Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 170635, 8 pages
http://dx.doi.org/10.1155/2015/170635

Research Article

The Application of Data Mining Technology to Build a Forecasting Model for Classification of Road Traffic Accidents

Yau-Ren Shiau,[1] Ching-Hsing Tsai,[1] Yung-Hsiang Hung,[2] and Yu-Ting Kuo[2]

[1]*Department of Industrial Engineering and System Management, Feng Chia University, No. 100 Wenhua Road, Taichung 40724, Taiwan*
[2]*Department of Industrial Engineering and Management, National Chin-Yi University of Technology, No. 57, Section 2, Zhongshan Road, Taiping District, Taichung 41170, Taiwan*

Correspondence should be addressed to Yung-Hsiang Hung; hys502@ncut.edu.tw

Received 24 March 2015; Revised 27 June 2015; Accepted 28 June 2015

Academic Editor: Aime' Lay-Ekuakille

With the ever-increasing number of vehicles on the road, traffic accidents have also increased, resulting in the loss of lives and properties, as well as immeasurable social costs. The environment, time, and region influence the occurrence of traffic accidents. The life and property loss is expected to be reduced by improving traffic engineering, education, and administration of law and advocacy. This study observed 2,471 traffic accidents which occurred in central Taiwan from January to December 2011 and used the Recursive Feature Elimination (RFE) of Feature Selection to screen the important factors affecting traffic accidents. It then established models to analyze traffic accidents with various methods, such as Fuzzy Robust Principal Component Analysis (FRPCA), Backpropagation Neural Network (BPNN), and Logistic Regression (LR). The proposed model aims to probe into the environments of traffic accidents, as well as the relationships between the variables of road designs, rule-violation items, and accident types. The results showed that the accuracy rate of classifiers FRPCA-BPNN (85.89%) and FRPCA-LR (85.14%) combined with FRPCA is higher than that of BPNN (84.37%) and LR (85.06%) by 1.52% and 0.08%, respectively. Moreover, the performance of FRPCA-BPNN and FRPCA-LR combined with FRPCA in classification prediction is better than that of BPNN and LR.

1. Introduction

As the demand for vehicles rises, the number of vehicles on the road increases greatly and traffic jams worsen, especially during rush hours; thus, traffic accidents are more likely to occur. Faced with more severe accidents, the traffic problem has become a topic of concern in Taiwan. The statistics of the Ministry of Health and Welfare (2013) indicated that accidental injury is the sixth major cause of death in Taiwan, with 6,873 deaths from accidental injuries. Most traffic accidents are caused by improper driving behaviors, and one of the major reasons is that drivers failed to pay attention to the road ahead (Ministry of Transportation, 2008).

According to the statistical data of the National Police Agency (2012), the number of road traffic accidents with death in Taichung City was next to Kaohsiung City. In 2012, the number of traffic accidents causing death was 208, and the death toll was 210. In 2012, the number of traffic accidents causing death was 198, the death toll was 203, and the gradient of number of accidents was −3.41%. Due to the increase in urban population, at a growth rate of 0.76% in 2012, the occurrence rate of road traffic accidents increased accordingly. In 2011, accidental injury ranked sixth among the ten major causes of death in Taiwan, and the death toll from motor vehicle accidents was about 30, accounting for 17.3% of the death rate per 100,000 persons (MOHW, 2012). This study uses the traffic accident data from the NPA of the region from January to December 2012 as the data source. The data content includes 17 items, such as weather, light rays, road category, speed limit, road type, accident site, road conditions, and roadblocks. There are 2,471 original observations.

According to previous transportation research [1–5], the causes of traffic accidents are mostly human factors, such as speeding, violation of signals, and drunk driving,

as well as the interaction between road environments and traffic engineering facilities. This study identifies the key factors that affect traffic accidents using Feature Selection and establishes models to analyze traffic accidents and their types with various methods, such as Fuzzy Robust Principal Component Analysis (FRPCA), Back Propagation Neural Network (BPNN), and FRPCA-Logistic Regression (LR). The environments of traffic accidents and the variables of road designs identified by the model could serve as reference for the police force and regulatory authorities to design and plans and improve traffic safety, thus decreasing the ratio of traffic accidents, damage to property, and loss of lives.

