IEEE Open Journal of
Intelligent Transportation
Systems

Received 17 December 2020; revised 20 May 2021, 13 September 2021, and 21 December 2021; accepted 15 March 2022. Date of publication 18 March 2022; date of current version 25 March 2022.

Digital Object Identifier 10.1109/OJITS.2022.3160404

Fatality Prediction for Motor Vehicle Collisions: Mining Big Data Using Deep Learning and Ensemble Methods

MAHZABEEN EMU[1] (Member, IEEE), FARJANA BINTAY KAMAL[2], SALIMUR CHOUDHURY[2] (Senior Member, IEEE), AND QUAZI ABIDUR RAHMAN[2]

[1]School of Computing, Queen's University, Kingston, ON K7L 3N6, Canada

[2]Computer Science Department, Lakehead University, Thunder Bay, ON P7B 5E1, Canada

CORRESPONDING AUTHOR: S. CHOUDHURY (e-mail: schoudh1@lakeheadu.ca)

This work was supported by the Vector Institute, Toronto, ON, Canada, under the Support Program of Vector Scholarships in Artificial Intelligence (VSAI).

ABSTRACT Motor vehicle crashes are one of the most common causes of fatalities on the roads. Real-time severity prediction of such crashes may contribute towards reducing the rate of fatality. In this study, the fundamental goal is to develop machine learning models that predict whether the outcome of a collision will be fatal or not. A Canadian road crash dataset containing 5.8 million records is utilized in this research. In this study, ensemble models have been developed using majority and soft voting to address the class imbalance in the dataset. The prediction accuracy of approximately 75% is achieved using Convolutional Neural Networks. Moreover, a comprehensive analysis of the attributes that are important in distinguishing between fatal vs. non-fatal motor vehicle collisions has been presented in this paper. In-depth information content analysis reveals the factors that contribute the most in the prediction model. These include roadway characteristics and weather conditions at the time of the crash, vehicle type, time when the collision happen, road user class and their position, any safety device used, and the status of traffic control. With real-time data based on weather and road conditions, an automated warning system can potentially be developed utilizing the prediction model employed in this study.

INDEX TERMS Deep learning, collision severity prediction, ensemble methods, information content analysis.

I. INTRODUCTION

A MOTOR vehicle collision or traffic crash usually occurs when a vehicle collides with other vehicles, pedestrians, road debris, animal, or some stationary object, such as building, tree or pole, which might result in a fatality, serious injuries and/or damage of resources. Possible factors associated with the prospect of collisions and its severity include vehicle type, roadway configuration, weather condition, road surface, road alignment, age and gender of the person involved in the collision, traffic control system, and safety measurements taken by the person.

Recent research has focused on the prediction of collision severity using machine learning methods [1], [2]. Moreover, some works have also identified and analyzed multiple factors associated with the severity of the crash [3], [4], [5]. In general, the research concerning road safety has one of these two aims: planning and prediction. The planning perspective tries to identify the causes of crashes and then prioritize the countermeasure that can be taken concerning these. For example, many statistical models concluded that the road surface, weather condition, reckless driving, and traffic condition lead to a severe crash [3], [4], [5]. Studying these factors and taking countermeasures, such as installing cable medians, rumble strips, etc. can reduce the crash severity. On the other hand, the prediction of a crash and giving people an alert about the road crash with real-time road

The review of this article was arranged by Associate Editor Maria Vanderschuren.

FIGURE 1. Possible machine learning enabled real-time risk prediction system.

information, can potentially save a life [6]. However, there has not been significant research that simultaneously focused on utilizing big data to predict the crash severity using deep learning and identify important contributors to fatal crashes. Previous works with deep learning have used the dataset with four thousand to four hundred and ten thousand records from a 2-8 years period [2], [7], [8], [9]. As a contrast, the study in this paper predicts the crash severity by employing traditional and deep learning methods on a large data set with approximately 5.8 million records collected over 20 years. In addition, a thorough information content analysis of important predictors has been conducted. As such, the main objective of this paper is developing machine-learning-based models to facilitate the prediction of the severity of road crashes. Two levels of severity are distinguished in the proposed prediction model: fatal vs. non-fatal. Fig. 1 illustrates a case where the developed machine-learning model can be integrated with a mobile applications (e.g., Google Maps, Apple Maps, Waze, Sygic GPS Navigation) as API (Application Programming Interface) to promote conscious driving.

The findings of this paper can be highly beneficial to potential road safety software/application developers. Specifically, the in-depth attribute selection analysis and the ensemble models proposed in this paper can be a subject of interest for road safety application developers, especially in the Canadian context.

The rest of the paper has been structured as follows. Section II gives an overview of the related works. Next, Section III provides the description of the dataset. Section IV defines the methods used for the prediction and attribute extraction. Section V represents the experimental results and a computational analysis on time and memory requirements of the proposed real-time risk prediction system. Finally, the summary of the findings and conclusions are discussed in Section VI.

II. RELATED WORK

Research has been done using clustering based regression approaches to distinguish the principal factors associated with the levels of pedestrians' physical damage consequences based on the context of New York, the United States, and Montreal, Canada [10]. The study established heavy weighted vehicles, dark lighting conditions, and mixed land use increase the probability of crashes being fatal. They provided recommendations, such as the need for training programs for the drivers, improved road lighting, and retrofitting of significant streets into complete lanes. Another work with data mining strategies (Decision tree, Naive Bayes, and K-nearest neighbor) has associated reported road attributes to crash severity in Ethiopia and produced a set of rules that the Ethiopian Traffic Agency could practice to enhance road safety [11]. Kumar and Toshniwal [12] identified the hazardous crash location. Data mining techniques are used, and the authors then analyzed the locations to identify the factors that are responsible for the crash at those locations. At first, the locations are divided into k-groups by using the k-means clustering technique. Then, to find the relationship between individual attributes, the association rule mining algorithm was applied. The authors concluded that highways with intersections are the high-frequency crash location.

Bedard *et al.* [13] implemented a multivariate logistic regression to identify the independent augmentation of the driver, collision, and vehicle attributes to drivers' inevitability risk. They discovered that growing seat belt usage, diminishing speed, and decreasing the amount and severity of driver-side influences might limit casualties. Another real-time analysis applied logistic regression to explore the crash contributing factors [14]. The research concluded that visibility issues, bad roadway surface, and heavy rainfall impact the severity of the crash. The regression model is easily interpretive as it provides a coefficient value for every vital attribute. However, the error term has a standard logistic distribution, which may not be valid in a real scenario.

