Accident Analysis and Prevention 157 (2021) 106164

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

A performance analysis of prediction techniques for impacting vehicles in hit-and-run road accidents

Alok Nikhil Jha [a,*], Niladri Chatterjee [b], Geetam Tiwari [a]

[a] TRIPP, Indian Institute of Technology, Delhi, India
[b] Department of Mathematics, Indian Institute of Technology, Delhi, India

ARTICLE INFO

Keywords:
Hit-and-run
Road accident
Vehicle type prediction
Support vector machine
Linear discriminant analysis
Naïve Bayes
Classification and regression tree
k-Nearest neighbor
Cross validation
Road user safety

ABSTRACT

Road accidents are globally accepted challenges. They are one of the significant causes of deaths and injuries besides other direct and indirect losses. Countries and international organizations have designed technologies, systems, and policies to prevent accidents. However, hit-and-run accidents remain one of the most dangerous types of road accidents as the information about the vehicle responsible for the accident remain unknown. Therefore, any mechanism which can provide information about the impacting vehicle in hit-and-run accidents will be useful in planning and executing preventive measures to address this road menace. Since there exist several models to predict the impacting unknown vehicle, it becomes important to find which is the most accurate amongst those available. This research applies a process-based approach that identifies the most accurate model out of six supervised learning classification models viz. Logistic Reasoning, Linear Discriminant Analysis, Naïve Bayes, Classification and Regression Trees, k-Nearest Neighbor and Support Vector Machine. These models are implemented using five-fold and ten-fold cross validation, on road accident data collected from five mid-sized Indian cities: Agra, Amritsar, Bhopal, Ludhiana, and Vizag (Vishakhapatnam).This study investigates the possible input factors that may have effect on the performance of applied models. Based on the results of the experiment conducted in this study, Support Vector Machine has been found to have the maximum potentiality to predict unknown impacting vehicle type in hit-and-run accidents for all the cities except Amritsar. The result indicates that, Classification and Regression Trees have maximum accuracy, for Amritsar. Naïve Bayes performed very poorly for the five cities. These recommendations will help in predicting unknown impacting vehicles in hit-and-run accidents. The outcome is useful for transportation authorities and policymakers to implement effective road safety measures for the safety of road users.

1. Introduction

Transportation is the lifeline of a country. Road transportation holds a vital significance as it serves as an ecosystem of an inter-related system for all other transportation modes ensuring the last mile connectivity for road users. Still, safety remains a common challenge for transportation. The fast-growing population, increasing motorization, and rapid urbanization has made the road users more vulnerable to accidents which have proved to be a critical challenge globally. The World Health Organization uses the term road traffic injury (WHO Violence and Injury Prevention, 2004), motor vehicle accidents (MVA) is used by U.S. Census Bureau (Statistical Abstract: Motor Vehicle Accidents and Fatalities, 2008) for road accidents. Countries have been planning and designing solutions to prevent accidents and to ensure road user's safety. Traffic accidents lead to fatalities, injuries, and other indirect impacts on the victims, e.g. post-accident traumas, property damage, and wastage of resources. Past researches reveal that deaths in traffic accidents are second highest after cardiovascular diseases (Crundall, 2005). Accidents occur by vehicle(s) crashing on a road user or vehicle(s) or a static fixture such as a pole or tree or road divider. Vehicles in a road accident are categorized as the impacted vehicle or victim vehicle and impacting vehicle or accused vehicle.

In general, details of the vehicle(s) involved in accidents are available. However, there are cases where one does not know the impacting vehicle as the driver flees from the accident scene. These types of accidents are hit-and-run accidents. Hit-and-run accident types, where impacting vehicles are not known, become a bigger problem to solve. Lack of information in an accident is a challenge as this missing

* Corresponding author.
E-mail addresses: alok.nikhil@gmail.com, trz148020@tripp.iitd.ac.in (A.N. Jha).

https://doi.org/10.1016/j.aap.2021.106164
Received 7 November 2019; Received in revised form 12 January 2021; Accepted 27 April 2021
Available online 3 May 2021

information has a significant impact. The impacting vehicles in hit-and-run vehicles are referred to as "unknown".

For an efficient safety system for road accidents, knowledge of unknown vehicles becomes essential. There are two possible cases when an impacting vehicle after the crash flees away. In one case, information about impacting vehicle type is known. In other case, no such impacting vehicle's information is available. In which case, a method is required for finding unknown vehicles. A process-based approach using six machine learning models on accident data is used to identify an accurate model.

2. Literature

With a large number of road transportation users, safety remains a vital concern as road accidents are a leading cause of fatalities and injuries. As per the Global Burden of Diseases (2017), road accidents fatalities are 12th leading cause of deaths in the USA, India, Germany and many other countries notwithstanding the fact that many accidents go unreported. A hit-and-run accident is a matter of global concern for the safety of road users. As per the National Highway Traffic Safety Administration, USA (NHTSA, 2012), fatalities caused by hit-and-run crashes increased by 13.7 %, from 1274 in 2009 to 1449 in 2011. In India, 57089 hit-and-run accidents were reported in 2015 and figure rose to 65000 in 2017 with 25 % increase in fatalities from 20709 to 26000 in 2015 and 2017 respectively (MoRT&H TRW, 2015; MoRTH Accident Trends, 2018). The increase shows the vulnerability of road users to hit-and-run accidents. Many countries have declared hit-and-run accidents as criminal offenses. Much research has been published to understand the dynamics of hit-and-run accidents. Kim et al. (Kim et al., 2008) identified a model combining logistic regression and rough sets to identify factors affecting hit-and-run crashes in Hawaii. Qin (2013) studied factors contributing to hit-and-run crashes in China. The researchers also proposed to control the ways in which hit-and-run accidents can possibly occur. Macleod (2012) concluded that driving under the influence of alcohol in the early morning increases the probability of hit-and-run accidents. Tay (2009) proposed that education, awareness campaigns and traffic enforcements can reduce hit-and-run accidents and is complemented by Aidoo (2013). USA and EU countries have improved accident recording mechanisms (Jha et al., 2020) and data is used in various studies on causes. However, not much research is done in finding the unknown vehicles.

