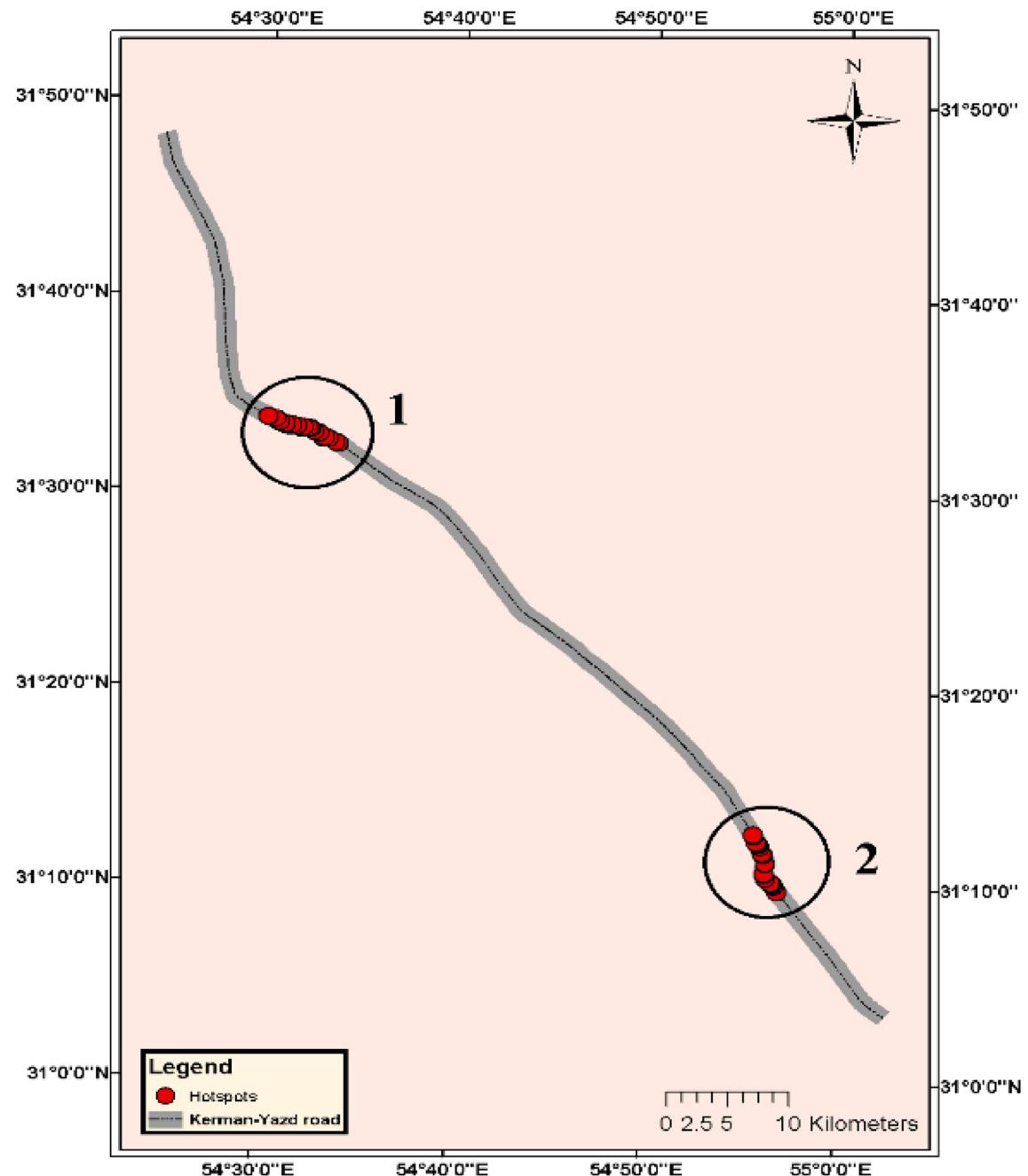


Fig. 12. Final accident prone areas (created by ArcGIS 10.4 software <http://www.esri.com>).

accident can be predicted. Accidents that occur near each other tend to have similar attributes. Therefore, by analyzing the number of accidents surrounding a specific accident, it is possible to predict its attributes. The results of the confusion matrix in Tables 4 and 6 demonstrate that both methods accurately predicted accident-free points and had desirable attributes, such as appropriate climate (clear and cloudy), standard slope (within $\pm 3\%$), level alignment, and sufficient lighting (Day). This result highlights the essential impact of these four attributes on accidents on the Yazd-Kerman road. However, due to the low number of training data, the test data model struggled to predict fatal accidents.

