IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Identification Of Black Clusters To Prevent Road Accidents Using Machine Learning

[1]MULLANGI VENKATA RAHUL VISHNU, [2]SAKAMURI JAYANTH
[12]Vellore Institute of Technology, Chennai, Tamilnadu

ABSTRACT

The total number of motor vehicles is still increasing at a high rate due to the social economy's rapid expansion and the speed at which cities are becoming more populated. Large and medium-sized cities' roads are getting more and more congested, which increases the frequency of traffic accidents. Finding accident hotspots early on is crucial to improving road safety and lowering the number of traffic accidents. Eight impact factors (holiday, day of week, time, rush hour traffic, accident location type, accident type, weather, responsibility, and black spot) were set for the analytic dataset in this study, which included data from traffic accidents on the Lianfeng Middle Road, Yinzhou District, Ningbo City. The classic K-means clustering algorithm's drawbacks were addressed by the proposal of the enhanced algorithm, which is vulnerable to early clustering centres and outliers. The dataset's traffic accidents were split into two groups using this algorithm: black spots and non-black spots. After that, we used the new dataset to build a black spot recognition model using a Bayesian network, and we compared it with other popular methods including the ID3 decision tree, logistic regression, and support vector machine. As demonstrated by the values of the ROC area, TP rate, FP rate, accuracy, precision, recall, F-measure, and F-measure reaching 0.618, 0.668, 0.580, 0.650, 0.668, 0.590, and 0.668, respectively, the Bayesian network was the most successful model for locating black spots associated with traffic accidents. In addition, a bivariate correlation model was used to confirm the impact factors and black spots' link. The findings showed that Black spots, which had a value of sig<0.05, showed significant associations with the accident location type, accident type, time, and responsibility. In order to greatly improve road safety, the conclusions may offer reference data for the detection and avoidance of high-risk areas for accidents.

INDEX TERMS Traffic safety, K-means clustering, Bayesian networks, black spot identification.

I. INTRODUCTION

The World Health Organization's global status reports on road safety, which were published in December 2018, highlighted that 1.35 million people die in traffic accidents each year [1]. Injuries from traffic are now the main cause of death for those between the ages of 5 and 29. The research indicates that the majority of road traffic fatalities and injuries occur in low- and middle-income nations. The overall number of traffic accidents in 2017 decreased by 4.6% from 2016 to 2017, but the mortality and injury rates remained high at 4.59 and 15.08 per 100,000 people, respectively, according to the Chinese annual statistics report on road traffic accidents [2]. Using Yinzhou, Ningbo as an example, traffic is growing at a very rapid rate due to the increase in motor vehicles This district has a high accident rate, which is detrimental to urbanisation and economic growth. Utilising traffic accident data resources to create focused traffic safety measures that will successfully lower the accident rate is crucial for enhancing road safety. Since it is not feasible to increase road safety at every point of collision due to financial and time constraints, black spot recognition technology is being researched extensively to pinpoint accident-prone areas. Analysing the cause of black spots is also a cost-effective and efficient technique to reduce traffic accidents.

Black spots for traffic accidents, sometimes referred to as dangerous road spots, are areas of the road where there has been a noticeable trend over time in both the frequency and severity of traffic accidents. As of yet, the term "black spot" in relation to traffic accidents lacks a universally recognised definition. Here are a few traditional definitions. According to UK definitions, a black spot is a 100-meter stretch of road where four incidents take place annually. A 100-meter stretch of road that has seen more than four fatal traffic incidents in the previous four years is referred to as a "black spot" in Norway. Divergent views exist on the definition of the black spot recognition targets in earlier research as well. Similar kinds of auto accidents usually happen regularly anywhere on a highway. These black patches were dubbed "accident black spots" by Shen et al. [3]. A site with one or more accidents was considered a "black spot" by Meuleners et al. [4]. A dense area surrounding a cluster centre is known as a black spot, according to Murat and Cakici [5]. The official definition of black spots, which is where the accident data came from, will serve as the basis for this study's definition of the term.