The knowledge of unknown vehicles of hit-and-run accidents has been found to have a vital relevance in understanding the dynamics of road, road users, environmental conditions, or any other related parameters and plan precautionary measures and reduce the vulnerabilities to road users. However, this subject is a matter of utter importance. We have proposed an approach to identify the best model to predict the unknown vehicles. An unknown vehicle is a missing value in accident record in hit-and-run accidents. There are three types of missingness (Jha et al., 2018) in accident data viz. missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). MCAR is an unsystematic missingness where the probability of a value being missing is a random event and unrelated to other records in the data. MAR is a systematic missingness where the probability of being missing is justifiable by variables recorded in the data. It is not random at all, and missing data can be correlated and predicted. MNAR (or NMAR) type is neither MAR nor is MCAR; the value of the variable that is missing is directly related to the reason for its missingness. The accident data recorded with missing impacting vehicles is an MAR (missing at random) (Jha et al., 2018) variable. A supervised learning based predictive analysis is followed for the prediction of these missing unknown vehicles. Six models have been tested for the prediction of missing data viz. (A) Logistic Regression (or LR), (B) Linear Discriminant Analysis (or LDA), (C) k-nearest neighbor (or KNN), (D) Classification and Regression Trees (or CART), (E) Support Vector Machine (or SVM), (F) Naive Bayes (or NB). LR is a regression-based statistical model to predict the dependent variable. KNN is based on nearest values of existing labels in class and used for the test sample data. SVM finds a separating hyperplane or line between data of two classes and takes data as an input and output. LDA searches in a given data set for a linear combination of predictors and model them in two targets. NB uses Bayes theorem of probability and predicts the class of unknown data set. CART follows a recursive partitioning approach for prediction. Several independent studies have been done on these supervised learning models; however, we are working on cross-pollination of these models with the road accidents and evaluate the best model that can be further used to predict unknown vehicles. This paper employs a process designed using applied prediction models on accident data from six Indian cities, i.e., Agra, Bhopal, Amritsar, Ludhiana, and Vizag (or Vishakhapatnam). The most accurate model is ascertained using the prediction accuracy of six models.

2.1. Prediction models

There are several ways to analyze data depending on the type of data and goal of the analysis. There are three most commonly used data analysis techniques applicable on a given data set:

A Exploratory
B Descriptive
C Predictive

The exploratory data analysis technique explores data to identify unexpected structured observations from the data. It includes graphical representations such as histograms, scatters and pie charts. The descriptive approach identifies the relationships within the data to discover properties. The approach finds the features of given data and transforms data into an expressive form for reporting. The predictive approach, is implemented in prediction by esvaluating patterns of available data. The predictive techniques are based on several classifications and regression-based models of Artificial Intelligence (AI) performed by machine learning. The models learn data patterns and can predict missing information. These techniques provide accurate computation for prediction and forecasting. These techniques have several models and categorized as:

A Supervised Learning,
B Unsupervised Learning,
C Semi-supervised Learning and
D Reinforcement Learning.

Supervised learning techniques are a data mining task, where statistical models with correct instances collect the knowledge called training step. Based on the knowledge and existing structure of training, the Supervised Machine Learning algorithms are applied when the learning of the model from the available or past data is applied to the new data to predict future events or data points. Classification and regression models are examples of supervised learning. In an unsupervised learning, there is no pattern available in data. Hence training is not possible as in the case of supervised learning. The organization of underlying data is the goal. It uses clustering, dimensionality reduction and association techniques. Semi-supervised techniques handle the mix of supervised and unsupervised data where both labeled data and unlabeled data exist. There are systems where some data parameters follow a structure and some do not have a fixed pattern. Segregating these kinds of data is either too complicated or costly or not at all feasible. In Reinforcement Learning techniques, the output depends on, the current input and output of previous inputs.

Data about road traffic accidents of different cities are collected that include missing attributes. The data collected has a structure, and input variables follow the structure within their boundary values. The data has some missing values, and we have focused on those missing values of

identifying the impacting vehicle involved in a hit-and-run accident. To predict the unknown vehicle type or missing values, which falls in the category of MAR, experimentation with supervised learning and selected multiple models has been conducted. Further, the performance of these techniques over the given data set was evaluated. The various methods studied for the prediction of missing data based on available training sets are: (A) Logistic Regression (LR), (B) Linear Discriminant Analysis (LDA), (C) k-Nearest Neighbor (KNN), (D) Classification and Regression Trees (CART), (E) Support Vector Machine (SVM), (F) Naive Bayes (NB).

