Paper No: SE-05 **Systems Engineering**

Identification of factors and classifying the accident severity in Colombo - Katunayake expressway, Sri Lanka

M.A.K. Kushan*
Department of Statistics & Computer Science
Faculty of Science
University of Kelaniya, Sri Lanka
kalanakushan290@gmail.com

N.V. Chandrasekara
Department of Statistics & Computer Science
Faculty of Science
University of Kelaniya, Sri Lanka
nvchandrasekara@kln.ac.lk

Abstract: **Sri Lanka's expressway system was launched in 2011 and now owns three major expressways. Many peoples choose expressways rather than normal ways due to the reasons of time, traffic, easy of driving, etc. According to police reports of highway main traffic police branch, in recent years the number of accidents occurring in expressways is increasing drastically. Nowadays, the rate of accident occurrence in Colombo-Katunayake Expressway is high compared to the other two expressways and there was no previous research has been done in Sri Lanka regarding accidents on Colombo-Katunayake expressway. Therefore, the objective of the study was to identify the factors contributing to accidents on the Colombo-Katunayake Expressway and to develop appropriate machine learning models to classify the severity of the accidents. In this study, 704 total accident cases were considered during the period 2013-2019. Chi-square test, logistic regression, and Kruskal–Wallis tests were used to identify the association between the accident severity and other influential variables found from the literature. Finally, seven variables: time category, driver's age category, vehicle type, the reason for the accident, number of vehicles involved, cause for accident and rainfall were identified as influencing variables to accident severity under 5% level of significance. Naïve Bayes classification algorithm and probabilistic neural network (PNN) were used in the study to forecast accident severity. A random under-sampling technique was used to overcome the class imbalanced problem persists in the data set considered in the study. The final models developed using the Naïve Bayes algorithm and PNN exhibit 72.14% and 74.29% overall classification accuracy respectively. Both aforementioned models can be considered as suitable models to forecast accident severity in the Colombo-Katunayake expressway where the PNN model exhibits slightly higher accuracy. The final models developed by this study can be used to implement safety improvements against traffic accidents in expressways of Sri Lanka.**

Keywords: *Classification, Class imbalance problem, Naïve Bayes algorithm, Probabilistic Neural Network (PNN), Road accident*

I. INTRODUCTION

When we look back at the transport history of Sri Lanka, the British era comes to mind in a very specific way. They have made several changes by expanding the road infrastructure, increasing the vehicle population and by developing the railways to transport goods. After the independence and the new government, Sri Lankan roads were further improved. With the rapid growth of roads, the risk of accidents was increased. By considering the situation, the first Traffic Act was enacted in Sri Lanka in 1934 and since 1938 the police have reported traffic accidents [1].

In Sri Lanka, there are 40,887 road accidents every year and an average of six deaths a day. Further, for a year nearly 9,000 motor vehicle accidents, over 5400 accidents due to speed occurred. The records reveal that over 740 pedestrians die every year and an average of two pedestrian deaths per day. Annually among the total road accidents, over 2470 people are faced with accidental deaths, 700 are left in serious injuries while some people with lifelong consequences [2].

With the introduction of the expressway system in Sri Lanka, people tend to use the expressway system apart from main roads for their long commutes due to the time, traffic and ease of driving. This leads to an increase the number of vehicles in expressways of Sri Lanka. As a consequence, the number of traffic accidents on highways also has been increasing dramatically in recent years. When considering all three expressways in Sri Lanka, there were 59 accidents in 2011, 449 accidents in 2012, 409 in 2013 and similarly, in 2018, 597 accidents have been reported. Overall, 3724 accidents were reported on the three major Expressways between 2011 and 2018. During this period 40 peoples died in fatal accidents [3]. Data reveals that the number of accidents has been increased over the last years s.

There are currently three major highways in Sri Lanka: The 'Southern Expressway (E01)', which was opened in 2011 and was the first expressway in Sri Lanka to cover 126.9 kilometers. The 'Colombo-Katunayake Expressway (E03)' was opened in 2013 and which connects the Colombo and Bandaranayake International Airport with a length of 25.8 kilometers and 'Outer-circular Expressway (E02)' was opened in 2014 from Kaduwela to Kottawa covering a distance of 18.9 kilometers [3].

In recent years, the incidence of traffic accidents on the Colombo-Katunayake Expressway is high compared to the other two expressways [3]. Apart from ordinary factors, climate factors may increase the number of accidents in the Colombo-Katunayake expressway. Therefore, it is imperative to identify the most likely causes of highway accidents in the Colombo-Katunayake expressway.