Three primary techniques can be employed in the identification of black spots: crash prediction methods, clustering techniques, and screening techniques [6]. Based on various threshold values for road safety, black patches can be simply screened. However, this approach of black spot recognition cannot provide recommendations for the projects needed to lower the safety concerns because in practise it is simple to overlook some trigger-ing aspects (road conditions, accident severity, etc.). As a result, the prediction of traffic accidents has made extensive use of machine learning algorithms in recent years. They are able to effectively categorise datasets and create a connection between the variables and the intensity of the traffic incidents. Classification, regression, clustering, and dimensionality reduction are examples of machine learning algorithms. K-means is an algorithm that is algorithm for dynamic clustering that relies on partitioning. There is no specific restriction on the clustering range for this widely used data-identification technique, and many datasets can be performed in parallel as long as they are independent of one another. The effectiveness of utilising the K-means algorithm for black spot detection in lanes was demonstrated by Gupta et al. [7].

A K-means spatial clustering technique was introduced by Zhong and Liu [8] in order to unify spatial similarity and entity attribute similarity. Black spots for traffic accidents have also been identified using alternative clustering algorithms. For instance, Murat and Cakici [5] evaluated the characteristics and employed entropy to locate accident dark patches using fuzzy clustering. When it comes to classification, the Bayesian network (BN) algorithm is adaptable. algorithm for dynamic clustering that relies on partitioning. There are no paaccident datasets because this is a widely used data-identification technique. It incorporates the decision-makers' experience with the previously collected data. Furthermore, without the need for assumptions, BNs can be used to examine the relationships between variables in order to provide predictions.

To illustrate the value of the Bayesian model in forecasting the severity of injuries, Deublein et al. [9] examined the Austrian Highway Research Network scenario. BNs were employed by Mujalli et al. [10] to enhance the imbalanced accident dataset classification. Delphi technology and BNs were used by Mbakwe et al. [11] to forecast traffic accidents in underdeveloped nations. Additional classification formulas, such as the logistic regression (logistic), support vector machine (SVM), and ID3 decision tree (ID3), have been used in earlier research to categorise various dataset types.

In recent years, scholars have attempted to approach the true definition of black spots while taking into account their unpredictability and uncertainty. To improve identification accuracy, they have combined multiple methods. Regression analysis and artificial neural networks were employed by Ali and Tayfour [12] to forecast traffic accident fatalities. Positive results were obtained when comparing the collected data with the predictions. To ascertain the threshold of black spot recognition, Lu [13] used a quality control technique based on hierarchical storage management and the empirical Bayesian model. Dereli and Erdogan [14] used spatial statistics based on models to ascertain the accident black spots. By comparing the negative binomial regression, Poisson regres- sion and empirical Bayesian models used in the study, it was found that the empirical Bayesian approach provides the best results in terms of consistency and accuracy. Xiao [15] proposed an SVM and K-nearest neighbour ensemble learn- ing method to improve the robustness of traffic incident detection. Debrabant *et al.* [16] used the optimized kernel density clustering method and the buffer method to determine black spots for accident occurrence points, road sections, and regions.

In the current study, K-means clustering has been appliedto determine the road traffic accident black spots, and BNs have been mostly used to classify and predict traffic acci- dents. Few studies have combined the two methods to study the classification and prediction of black spots. Thus, a unique identification model fusing BNs and K-means was put out in this work. To demonstrate that the approach is efficient and ideal, three contrast models were developed.

The steps involved in this study are as follows. The collecting and processing of data on traffic accidents, together with data filtering and variable setup, are first presented. The accident black spot recognition models are then established using an upgraded K-means clustering approach in conjunction with BNs, ID3, logistic, and SVM algorithms. Third, a receiver operating characteristic curve (ROC) and additional evaluation indicators are used to assess the accuracy of four models. Ultimately, a model for correlation analysis is presented in order to identify the critical variables that have a strong association with black spots for traffic accidents.

As a result, this work provided a unique identification model that blends BNs and K-means. Three comparison models were created to show the methodology's effectiveness and optimality.

The organisation of this study is comprised of the subsequent steps. The collection, processing, filtering, and variable configuration of data related to traffic incidents are all covered in the first section. Next, an improved K-means clustering strategy is combined with BNs, ID3, logistic, and SVM algorithms to build the accident black spot identification models. Third, the accuracy of four models is evaluated by means of a receiver operating characteristic curve (ROC) and further evaluation signals. Finally, a correlation analysis model is offered to find the essential factors that show a strong correlation with black areas for traffic accidents.

II. METHODOLOGY

A. K-MEANS ALGORITHM

The K-means algorithm was first proposed by MacQueen [17]. Its effectiveness and simplicity have led to its widespread application and study in recent years. The K-means algorithm works specifically as follows.