Some research focused on the prediction performance of various classification models and concluded with their comparison. Chong *et al.* [1] compared the performance of four statistical and machine learning methods including, Multinomial Logit (MNL), Nearest Neighbor Classification (NNC), Support Vector Machines (SVM) and Random Forests (RF) predict crash severity. The authors also analyzed the effect of clustering methods in these four prediction models. Among the four methods, NNC performed better than the others, and MNL was the worst. Further research was conducted to observe the performance of four machine learning models implemented to represent the severity of an impairment that happened during traffic collisions [15]. The prediction is made by using Hybrid Learning methods, Support Vector Machines (SVM), Decision Trees, and a simultaneous Hybrid model concerning Decision Trees and Neural Networks. Analysis results uncovered that, among the machine learning paradigms considered, the hybrid decision tree-neural network method outperformed others. To classify the crash severity involving Powered Two Wheelers, another work used Logistic Regression, Random Forest, Support Vector Machines, and Deep Neural Network [2]. The authors

TABLE 1. Summary of related research.

Reference	Dataset Size	Target of Concern	Location	Limitation
[10], [3]	6896 (U.S.A.) & 5,820 (Canada) instances [10]	Identifying critical factors for severity outcomes of crashes	New York, U.S.A and Montreal, Canada [10], Greece [3]	Analysis applicable for pedestrian road users only [10], require huge capital and maintenance expenses, dependency on sensors [3]
[11]	18,288 instances	Building a support tool for Ethiopian traffic policymakers	Ethiopia	Not suitable for road users
[12]	15,574 instances	Identifying locations with high-frequency severe crashes	Dehradun, India	Missing important attributes (e.g., weather, speed, and road surface related information)
[13], [14]	110813 instances [13], 5181 instances [14]	Analysing crash contributing factors	U.S.A	Involve single fixed object crashes only [13], require additional expenditure for underground sensors [14], and fault intolerance [14]
[1], [15], [2]	49,068 instances [1], 417,670 instances [15], 16,463 and 7,424 instances [2]	Prediction of accident severity	United States [1], [15], [2]	Exclusion of potentially significant attributes (e.g., speed) [1], heavily dependent on driving behavioural pattern [15], fault intolerance [2]
[16], [17]	35275 instances [16], 100 instances [17]	Distinguishing drivers' behavioural pattern towards fatality	Netherlands [16], England [17]	Missing potentially significant features (e.g., alcohol consumption frequency)
[4], [5], [18]	4112 instances [4], 13,775 instances [18]	Identifying relationship between accident notification and severity [4], traffic flow and severity [5], weather/road condition and severity [18]	U.S.A. [4], [5], Canada [18]	Dependency on sensors [4], [5], lack analysis on geometrical features of road configurations [18]

analyzed the prediction of the models in the full set of attributes and some reduced attributes. The performance of the models is high in both the attributes set. The study concluded that by obtaining attributes of the crash with the sensor, it is conceivable to build an efficient, intelligent system capable of identifying crash severity.

Another group of works aimed to find some critical factors that are related to the road crash. Multivariate analysis performed on a Dutch database that includes data on drivers' qualifications, yearly mileage, regular driving performance in the figure of fines, and crash involvement [16]. The study concluded that collisions raised as yearly mileage increased. Furthermore, multivariate investigations determined that drivers' gender does not contribute to the crash association. It is also pointed out that more recent drivers have the most significant rate of collisions, and education qualification is not correlated to crash engagement. Another Multivariate Investigation was done to find the association among drivometer Variables, drivers' crash history, gender, and exposure status [17]. Theofilatos [3] also used real-time traffic and weather data on two urban arterials in Athens, Greece, to study the road crash probability and seriousness. To get the potential significance of the variable, Random Forest is used firstly, and then a Bayesian logistic regression for predicting the crash occurrence. This study identified the flow per lane, the speed, and the occupancy as the most important factors that work as influencers towards the crash likelihood and severity. Evanco [4] carried a multivariate population-based analytical investigation to learn the association between deaths and crash notification conditions. The study illustrated that crash announcement time is an influential determinant of the number of inevitability for collisions on rural road configuration. Another study concluded that the characteristics of the traffic flow contributing to the crash severity [5]. They suggested that, due to the high density of traffic, high variation in the speed and frequent changes of lane lead to the low severity crashes. In contrast, less congested traffic leads to many severe crashes. Based on the data collected from 31 different highway routes across Ontario, Canada, the authors of [18] stated that low visibility, storm hour, and low temperatures are factors associated with the high risk of a crash. Moreover, they identified the surface condition of the road as the crucial factor. Table 1 includes the summary of the surveyed research papers.

From this detailed literature review, it is observed that the prediction of crash severity by applying deep learning methods on a large dataset along with attribute-importance analysis is still under-explored. Moreover, most of the existing literature either emphasize on finding the important attributes liable for road crash fatalities or generate risk prediction models exclusively. Following are the major contributions of this study:

- To bridge the research gap, a large dataset with significant number of observations and attributes have been considered to identify critical factors at a fine granularity-level. In addition, the use of ensemble and deep-learning techniques to develop a prediction model have been simultaneously proposed for fatality prediction.
- Moreover, the computational complexity of time and memory requirements of the proposed system has been

TABLE 2. Attributes list with their actual values.

Attributes Name	Values
1. Collision level data elements	
C_YEAR	19yy-20yy, where yy=last two digits of the calendar year; 2020 (7%)
C_MNTH	January, February, March, April, May, June, July, August (9%), September, October, November, December, Unknown, and Jurisdiction does not provide this data element
C_WDAY	Monday, Tuesday, Wednesday, Thursday, Friday (17%), Saturday, Sunday, Unknown, and Jurisdiction does not provide this data element
C_HOUR	Midnight to 0:59, 1:00 to 1:59, 2:00 to 2:59, 3:00 to 3:59, 4:00 to 4:59, 5:00 to 5:59, 6:00 to 6:59, 7:00 to 7:59, 8:00 to 8:59, 9:00 to 9:59, 10:00 to 10:59, 16:00 to 16:59 (45.79%), 17:00 to 17:59, 18:00 to 18:59, 19:00 to 19:59, 20:00 to 20:59, 21:00 to 21:59, 22:00 to 22:59, 23:00 to 23:59, Unknown, and Jurisdiction does not provide this data element
C_SEV	Collision producing at least one fatality, Collision producing non-fatal injury (98%), Unknown, and Jurisdiction does not provide this data element
C_RCFG	Non-intersection, At an intersection of at least two public roadways (46%), Intersection with parking lot entrance/exit, private driveway or laneway, Railroad level crossing, Bridge, Overpass, Viaduct, Tunnel or underpass, Passing or climbing lane, Ramp, Traffic circle, Express lane of a freeway system, Collector lane of a freeway system, Collector lane of a freeway system, Choice is other than the preceding values, Unknown, and Jurisdiction does not provide this data element
C_WTHR	Clear and sunny (69.52%), Overcast, cloudy but no precipitation, Raining, Snowing, Freezing rain, sleet, hail, Visibility limitation, Strong wind, Choice is other than the preceding values, Unknown, and Jurisdiction does not provide this data element
C_RSUR	Dry-normal (65.62%), Wet, Snow, Slush, Wet snow, Icy, Sand/gravel/dirt, Muddy, Oil, flooded, Choice is other than the preceding values, Unknown, and Jurisdiction does not provide this data element
C_RALN	Straight and level (77.85%), Curved and level, Curved with gradient, Top of hill or gradient, Bottom of hill or gradient, Unknown
C_TRAF	Traffic signals fully operational or flashing mode, Stop, Yield, Warning sign Yellow, Diamond shape, Pedestrian crosswalk, Police officer, School guard, flagman, School crossing, Reduced speed zone, No passing zone, Markings on the road e.g. no passing (30.71%), School bus stopped with school bus signal lights flashing or not, Railway crossing with signals, or Signals and gates, Signs only, Control device not specified, No control present
2.Vehicle level data elements	
V_TYPE	Light Duty (90.34%), Panel/cargo van, Other trucks and vans, Unit trucks, Road tractor, School bus, Urban and Intercity, Motorcycle and moped, Off road vehicles, Bicycle, Purpose-built motor, Farm and Construction equipment, Snowmobile, Street car
V_YEAR	19yy-20yy, Data element is not applicable, Choice is other than the preceding values, Unknown, Jurisdiction does not provide this data element
3.Person level data elements	
P_SEX	Male (54.09%), Female, Data element is not applicable, Choice is other than the preceding values, Unknown, Jurisdiction does not provide this data element
P_PSN	Driver (67.14%), Front row and center, Front row and right outboard, Second row and left outboard, Second row and center, Second row and right outboard, Third row and left outboard, Third row and center, Third row, Right outboard, Position unknown, but the person was definitely an occupant, Sitting on someone's lap, Outside passenger compartment, Pedestrian
P_SAFE	No safety device or child restrain used (93.88%), Safety device used, Helmet worn, Reflective clothing worn, both helmet and reflective cloth, Other safety device
P_USER	Motor Vehicle Driver (65.92%), Motor Vehicle Passenger, Pedestrian, Bicyclist, Motorcyclist