A motor traffic accident is any kind of vehicle accident that happens on a public highway. It includes a clash between vehicles and animals, vehicles and fixed obstacles, vehicles and pedestrians and also a single-vehicle accident without other road users [4].

Road accidents can be identified as a growing problem nationally or internationally. According to the World Health Organization (WHO) report, the number of accidents from non-communicable diseases increased from 28.1 to 49.7 in 1990 to 2020, with road accidents being the main cause of this increase [4]. Between 1938 and 1997, the number of road traffic fatalities in Sri Lanka increased tenfold and in 1997 there were 1835 deaths. Despite the need for prompt action to reduce this growing burden, successive governments at that time did not trust the country's road safety strategies and road safety research [5].

The main objective of most previous studies was to identify factors that significantly affect the severity of an accident [1,2,5]. Highway characteristics, atmospheric factors, and driving characteristics have been identified as factors contributing to the severity of a traffic accident, and usually one or more of the above-mentioned factors can contribute to the severity of the accident. [6].

According to the Institution of Highways and Transportations (IHT), there are numerous factors for road traffic accidents such as light conditions, weather, faulty design, vehicles with mechanical defects and Inadequate road infrastructure development [4].

While many researchers have not considered all the factors that contribute to an accident, but some believe there is a greater potential for the severity of accidents, such as time of the accident, age of the driver, gender of the driver, type of the vehicle, type of the accident, place, cause for accident. Other factors were not investigated due to significant limitations in the data obtained from accident reports.

In recent years, researchers have increasingly focused on identifying the factors that significantly affect the severity of road accidents [8, 9]. There are many techniques that researchers have used to study this issue including artificial neural network, cage logic processing, fuzzy ART maps, log-linear modeling, etc. [7, 9, 10]

The use of artificial neural network techniques to model traffic accident data reports helps to understand the drivers' behavior, road conditions, and weather conditions associated with the severity of various injuries. This will help decision-makers to develop better traffic safety control policies.

Many researchers used an artificial neural network to analyze the frequency of freeway accidents and pointed out that the artificial neural network system does not require a predefined underlying relationship between dependent and independent variables [8]. Studies have shown that artificial neural networks are a consistent alternative method for analyzing the frequency of freeway accidents [8,11].

Non-linear relationships between injury severity levels and crash-related factors were modeled by using a series of artificial neural. From those factors, it shows that artificial neural networks have better predictive power compared to other methods such as logistic regression, ARIMA, VAR, etc. [12]

Another research that focused on traffic accidents at signalized intersections, is divided the severity of the injury into three classes such as no injury, possible injury and disabling injury. And also founded that the multi-layered perceptron (MLP) classification accuracy was higher than the Fuzzy ARTMAP [11].

Today road accidents become a vast problem not just for Sri Lanka but for all countries in the world. Therefore, it is imperative to take immediate steps to prevent accidents. The severity of the road accident on this study road is the greatest in property damages only accidents (86.22%) when compared with the Slight Injury (9.52%), Serious and Fatal injury (4.29%) [3]. Not a great deal of research has been conducted on expressways in Sri Lanka, and most of the research has been done using traditional methods such as logistic regression. Hence, this study is a novel approach to road accidents. The result of this research helps to place a policy of prevention for the police maker analyzing the road accident severity. This paper identifies the most likely causes of expressway accidents, builds the Naïve Bayes algorithm and Probabilistic Neural Network (PNN) to classify the severity of accidents on the Colombo-Katunayake expressway and compare the classification accuracy of these two models.

The organization of the paper is as follows: Section 2 consists of theories & techniques used for the study. Section 3 describes data pre-processing and preliminary analysis. Model building process and Classification of Accident severity are in Section 4 and Section 5 respectively. Performance measures explain in Section 6 whereas Section 7 contains results and discussion. Section 8 consists of conclusions followed by the acknowledgment and references conclude the article.

II. THEORIES & TECHNIQUES

Classification is an important data mining technology with a wide range of applications for classifying the various types of data used in almost every area of our lives. Classification analysis is the organization of data in a given class. This also known as supervised classification, uses the class labels provided to order data collection objects. Classification approaches usually use a training set that is associated with class labels that all objects are already known. The classification algorithm learns from the training set and builds a model. The trained model is used to classify new objects [13].