2.1.1. Logistic Regression

Logistic regression is a statistical model based on regression where features of available inputs are used to predict the dependent or categorical variables. David Cox (1972) developed the model in 1972, based on regression analysis where the categorical dependent variable or class variable can be a binary having two or otherwise many classes. Linear regression being unbounded in terms of its values is not recommended and not suitable for classification problems. However, in logistic regression, the data is fit into the linear regression model, and logistic function acts on the model, thereby predicting the categorical dependent variables.

2.1.2. Linear Discriminant Analysis

Linear Discriminant Analysis, also called 'Discriminant Function Analysis' (Han and Kamber, 2006) is a generalization of Fisher's Linear Discriminant method. The analysis finds different linear combinations of features that can characterize or can separate two or more classes of events. The concept of searching a linear combination of variables or predictors in a data set and separately modeling those predictor variables in classes or targets is the basis of this model. The analysis guarantees maximum class separability by transforming features into lower dimensional space (Mitchell, 2017). Bayes' theorem is applied to the classes to estimate the probability. LDA is of two types. Class-dependent, where space computed is for each class on its data. The other type is class independent, where each class is separated against the other classes (Viszlay et al., 2014). LDA technique performs better than logistic regression when the sample size is large and when data sets are well separated (Kumar, 2016).

2.1.3. k-Nearest (MoRT&H TRW, 2015) Neighbor

KNN or k-nearest neighbor algorithm is a non-parametric classification technique (Soucy, 2001) which computes class label for a given test data set by k-nearest neighbor of any given test sample or test data. The technique initially selects appropriate k nearest samples i.e., k-nearest neighbor from all the training data samples for a given data set. The algorithm predicts the possible class a data belongs to, with a simple classifier (Qin, 2013). It is based on the principle that most similar data samples belonging to the same class have a high probability for the basis for predicting and designating a test sample in that class or neighbor group. When the instances are less with much training, KNNs are well suited. There is no data loss in KNN and it has speedy training, however its slow query time is major drawback.

2.1.4. Classification and Regression Trees

Decision trees are one of the efficient tree models due to its simplicity and ease of result interpretation. Trees are suitable for discrete and continuous variables in classification. CART works for both categorical and continuous input and output variables. Classification and Regression Trees or CART was introduced in 1984 (Kumar, 2016). CART is a non-parametric technique. Thus, the same variable can be used multiple times at different stages to handle multi-collinear problems in data (Thomas and Cover, 1967). CART behaves like a structure of trees with branches dividing a data set into smaller subsets until the final result, which could be a tree with decision nodes and leaf nodes. CART follows a recursive portioning approach to create the tree and a simple prediction model is fit into each tree. CART operates on three main rules, i.e., splitting rule (it splits), termination rule (branch terminates), and prediction rule (predicts variable). These decision rules to discover behavior within a data set, are well-supported by the structure of trees. The said rules are used for analysis and predictions in several classifications and regression problems.

2.1.5. Support Vector Machine

The basic idea of Support Vector Machines or SVM is to map the original data into an optimal hyperplane that categorizes them into a feature space through a nonlinear mapping function. SVM finds a separating hyperplane or line between data of two classes and takes data as an input and output. SVM is a supervised machine learning algorithm proposed by Cortes (1995) and later, Soucy (2001) also researched and extended the work. The hyperplane in the two-dimensional space is a line categorizing and dividing the data set in two parts such that each class lies on either side. There could be multiple lines/planes possible. Points closest to the line from both the classes are called support vectors. SVM computes the distance between the line, support vectors, maximize the margin and gets optimal hyperplane. The function of hyperplane constructed by the training set for classifying the test samples is:

$$f(x) = m(x) + c$$

Equation 1: Function for hyperplane

where $f(x)$ is a hyperplane function, m is a normal vector of the hyperplane, and c is a variable.

SVM approach works well with high-dimensional spaces and has several features in the feature vector. SVM can be used with small data sets excellently as well. The performance of SVM is affected by a considerable dataset having noise. The technique has better memory efficiency.

2.1.6. Naive Bayes

Naïve Bayes uses Bayes theorem of probability and predicts the class of unknown data set. NB is based on the assumption that a feature class within a data set is not at related to other data set. For example, a green, round, with 3-inch diameter object can be a guava or a lawn tennis ball. These features may depend on each other or other features if existed or considered. Features always contribute independently to the probability that an object could be a tennis ball or a guava, though the probability may be less with a single feature. NB models read data from the training set and compute the mean and standard deviation of the predictor variables for each class. Then probability feature is calculated in each class for all predictor variables and identifies the highest probability out of each class. Naive Bayes classifiers require several parameter forms of linearity (Mitchell, 2017) and are highly scalable. The Naïve Bayes is used in real-time predictions as it is fast and often used in recommendation systems and multi-class predictions.

3. Methodology

Supervised learning techniques have their pros and cons and suitability on different sizes and types of data set. For example, regression techniques used in analyzing traffic accident severity (Kashani, 2011) by deploying various assumptions and predefined underlying relationships criteria among all the variables, i.e., dependent and independent variables (Mujalli, 2011). The different data mining (DM) techniques have other possible ways to analyze data and extract useful knowledge and information (Han and Kamber, 2006). These DM techniques have been applied in the analysis of crash severity analysis and provided satisfactory results. Abdel (1997) researched on predictions by neural networks in the severity of injuries in two-wheeler crashes. The Bayesian analysis is equally practiced in analyzing traffic accidents and understanding accident-severity. With much classification and regression models for prediction; accuracy estimation becomes an essential aspect of finding the most appropriate model that can be applied to get an accurate result

and making the right decision. One way to get the most accurate model for prediction is by applying the six models.