analyzed so that the developed model can be integrated in a real-time mobile application without having any additional sensor requirements. The dataset used in this study to develop the model is solely focused in Canadian context. However, the proposed model can be generalized and adapted for other communities by training those on appropriate region-based datasets using the methodology employed in this study.

III. DATA

The data set [19] includes information from 1995-2014, given by Transport Canada collected for crash statistics of car crashes in Canada. The initial data set included 22 attributes, and all of them were not considered for the study. In the data set the collision configuration and the number of vehicle attributes provide information about how many vehicles were involved and whether they had hit on an object or other vehicles. Another attribute for the requirement of medical treatment represents the people's injured condition after the crash. These three attributes contribute to the afterward occurrence of the fatality. Therefore, these three attributes have been removed during the prediction of a crash's fatality. Next, the person age attribute can range among three categories that are 00 (Less than 1 Year old), 1-98 (1 to 98 Years old), and 99 (99 Years or older). These are not insightful and excluded from the attribute set during the prediction. Moreover, the value of the vehicle sequence number and person sequence number were excluded more not being relevant during prediction analysis. In total, 16 attributes have been considered for this study. In the data set, the coded values of the records are used. To give a more precise idea of all the attributes considered for this study, Table 2 represents the attributes' original categories. Each attribute is purely categorical in nature. Some attributes consist of two categories, while others have at most 30 categories. The most frequently appearing values of each attribute have been highlighted in Table 2 with their respective percentages against all the records. The attribute C_YEAR represents the collision

year. Next, the month when the crash occurred is indicated by C_MNTH. Among the categorical values of days (C_WDAY) and hours (C_HOUR) ratio of Friday and the time of collision between 16:00 to 16:59 are higher. The attribute that provides the severity information of the collision is C_SEV.

The data set also contains information about weather, traffic, and road description. C_WTHR explains various weather conditions where majority records favor a clear and sunny category. Furthermore, there are three variables related to road conditions. Road surface characteristics are reported with the C_RSUR, where most of the records appear as dry and normal ground. Next, the information on different kinds of the lane is indicated with C_RCFG, where it represents mainly the road configuration. Other details, such as the road alignment and traffic control when the collision happened, are expressed with C_RALN & C_TRAF. The vehicle information is given by the variables V_TYPE and V_YEAR. The values of V_TYPE and V_YEAR reflect the type of engine and the design specifications.

Next, some other attributes describe personal level information such as gender, any safety measure used, the person's position, and class of the person. P_SEX is defining the gender of victims. The position of the person is represented with P_PSN. Another attribute P_USER depicts whether the person is driver, passenger, or pedestrian. Lastly, the information about any safety measures were used by the riders is given by the P_SAFE. In total, the data set contains 58,60,405 records, where each record includes information about the observed crash type, road type, vehicle type, and some victim information. For the prediction analysis, C_SEV (Collision severity) has been considered as the target class label. In the target class, the number of records for minority class (fatal) is 98,633, and the number of records for majority class (non-fatal) is 57,61,772.

IV. METHODOLOGY

A. DATA PREPROCESSING

Some of the records are present in the dataset, where the attributes have missing data. Therefore, at the initial step of data preprocessing, the rows in the data set containing any missing value were dropped. Then, label encoding to the person gender attribute (P_SEX) was applied. In this process, the labels were converted into the numeric form, where "M" and "F" represent 1 and 2, respectively.

1) CATEGORIZING ATTRIBUTE VALUES

Two attributes are further categorized, which are collision month (C_MNTH), and collision time (C_HOUR). The categorization has been done to better organize the data and distinguish more meaningful and critical attributes responsible for crashes. The data set contains information about the month of a crash that occurred, but information about the relation of a crash and the season is more useful. Therefore, the collision month attribute categorized into four categories. The four categories are Spring (March-May), Summer (June-August), Fall (September-November), and Winter (December-February). From these categories, the information about which seasons are more prone to crashes can be identified during attribute selection. Another attribute is the collision time that provides the hour of the collision. This attribute has been categorized into five types, such as Midnight (0.00 - 3.59), Dawn (4.00 - 6.59), Morning (7.00 - 11.59), Afternoon (12.00 - 17.59), and Evening (18.00 - 23.59). Rather than a specific hour of a crash, one of these categories for the crash hour is more meaningful.

2) ATTRIBUTE ENCODING

The dataset used in this paper has attributes with different categories. One-hot encoding has been applied to perform in-depth attribute analysis of every factor present In this procedure, categorical data are converted into binary vectors. Firstly integer encoding is applied to the categories, then the individual integer is represented as a binary vector where the index of each integer is 1, and the remaining index is marked as 0. In this dataset, all the categorical attributes are already integer encoded, therefore, one-hot encoding has been applied only to the following 13 attributes from the total of 16 attributes: C_MNTH, C_WDAY, C_HOUR, C_RCFG, C_WTHR, C_RSUR, C_RALN, C_TRAF, V_TYPE, P_SEX, P_PSN, P_SAFE, and P_USER. Finally, the dataset has 104 attributes to analyze after one-hot encoding.