A. Naive Bayes classifier

The Naive Base algorithm is a simple probabilistic classifier that computes a set of probabilities by computing a combination of frequencies and values in a given dataset. The algorithm uses the base theorem and the assumption made in this classifier is that the predictors/features are independent. That is the presence of one particular feature does not affect the other. This conditional independent assumption is rarely true in real-world applications, and hence is characterized as naive, but the algorithm tends to perform well and fast in various classification problems [14].

B. Artificial Neural Networks (ANN)

When it comes to machine learning, it can train the computer to detect data patterns in different ways. This includes learning from experience, learning by examples, analogues learning, etc. With an increase in training time, the performance of any machine learning algorithm can be increased. The basis of the adaptive system is the machine learning mechanism, and Artificial Neural Networks (ANNs) are the most popular approach to machine learning [15].

The Artificial Neural Network (ANN) can be described as a technology that functions similar to that of the human brain. The human brain consists of a set of basic information processing units called neurons and there were approximately 10 billion neurons and 60 trillion connections between synapses [16]. By using multiple neurons simultaneously, the human brain can perform calculations, tasks, and decisions much faster than any high-performance computer present in today [15].

ANN models have the ability to mimic human information processing processes such as knowledge processing, prediction, and control. ANN system has gained much attention compared to other existing technologies, due to the reasons for learning spontaneously from examples, arguing about vague data, and responding to new information not previously stored in memory. ANN technology is now widely used in engineering, medicine, and many other fields. Therefore, this technology is gaining greater acceptance worldwide [15].

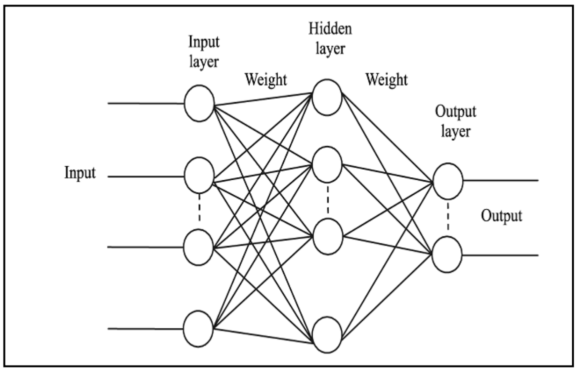

Fig. 1. Architecture of ANN

ANN is relatively new in the field of traffics engineering and accident analysis [15]. This new approach has rarely been shown in areas such as predicting traffic congestion [15, 16], determining truck attribute, and a few other applications. Further, the artificial network has been used to solve problems that have not been solved by statistical methods [8,12]. Researchers have done comparative studies of statistical methods with ANNs [15,17]. Their studies show that training in medium and large datasets can be very useful for ANN prediction [15, 16].

A neural network can learn both at the presence of teachers (Supervised learning) and without teachers (Unsupervised learning). Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses. Supervised learning techniques mainly divide into two parts. They were regression and classification.

C. Probabilistic Neural Network (PNN)

The probabilistic neural network (PNN) is one of the most commonly used feedforward neural network techniques for classification and pattern recognition problems. Due to the ease of training and good statistical basis of Bayesian estimation theory, the PNN tool is a widely used technique to solve many classification problems. When constructing a more accurate PNN model, care should be taken to determine the size of the network, the location of the pattern stratified neurons, and the value of the smooth parameter [18].

With the presence of a categorical response variable (accident severity), two main classification techniques (Naïve Bayes algorithm and PNN) were used in this research to classify the severity of the accident. Accident severity classification and performance analysis are described in the next sections.

D. Class Imbalance Problem

Data distributions are said to suffer from class imbalance when class distributions are very unbalanced. Most of the classification algorithms are more focusing on the classification of majority sample cases correctly while ignoring or misclassifying minority sample cases. Therefore, most classification learning algorithms provide low predictive accuracy for a class that has rarely occurred. Minority samples, except for the majority class samples, rarely occur but are very relevant to many real-world problems. There are different methods available to overcome the class imbalance in the dataset, which were the algorithmic approach, data-preprocessing approach, feature selection approach, etc. [19]. The multiple class imbalance problem arose because there were three classes (property damage, minor injury, serious injury) of the response variable in this study.

E. Random Under-Sampling

Under-sampling is one of the most used methods by many researchers to solve the class imbalance problem in data level approach. The most recently identified under-sampling method is random under-sampling, which is a non-heuristic method that attempts to balance class distributions by randomly deleting data from majority classes [20].