A good model generalizes well from the training data to the test data from the problem domain, which is finding an unknown vehicle. This generalization of the models, while finding the unknown in the data set, can be achieved with a good statistical fit. The goodness of fit implies that a model, while learning from training data, is neither over-fit nor under-fit. Over-fitting happens when a model considers the details and the noise in the data. On the other hand, when a statistical model cannot adequately capture the underlying structure of training data, it is an under-fit model and affects to the extent that it has negative impacts on the performance of the model. The other procedure could be to identify the most accurate model on the given data and use it to predict the missing values by managing the over-fitting and under-fitting issues in learning, thus reducing the bias. A process-based approach framework is proposed using K-fold cross validation, which is applied to all the modes while learning from training and finding the test set.

3.1. Cross validation

Cross Validation techniques give better accuracy by minimizing the over-fitting issues and reducing the bias in the result (Lingraj et al., 2018). Cross validation is a resampling procedure used for model evaluation for computation of accuracy. Initially, the residual evaluation methods were in use for the assessment of the models with the entire data set for training the learner. This method does not give any indication of the performance of learning due to a lack of data observations. To solve this problem, cross validation techniques are used. In cross validation (Stone, 1974), a data subset each is used for training and testing the model. The cross validation is of three types viz. holdout cross validation, k-fold cross validation, and leave one out cross validation. The holdout method is a simple cross validation technique, also known as the test sample estimation method. The method, partitions the data into a training set and test set or holdout set, which are mutually exclusive. Out of this, two-third of the data is a training set and one-third a test set. The training set learns the pattern and applies it to the test set. This method is preferred over the classical residual method, as it takes lesser time to compute. The evaluation of this technique can have a high variance. *k*-fold cross validation (Stone, 1974) involves recursive execution of the holdout process for k times (or k passes). The holdout method is applied on a given data, divided into k subsets and repeated on those k subsets for k times. This process significantly reduces bias and variance as maximum data is used for fitting and validation set. The third cross validation method, Leave-one-out cross validation, is a *k*-fold cross validation at the extremities when K equal to N, i.e., the number of data points in the set. The model is trained on data for N times, leaving one subset, and a prediction made accordingly based on the training and applied to the leftover subset. The model is approximately unbiased because the difference in size between the training set used in each fold and the entire dataset is only a single pattern. The model tends to have a high variance with every repeat, the estimate with different initial samples of data from the same distribution varies. The k-fold cross validation was used for this work. The steps for *k*-fold cross validation are:

Step-I: Random shuffling and Splitting of the data set into k groups
Step-II: For the data in k groups:

(a) One group is taken as a test set and remaining are training set
(b) The model is fit on the training set and evaluated on the test set
(c) The evaluation score is retained, and the model is discarded

Step-III: The performance of the models computed in the data.

Fig. 1 shows an illustration of a test and training subset generated in *k*-fold cross validation. In the above example, a total of twenty-five instances are considered. Eighty percent of the data is for training and rest for testing the model after training. The process is repeated k times so that all data groups are used for testing once. For any classification and regression model, accuracy is, the total variables correctly classified in that data set. The accuracy is computed using the classical model where there is true positive as TP, true negative as TN, false-positive as FP, and false-negative as FN and defined as:

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})}$$

Equation 2: Accuracy Computation

The accuracy is computed in k-fold using the standard accuracy computation stated in Equation 2. As a general rule and empirical evidence, K = 5 or K = 10 is used.

3.2. Approach

In this experiment, a process is proposed for computation of the performance of six different models together on the accident data. Fig. 2 illustrates the process for the followed approach. It consists of the following steps:

Step –I: Input the data, organize and clean,
Step –II: Perform K-fold cross validation on the learning models with the accident data, and
Step –III: Publish the result for that data.

The accident data is cleaned and processed for any errors or any

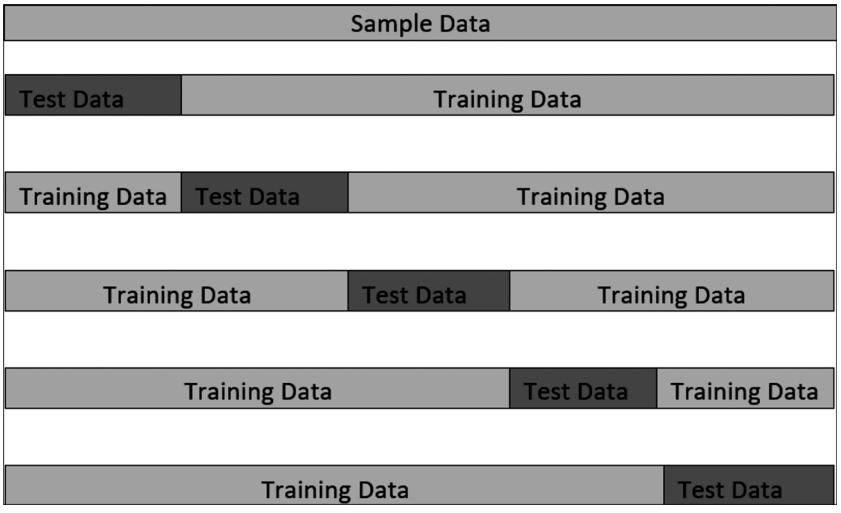

Fig. 1. Illustration of K-Fold Process.