The KNN algorithm is particularly notable for its capacity to predict the damage and injury levels of accidents. Compared to the RF algorithm, the KNN algorithm provides a more accurate prediction of injury and damage levels in accidents. Several similar studies, including²¹ and²², have also reported that the KNN method outperforms other algorithms when it comes to predicting accident severity. Therefore, KNN has the potential to become a promising tool for predicting the severity of accidents on roads.

However, the KNN algorithm does not necessarily have a better performance than the RF algorithm in accident analysis. The performance of each algorithm depends on the specific dataset and the problem being

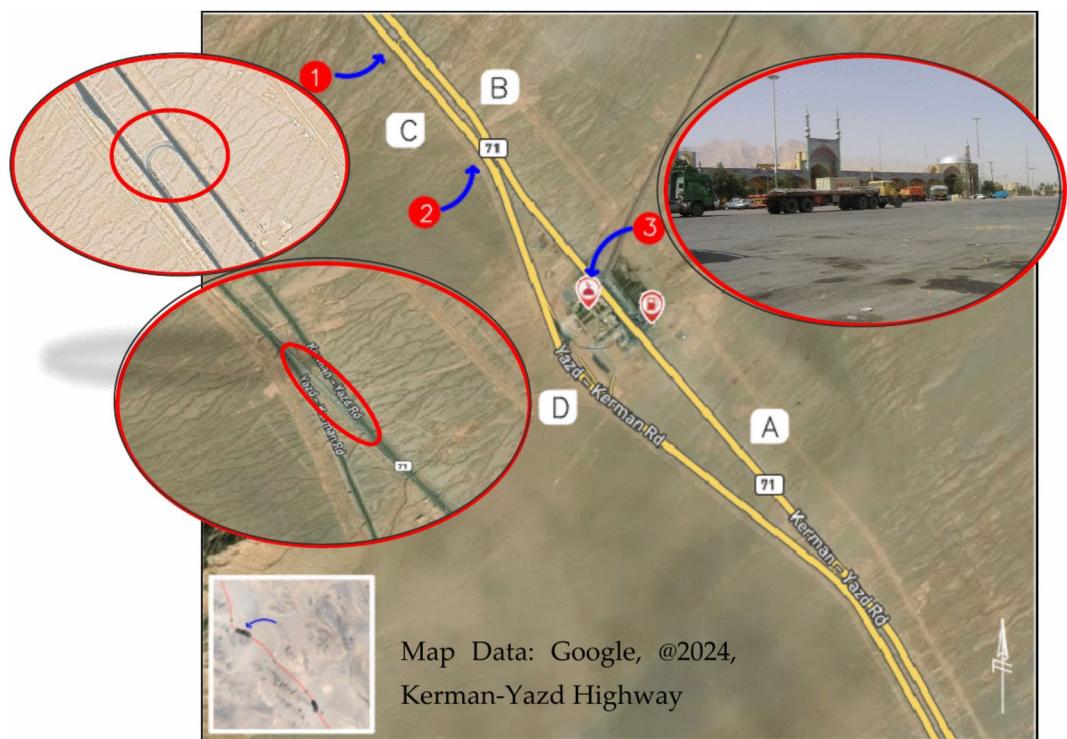


Fig. 13. Field investigation of the accident prone area, cluster one (created by <http://www.maps.google.com>).