III. DATA PRE-PROCESSING & PRELIMINARY ANALYSIS

The accident data set used in this study was obtained from the expressway main traffic police branch in Kaduwela, Sri Lanka. Meteorological data (daily rainfall, daily average temperature) for the date of the accident have been obtained from the Department of Meteorology. All accidents on the Colombo-Katunayake Expressway from 2013 to 2019 were considered in this study which covers 704 cases. The police Accident Reporting System on the Colombo-Katunayake Expressway is poor and collecting data on the correct road accidents has been a great challenge in this study.

According to the variable definitions for the records of the general data estimation system, the dataset has drivers' records and vehicles' information. Some attributes which can be considered as input variables to models are time, driver's age, driver's gender, the ethnic group of driver, the total vehicle involved, cause for the accident, vehicle type, reason for the accident, injury severity, place category, rainfall, temperature. Additionally; the response variable of this research (accident severity) is also another attribute of a given accident record. The target variable, accident severity has three classes: property damages only (no injury), slight injury and serious and fatal injury. and Figure 2 illustrates the percentage of cases under each category.

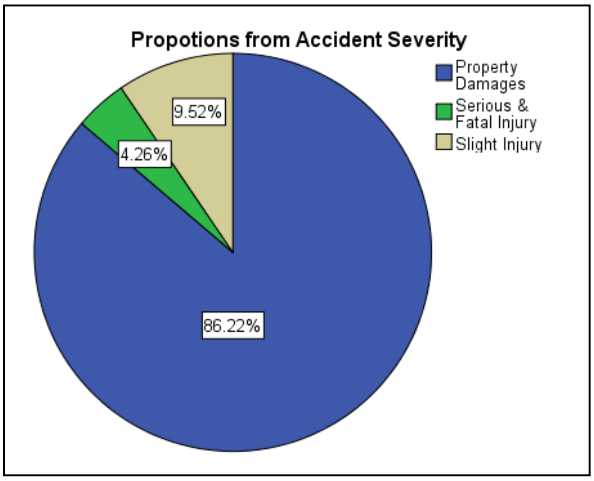

Fig. 2. Percentage of cases under each category of accident severity on the Colombo-Katunayake Expressway, Sri Lanka

In the original dataset, 86.22% of the cases belong to the class no injury (only property damages), 9.52% of the cases for slight injury and 4.29% of the cases for serious and fatal injury class.

The main objective of this research was to identify the major causes of accidents on the Colombo-Katunayake Expressway in Sri Lanka and to classify the severity of accidents. Developed machine learning models can be used to understand the relationship between driver characteristics, vehicle type, road conditions, and the environment and the severity of the accident. The results from this data analysis can provide critical information for the adoption of high technology in road accident prevention policy, especially in accident control mechanisms. The records in the dataset are arranged in pairs of inputs and outputs and each record has its own output. The supervised learning algorithms will map the input vector to the desired output for the given set of values.

The number of deaths due to road accidents on Colombo-Katunayake Expressway from 2013 to 2019 is very low. Hence, one response category was created as serious and fatal injuries which include all very serious accidents. Further, Expressway's main traffic police branch classifies the reason for an accident into eight main categories.

They are drink and drive, careless driving, rain, technical errors, hitting animals, sleeping, high speed and by other reasons.

Figure 3 illustrates the bar chart of the reason for accidents where the highest reason is other and the second highest reason is careless driving. Further, high speed and rain can be considered as potential reasons for an accident.

Eleven independent variables have been used to explain the variation in the dependent variable (accident severity). as: time category, driver's age category, driver's gender, the ethnic group of driver, number of vehicles involved, cause for the accident, vehicle type, reason for accident, place category, rainfall, the temperature in this study.

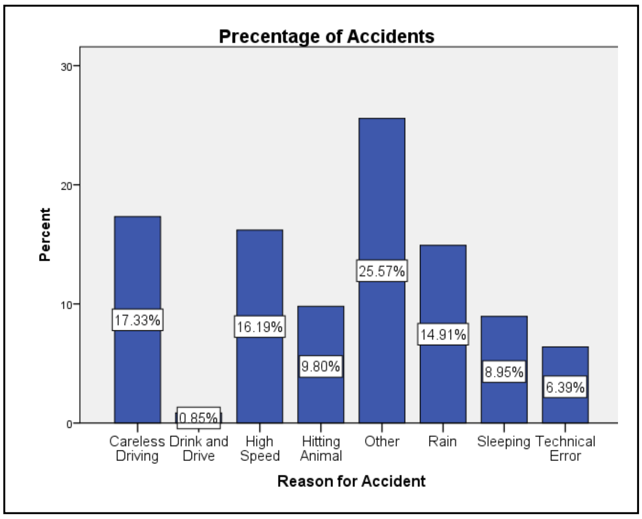

Fig. 3. Chart of the percentages of reasons for the accident

Among all independent variables, nine were categorical variables (non-climate factors): time category, driver's age category, driver's gender, ethnic group of driver, number of vehicles involved, cause for accident, vehicle type, reason for accident, place category and two were continuous variables (climate factors): rainfall and temperature.