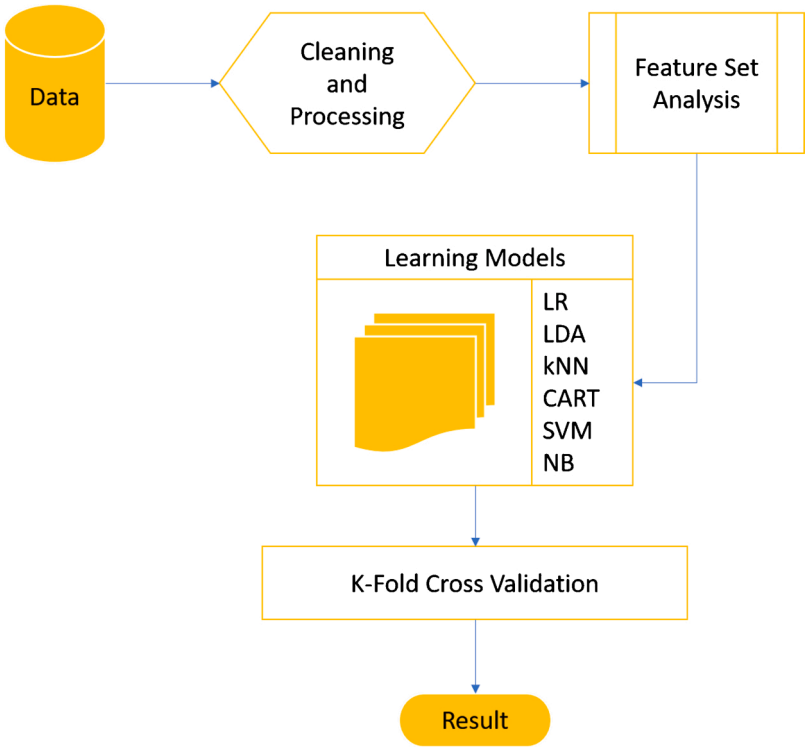

Fig. 2. Process for the approach followed in the research.

other outlier values and passed to the feature analysis step. A feature selection and ranking step is performed on the cleaned data for the identification of most important features to be considered for the experiment and the features which do not have any impact are relieved by understanding the relationship between the features. The performance of each learning model is computed using *k*-fold cross validation over the six models. viz. LR, LDA, KNN, CART, SVM and NB separately with the collected accident data. The computing device used in the experiment had a core i3-4170 CPU@ 3.17 GHz processor and 12 GB primary memory.

4. Experiment and discussion

4.1. Data collection

Road accident data was collected from accident records of five cities in India: Agra, Amritsar, Bhopal, Ludhiana, and Vizag for the period from year 2007 to 2011. The road accident data from the said five cities was obtained from TRIPP, IIT Delhi, that has following six variables: (A) Month, day and time of the accident: this has Date, Day, Time, Holiday Status, (B) Type of accident: this has Collision Type, Hit-and-run status, (C) Accident Impact: this has Total Fatalities, Total Injuries, Severity, Total Vehicles impacted, (D) Vehicle-related information: this has Victim Vehicle, Impacting Vehicle, (E) Collision type: this has head-on, sidewise, back hit, hit-and-run and (F) Type of road: this has straight road, junction, etc. and if the divider is there on the road. Table 1 lists the number of fatal accidents in the five cities for the period from year 2007 to year 2011.

Table 1
Fatal accident record for the five cities for the period 2007 to 2011.

City	Accidents Count
Agra	674
Amritsar	263
Bhopal	685
Ludhiana	581
Vizag	1164

The various road users in the data are: (A) Pedestrians, (B) NMT: Non-motorized transport viz. cycle rickshaw, cyclist, (C) MTW: Motorized two-wheelers viz. Scooters, motorcycles, (D) Cars: Cars, Jeeps, taxi, cabs, (E) Bus: Different bus types viz. city bus, school bus, intercity bus, (F) LMV: Light motor vehicles viz. tempo, three-wheeled scooter taxi (TST), (G) Truck and (H) HMV: Heavy motor vehicles viz. trailer, lorry, tractor, cranes.

5. Results and discussion

With the models applied on accident data with unknown variables as the target, there are three possible cases. In Case 1, the impacting vehicles are entirely unknown throughout the data, which means all the accidents are hit-and-run accidents. The prediction of unknown vehicles is impossible as there is not a single correct classified pattern to learn, and without learning, the models cannot predict. In Case 2, the impacting vehicles are known, i.e., none of the accidents are hit-and-run, and then there is no sign of this experiment. In Case 3, which is a mix of Case 1 and Case 2, i.e., data contains unknown and known impacting vehicles, and can be used. The computation is done with the 5-fold and 10-fold cross validation to get the patterns. The performance of 10-fold cross validation is used in the results.

5.1. Results

The 10-fold cross validation and 5-fold cross validation is applied on the six models viz. LR, LDA, KNN, CART, SVM and NB; and accuracy of these models are computed on accident data from the five cities. The result provides accuracy of these models computed using both 5-fold and 10-fold cross validations and summarized in Tables 2 and 3 , respectively.