Algorithm 1 Traditional K-Means Algorithm

```
Input: Dataset $D = \{x_1, x_2, \ldots, x_m\}$; Cluster number, $K$.
Output: Cluster division $C = \{C_1, C_2, \ldots, C_K\}$
Select K samples from D as the initial mean vector randomly $\{\mu_1, \mu_2, \ldots, \mu_K\}$
Initialization: $C_i \leftarrow \emptyset (1 \leq i \leq k)$
Repeat
  For $j = 1, 2, \ldots, m$ do
    $d_{ji} = \|x_j - \mu_i\|_2$ /* Calculate the distance between the sample $x_j$ and each mean vector $\mu_i (1 \leq i \leq k)$*/
    $\lambda_j = \arg\min_{i \in \{1,2,\cdots,k\}} d_{ji}$ /* Determine the cluster tag of $x_j$ based on the closest mean vector */
    $C_{\lambda_j} = C_{\lambda_j} \cup \{x_j\}$ /* Divide the sample $x_j$ into the corresponding cluster */
  End for
  For $i = 1, 2, \ldots, K$ do
    $\mu_i' = \frac{1}{C_i} \sum_{x \in C_i} x$ /* Calculate the new mean vector */
    If $\mu_i' \neq \mu_i$ then
      Update the current mean vector $\mu_i$ to $\mu_i'$
    Else
      Keep the current mean vector unchanged
    End if
  End for
Until the current mean vector is not updated
```

The conventional K-means method has a few shortcomings. It is necessary to specify the starting cluster number (K) in advance. In actuality, though, the value of K is challenging in order to decide. Furthermore, the initial clustering centre has a significant impact on the outcome. Distinct clustering outcomes will ultimately arise from the random selection of the initial clustering centres. A local minimal solution is simple to get into. Furthermore, the K-means algorithm's application to big datasets becomes problematic due to its susceptibility to isolated points, commonly referred to as noise data. An enhanced K-means algorithm is explained as follows in order to hasten data clustering and boost black spot recognition accuracy: First, the individual traffic accident spots are eliminated based on distance; second, the cluster center's coordinates are initialised; and third, the conventional K-means algorithm is invoked. Among the benefits of the enhanced algorithm is the removal of noise data, which has an impact on the identification accuracy and the absence of a requirement to predetermine K's value. These upgrades may improve clustering's precision and efficiency.

B. BAYESIAN NETWORKS

Judea Pearl (1988) made the initial proposal for the Bayesian network. It is made up of a conditional probability table (CPT) and a direct acyclic graph (DAG). There are two phases to the construction of BNs: 1) Create a network structure with all attribute and class variables using the BN learning technique; 2) Determine the probability of the class variables when the attribute values variables are given the initial cluster centre coordinates based on Naïve Bayes assumes that attributes are independent and equally important. However, this assumption might not be correct in reality.

Algorithm 2 Improved K-Means Algorithm

```
Input: Dataset $D = \{x_1, x_2, \ldots, x_m\}$; Screening distance, $r$; Threshold, $N_0$; Mutation distance, $M \gg m$.
Output: Cluster division $C = \{C_1, C_2, \ldots, C_K\}$
For $x_j \in D$ do
  Calculate the number of data points ($N$) with $x_j$ as the centre and $r$ as the radius
  If $N \leq N_0$ then
    Remove $x_j$ from $D$
  Else
    $D' \leftarrow x_j$
  End if
End for
For $x_j \in D'$ do
    $m_{X_j} = |X_j - X_{j-1}|$; $m_{Y_j} = |Y_j - Y_{j-1}|$ /* Arrange the samples $x_j(X, Y)$ by coordinate size; calculate the coordinate distance between two adjacent data points */
    If $m = M$ then
      $x_j$ and $x_{j-1}$ implement jumps between the clusters
      Mark the number of times $m$ equals $M$ as $k$, and the number of clusters is $K = k + 1$
      $C_i = \left[\frac{X_j+X_j'}{2}, \frac{Y_j+Y_j'}{2}\right]$, $(1 \leq i \leq k)$ /* Calculate the initial cluster centre coordinates based on the two data points at the edge of each cluster */
        $C \leftarrow C_i$
      Call the traditional K-means algorithm
    End if
  End for
```

1) THE BN LEARNING BASED ON THE TAN CLASSIFIER

The Naïve Bayes classifier serves as the foundation for the TAN (tree-augmented naïve Bayes) classifier. A BN's structure based on the TAN classifier is displayed in Fig.1.