Association between the response variable (accident severity) and categorical predictor variables were assessed using Pearson's chi-squared parametric test. Among nine categorical predictor variables driver's age category, time category, vehicle type, reason for the accident, number of vehicles involved and cause for the accident were identified as significant variables at a 5% level of significance. Furthermore, Logistic regression (parametric) & Kruskal–Wallis tests (non-parametric) were used to identify the association between the response variable (accident severity) and continuous predictor variables. Rainfall was found to be a significant variable in both tests.

Finally, nine variables: time category, driver's age category, vehicle type, reason for the accident, number of vehicles involved, cause for accident and rainfall were identified as an associated variable to accident severity under 5% level of significance.

IV. MODEL BUILDING PROCESS

The entire dataset was divided into two parts, 80% of data was used for model building and 20% of the data was used to validate the model. Since this is not a time series data, 80% of the data were selected in various ways for model building and the remaining 20% were used to test the built model.

Method 1: 80% (564) of the randomly selected data was used for model building and the rest 20% (140) data were used to test the model

Method 2: The first 80% (564) of the dataset was used to build the model and rest 20% (140) data were used to test the model.

Method 3: Considering the class ratio (property damage: slight Injury: the ratio of serious & fatal injuries is 86: 10: 4) of the original dataset, (86/100) *564 = 485 ‘property damages only ’accidents

(10/100) *564 = 56 ‘slight injury ‘accidents

(4/100) *564 = 23 ‘serious & fatal injury ’accidents

(485+56+23) = 564 (accident records)

Altogether, 80% (564) of the data is randomly selected to build the model. The rest 20% (140) data were used to test the model.

Method 4: This was an improved version of the Method 3 with the under-sampling technique. By using the under-sampling technique considering the class ratio between property damages: slight injury: serious & fatal injuries which were started in *Method 3*, Out of 564 of data, 176 of the data was randomly selected as follows:

(86/100) *564*(20%) = 97 ‘property damages’ accidents

(10/100) *564 = 56 ‘slight injury’ accidents

(4/100) *564 = 23 ‘serious & fatal injury’ accidents

(97+56+23) = 176 (accident records)

After the under-sampling technique, the original dataset ratio was changed to a new ratio (Property Damage: Slight Injury: the ratio of serious & fatal injuries is 55: 31: 14). The rest 20% (140) data which were used in *Method 3* were used to test the model.

Above data splitting methods were used in the model building process for both the Naïve Bayes algorithm and Probabilistic Neural Network (PNN).

V. CLASSIFICATION OF ACCIDENT SEVERITY

To clarify the severity of accidents, the above mentioned two classification techniques have been used and two techniques were compared using model accuracy.

A. Naive Bayes classifier

In Model 1, the model was constructed using the *Method1* data partitioning technique considering all predictor variables.

The key assumption of the Naïve Bayes Algorithm is all explanatory variables are needed to be independent with each other. Hence, it evaluated the multi-collinearity of all the above variables using the Variation Inflation Factor (VIF) value. Multi-collinearity was detected from two variables ‘Vehicle type’ and ‘Number of vehicles involved’. Hence, Model 2 was created by dropping the two variables which exhibit multi-collinearity and using the *Method1* data splitting technique.

B. Probabilistic Neural Network (PNN)

Since the dataset is not a time series data, four processes described under Section 4 were carried out to select the 80% data to fit the model. When selecting the variables, the first evaluation was carried out using all eleven explanatory variables, but no significant classification accuracy in the built-in model which has been fit for the data. Then, the identified seven significant variables at a 5% significance level from Pearson’s chi-squared test, logistic regression, and Kruskal –Wallis tests (as mentioned in Section 2) were used for further models.

In Model 3, *Method 1* data splitting technique was used to construct the model.

In Model 4, *Method 2* data splitting technique was used to construct the model.