The performance of the models using both folds shows an approximately similar pattern but minor difference in accuracies. SVM has highest accuracy for all the cities except Amritsar using both cross-

Table 2
Performance (in %) of prediction models for each city using 5-Fold cross validation.

City	LR	LDA	KNN	CART	SVM	NB
Amritsar	23	**25**	20	**26**	21	12
Ludhiana	42	36	42	36	**44**	9
Bhopal	38	32	38	39	**40**	5
Vizag	36	33	33	33	**37**	4
Agra	41	38	43	34	**44**	9

Table 3
Performance (in %) of prediction models for each city using 10-Fold cross validation.

City	LR	LDA	KNN	CART	SVM	NB
Amritsar	23	24	20	**26**	23	12
Ludhiana	44	40	41	38	**45**	7
Bhopal	37	33	37	33	**38**	5
Vizag	**36**	34	34	34	**37**	4
Agra	42	39	43	36	**44**	8

validation. In Amritsar, CART has the highest accuracy. NB's performance is lowest for all the cities. The summary of results with accident records of cities and unknown vehicles is shown in Table 4.

5.2. Discussions

The pattern of the accuracy is same for both 5-fold and 10-fold as shown in Table 2 and Table 3. Ludhiana has reported 581 road accidents with 21 % by unknown vehicles. SVM showed the highest accuracy of 45 % followed by LR and KNN with 44 % and 41 % respectively. NB has lowest performance. Amritsar recorded 263 accidents with 10 % caused by unknown vehicles. The number of accidents reported is lowest in Amritsar. CART is the best model with 26 % accuracy. The accuracy is comparatively low but with limited number of records this is the best possible outcome from the studied models. Agra and Bhopal have recorded 674 and 685 accidents, respectively. SVM has the best performance for both the cities with 44 % for Agra and 38 % for Bhopal. There could be several reasons for the variation in performance even though the data samples size is approximately same. Performance of all the other models for these cities shows a similar variation for both the folds as shown in Table 3. KNN and LR have 43 % and 42 % accuracy for Agra but 37 % for Bhopal. The performance of NB is the lowest for both the cities. Vizag has the highest sample size of 1164 records which is approximately double to the accidents recorded for Ludhiana, Agra and Bhopal and four times that of Amritsar. Performance of SVM is the highest for Vizag with 37 % accuracy. Performance of LDA, KNN and CART is 34 % and accuracy of NB is 4%.

We observed that, the same pattern of variation in accuracy exists in both the cross validation with slight differences in the values. This gives a clear indication that performance of the models is correctly evaluated.

We observed that NB's performance is always the lowest. For Amritsar with only 263 records, NB has 12 % accuracy while Vizag having 1164 records, has 4% accuracy. NB is a parametric method with high-bias/low-variance. NB has a possibility to miss the relevant relations between features because of under-fitting resulting in poor performance by NB.

The performance of the best models is between 26 % for Amritsar to 45 % for Ludhiana. Amritsar has the lowest sample size. Vizag has maximum accident sample size but accuracy of SVM is 37 %. This indicates that the data sample size is not the only factor affecting the performance though it is one of the most important factors. The performance of any supervised learning model depends on the data size and variation among underlying variables of that data. The accident share of various impacting vehicle for each city is summarized in Figs. 3–7.

Impacting vehicle share for Ludhiana shows that road users are most vulnerable to trucks followed by unknown vehicles as shown in Fig. 3. Trucks, unknown vehicles and cars impose highest risks with 36 %, 21 % and 18 % share in road accidents, respectively.

LMVs have lowest accident share. HMVs and MTWs have an equal share of 6%. Impacting vehicles share for Amritsar in Fig. 4 shows that MTW is most vulnerable for accidents. It is almost four times compared to Ludhiana. Trucks and cars are second and third vulnerable vehicles in Amritsar. The number of trucks involved in accidents in Amritsar is approximately half of those reported in Ludhiana. One possible reason could be Amritsar's population density. SVM produced the highest accuracy for Ludhiana, as Ludhiana has higher number of accident records data and are separable easily in the hyperplane. Alternatively, Amritsar has a lesser data and is a reason for low performance of SVM. The low data size and data distribution among impacting vehicles makes CART more suitable for Amritsar.

The concentration of unknown vehicle also affects performance of models for a city. Agra and Bhopal have similar number of accident instances; however, the proportion of impacting vehicles varies and that could be one of the reasons for variation in performance of SVM. Agra has19 % unknown vehicles (Fig. 5) and it is second most vulnerable city. Bhopal has 15 % unknown vehicles (Fig. 6) causing accident and is third most vulnerable after trucks and MTWs. The variation in impacting vehicles for Agra and Bhopal are shown in Figs. 5 and 6 respectively. The variation of trucks and MTWs is quite significant in both cities. Trucks are most vulnerable for both Agra and Bhopal accounting for 34 % and 26 % accidents, respectively.

In Vizag, having most impacting vehicles being HMV and MTW as shown in Fig. 7 resulting in a higher bias in data, hence even with the larger data size, the performance is lower compared to other cities. Furthermore, the unknown vehicle has only 5% proportion compared to overall accidents. This smaller concentration also impacts the SVM performance even though it has the highest data instances. For a better learning and performance, data size, data density and proportion of unknown vehicles are most important parameters in identifying the best model for predicting the unknown vehicles.

We also observed that the vehicle shares in the accident data, for the five cities are not evenly distributed except for Amritsar. Amritsar has

Table 4
City-wise summary of total accidents, hit-and-run accidents, and best accuracy with the model.