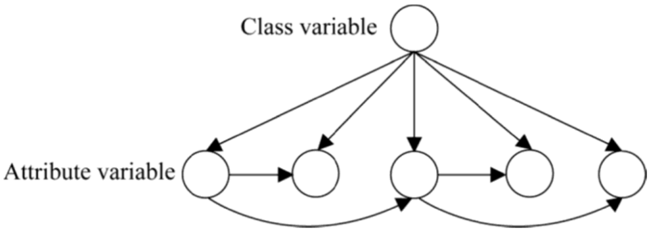
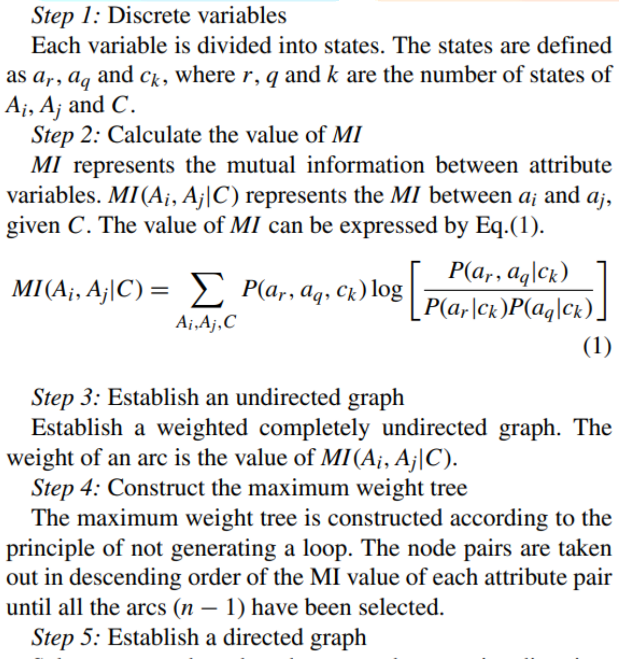


FIGURE 1. The structure of a BN based on the TAN classifier.

i Mutual information (MI) is a feature selec- tion algorithm used to represent the correlation betweentwo variables. Using the MI value as a weight reflects that different attribute variables have dissimilar effects on the classification and largely eliminates the influence of the A feature selection approach called mutual information (MI) is used to show how two variables are correlated. The influence of the MI value is essentially eliminated when using it as a weight, reflecting the fact that different attribute variables have varying effects on the categorization.

Assumption of independence regarding the consequences of classification. A typical TAN is a learning algorithm that uses the conditional mutual information to build a TAN classifier. The hypothesis for training set D is finished. Ai and Aj(i, j 1, 2... n) are the variables holding attributes. The class variable is C. So, the following formula can be used to determine the learning of the BN structure.



Step 1: Discrete variables

Each variable is divided into states. The states are defined as a_r, a_q and c_k, where r, q and k are the number of states of A_i, A_j and C.

Step 2: Calculate the value of *MI*

MI represents the mutual information between attribute variables. $MI(A_i, A_j|C)$ represents the *MI* between a_i and a_j, given C. The value of *MI* can be expressed by Eq.(1).

$$MI(A_i, A_j|C) = \sum_{A_i,A_j,C} P(a_r, a_q, c_k) \log \left[\frac{P(a_r, a_q|c_k)}{P(a_r|c_k)P(a_q|c_k)} \right] \tag{1}$$

Step 3: Establish an undirected graph

Establish a weighted completely undirected graph. The weight of an arc is the value of $MI(A_i, A_j|C)$.

Step 4: Construct the maximum weight tree

The maximum weight tree is constructed according to the principle of not generating a loop. The node pairs are taken out in descending order of the MI value of each attribute pair until all the arcs $(n - 1)$ have been selected.

Step 5: Establish a directed graph

where TP is the quantity of samples that both predictably and truly test positive; FP is the proportion of samples that are expected to be positive but are in fact negative; The number of samples that are predicted to be positive but are actually negative is known as FN, while the number of samples that are actually negative but are predicted to be negative is known as TN.