B. PREDICTION METHODS

1) CLASSIFIERS

For predicting whether a crash is fatal or non-fatal four methodologies have been employed: K-Nearest Neighbor [20], Random Forests [21], SVM [22], and Convolutional Neural Network (CNN) [23].

In K-NN, for each test record, the distance between the point considered and each training sample is calculated. Among the distances from all other records, k minimum distances from the test sample is taken where each of these distances corresponds to an already classified data point. The final classification decision is made by taking a majority vote over k nearest neighbor. The experiments for k-NN has been performed with various k values ranging from 3 to 10. Apart from two different values of k that are 7 and 9, others lead to very poor prediction accuracy rates. Among these two, the value of k being 7 approaches towards the best prediction performance. Therefore, the value of k was selected as 7. The standard KNeighborsClassifier package [24] has been applied using python with the minkowski as default distance measure.

The Random Forests classifier is the combination of multiple decision trees as classifiers, where the decision trees are trained on the random subset of training data. To generate the decision tree, bagging is used by randomly selecting N samples with replacement from the original training set. Only a subset of h attributes of all H attributes is tested for selecting the most informative attributes at each node of the decision tree. To classify records from the test

data set, each sample is passed to the model with M trees. The class that obtains majority vote among the M trees is selected as the predicted class for the considered record. The RandomForestClassifier package [24] has been implemented using python, with 250 trees in the Random Forests.

Support Vector Machine (SVM) creates a line or hyperplane to separate the dataset into the desired class labels. For the classification procedure, the hyperplane is learned from the training, such as maximize the margin. Simply, the support vector machine aims to find the best decision boundary that separates the classes as much as possible. If the data set is not linearly separable and leads to poor classification results, then the kernel trick is used. This method mapped the non-linear separable data into higher dimensional space, therefore, it becomes linearly separable. In the considered dataset, the number of observations is larger than the number of attributes; therefore, the RBF kernel trick has been used. In this experiment, to implement the support vector machine the svm package [24] has been used using python.

CNN is a multi-layer supervised learning neural network. It is comprised of convolutional layers, pooling layers, and fully connected layers. The core modules of the network are the convolutional layer and the pooling layer. They work as an attribute extraction function. Here the network improves the accuracy by frequent iterative training. During this iteration, gradient descent is used that minimize the loss function, and adjust the weights of the network each time. In this research, 1-D convolutional layer and three fully connected layers have been employed with 100 nodes each. For the pooling layer, max pooling has been considered. A dropout rate of 0.5 has been considered to prevent overfitting in neural networks. Finally, batch normalization has been performed to effectively reduce training epochs and stabilize the overall training process. There was two output of CNN, fatal or non-fatal crash. To implement the CNN, the Keras package [25] with TensorFlow [26] running has been used in the backend.

2) ENSEMBLE MODELS

In the data set, the fatal crash to non-fatal crash ratio is around 1:60, which clearly indicating a significant degree of class imbalance. Traditionally, the imbalance problem is solved with any of the following methods, such as over sampling, under sampling, and SMOTE (Synthetic Minority Over-Sampling Technique). Oversampling the minority class records can lead to overfitting [27]. Therefore, to solve the class imbalance problem of the dataset, under-sampling method has been considered in this study. In the under-sampling method, the records from the majority class (non-fatal) are reduced randomly.

The process of predicting the class label started with dividing the dataset into training sets and testing sets using the stratified 5-fold cross-validation. Instead of running the model into a single fold, under-sampling has been performed five times to the training sets to ensure the results' stability. Following this, the classification methods (e.g., SVM, Random Forests, K-NN, and CNN) has been applied to each of the training sets and generated multiple models for the prediction. The significant benefit of using the ensemble method is that multiple models are more reliable than a single prediction model and provide more accurate results. Moreover, the issues regarding highly imbalanced data set gets resolved as well with the aid of this sophisticated ensembling approach. The prediction results of these multiple models consolidated by using two different voting approaches that are majority voting approach and soft voting approach. In the hard voting or majority voting, every model provides a vote towards a class label for each test record. The final prediction result is the one that gets the majority of the votes. On the other side, in the soft voting method, every model provides the probability of class label, and then the average of the probability for each class is calculated. The class, which has higher prediction probability, is assigned as the class label for the test record.

C. ATTRIBUTE SELECTION

To get the attributes that are contributing to the prediction of crash severity, information-content analysis has been conducted. Information gain quantifies how much information an attribute is giving about the class. Information gain is basically a reduction in entropy. It is evaluated from the difference between the conditional and unconditional entropy of the outcome. The higher the value of the information gain, the more significant the attributes during the prediction of the class labels.

V. RESULTS & DISCUSSION

To get the important attributes that are contributing to the severity of the crash, the aforementioned methods were applied to each of the training sets. By using the stratified 5-fold cross-validation, five different training sets were used. Then, under-sampling was performed 5 times to each of the training sets, therefore, in total, the number of the training set was 25.The information gain of all the attributes in all different training sets were plotted for analyzing the important attributes. For getting the most contributing attributes set, some cut-offs were considered from the plotted results. Using different cut-offs, the contribution of various attributes in predicting the severity was analyzed. In the result section, the cut-offs and the contributing attributes have been elaborately discussed.

A. PERFORMANCE EVALUATION

Evaluating the performance based only on the accuracy is not a good measure if the data set is imbalanced. Therefore, the precision, recall, and F1-score were calculated along with the accuracy for the performance evaluation of different models. The values of accuracy, precision, recall, and F1- score were calculated as follows.

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \times 100\% \qquad (1)$$

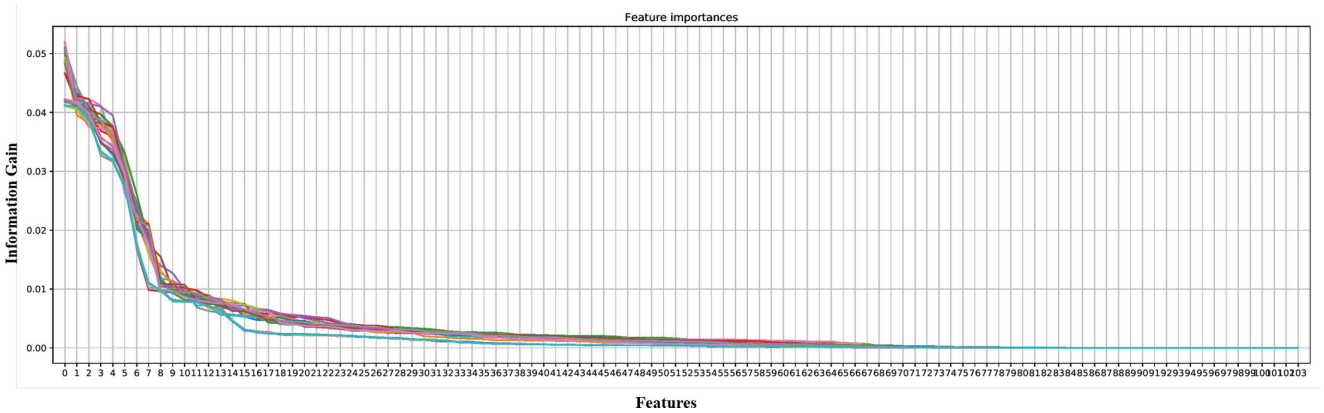

FIGURE 2. Attributes ranking by calculating the information gain for all the 25 different training sets. In this graph, each line illustrates the information gain of attributes in particular one out of 25 training sets.