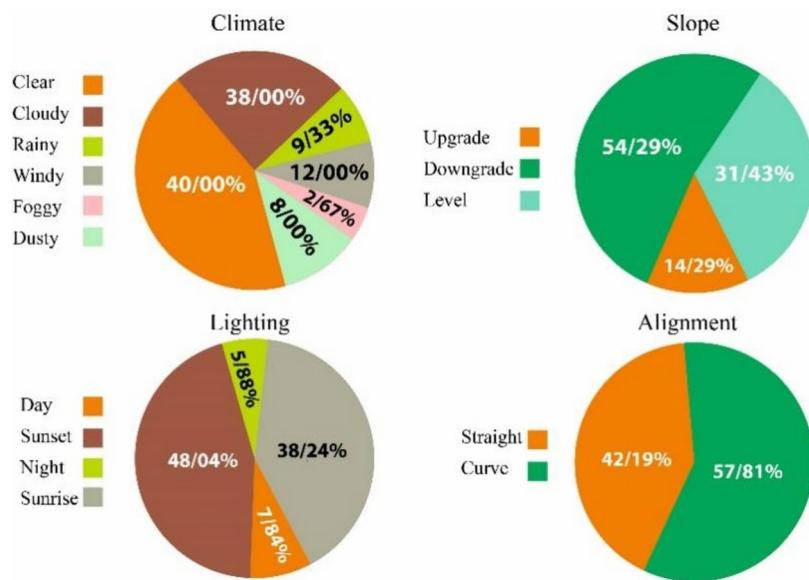


Fig. 14. Statistics of cluster one's parameters (created by Excel software <http://www.microsoft.com>).

solved. Nevertheless, KNN is a simple and intuitive algorithm that can work well for certain types of accident analysis problems. It operates by finding the K-nearest neighbors to a given data point and using their labels to predict the label of the new point. This can be effective when the data has a clear structure and the nearest neighbors are likely to have similar labels. On the other hand, RF is a more complex algorithm capable of accommodating a broader array of data structures and relationships among variables. It functions by building a large number of decision trees and combining their predictions to develop a final prediction. This can be effective when there are data-related complexities in the relationships between variables and a large number of features. Ultimately, the choice of algorithm depends on the specific problem being solved and the characteristics

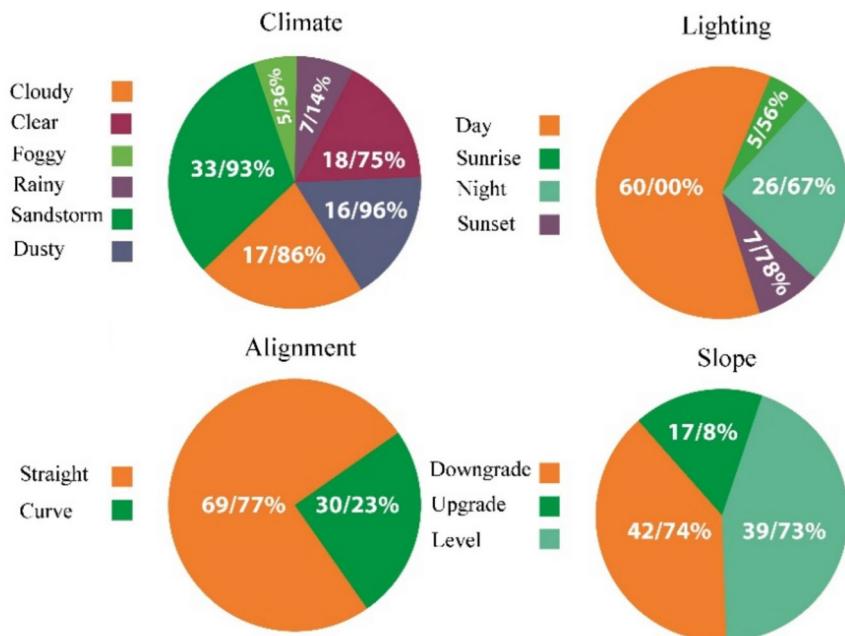


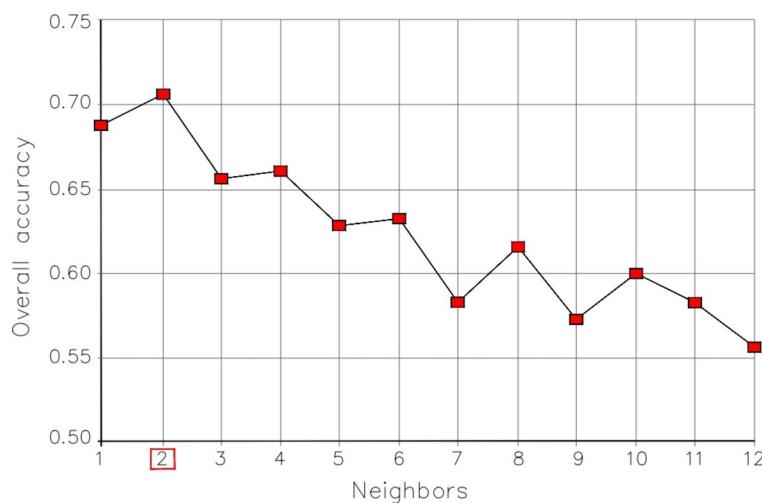
Fig. 15. Statistics of cluster two's parameters (created by Excel software <http://www.microsoft.com>).