In Model 5, *Method 3* data splitting technique was used to construct the model.

VI. PERFORMANCE MEASURE

A. Classification

Classification is one of the supervised learning techniques and means to group the output (response variable) inside a class. If the algorithm tries to label two different classes of input, it is called binary classification, and selecting more than two classes is called multiclass classification. The performance of the classification algorithm is usually checked by classification accuracy. Determining the best of the best classification depends on the user interpreting the problem. The classification accuracy depends on the number of data correctly classified into the relevant classes. This ignores the fact that the wrong class can also have costs associated with the wrong assignment and which should need to be determined [21,22].

B. Confusion matrix

The confusion matrix represents the accuracy of the solution to a classification problem. The n-classes classification problem generates a n x n confusion matrix, the number of values in Cij class indicates that the number of observations of the dataset that were predicted to class Cj but where the correct class is Ci. Obviously, the best solution is only when there are zero values outside the diagonal [13].

The confusion matrix contains information about actual and predicted classifications performed by a classification system. The following figure 4 shows the confusion matrix for a binary classification problem and the equation for calculating classification accuracy is displayed in Equation 1. This procedure is also applicable to the multi-class classification problem.

Figure 4 below shows the structure of the classification table and the equation for calculating classification accuracy.

		True Class	
		P	N
Hypothesized class	Y	True Positive (TP)	False Positive (FP)
	N	False Negative (FN)	True Negative (TN)

Fig. 4. Structure of classification table / confusion matrix

Equation 1:

Classification accuracy of the table = (TP +TN) / (P+N)

VIII. RESULTS & DISCUSSION

The following tables illustrate the classification accuracy of accident severity from all 5 models discussed above.

TABLE I. SUMMARY OF CLASSIFICATION ACCURACY FOR MODEL FITTING FOR MODEL 1-MODEL 5

Model	**Overall Accuracy**	**Accuracy of the class property damages**	**Accuracy of the class slight injuries**	**Accuracy of the class serious & fatal injuries**
1	83.13%	93.76%	21.31%	19.05%
2	83.84%	94.39%	22.95%	19.05%
3	90.24%	98.76%	40.98%	38.10%
4	92.90%	98.82%	40.90%	38.46%
5	90.78%	98.35%	46.43%	39.13%

TABLE II. SUMMARY OF CLASSIFICATION ACCURACY FOR MODEL TESTING FOR MODEL 1-MODEL 5

Model	**Overall Accuracy**	**Accuracy of the class property damages**	**Accuracy of the class slight injuries**	**Accuracy of the class serious & fatal injuries**
1	80.85%	91.87%	11.11%	0.00%
2	82.26%	93.50%	11.11%	0.00%
3	70.0%	77.60%	16.67%	0.00%
4	60.0%	79.00%	17.39%	5.89%
5	65.71%	73.77%	18.18%	14.29%

By evaluating the results from the above tables, the discussion point was raised about the accuracy from slight injuries and serious and fatal injuries are been much lower than the accuracy from property damages. Hence, it can be identified that there is a problem with the dataset which has been used for this research. That problem is defined as the class imbalance problem.

According to the above dataset, the imbalanced ratio of classes of property damages: slight injuries: serious and fatal injuries is 86:10:4. The random under-sampling technique considered under the data processing approach was used to overcome the imbalance problem in this study.In Model 6, the Naïve Base classification algorithm was used to construct the model using the *Method 4* data splitting technique with significant independent variables.

In Model 7, the Probabilistic Neural Network (PNN) classification technique was used to construct the model using the *Method 4* data splitting technique with significant variables that were identified from the statistical tests. Following figure 5 show the optimum network architecture of the PNN network.

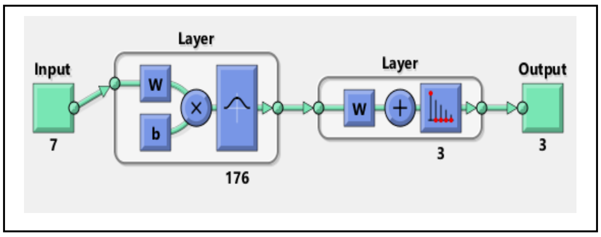

Fig 5: Network architecture of the Model 7 PNN

Seven input variables were used to build the PNN model with 0.1 spread. There were two hidden layers with 176 and 3 neurons respectively to classify the output category (Accident severity). The following tables are illustrating the classification accuracy of accident severity from Model 6 and Model 7.