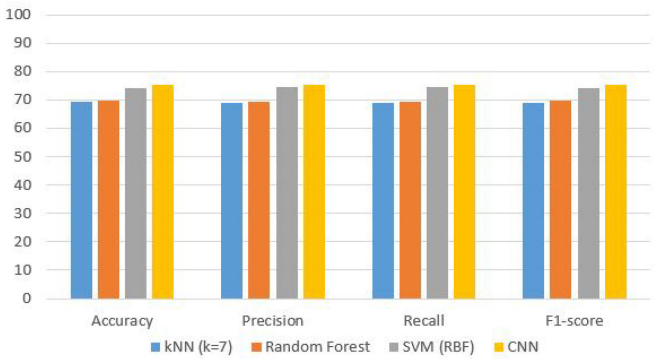

FIGURE 3. Comparison of the prediction performance using KNN, Random Forest, and CNN with soft voting on the complete attribute set (104 attributes).

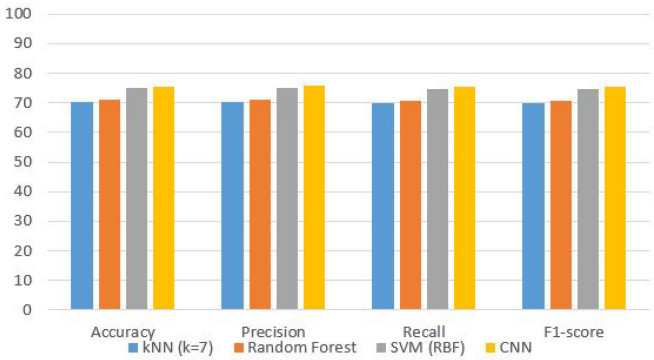

FIGURE 4. Comparison of the prediction using KNN, Random Forest, SVM and CNN with majority voting on the full attribute set (104 attributes).

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{3}$$

$$F1 - score = \frac{2 \times P \times R}{(P + R)} \times 100\%. \tag{4}$$

Here, TP = True Positive, TN = True Negative, FP = False Positive, and FN = False Negative.

To predict whether a crash is fatal or non-fatal, both soft voting and majority voting techniques were applied to the developed models.

In Fig. 3, soft ensembling was employed on the KNN, Random Forests, and CNN for the prediction. The soft voting technique provides the prediction of a crash being fatal or non-fatal, based on which class has the highest average probability. It can not give probabilities using the SVM, therefore, soft ensembling could not be applied to the SVM. Only the majority voting was used for the experiment on the SVM. The overall accuracy ranged from 69.38% to 75.56% for KNN, Random Forests, SVM, and CNN. Here, the prediction accuracy of KNN is lower (69.38%) than the rest of the three methods. Among all the methods, CNN performed significantly higher than others with 75.56% accuracy. In addition to this, while taking both precision and recall consideration, CNN also performed better than all other methods, with 75.74% precision, and 75.45% recall.

In Fig. 4, majority voting was used on the KNN, Random Forests, SVM, and CNN for the final prediction. After using majority voting, the overall accuracy ranged between 70.38% and 75.56%. The performance of KNN and Random Forests improved in terms of all the performance metrics after using the majority voting. For KNN, the accuracy increased steadily from 69.38% to 70.38%, and the improvement reached to 71% for the Random Forests. However, CNN and SVM achieved approximately the same performance by using both majority voting and soft ensembling. Thus, from the experimental results, it is clear that CNN performed best in predicting crash severity by using both majority voting and soft voting.

B. ATTRIBUTE SELECTION ANALYSIS

Initially, the data set includes a total of 104 attributes. Next, this research study focused to identify the most critical attributes from the 104 encoded attributes set to get a precise idea about which attributes are more contributing to the fatality. In the stratified 5-fold cross-validation, under sampling was applied five times for each of the training sets. Therefore, a total of 25 training sets were considered. Next,

the information gain for each of the attributes in each of the training sets were calculated.

Fig. 2 represents the information gain of all 104 attributes in different training sets. After plotting the information gain shown in Fig. 2, it is noticeable that when the information gain is less than 0.01, there is a drastic fall in the information gain of the attributes. Then from the top 15 attributes, the information gain gradually decreases until the top attributes contributing towards fatal crashes. After the top 67 attributes, there is not much significance in information gain for the following attributes. Therefore, three cut offs were considered to identify the most critical attributes from the attributes set. The cut-offs are top 67 attributes, top 15 attributes, and attributes with information gain ≥ 0.01. However, top 15 attributes and attributes with information gain ≥ 0.01 decreased the prediction performance drastically. Therefore, the top 67 attributes were considered for prediction analysis.

C. PREDICTION ANALYSIS

Then, KNN, SVM, RF, and CNN were applied to the 67 attributes set to observe their collective behavior in the prediction. Fig. 5 and Fig. 6 show the experimental results using the reduced 67 attributes set. Soft voting was used to the aforementioned models (KNN, RF, SVM, and CNN) in Fig. 5. CNN achieved the highest prediction performance (75.16%) than the other three prediction models. On the other hand, with the majority voting in Fig. 6, the highest prediction accuracy is 75.46%. The prediction results of a crash being severe or not is pointing towards the fact that though the individual information gain of attributes was not high (≤ 0.05), but by using the top 67 attributes, the prediction accuracy was approximately equivalent to the prediction results with 104 attributes.

Table 3 lists the categories of top 67 attributes and their probable data source in the first two columns that are suitable for building a real-time risk prediction application. The third column lists all the attributes associated with individual categories. The attributes having information gain ≥ 0.01 have been registered with "*" markers in Table 3. The last two columns of the table include a comparison on the percentage of fatal and non-fatal crashes in the presence of each specific attribute. The percentage estimations can give an overall idea regarding the presence of an attribute being associated with fatal and non-fatal crashes. This percentage, R_f is calculated using the following formula:

$$R_f = \frac{X_f}{C} \times 100; \tag{5}$$

where, X_f corresponds to the number of fatal/non-fatal crashes occurring in the presence of a specific attribute f, while C represents the total count of fatal/non-fatal crashes in the dataset. In order to determine the statistical significance of the differences between fatality and non-fatality rates due to the presence of an attribute, the two proportions Z test was performed. It was found out that all of the differences between percentages for every attribute listed in Table 3 were indeed statistically significant (p-value < 0.05).