Fig. 16. Field investigation of the accident prone area, cluster two (created by <http://www.maps.google.com>).

Class	Name	Training	Testing	Sum
Class 0	Accident-free	81	23	104
Class 1	Damage	162	51	213
Class 2	injury	174	33	207
Class 3	Fatal	35	10	45

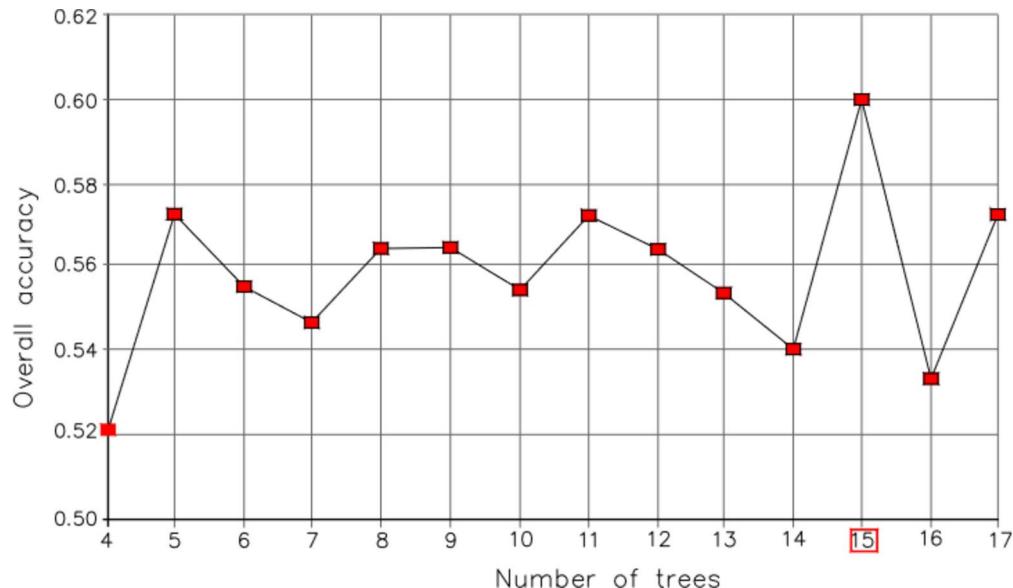
Table 3. Number of accidents per class.

**Fig. 17.** Overall Accuracy of KNN (created by Excel software <http://www.microsoft.com>).

Model	Actual	Predicted				
		Class 0	Class 1	Class 2	Class 3	Sum
KNN	Class 0	20	0	0	0	20
	Class 1	3	32	13	0	48
	Class 2	0	18	19	3	40
	Class 3	0	1	1	7	9
	Sum	23	51	33	10	117

Table 4. Confusion matrix of KNN.

Model	Precision	Recall	Overall Accuracy	Kappa
KNN	75%	72%	71%	60%

Table 5. Accuracy of KNN.**Fig. 18.** Overall Accuracy of RF (created by Excel software <http://www.microsoft.com>).

Model	Actual	Predicted				
		Class 0	Class 1	Class 2	Class 3	Sum
RF	Class 0	21	0	1	0	22
	Class 1	1	25	12	3	41
	Class 2	1	25	18	5	48
	Class 3	0	1	2	2	5
	Sum	23	51	33	10	117

Table 6. Confusion matrix of RF.

Model	Precision	Recall	Overall Accuracy	Kappa
RF	60%	56%	60%	41%

Table 7. Accuracy of RF.

of the dataset. It is important to try multiple algorithms and compare their performance to determine which one works best for a given problem.