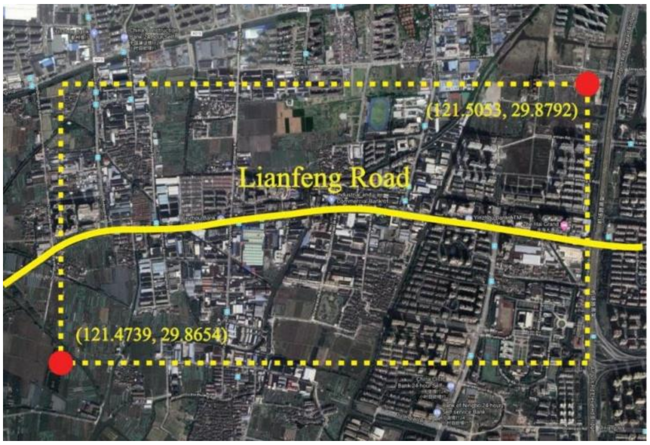
D. THE INDICATORS OF EVALUATION

Receiver operating characteristic curve, often known as the sensitivity curve, or ROC curve: The identical signal stimulation is reflected by the spots on the curve. It is a curve created by utilising a number of distinct two-category techniques (the decision threshold or demarcation value). The true positive rate, or sensitivity, is the ordinate. A false positive rate (1−specificity) is the abscissa. A statistical technique that can accurately characterise a classification model's overall test performance is ROC curve analysis.

The AUC, or area under the curve: The area enclosed by the ROC curve and the coordinate axis is referred to as the AUC. The AUC value ranges from 0.5 to 1 since the reference line (y = x) is usually above the ROC curve. In terms of value, the AUC

The remaining five comprehensive evaluation indicators are described as follows. The true positive rate (TPR), which is used to measure sensitivity, is the percentage of samples that are both anticipated to be positive and are actually positive. The false positive rate (FPR), which is equivalent to 1-specificity, is the percentage of samples that are actually negative but are predicted to be positive. As the ratio between the TP and FP increases, the categorization effect of the approach gets better. Precision (P) is a measure of the proportion of samples that are predicted to be positive in a positive scenario. The coverage indicator that is comparable to sensitivity is called recall (R). The F-measure (F) can recall indicators and evaluate precision in a complete manner. The classifier works better when the value of F is higher. One popular indicator is accuracy (ACC). This is the proportion of accurately identified samples to the total number sampled. Generally speaking, a higher ACC value indicates greater classifier performance. Equations (6-11) are used to obtain the indications.

For this reason, Lianfeng Middle Road was selected as the object road. The area was selected by taking into account the longitude and latitude (east longitude: 121.4739◦ to 121.5053◦, north latitude: 29.8654◦ to 29.8792◦). throughout all, 1,005 accidents occurred throughout the region. Figures 2 and 3 show.



III. DATA COLLECTION AND PROCESSING

A. DATA COLLECTION

Prior to the study, the traffic accident alert data were gathered via the "Yinzhou Traffic Police" application's big data platform. There are twelve items in all: the event number, the location in terms of latitude and longitude, the event time, the squadron event, the alarm, the event position, the type of event, the reason of the event, the weather, the event environment, and pictures of the scene. 37,654 genuine occurrences from the fourth quarter of 2016 were kept after removing the duplicate data.

One of the busiest areas in the Yinzhou District is Lianfeng Middle Road. Approximately 1000 traffic events every quarter, with a variety of complex event kinds, have happened on Lianfeng Road and its adjacent areas in recent years, according to Yinzhou traffic accident data reports.