City	Total Accidents recorded	Unknown Vehicles (Hit-and-run cases)	Accuracy for Prediction [a]	Model
Amritsar	263	10 %	26 %	CART
Ludhiana	581	21 %	45 %	SVM
Bhopal	685	15 %	38 %	SVM
Vizag	1164	5%	37 %	SVM
Agra	674	18 %	44 %	SVM

[a] Based on 10-fold cross validation.

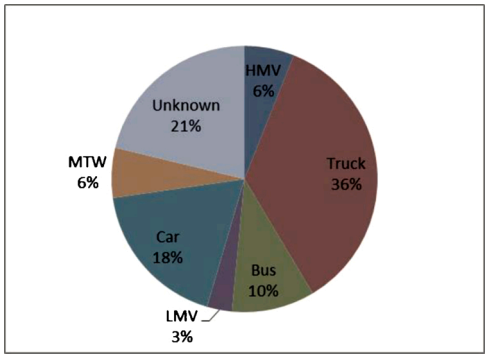

Fig. 3. Accident share by impacting vehicles in Ludhiana.

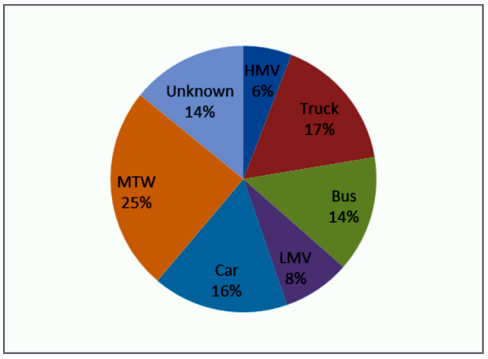

Fig. 4. Accident share by impacting vehicles in Amritsar.

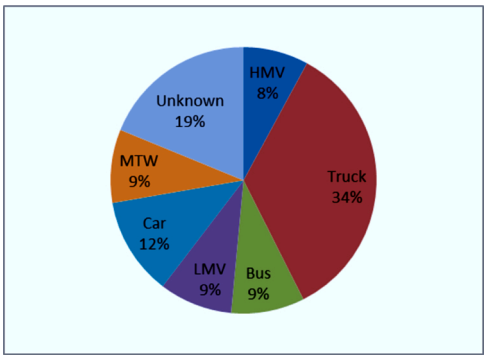

Fig. 5. Accident share by impacting vehicles in Agra.

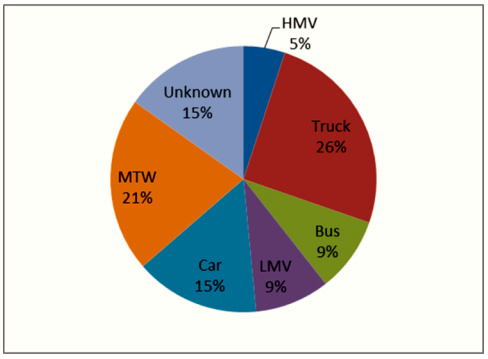

Fig. 6. Accident share by impacting vehicles in Bhopal.

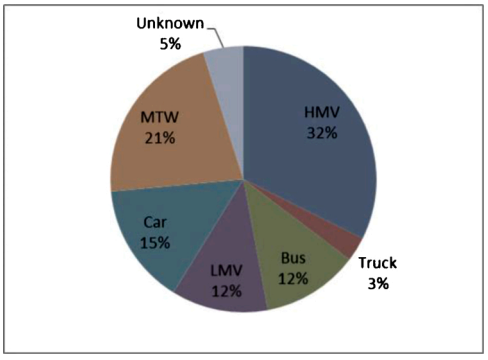

Fig. 7. Accident share by impacting vehicles in Vizag.

approximately uniform share of impacting vehicles with car, truck, bus and unknown. This uniformity does not exist in other cities. CART gives best performance over other models for small data size. For other models, a larger data set is required. CART is a non-parametric model and does not get impacted by outliers in the input variables and compared to other models with high variance, the learning becomes ineffective and in case of Amritsar, the even distribution makes it effective over other models.

With a more extensive data set, the performance can be improved and the dependency between unknown vehicle and prediction accuracy may change. It is also possible that with multiple iterations and multi-core processors, there could be a slight change in performance of the models.

6. Conclusion and future work

Road accidents are one of the major causes of deaths and injuries. There are property losses and other irreversible direct and indirect impacts. It is a well-identified global problem, and many international bodies have been working in designing policies, technologies to prevent these accidents. For the prevention of road accidents, knowledge of participating vehicles i.e. victim vehicles and impacting vehicles, is essential. Hit-and-run type of accidents has no information on impacting vehicles, as, after the accident, impacting vehicle flees the accident scene. These are the most dangerous type of accidents. The vehicles responsible for accidents are unknown and quoted as 'unknown vehicles' in recording any accident. Knowledge of unknown vehicles is crucial to set up prevention plans. There are several models for prediction, and therefore identification of the most accurate model is important.

The vulnerability of hit-and-run accidents on road users is described and an approach is proposed to identify the best way that may be used to predict unknown vehicles. The approach created a process cycle and implemented various learning models over accident data using K-fold cross validation technique to get the best performing models.