With the advancements of information technology, data mining becomes increasingly mature, and useful information without preconditions can be found in databases. Relational models can be built to determine the correlation between characterization factors of traffic accidents and casualties. This study uses the Recursive Feature Elimination (RFE), FRPCA, BPNN, and LR of Feature Selection to determine important factors influencing traffic accidents. The results can provide suggestions for improving the occurrence of traffic accidents. Finally, the model was statistically evaluated.

The remainder of this paper is organized as follows. Section 2 reviews the literature concerning the severity of injuries in traffic accidents; Section 3 presents the FRPCA; Section 4 discusses the research data; Section 5 offers conclusions.

2. Material and Methods

Many studies have focused on forecasting and modeling traffic accidents and analyzed the results. The results suggest that the significant factors influencing the occurrence of accidents must be eliminated or controlled in order to prevent traffic accidents and reduce injuries and deaths.

In terms of research methods, most studies use BPNN or LR to forecast or model analysis results [6–9]. Gang and Zhuping [10] suggested that the PSO-SVM is better than BPNN in traffic safety forecasting. Chang et al. [6] used the established modeling method and LR to discuss the contributing factors and conditions of driving after drinking. The analysis results showed that law enforcement, drivers' drinking habits, and regulatory knowledge of drunk driving apparently influenced drivers' selecting drunk driving behaviors. Kong and Yang [8] used LR for casualties and driving speed in traffic accident survey data and found that, regarding the correlation of collisions between vehicles and pedestrians, the risk of pedestrian death was 26% when the vehicle's speed was 50 km/h, 50% when the speed was 58 km/h, and 82% when the speed was 70 km/h. However, the analysis result showed that age was not a major risk factor in death. Fu and Zhou [7] pointed out that the traditional BPNN has some defects, such as local minima, too many iterations, and too slow training. Therefore, the improved LM-BP neural network was used for forecasting. The forecast results of traffic accidents, death toll, and amount of direct economic loss were significant; thus, the BP network is applicable to traffic accident forecasting.

A number of recent studies have used data mining or statistical methods [11–15]. Karacasua and Er [14] used chi-square significance testing to analyze whether the same age and gender have similar traffic accidents, as well as the correlation among education, age, gender, and psychology. The findings showed that (1) males were more prone to traffic accidents than females; (2) driving while being intoxicated and speeding were major causes. Kanchan et al. [13] used statistical software for analysis and found that the injured were mostly male, and the major causes of death were head and abdominal injuries. Traffic accidents are a significant public health hazard; thus, first aid should be strengthened, and traffic regulations and health education should be strictly implemented. Kashani et al. [16] used classification and CART to analyze traffic collision data. The results showed that improper passing and not using seat belts were the most important factors influencing the severity of injuries. De Oña et al. [15] used Latent Class Cluster (LCC) to reduce the heterogeneity of traffic accident data and combined it with Bayesian Networks (BNS) to recognize major factors. The results indicated that weather factors, pavement markings, and road width were significant factors.

Based on the above discussion, this study uses BPNN, LR, and statistical methods differing from previous studies, which aim at accident patterns and types. The FRPCA is used for data preprocessing, which is combined with BPNN and LR models to analyze the performance of the aforesaid four classification models (BPNN, LR, FRPCA-BPNN, and FRPCA-LR) in forecasting.

2.1. Research Method. This study uses four constructs, namely, (1) natural factors; (2) environmental factors; (3) road design; and (4) accident types and patterns of road traffic accident cases in Taichung City, to discuss the factors influencing the occurrence of road traffic accidents. The research structure is as shown in Figure 1.

2.2. Data Preprocessing of Feature Selection. This study uses RFE as the Feature Selection method, which is a Feature Selection algorithm, with the principle proposed by Guyon et al. [17]. Guyon used RFE to select the key and important feature set, which not only shortens classification computing time but also improves the classification accuracy rate. The purpose of RFE is to calculate the weight vectors of each feature, which are ordered according to the calculated weight vectors as the basis of classification. RFE is an iterative process that eliminates features backwards, and its feature set screening procedure is described as follows:

(1) Use current data set to train classifier.

(2) Calculate the weight of each feature.

(3) Delete the feature with minimum weight.

The iterative process is ended when there is one feature remaining. A list of features ordered according to the weights is obtained as a result of execution, and unimportant or uncorrelated features are eliminated from the list first; thus, they are listed at the end, whereas, the most important features are eliminated last and are listed at the front [18, 19].