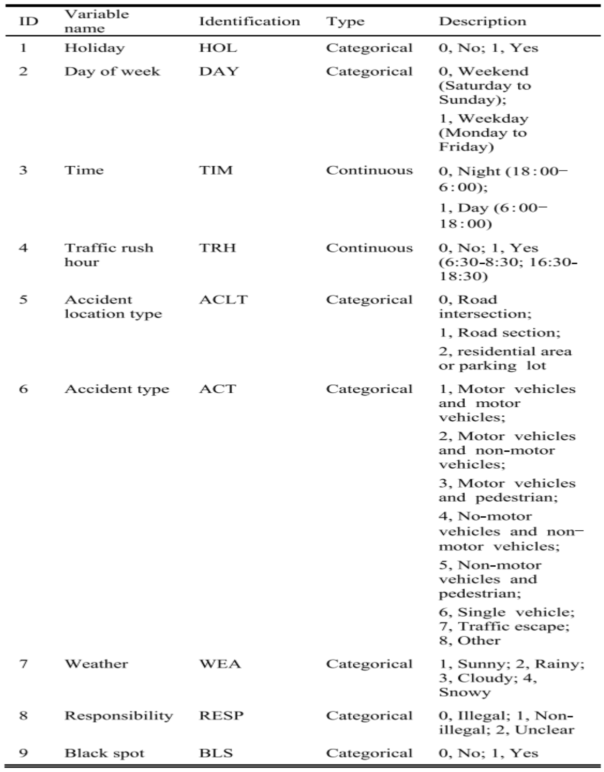
B. DATA PROCESSING

The This study examined data on traffic accidents for the Lianfeng Middle Road area from the fourth quarter of 2016. Based on the idea of lowering model complexity and raising model accuracy, nine variables were chosen. As attribute variables, eight impact factors were defined: DAY, TRH, ACLT, ACT, WEA, RESP, and HOL. The class variable was thought to be HOS.

Before being used, the data must be processed in order to ensure the accuracy of the analysis results. Two categories of traffic accidents were established: zero-hot spot and one-non-hot spot. Daytime hours were from 6:00 to 18:00, and nighttime hours were from 18:00 to 6:00. The definition of rush hour for traffic in China was followed. Eight different sorts of accidents occurred in three different types of accident locations. Table 1 displays the general data for the chosen variables.

IV. RESULTS

TABLE 1. Variable descriptions.



ID	Variable name	Identification	Type	Description
1	Holiday	HOL	Categorical	0, No; 1, Yes
2	Day of week	DAY	Categorical	0, Weekend (Saturday to Sunday); 1, Weekday (Monday to Friday)
3	Time	TIM	Continuous	0, Night (18:00–6:00); 1, Day (6:00–18:00)
4	Traffic rush hour	TRH	Continuous	0, No; 1, Yes (6:30-8:30; 16:30-18:30)

5	Accident location type	ACLT	Categorical	0, Road intersection;
				1, Road section;
				2, residential area or parking lot

6	Accident type	ACT	Categorical	1, Motor vehicles and motor vehicles; 2, Motor vehicles and non-motor vehicles; 3, Motor vehicles and pedestrian; 4, No-motor vehicles and non–motor vehicles; 5, Non-motor vehicles and pedestrian; 6, Single vehicle; 7, Traffic escape; 8, Other
7	Weather	WEA	Categorical	1, Sunny; 2, Rainy; 3, Cloudy; 4, Snowy
8	Responsibility	RESP	Categorical	0, Illegal; 1, Non-illegal; 2, Unclear
9	Black spot	BLS	Categorical	0, No; 1, Yes

C. THE K-MEANS CLUSTERING RESULTS

The Yinzhou traffic police department defines an accident black site as follows. Considering the quarterly. A possible black spot is defined as an area that, on average, has more than 25 traffic accident points within a 50-meter radius. First, the black spots on Lianfeng Middle Road were located using the refined K-means method. After carrying out the experiment multiple times, the cluster centres remained in the same locations and had a regular distribution. Six clusters were created from the dataset after 334 black spots were chosen from 1005 accidents, as illustrated in Fig. 4. It is challenging to calculate the number of classes (K) when using the accident data from Lianfeng Middle Road. In light of this, the enhanced algorithm's result above indicates that it was simple to get the where the dark dots are located using the conventional K-means technique. On the other hand, the cluster centres' dispersion is erratic and random. Repeated trials could reveal that the clustering is erroneous. Figure 4 displays the distribution of the black spots based on the conventional K-means algorithm. As a result, the outcomes demonstrate that the enhanced algorithm was able to locate the cluster centres with accuracy and efficiency.