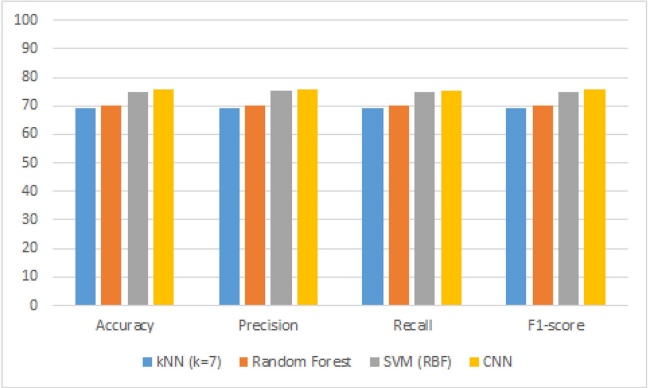

FIGURE 5. Comparison of the prediction performance using KNN, Random Forest, and CNN with soft voting after selection of top 67 attributes.

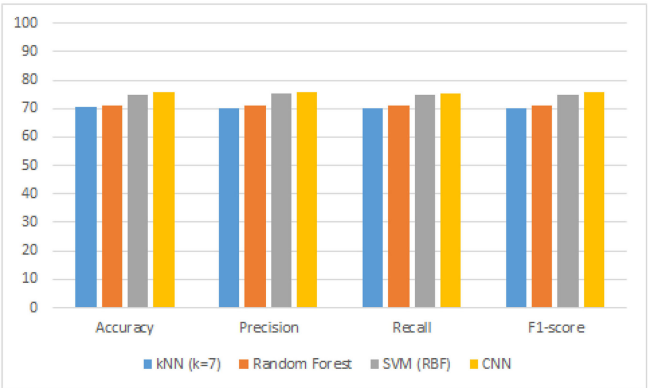

FIGURE 6. Comparison of the prediction performance using KNN, Random Forest, SVM, and CNN with majority voting after selection of top 67 attributes.

A few number of the attribute and their respective fatality versus non-fatality rates may not seem to be meaningful. For example, the presence of "clear and sunny" weather is indicating towards higher percentage of fatal crashes compared to non-fatal crashes. The reason behind this may be due to the limitation of dataset which solely depends on the Canadian context. Further investigation of these attributes using comprehensive datasets from other countries will potentially be helpful in resolving such issues.

To summarize, Figs. 3 and 4 illustrate the performances of each methodologies before attribute selection, whilst Figs. 5 and 6 demonstrate the results after the attribute selection process on reduced 67 dominant attribute set. These four figures demonstrate that, even after the attribute set was reduced by 35.6% compared to the original attribute set, the loss of performance in terms of prediction aptitude was quite trivial. Moreover, this prediction results give the insight that individually, the attributes may not give high information (< 0.055) about the class labels, but collectively they are useful for predicting the fatality or non-fatality of a crash.

TABLE 3. Top 67 attributes with their probable data source and contribution statistics towards fatal and non-fatal road crashes. The last two columns represent the percentages of fatal and non-fatal crashes in the presence of these attributes.

Category	Probable data source	Attributes	Non-fatality Rate	Fatality Rate
Roadway configuration	Real-time road information API (e.g., Ontario 511) or road database by Government (e.g., Road Network File)	Intersection with parking lot entrance/exit, private driveway/laneway*,	**53.42%**	30.09%
		At an intersection of at least two public roadways,	38.96%	**64.96%**
		Bridge, overpass, viaduct,	**6.04%**	3.00%
		Passing or climbing lane,	**0.18%**	0.04%
		Ramp	0.008%	**0.07%**
Road user	Manual configuration	Motorcyclist,	32.32%	**37.76%**
		Bicyclist,	**66.04%**	58.51%
		Motor vehicle passenger,	1.35%	**2.74%**
		Motor vehicle driver*	0.41%	**0.96%**
Road alignment	Real-time road information API (e.g., Ontario 511, World Street Map) or road database by Government (e.g., Road Network File)	Straight and level*,	**78.13%**	61.14%
		Straight with gradient,	3.80%	**9.48%**
		Curved with gradient,	6.30%	**14.74%**
		Top of hill or gradient,	0.53%	**1.15%**
		Bottom of hill or gradient,	0.37%	**0.90%**
		Curved and level,	10.84%	**12.57%**
Person position	Manual configuration	Front row, right outboard, including motorcycle passenger in sidecar*,	53.70%	**66.01%**
		Second row, right outboard*,	**67.24%**	61.26%
		Driver,	**46.29%**	33.99%
		Third row, right outboard etc,	0.63%	**1.71%**
		Third row, center,	1.83%	**2.44%**
		Second row, center,	4.50%	**5.33%**
		Second row, left outboard, including motorcycle passenger,	1.39%	**1.89%**
		Front row, center	18.34%	**19.84%**
Safety device used	Manual configuration	Unusual safety device used,	**94.22%**	7 3.84%
		No safety device equipped e.g. buses*,	2.63%	**19.08%**
		Helmet worn for motorcyclists, bicyclists, snow mobilers, all-terrain vehicle riders	1.37%	**3.36%**
Collision hour	Calender APIs	Evening,	3.28%	**6.21%**
		Morning,	4.72%	**11.56%**
		Afternoon,	**45.95%**	36.46%
		Midnight*,	23.41%	**26.49%**
		Dawn	**22.61%**	19.25%
Collision month	Calender APIs	Winter,	**24.49%**	22.01%
		Summer	27.45%	**30.43%**
Vehicle type	Manual configuration	Light Duty Vehicle (e.g., Passenger car, Passenger van, Light utility vehicles, and light duty pick up trucks)*,	**90.50%**	80.39%
		Unit trucks $>$ 4536 KG GVWR All heavy unit trucks, with or without a trailer,	1.04%	**4.91%**
		Urban and Intercity Bus,	1.16%	**3.07%**
		Panel/cargo van $\leq$ 4536 KG GVWR Panel or window type of van designed primarily for carrying goods,	1.53%	**3.68%**
		Motorcycle and moped motorcycle and limited-speed motorcycle,	0.82%	**1.44%**
		School bus standard large type,	0.36%	**0.84%**
		Road tractor with or without a semi-trailer,	3.10%	**4.19%**
		Smaller school bus smaller type, seats $<$ 25 passengers	0.04%	**0.16%**
Traffic control	Real-time traffic information application (e.g., Waze, Traffic Spotter)	Markings on the road e.g. no passing*,	53.66%	**77.73%**
		Traffic signals fully operational*,	**31.08%**	8.73%
		School guard, flagman,	**0.47%**	0.23%
		Railway crossing with signs only,	0.03%	**0.21%**
		School crossing,	0.01%	**0.08%**
		Yield sign,	**0.46%**	0.14%
		Pedestrian crosswalk,	**1.15%**	0.48%
		Stop Sign	0.05%	**0.12%**
Person sex	Manual configuration	Male	1.53%	**3.68%**
Day of week	Calender APIs	Sunday,	12.25%	**16.37%**
		Saturday,	15.02%	**18.18%**
		Wednesday,	**13.93%**	11.51%
		Tuesday,	**13.86%**	11.84%
		Monday	**13.16%**	11.84%
Weather condition	Crowdsourcing from various weather apps (e.g., WeatherCAN, Dark Sky, The Weather Channel, AccuWeather, and so forth)	Clear and sunny,	**71.13%**	68.18%
		Freezing rain, sleet, hail,	1.41%	**3.68%**
		Snowing, not including drifting snow,	**11.31%**	8.24%
		Raining,	9.32%	**11.53%**
		Overcast, cloudy but no precipitation,	6.01%	**6.99%**
		Strong wind,	0.24%	**0.50%**
		Visibility limitation	0.53%	**0.84%**
Road surface	Smartphone's accelerometers and GPS sensor based platform (e.g., SmartRoadSense)	Wet,	**20.22%**	16.05%
		Muddy,	5.48%	**6.78%**
		Slush, wet snow,	0.42%	**1.21%**
		Icy includes packed snow,	4.52%	**5.44%**
		Dry, normal	67.80%	**68.47%**