Conclusions and recommendations

The present study examined the application of machine learning algorithms and spatial analyses to identify the accident prone areas along the Yazd-Kerman road. This research successfully assessed the results of clustering, revealing the fundamental factors contributing to accidents in order to guide future practical interventions. Additionally, a framework was established to predict the severity of accidents along the road and to suggest strategies for risk reduction. The following conclusions were drawn from the current research.

Agglomerative Hierarchy and BIRCH clustering algorithms were able to identify clusters with a significant number of accidents. Each method identified two clusters with a high concentration of accidents. By identifying the intersection of these clusters, two areas were determined as being particularly susceptible to accidents.

In the first area, the main reasons for accidents were the presence of a resting area (Abolfazl mosque) that caused traffic chaos, insufficient lighting at curves, and improper road signage. Most accidents occurred during sunrise and sunset, which coincided with the time for praying, emphasizing the mosque's role in being a possible cause of accidents in this area. In the second accident prone area, the main cause of accidents was the downward slope and the presence of a high-speed wind corridor perpendicular to the road direction, which limited the driver's visibility.

The intelligence techniques applied in the present research were able to predict accidents or their severity, which is an effective approach to mitigating the severity of accidents and their outcomes. Two classification algorithms, Random Forest and K-Nearest Neighbor, were able to classify and predict accidents' severity. In comparison to the Random Forest technique, the K-Nearest Neighbor approach exhibited a higher level of effectiveness, recording an overall accuracy of 71% as opposed to 60%.

The results indicated that both algorithms were capable of correctly identifying more than 80% of accident-free points, demonstrating a strong correlation between road and environmental characteristics and the occurrence of accidents. These results highlight the importance of using classification algorithms to predict the severity of accidents, leading to the implementation of preventive measures to reduce the occurrence and severity of accidents, in addition to enabling the efficient allocation and use of construction resources.

Accurately identifying the locations of accidents poses a challenge for studies in this field. To achieve reliable results, the inaccuracy of the coordinates of accidents should not be more than a few meters. Even in practice, the police should pay attention to the fact that some accidents may be misleading as the precise position of the accident cannot be determined. Thus, the police can only record the approximate position of the accident. This inaccuracy is also true for the time of the accident.

Research limitations and future research

The results of this study could have been more satisfactory if additional attributes were available. Validating the number of injury accidents with the help of hospital reports may be a good idea for future research if hospital records are available. Furthermore, by including additional attributes, such as traffic density and road surface conditions, the classification algorithms might yield more precise predictions and could potentially identify the factors that contribute to accidents' severity with higher accuracy. Also, due to the limitation of data, this study did not consider the type of vehicles (such as trucks, sedans, motorbikes, etc.); however, taking vehicle types into account is recommended. It is worth mentioning that the number of accidents for each vehicle type should be considerable, otherwise, the prediction will not be thoroughly accurate.

Data availability

Data are contained within the article. To ask about the data, any of the corresponding authors, Farhad Hosseinali (F.H., f.hosseinali@sru.ac.ir), and Mostafa Adresi (M.A., m.adresi@sru.ac.ir), are responsible.