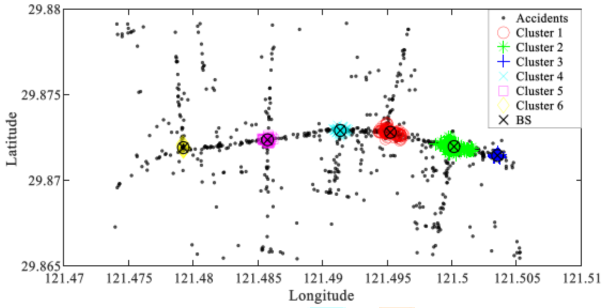


FIGURE 4. Distribution of the black spots based on the improved K-means algorithm

D. BAYESIAN STRUCTURE OF NETWORKS

The BNs based on the TAN algorithm were created using the Weka software (Witten and Frank, 2013). Thirty percent of the information was used for testing, and the remaining seventy percent was kept for training the BNs. Fig. 6 depicts the structure of a BN, which consists of 15 directed arcs and 9 nodes. Table 1 displays nine nodes that correspond to eight attribute variables and one class variable.

E. THE OUTCOMES OF CLASSIFICATION

A. THE RESULT OF THE PEARSON CORRELATIONCOEFFICIENT TEST

The Pearson correlation coefficient was chosen to investigate the connection between the eight impact factors and the black patches using SPSS software. The findings are shown in Table 3 and show a significant association (p<0.05) between traffic accident dark patches and four criteria (accident type, timing, location, and culpability). 0.05 denotes the statistical significance of the correlation between the two selected variables. However, there is no discernible relationship between black patches and WEA, PEP, DAY, HOL, or DAY. Additionally, Table 3 displays the results of the computation of the correlation coefficient between the selected factors and the black spots. The correlation coefficient's absolute value for the accident site type test result is the greatest in the dataset, indicating that the accident site's outcome A key factor in determining whether black spots are present is the type test. Factors 3, 6, and 8 are also associated with accident dark markings.

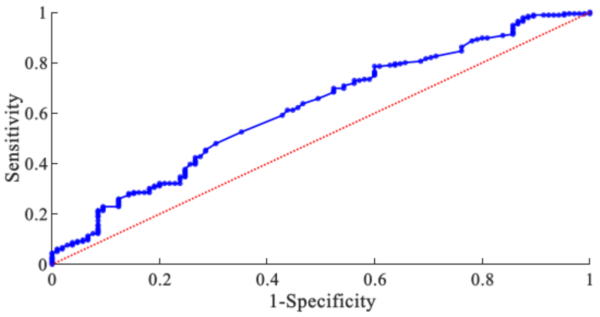
A detailed description of each of the four significant correlation variables can be found below.

1) ACCIDENT LOCATION TYPE

Between ACLT and BLS, the correlation coefficient has a value of 0.266. We can plainly see that more black marks emerged at road intersections when compared to the other accident location types (road segment, residential area, and parking lot). This result is in line with the findings of Al-Ghamdi [19] and Savolainen et al. [20], who discovered that road intersections are the site of more incidents than other places and that location has a substantial impact on traffic accidents.

FIRST, THE ROC CURVE

To be compared with the BN model mentioned above, the ID3, logistic, and SVM models were also developed. Figure 7 displays the ROC curves for the four classification models, with sensitivity shown on the Y-axis and 1-specificity on the X-axis. Every single



(a) The ROC curve for the BN model

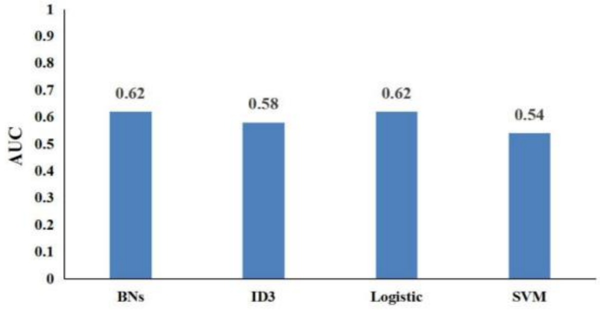


FIGURE 8. AUC for the four models.

2) ACCIDENT TYPE

It was discovered that ACT and BLS were highly correlated. The correlation coefficient between them is 0.141, meaning that compared to other accident types, a collision involving a motor vehicle has the biggest effect on the black areas. This is in line with the results of Yang et al. [21], who discovered that a significant contributing factor to urban road traffic accidents is the quantity of motor vehicles. We might conclude that further research is needed to fully understand motor vehicle collisions.

3) TIME

The accident time was split into two periods for this study: daytime (6:00–18:00) and nighttime (18:00–6:00). The correlation coefficient between TIM and BLS has a value of 0.104. It was previously established in a study that nighttime visibility is worse than during the day, and that drivers are more prone to distractions.