TABLE III. SUMMARY OF CLASSIFICATION ACCURACY FOR MODEL FITTING FOR MODEL 6 AND MODEL 7

Model	**Overall Accuracy**	**Accuracy of the class property damages**	**Accuracy of the class slight injuries**	**Accuracy of the class serious & fatal injuries**
Naïve Bayes (Model 6)	60.23%	87.76%	27.78%	20.83%
PNN (Model 7)	99.43%	100%	100%	95.83%

TABLE IV. CONFUSION MATRIX FOR A TESTING DATASET FOR THE MODEL 6 (NAÏVE BAYES UNDER-SAMPLING)

Observed	**Predicted**			
	Property Damages Only	**Slight Injury**	**Serious & Fatal Injury**	**Correct Percentage**
Property Damages Only	99	18	4	81.82%
Slight Injury	11	2	0	15.38%
Serious & Fatal Injury	3	3	0	0.00%
Overall Percentage	80.71%	16.43%	2.86%	72.14%

TABLE V. CONFUSION MATRIX FOR TESTING DATASET FOR THE MODEL 7 (PNN UNDER-SAMPLING)

Observed	**Predicted**			
	Property Damages Only	**Slight Injury**	**Serious & Fatal Injury**	**Percent Correct**
Property Damages Only	99	18	4	81.82%
Slight Injury	9	4	0	30.77%
Serious & Fatal Injury	5	0	1	16.67%
Overall Percentage	80.71%	15.71%	3.57%	74.29%

In Naïve Base classification algorithm, 99 of the 121 (82%) ‘Property Damages Only’ are classified correctly. 2 of the 13 (15%) ‘Slight Injury’ are classified correctly and none of the 6 (0%) ‘Serious & Fatal Injury’ are classified correctly.

In Probabilistic Neural Network (PNN) 99 of the 121 (82%) ‘Property Damages Only’ are classified correctly. 4 of the 13 (31%) ‘Slight Injury’ are classified correctly and 1 of the 6 (17%) ‘Serious & Fatal Injury’ are classified correctly.

The Naïve Base classification algorithm predicts the severity of accidents with an accuracy of 72.14%, while the

Probabilistic Neural Network (PNN) classifies the severity of accidents with an accuracy of 74.29%.

Table 6 compares the performance of two techniques considered thought the study.

TABLE VI. CLASSIFICATION ACCURACY RELATED TO MODEL 6 & MODEL 7

Model	**Technique used**	**Overall classification accuracy**
6	Naïve Bayes	72.14%
7	PNN	74.29%

The classification accuracy of the final models (Naïve Bayes and PNN) which were created using the under-sampling data technique is greater than the accuracy derived from the original dataset.

By considering both models in Table 6, it can be said that the accuracy of both models is higher than 70%, which could be considered as approximately good models for classifying the accident severity of Colombo-Katunayake Expressway, Sri Lanka.

From the above two techniques, the overall accuracy of the Naïve Bayes classification model is similar to that of the probabilistic neural network classification model. By considering the accuracy of minority classes (Slight injuries, Serious and fatal injuries), the predictive accuracy of the Probabilistic Neural Network (PNN) classification model is better than the Naïve Bayes classification model.

Both models perform in good accuracy with the presence of significant variables. Hence, we can come up with a conclusion that the significant variables could be an effecting cause for being an accident of Colombo-Katunayake expressway. Further, it can be said that both the Naïve Bayes classification model and the PNN models are good for classifying accident severity of Colombo-Katunayake expressway while PNN exhibit slightly higher accuracy in classification when compared to Naïve Bayes classification model.

Some limitations of the study are as follows: Data collecting technique is more or less poor in Sri Lankan expressways, the study could have gone to a more realistic and accurate model if there are more variables to consider the accident severity; data collecting method of all the expressways of Sri Lanka are done in written format by human hand rather than computer systems. So it can take a lot of time converting data from written format to any computer format (E.g. Excel) for a person who is willing to do research on expressways of Sri Lanka and by using a multinomial regression model or any other statistical method, one can identify most influential factors for accident severity.

VII. CONCLUSION

This study aimed to identify the major causes of accidents on the Colombo-Katunayake Expressway in Sri Lanka and to classify the severity of accidents. After building the appropriate models to achieve that objective, the following conclusions were drawn.