Received: 3 August 2024; Accepted: 25 November 2024

Published online: 01 December 2024

References

- Adresi, M., Baghalishahi, A. M., Zeini, M. & Khishdari, A. Impact of speed limit method on motorway safety. *Gradevinar* **68**, 705–713 (2016).
- WHO. *Save lives: a road safety technical package*. (2017).
- WHO. *Global status report on road safety*. (2018).
- Soltani, A., Mansouriyanis, O., RoohaniQadikolaei, M. & Zaroujtaghi, A. Two decades of geospatial evolution: tracing the analytical journey towards data-driven road crash prevention. *Appl. Spat. Anal. Policy* **17**(3), 1301–1334 (2024).
- Wu, B.-J., Jin, L.-H., Zheng, X.-Z. & Chen, S. Coupling analysis of crane accident risks based on Bayesian network and the N-K model. *Sci. Rep.* **14**(1), 1133 (2024).
- Yan, M., Jia, J. & Chen, Y. Research on accident early warning of metallurgical enterprises based on grey DEMATEL/ISM and Bayesian network. *Sci. Rep.* **14**(1), 18312 (2024).
- Ghaffari, A., Tavakoli Kashani, A. & Moghimdarzi, S. Identification of black spots based on reliability approach. *Promet Traffic Transp.* **25**(6), 525–532 (2013).
- Xu, Q. & Tao, G. Traffic accident hotspots identification based on clustering ensemble model. In *2018 5th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2018 4th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom)*. (2018).
- Tavakoli Kashani, A., Rakhshani Moghadam, M. & Amirifar, S. Factors affecting driver injury severity in fatigue and drowsiness accidents: a data mining framework. *J. Inj. Violence Res* **14**(1), 75–88 (2022).
- Agrawal, K., Ruth, V. M., Sravani, K. & Bharathi, C. Analysis of road accident locations using DBSCAN algorithm. In *International Journal of Scientific Research in Science and Technology* (2018).
- Hajela, G., Chawla, M. & Rasool, A. A clustering based hotspot identification approach for crime prediction. *Procedia Comput. Sci.* **167**, 1462–1470 (2020).
- Mohammed, A. F. & Baiee, W. R. The GIS based criminal hotspot analysis using DBSCAN technique. *IOP Conf. Ser. Mater. Sci. Eng.* **928**(3), 032081 (2020).
- Karami, A. & Johansson, R. Choosing DBSCAN parameters automatically using differential evolution. *Int. J. Comput. Appl.* **91**, 1–11 (2014).
- Akbari, Z. & Unland, R. Automated determination of the input parameter of DBSCAN based on outlier detection. In *Artificial Intelligence Applications and Innovations*. (Springer International Publishing, 2016).
- Puspitasari, D., Wahyudi, M., Rizaldi, M., Nurhadi, A. & Ramanda, K. K-means algorithm for clustering the location of accident-prone on the highway. In *Journal of Physics: Conference Series*. (IOP Publishing, 2020).
- Wan, X., Wang, J., Du, Y. & Zhong, Y. DBH-CLUS: a hierarchical clustering method to identify pick-up/drop-off hotspots. In *2015 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*. (IEEE, 2015).
- Khosravi, Y., Hosseinali, F. & Adresi, M. Assessing road accidents in spatial context via statistical and nonstatistical approaches to detect road accident hotspot using GIS. *Geodetski Vestnik* **66**(3), 412–431 (2022).
- Soltani, A., Edward Harrison, J., Ryder, C., Flavel, J. & Watson, A. Police and hospital data linkage for traffic injury surveillance: A systematic review. *Accid. Anal. Prev.* **197**, 107426 (2024).
- Beshah, T. & Hill, S. Mining road traffic accident data to improve safety: role of road-related factors on accident severity in Ethiopia. In *AAAI Spring Symposium: Artificial Intelligence for Development* (2010).
- Tamije Selvy, P., Ragul, M., Vignesh, G. N. & Anitha, V. M. Real time efficient accident predictor system using machine learning techniques (KNN, RF, LR, DT). *Int. J. Eng. Adv. Technol.* **10**(2), 108–111 (2020).
- Xing, L. et al. Comparison of different models for evaluating vehicle collision risks at upstream diverging area of toll plaza. *Accid. Anal. Prev.* **135**, 105343 (2020).
- Iranitalab, A. & Khattak, A. Comparison of four statistical and machine learning methods for crash severity prediction. *Accid. Anal. Prev.* **108**, 27–36 (2017).
- 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. *Accid. Anal. Prev.* **154**, 106094 (2021).
- Amiri, A. M., Sadri, A., Nadimi, N. & Shams, M. A comparison between artificial neural network and hybrid intelligent genetic algorithm in predicting the severity of fixed object crashes among elderly drivers. *Accid. Anal. Prev.* **138**, 105468 (2020).
- Curry, A. E., Pfeiffer, M. R., Carey, M. E. & Cook, L. J. Catalyzing traffic safety advancements via data linkage: Development of the New Jersey Safety and Health Outcomes (NJ-SHO) data warehouse. *Traffic Inj. Prev.* **20**(sup2), S151–S155 (2019).
- Soltani, A. & Roohani Qadikolaei, M. Space-time analysis of accident frequency and the role of built environment in mitigation. *Transp. Policy* **150**, 189–205 (2024).
- Sasirekha, K. & Baby, P. S. Agglomerative hierarchical clustering algorithm—a review. In *International Journal of Scientific and Research Publications* (2013).
- Silva, M. et al. Agglomerative concentric hypersphere clustering applied to structural damage detection. *Mech. Syst. Signal Process.* **92**, 196–212 (2017).
- Szekely, G. J. & Rizzo, M. L. Hierarchical clustering via joint between-within distances: extending ward's minimum variance method. *J. Classif.* **22**(2), 151–183 (2005).
- Zhang, T., Ramakrishnan, R. & Livny, M. BIRCH: an efficient data clustering method for very large databases. *ACM SIGMOD Rec.* **25**(2), 103–114 (1996).
- Shirkhorshidi, A. S., Aghabozorgi, Sr., Wah, T. & Herawan, T. *Big data clustering: a review*. In *Computational Science and Its Applications—ICCSA 2014: 14th International Conference*. 2014. Portugal: Springer International Publishing.
- Lv, Y., Tang, S. & Zhao, H. Real-time highway traffic accident prediction based on the k-nearest neighbor method. In *2009 International Conference on Measuring Technology and Mechatronics Automation* (2009).
- Guerrero-Estrada, A.-Y., Quezada, L. F. & Sun, G.-H. Benchmarking quantum versions of the kNN algorithm with a metric based on amplitude-encoded features. *Sci. Rep.* **14**(1), 16697 (2024).
- Pulungan, A., Zarlis, M. & Suwilo, S. *Performance Analysis of Distance Measures in K-Nearest Neighbor* (2020).
- Wu, Y. et al. Risk assessment of bridge construction investigated using random forest algorithm. *Sci. Rep.* **14**(1), 20964 (2024).
- Dieterich, T. G. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Mach. Learn.* **40**(2), 139–157 (2000).
- Piryonesi, S. M. & El-Diraby, T. E. Data analytics in asset management: Cost-effective prediction of the pavement condition index. *J. Infrastruct. Syst.* **26**(1), 04019036 (2020).
- Gupta, V. K., Gupta, A., Kumar, D. & Sardana, A. Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model. *Big Data Min. Anal.* **4**(2), 116–123 (2021).
- Ramo, R. & Chuvieco, E. Developing a random forest algorithm for MODIS global burned area classification. *Remote Sens.* **9**(11), 1193 (2017).
- Kulkarni, A., Chong, D. & Batarseh, F.A. *5 - Foundations of data imbalance and solutions for a data democracy*. In *Data Democracy*. In (eds Batarseh, F. A. & Yang, R.) 83–106 (Academic Press, 2020).