Note: The information gain of "*" marked attributes are greater than 0.01.

D. COMPUTATIONAL TIME AND MEMORY ANALYSIS

It is readily apparent that ensemble CNN emerges as the most suitable model to build an intelligent transportation system (ITS) deployed on different map navigation applications. The integration of pre-trained CNN model on mobile applications can serve the purpose of a real-time fatality risk prediction

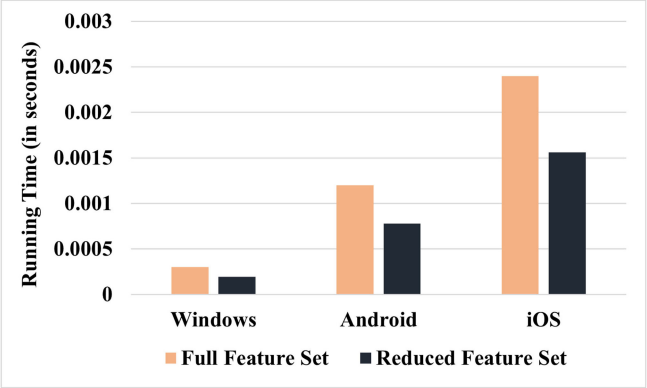

FIGURE 7. Running time comparison of CNN on windows, iOS, and android devices against full versus reduced attribute set.

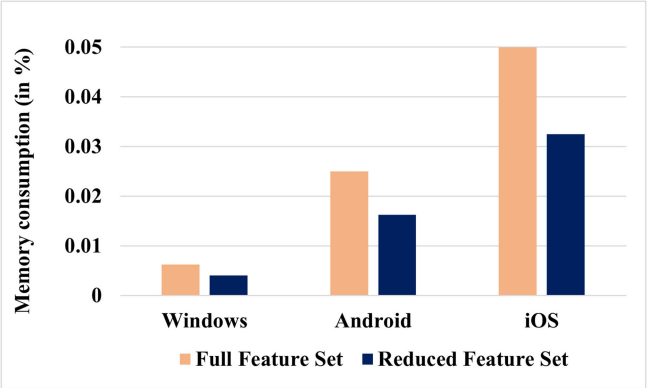

FIGURE 8. Memory consumption comparison of CNN on windows, iOS, and android devices against full versus reduced attribute set.

system without the need of frequent or online training. The reduced (top 67) attribute set can be specifically viable for real-time crash prediction analytics due to even lower computational burden. In support of evidence, time and memory analysis experiments were carried out on different mobilete operating systems.

At first, the required running time and memory consumption of the pre-trained CNN model were derived as baseline on windows through a simulation setup. For this purpose, the experiments were carried out on DELL ALIENWARE m15 R3 machine of Intel core i7-10750H CPU @2.6 GHz equipped with 16 GB RAM and Windows 10 Home. Afterwards, the required running time and memory were calculated on iOS and android devices using numerical analysis. As an example, the most commonly used iOS and android devices in 2020 were considered for the experiments, which were iPhone 7 and Samsung Galaxy A51 [28], [29]. The required time to collect data from different possible sources/APIs was considered negligible throughout this process.

Figs. 7 and 8 justify the applicability of the pre-trained CNN model based on ultra-low running time ($\leq$ 0.0024 seconds) and memory consumption ($\leq$ 0.05%) across various kinds of devices. Furthermore, the reduced set of attributes can even lower the computational overhead by approximately 35% in terms of time and memory without compromising the prediction accuracy. It is noteworthy that the reduced set of attributes can increase the potentiality of the ITS system by removing the obligation to collect 104 attribute values as input and eventually reducing it to 67 only. Hence, the chances of occurring missing input or attribute values leading to the system's failure can decline significantly. Consequently, the reduced set of attributes can be considered very useful in case of data unavailability of some relatively insignificant attributes for a live application to some extent.

VI. CONCLUSION & FUTURE WORK

In this study, a data set of 5.8 million records was considered for predicting the severity (fatal vs. non-fatal) of motor vehicle collisions, and identifying the most critical factors associated with it.

Deep learning and traditional machine learning methods were ensembled to develop better prediction models. ensemble methods using majority voting and soft voting techniques were employed to address class imbalance in the dataset. CNN performed the best with an accuracy of approximately 75% compared to the other prediction models considered in this study. Next, the number of attributes were reduced from 104 to 67 through information content analysis without any significant reduction in prediction performance. These important attributes belong to following 9 broad categories:

- roadway configuration
- road alignment
- collision hour
- weather condition
- person position
- road user class
- safety device used
- traffic control employment.

It is envisioned that this research will contribute positively towards increasing transportation safety. The most contributing attributes to fatal crashes at a fine granularity-level have been discussed in Table 3 of Section V. Unlike other existing similar research studies, this paper focuses on developing a prediction model with reasonable time and memory complexity. The proposed fatal road crash prediction model is suitable to be deployed into real-time mobile map navigation applications, while disregarding the use of any addition sensor.

Possible future work includes incorporating more driver specific behavioural patterns and spatial attributes from different countries and regions in the prediction model. Once more potentially significant attributes, such as, speed, location-specific detailed weather data, behavioural record, alcohol consumption, anti-lock braking system (ABS), and tire category related information are accessible, it is expected that the prediction performance will further improve. Further research is required to compare the performances of specialized versus generalized models trained on region-specific and multi-region datasets, respectively.

REFERENCES

[1] A. Iranitalab and A. Khattak, "Comparison of four statistical and machine learning methods for crash severity prediction," *Accid. Anal. Prevent.*, vol. 108, pp. 27–36, Nov. 2017.

[2] N. S. Hadjidimitriou, M. Lippi, M. Dell'Amico, and A. Skiera, "Machine learning for severity classification of accidents involving powered two wheelers," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 10, pp. 4308–4317, Oct. 2020.