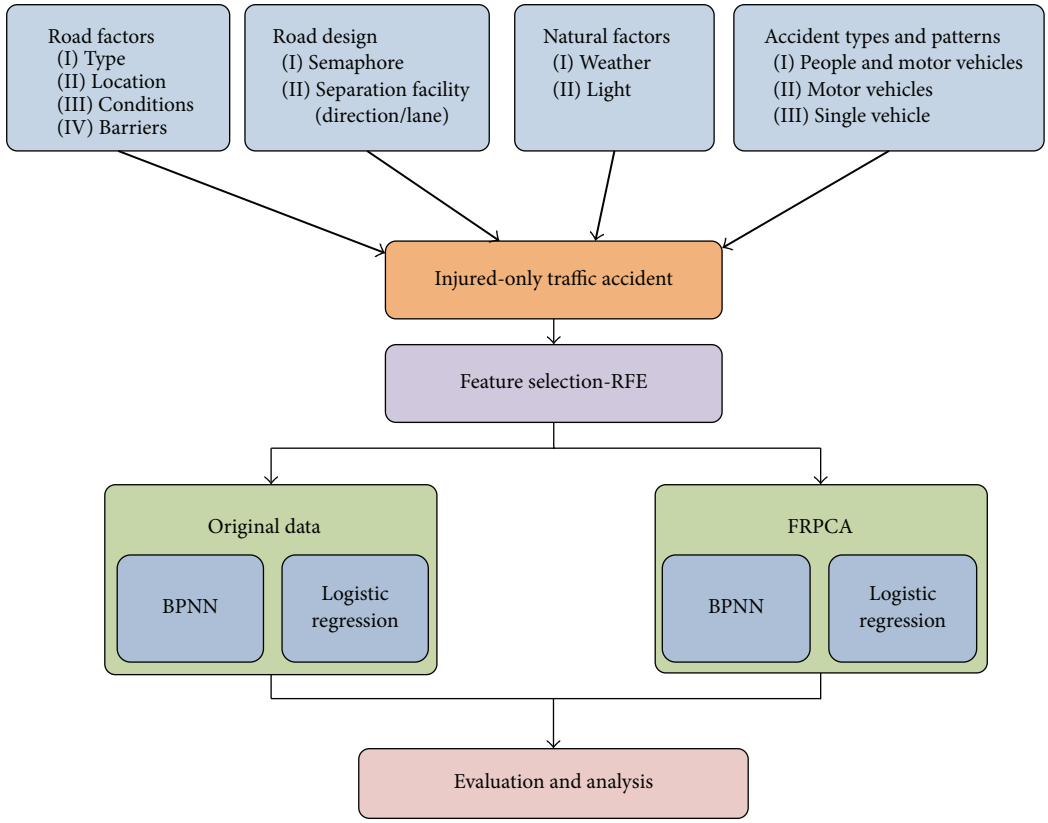

Figure 1: Research structure.

The RFE selects the feature set in three major steps, imports the data set for classification, calculates the weight of each feature, and deletes the feature with minimum weight. Feature ordering is obtained, the feature with minimum weight square is removed in each cycle, and then the remaining features are retrained to obtain a new feature ordering. RFE continuously executes this process, and a feature order list is obtained [20]. It is noteworthy that one of the features ordered in the front does not always enable the classifier to obtain the best classification performance; however, the combination of multiple features enables the classifier to obtain the optimum classification performance. Therefore, RFE algorithm can select the most complementary feature combination [4].

2.3. Backpropagation Neural Network (BPNN). BPNN is the most frequently used supervised learning among the neural networks and is highly effective in classification problems [21]. The parameters are divided into structural parameters and learning parameters. The structural parameters include the number of hidden layers, while the learning parameters include the learning rate, initial weight range, and momentum term. Generally, Trial and Error is used to determine the optimal parameter values when selecting structural parameters and learning parameters. The most used nonlinear transfer function in the hidden layer of BPNN is the log-sigmoid transfer function, whose output is between 0 and 1, in order to respond to the negative infinity to positive infinity input of neurons.

An alternative is the tangent sigmoid transfer function "tansig," as shown in the hidden layer. The linear transfer function purelin is in the output layer. If the sigmoid transfer function is used in the output layer, the network output is restricted to a very small range. If the linear transfer function is used in the output layer, the network output can be an arbitrary value.