41. Singh, P., Singh, N., Singh, K. K. & Singh, A. Chapter 5 - Diagnosing of disease using machine learning. In *Machine Learning and the Internet of Medical Things in Healthcare* 89–111 (Academic Press, 2021).
42. MomeniKho, S. M., Pahlavani, P. & Bigdeli, B. Classification and association rule mining of road collisions for analyzing the fatal severity, a case study. *J. Transp. Health* **23**, 101278 (2021).
43. Mohajer, B., Azmin, M., Mohebi, F., Ahmadi, N. & Farzadfar, F. Low-quality domestic automobiles continue to threaten lives in Iran: economic instability as the potential contributor. *Arch. Iran. Med.* **23**(11), 764 (2020).
44. Zhang, S., Li, X., Zong, M., Zhu, X. & Cheng, D. Learning k for kNN Classification. *ACM Trans. Intell. Syst. Technol.* **8**(3), 1–19 (2017).

Author contributions

Conceptualization, F.H. and M.A.; methodology, F.H. and M.A.; software, Y.K.; validation F.H and M.A.; formal analysis, Y.K.; investigation, F.H. and M.A.; resources, F.H. and M.A.; data curation, Y.K.; writing—original draft preparation, Y.K.; writing—review and editing, F.H. and M.A.; visualization, Y.K. and M.A.; supervision, F.H. and M.A.; project administration, F.H. and M.A.; funding acquisition, F.H. and M.A., All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by Shahid Rajaee Teacher Training University, grant numbers 5973.59 and 5973.63.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to F.H. or M.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024