- From the models considered in this study, both the Naïve Bayes classification algorithm and Probabilistic Neural Network (PNN) are good models for classifying the severity of accidents in Colombo-Katunayake Expressway with higher accuracy. Among them, PNN performs slightly better than the Naïve Bayes classifier.
- Considering the classification accuracies of the rarely occurred and more important 'serious & fatal injury' and 'slight injury' categories, the PNN model performed higher classification accuracy than the Naïve Base classification algorithm.
- The significant variables identified in this study: Time category, Driver's Age category, Vehicle type, Reason for the accident, Number of vehicles involved, Cause for accident and Rainfall can be considered as the most influential factors for met with an accident on the Colombo-Katunayake Expressway.
- Final Results obtained from this study can be used to implement safety improvements against the traffic accidents in expressways of Sri Lanka, Specifically in the Colombo-Katunayake Expressway.

REFERENCES

[1] S. D. Dharmaratne, A. U. Jayatilleke, and A. C. Jayatilleke, "Road traffic crashes, injury and fatality trends in Sri Lanka: 1938-2013," *Bulletin of the World Health Organization*, 2015, 93, 640-647.

[2] S. Renuraj, N. Varathan, and N. Satkunananthan, "Factors influencing traffic accidents in Jaffna," *Sri Lankan Journal of Applied Statistics*, 2015, 16(2).

[3] Annual reports of expressways accidents from 2011-2018 [According to Police Report of Highway Main Traffic Police Branch in Kaduwela, Sri Lanka]

[4] T. B. Tesema, A. Abraham, and C. Grosan, "Rule mining and classification of road traffic accidents using adaptive regression trees," *International Journal of Simulation*, 2005, 80-94.

[5] S. D. Dharmaratne, and S. N. Ameratunga, "Road traffic injuries in Sri Lanka: a call to action," *Journal of the College of Physicians and Surgeons--Pakistan: JCPSP*, 2004, 729-730.

[6] J. de Oña, R. O. Mujalli , and F. J. Calvo, "Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks," *Accident Analysis & Prevention*, 2011, 402-411.

[7] J. Maiti and A. Bhattacherjee, "Predicting accident susceptibility: a logistic regression analysis of underground coal mine workers," *Journal of the Southern African Institute of Mining and Metallurgy*, 2001, 203-208.

[8] M. M. Chong, A. Abraham, and M. Paprzycki, "Traffic accident analysis using decision trees and neural networks", 2004, arXiv preprint cs/0405050.

[9] Chong, M., A. Abraham, and M. Paprzycki, "Traffic accident data mining using machine learning paradigms," *Fourth International Conference on Intelligent Systems Design and Applications*, 2004, pp. 415-420.

[10] M. Chong, A. Abraham, and M. Paprzycki, "Traffic accident analysis using machine learning paradigms," *Informatica*, 2005.

[11] L. Y. Chang, "Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network," *Safety science*, 2005, 541-557.

[12] D. Delen, R. Sharda, and M. Bessonov, "Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks," *Accident Analysis & Prevention*, 2006, 434-444.

[13] M. M. Saritas and A. Yasar, "Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification," *International*

Journal of Intelligent Systems and Applications in Engineering, 2019, 88-91.

[14] G. Dimitoglou, J. A. Adams, and C. M. Jim, "Comparison of the C4. 5 and a Naïve Bayes classifier for the prediction of lung cancer survivability," 2012, arXiv preprint arXiv:1206.1121.

[15] G.A. Ali, and C.S. Bakhiet, "Comparative analysis and Prediction of Traffic accidents in sudan using artificial Neural Networks and Statistical methods," Sudan University of Science and technology.

[16] H.S. Stem, "Neural networks, in applied statistic (with discussions)," *Technometric,* 1995, 205-220

[17] S. Peltzman, "The effects of automobile safety regulation," *Journal of political Economy*, 1975, 83(4), 677-725.

[18] R. D. Romero, D. S. Touretzky, and R. H. Thibadeau, "Optical Chinese character recognition using probabilistic neural networks," *Pattern recognition*, 1997, 30(8), 1279-1292.

[19] R. Longadge and S. Dongre, "Class imbalance problem in data mining review", 2013, arXiv preprint arXiv:1305.1707.

[20] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, "Class imbalance problem," *Fourth international conference on natural computation*, 2008, pp. 192-201.

[21] M. H. Dunham, "Data mining: Introductory and advanced topics" Pearson Education India, 2006.

[22] J. COE, "Performance comparison of Naïve Bayes and J48 classification algorithms*,"* *International Journal of Applied Engineering Research*, 2012.