A 10-fold cross validation is used in the analysis. Support Vector Machine (SVM) had the highest accuracy for the five cities except Amritsar whereas Classification and Regression Trees (CART) had the highest efficiency. The difference in the accuracy of the models depends on data size, total number of variables, and variable values. The model finalized by the process shows that non-parametric methods are most suitable for predictions.

Data size is insufficient for proper learning, and hence performance is justified with such limited learning scope. With this much-limited data set, that is the best accuracy that can be achieved in the study. One possible way could have been to scale up the data with a random function. However, the purpose was to use real data from a city, and hence this scaled-up quantity would not have fulfilled the objective. The type of data, e.g., continuous or categorical also have an impact on the performance. The distribution of vehicles causing the accident in the city plays another vital role. The work can be extended by applying other classification and regression models, such as self-organizing maps, random forest, neural networks, clustering techniques, rough sets and deep learning techniques. The information of the unknown vehicles predicted with the best performing model will be very useful in preventing the hit-and-run accidents and designing better road user safety plans.

Author's statement

Alok Nikhil Jha: Conceptualization, methodology, writing, Niladri Chatterjee: Supervision, reviewing, concept validation Geetam Tiwari: Supervision, reviewing.

Declaration of Competing Interest

The corresponding author, on behalf of other authors for the manuscript titled "A Performance Analysis of Prediction Techniques for Impacting Vehicles in Hit-and-Run Road Accidents" certify authors have NO affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in any form from any organization is discussed in the manuscript.

Acknowledgments

The data set used in the experiment is provided by the Transport Research and Injury Prevention Program (TRIPP) at IIT Delhi.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.aap.2021.106164.

References

Abdel, A., 1997. Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. Transp. Res. Rec. 6–13.
Aidoo, E.N., 2013. The effect of road and environmental characteristics on pedestrian hit and run accidents in Ghana. Accid. Anal. Prev. 53, 23–27.
Cortes, C.V., 1995. Support-vector networks. Mach Learn 20. September, pp. 273–297.
Cox, D., 1972. Regression models and life-tables. J. R. Stat. Soc. Ser. B Method. 34, 187–220.
Crundall, D.B., 2005. Road traffic accidents in the United Arab Emirates compared to Western countries. Adv. Transp. Stud. 6.
Global Burden of Diseases, I, 2017. IHME USA. Retrieved January 2020, from. https://vizhub.healthdata.org/gbd-compare/.
Han, M., Kamber, 2006. Data Mining: Concepts and Techniques. Morgan kufman Publishers, San Francisco.
Jha, A.N., Tiwari, G., Chatterjee, N., 2018. Data recording patterns and missing data in road crashes: case study of five indian cities. BMJ Journals Injury Prevention 2018 (24), 195. https://doi.org/10.1136/injuryprevention-2018-safety.539.
Jha, A.N., Tiwari, G., Chatterjee, N., 2020. Road accidents in EU, USA and India: a critical analysis of Data Collection Framework. Strategic System Assurance and Business Analytics. Springer Nature.
Kashani, Aa., 2011. Analysis of the traffic injury severity on two lanes, two-way rural roads based on classification tree models. Saf. Sci. 49, 1314–1320.
Kim, K., Pant, P., Yamashita, E., 2008. Hit and run crashes: using rough set analysis with logistic regression to capture critical attributes and determinants. Transportation Research Board Annual Meeting. Transportation Research Board.
Kumar, S.R., 2016. Elm variants comparison on applications of time series 688 data forecasting. In: International Conference on Advances in Computing. Communications and Informatics 689 (ICACCI), pp. 1404–1409.
Lingraj, D., Sanjay, A., Ajith, A., 2018. Nested cross-validation based adaptive sparse representation algorithm and its application to pathological brain classification. Expert Syst. Appl. 313–321.
Macleod, K.E., 2012. Factors associated with hit and run pedestrian fatalities and driver identification. Accid. Anal. Prev. 366–372.
Mitchell, T.M., 2017. Generative and discriminative classifiers: naive Bayes and logistic regression. Textbook Machine Learning. McGraw Hill.
MoRT&H TRW, 2015. Road Accidents in India – 2015, Pg No 81. Government of India, New Delhi.
MoRTH Accident Trends, 2018. Accident Death Trends. Govt. of India.
Mujalli, R., 2011. A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks. J. Safety Res. 42, 317–326.
NHTSA, N.H., 2012. Fatality Analysis Reporting System (FARS). NHTSA, USA.
Qin, Z.W., 2013. Cost-sensitive classification with k-Nearest Neighbors. International Conference on Knowledge Science, Engineering and Management 112–131.
Soucy, P., 2001. A simple kNN algorithm for text categorization Data Mining. In: Proceedings IEEE International Conference. IEEE, pp. 647–648.
Statistical Abstract: Motor Vehicle Accidents and Fatalities, 2008. Statistical Abstract from National Data Book. The National Data Book.
Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc. 111–147.
Tay, R.B., 2009. Factors contributing to hit and run in fatal crashes. Accid. Anal. Prev. 41, 227–233.
Thomas, M., Cover, P.E., 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory 13 (1), 21–27.
Viszlay, P., Lojka, M., Juhar, J., 2014. Class-dependent two-dimensional linear discriminant analysis using two-pass recognition strategy (n.d.). Proceedings of the 22nd European Signal Processing Conference (EUSIPCO) 796–1800.
WHO Violence and Injury Prevention, 2004. World Report on Road Traffic Injury Prevention. WHO.