[3] A. Theofilatos, "Incorporating real-time traffic and weather data to explore road accident likelihood and severity in urban arterials," *J. Safety Research*, vol. 61, pp. 9–21, Jun. 2017.

[4] W. E. Evanco, "Impact of rapid incident detection on freeway accident fatalities," Mitretek, McLean, VA, USA, Rep. WN 96W0000071, 1996.

[5] C. Xu, A. P. Tarko, W. Wang, and P. Liu, "Predicting crash likelihood and severity on freeways with real-time loop detector data," *Accid. Anal. Prevent.*, vol. 57, pp. 30–39, Aug. 2013.

[6] T. Huang, S. Wang, and A. Sharma, "Highway crash detection and risk estimation using deep learning," *Accid. Anal. Prevent.*, vol. 135, Feb. 2020, Art. no. 105392.

[7] M. Zheng *et al.*, "Traffic accident's severity prediction: A deep-learning approach-based CNN network," *IEEE Access*, vol. 7, pp. 39897–39910, 2019.

[8] D. Delen, R. Sharda, and M. Bessonov, "Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks," *Accid. Anal. Prevent.*, vol. 38, no. 3, pp. 434–444, 2006.

[9] Q. Cai, M. Abdel-Aty, Y. Sun, J. Lee, and J. Yuan, "Applying a deep learning approach for transportation safety planning by using high-resolution transportation and land use data," *Transp. Res. A, Policy Pract.*, vol. 127, pp. 71–85, Sep. 2019.

[10] M. G. Mohamed, N. Saunier, L. F. Miranda-Moreno, and S. V. Ukkusuri, "A clustering regression approach: A comprehensive injury severity analysis of pedestrian—Vehicle crashes in New York, U.S and Montreal, Canada," *Safety Sci.*, vol. 54, pp. 27–37, Apr. 2013.

[11] T. Beshah and S. Hill, "Mining road traffic accident data to improve safety: Role of road-related factors on accident severity in ethiopia," in *Proc. AAAI Spring Symp. Series*, 2010, pp. 1–6.

[12] S. Kumar and D. Toshniwal, "A data mining approach to characterize road accident locations," *J. Modern Transp.*, vol. 24, no. 1, pp. 62–72, 2016.

[13] M. Bèdard, G. Guyatt, M. Stones, and J. Hirdes, "The independent contribution of driver, crash, and vehicle characteristics to driver fatalities," *Accid. Anal. Prevent.*, vol. 34, no. 6, pp. 717–727, 2002.

[14] M. A. Abdel-Aty and R. Pemmanaboina, "Calibrating a real-time traffic crash-prediction model using archived weather and its traffic data," *IEEE Trans. Intell. Transp. Syst.*, vol. 7, no. 2, pp. 167–174, Jun. 2006.

[15] C. Miao, A. Ajith, and P. Marcin, "Traffic accident analysis using machine learning paradigms," *Informatica*, vol. 29, pp. 89–98, May 2005.

[16] F. L. Peter, A. M. M. Vissers, and M. Jessurun, "Annual mileage, driving violations, and accident involvement in relation to drivers' sex, age, and level of education," *Accid. Anal. Prevent.*, vol. 31, no. 5, pp. 593–597, 1999.

[17] T. Wilson and J. Greensmith, "Multivariate analysis of the relationship between drivometer variables and drivers' accident, sex, and exposure status," *Human Factors*, vol. 25, no. 3, pp. 303–312, 1983.

[18] T. Usman, L. Fu, and L. F. Miranda-Moreno, "A disaggregate model for quantifying the safety effects of winter road maintenance activities at an operational level," *Accid. Anal. Prevent.*, vol. 48, pp. 368–378, Sep. 2012.

[19] "Canadian Car Accidents 1994-2014." Kaggle. Jul. 10, 2017. [Online]. Available: https://www.kaggle.com/tbsteal/canadian-car-accidents-19942014

[20] K. Hattori and M. Takahashi, "A new nearest-neighbor rule in the pattern classification problem," *Pattern Recognit.*, vol. 32, no. 3, pp. 425–432, 1999.

[21] M. Pal, "Random forest classifier for remote sensing classification," *Int. J. Remote Sens.*, vol. 26, no. 1, pp. 217–222, 2005.

[22] S. V. M. Vishwanathan and M. N. Murty, "SSVM: A simple SVM algorithm," in *Proc. Int. Joint Conf. Neural Netw.*, vol. 3, 2002, pp. 2393–2398.

[23] T. Roska and L. O. Chua, "The CNN universal machine: 10 years later," *J. Circuits Syst. Comput.*, vol. 12, no. 4, pp. 377–388, 2003.

[24] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011.

[25] F. Chollet *et al.* "Keras." 2015. [Online]. Available: http://keras.io

[26] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Oper. Syst. Design Implement.*, 2016, pp. 265–283.

[27] M. Altini, "Dealing with imbalanced data: Undersampling, oversampling and proper cross-validation," Aug. 2015, [Online]. Available: https://www.marcoaltini.com/blog/dealing-with-imbalanced-data-undersampling-oversampling-and-proper-cross-validation (Accessed: Mar. 19, 2022).

[28] *Leading Apple iPhone Shares Based on Web Usage Worldwide From 2019 to 2020, by Model*, Statista, Hamburg, Germany, 2020.

[29] Yordan. "Top 20 Most Popular Phones in 2020." GSMArena.com. 2020. [Online]. Available: https://www.gsmarena.com/top_20_most_popular_phones_in_2020-news-46737.php (Accessed: Mar. 19, 2022).

MAHZABEEN EMU (Member, IEEE) received the master's degree from the Department of Computer Science, Lakehead University. She is currently pursuing the Ph.D. degree with the School of Computing, Queen's University, Kingston, ON, Canada. Her research interests include optimization and artificial intelligence in the field of networking. She is a recipient of the Vector Institute AI Scholarship in 2019, OGS 2020–21, OGS 2021–22, and Mitacs accelerate grant. She also received the prestigious 2021 Governor General Gold Medal Award at 56th Convocation of Lakehead University, Canada, and another Gold Medal Award at the 10th Convocation of Ahsanullah University of Science & Technology, Bangladesh for the highest academic standing during master's and undergraduate studies, respectively.

FARJANA BINTAY KAMAL received the B.Sc. degree in computer science and engineering from the Chittagong University of Engineering and Technology, Bangladesh, in 2018. Her research interests include artificial intelligence and data science.

SALIMUR CHOUDHURY (Senior Member, IEEE) is an Associate Professor and a Graduate Co-Coordinator with the Department of Computer Science and leads the Optimization Research Group, Lakehead University. He has published more than 40 peer reviewed publications and received grants from various government sectors and industries as well. The primary research focus of his is network optimization. He is the co-founder of the conference, SGIoT.

QUAZI ABIDUR RAHMAN received the Ph.D. degree from Queen's University, Kingston, ON, Canada. He is an Assistant Professor with the Department of Computer Science, Lakehead University, Thunder Bay, ON, Canada. His research interests include data analytics and applied machine learning.