These are the primary reasons for nighttime traffic accidents [22]. This is consistent with the study's findings.

4) RESPONSIBILITY

The correlation coefficient only gets as high as 0.083. Nonetheless, it is one of the important variables and demonstrates that illegal activity has a bigger effect on black spots than non-illegal behaviour. The primary infractions that lead to accidents on Lianfeng Middle Road are speeding, driving while intoxicated, driving and driving while fatigued. Hordofa et al. [23], Zhang et al. [24] emphasised the influence of unlawful actions on traffic accidents, including speeding and driving while exhausted.

V. DISCUSSION AND CONCLUSION

This study applied updated K-means clustering and BN algorithms to identify traffic accident black areas using Ningbo traffic accident data. The logistic, SVM, and ID3 algorithms were compared to the BN. A number of measures were used to assess how well the four models performed. Based on the experimental data, a correlation analysis between the variables was also performed. The findings demonstrate that: 1) the enhanced K-means algorithm can screen out the Lianfeng Middle Road area's traffic accident black spots with a reasonable degree of accuracy; and 2) all four models can successfully identify black spots, with accuracy levels above 0.6, though each model's performance varies. The BN, however, is better than the other three.

The accident position type, accident kind, time, and responsibility are the factors that have the biggest effects on the black mark. There are several restrictions on this study. When there is no clear interval in the distribution of the accident spots, the enhanced K-means algorithm is not appropriate. Sifting out the black spots becomes more difficult when there is a tight gap between two types of cluster centres. There were no specific subjective accident factors, such as driver characteristics or speed, and the variables utilised in the BN were not comprehensive. Thus, next research will need to thoroughly examine the K-means and BN algorithms in order to enhance the black spot recognition model through additional development and experiments to carry out more research on road accidents.

REFERENCES

[1] *Global Status Report on Road Safety*, WHO, Geneva, Switzerland, 2018.

[2] *Chinese Annual Statistics Report on Road Traffic Accidents*, Ministry of Public Security, Beijing, China, 2017.

[3] X. Shen, X.-C. Guo, and J.-M. Song, ''Study on road traffic accident blackspot identification method,'' *J. Highway Transp. Res. Develop.*, vol. 20, no. 4, pp. 95–97, Aug. 2003.

[4] L. B. Meuleners, D. Hendrie, A. H. Lee, and M. Legge, ''Effectiveness of the black spot programs in western Australia,'' *Accident Anal. Prevention*,vol. 40, no. 3, pp. 1211–1216, May 2008.

[5] Y. S. Murat and Z. Cakici, ''An integration of different computing approaches in traffic safety analysis,'' *Transp. Res. Procedia*, vol. 22,pp. 265–274, Jan. 2017.

[6] M. Ghadi and Á. Török, ''Comparison different black spot identification methods,'' *Transp. Res. Procedia*, vol. 27, pp. 1105–1112, Jan. 2017.

[7] A. Gupta, R. Ajaykumar, and P. S. N. Merchant, ''Automated lane detection by K-means clustering: A machine learning approach,'' *Electron. Imag.*, vol. 14, pp. 1–6, Feb. 2016.

[8] Y. Zhong and D. Liu, ''The application of K-means clustering algorithm based on Hadoop,'' in *Proc. ICCCBDA*, Jul. 2016, pp. 88–92.

[9] M. Deublein, M. Schubert, B. T. Adey, J. Köhler, and M. H. Faber, ''Predic- tion of road accidents: A Bayesian hierarchical approach,'' *Accident Anal.Prevention*, vol. 51, pp. 274–291, Mar. 2013.

[10] R. O. Mujalli, G. López, and L. Garach, ''Bayes classifiers for imbalanced traffic accidents datasets,'' *Accident Anal. Prevention*, vol. 88, pp. 37–51, Mar. 2016.

[11] A. C. Mbakwe, A. A. Saka, K. Choi, and Y.-J. Lee, ''Alternative method of highway traffic safety analysis for developing countries using Delphi technique and Bayesian network,'' *Accident Anal. Prevention*, vol. 93,pp. 135–146, Aug. 2016.

[12] G. A. Ali and A. Tayfour, ''Characteristics and prediction of traffic accident casualties in sudan using statistical modeling and artificial neural net- works,'' *Int. J. Transp. Sci. Technol.*, vol. 1, no. 4, pp. 305–317, Dec. 2012.