2.4. LR. LR can be used to analyze one or several forecast values. These results have a binary (e.g., existence or nonexistence of an event) relationship [22, 23]. LR is derived from the cumulative probability function of the logistic model and is a linear probability model, which is similar to a linear regression model. The difference is that LR can test the dependent variable of a nominal scale, where the discussed dependent variable is discontinuous, especially in binary classification. The purpose of LR is to establish the simplest and fittest analysis result. Furthermore, it can be used in a practical model to forecast the relationships between dependent variables and a set of forecast variances, where the explanatory variable can be a categorical or continuous variable.

2.5. FRPCA. The nonlinear FRPCA algorithm is deduced from the linear fuzzy principal component analysis algorithm, as introduced by Yang and Wang [24], and the

nonlinear criteria in blind source separation of Karhunen et al. [25]. The robust principal component analysis, as proposed by Yang and Wang, is established on the principal component analysis learning rule and energy function, as proposed by Xu and Yuilles [26], and the objective function bias is proposed. These methods are briefly introduced as follows. Xu and Yuilles [26] proposed the optimal function of constraint $u_i \in \{0, 1\}$:

$$E(U, w) = \sum_{i=1}^{n} u_i e(x_i) + \eta \sum_{i=1}^{n} (1 - u_i). \quad (1)$$

The objective is to minimize $E(U, w)$, where $X = \{x_1, x_2, \ldots, x_n\}$ is the data set, $U = \{u_i \mid i = 1, \ldots, n\}$ is the membership set, η is the threshold, u_i is the binary variable, and w is the continuous variable, rendering gradient descent method optimization difficult to solve; thus, they transformed the minimization problem, where Gibbs distribution is maximized by the following equation:

$$P(U, w) = \frac{\exp(-\gamma E(U, w))}{Z}, \quad (2)$$

where Z is the separation function; ensuring $\sum U \int wP(U, w) = 1$, $e(x_i)$ can be one of the following functions:

$$e_1(x_i) = \left\| x_i - w^T x_i w \right\|^2,$$
$$e_2(x_i) = \|x_i\|^2 - \frac{\left\| w^T x_i \right\|^2}{\|w\|^2} = x_i^T x_i - \frac{w^T x_i x_i^T w}{w^T w}. \quad (3)$$

The gradient descent rule for minimizing $E_1 = \sum_{i=1}^{n} e_1(x_i)$ and $E_2 = \sum_{i=1}^{n} e_2(x_i)$ is

$$w^{\text{new}} = w^{\text{old}} + \alpha_t [y(x_i - u) + (y - v) x_i],$$
$$w^{\text{new}} = w^{\text{old}} + \alpha_t \left[x_i y - \frac{w}{w^T w} y^2 \right], \quad (4)$$

where α_t is the learning rate, $y = w^T x_i$.

Therefore, Yang and Wang [24] proposed a new objective function:

$$\text{Min } E = \sum_{i=1}^{n} u_i^{m_1} e(x_i) + \eta \sum_{i=1}^{n} (1 - u_i)^{m_1}. \quad (5)$$

The constraints are $u_i \in [0, 1]$, $m_1 \in [1, \infty)$, where u_i is the membership value belonging to the x_i data cluster, $(1 - u_i)$ is the membership value of the x_i disturbance cluster, and m_1 is the fuzzy variable. In this case, $e(x_i)$ is the error between the measured x_i and the cluster center, which is similar to the C-means algorithm [27].

As u_i is a continuous variable, it avoids the difficulty of an optimum mix of discrete types and continuous types; thus, the gradient descent method can be used. First, u_i equals 0 as calculated by the slope of (2), so

$$u_i = \frac{1}{1 + [e(x_i)/\eta]^{1/(m_1 - 1)}}. \quad (6)$$

u_i is replaced in (2), and the following equation is obtained:

$$E = \sum_{i=1}^{n} \left[\frac{1}{1 + [e(x_i)/\eta]^{1/(m_1 - 1)}} \right]^{m_1 - 1} e(x_i). \quad (7)$$

On the other hand, w gradient is

$$\frac{\partial E}{\partial w} = \left[\frac{1}{1 + [e(x_i)/\eta]^{1/(m_1 - 1)}} \right]^{m_1} \left[\frac{\partial e(x_i)}{\partial w} \right]. \quad (8)$$

m_1 is the fuzzy variable. If $m_1 = 1$, the fuzzy membership is demoted to a fixed membership and can be determined by the following rule:

$$u_i = \begin{cases} 1 & \text{if } (e(x_i)) < \eta, \\ 0 & \text{otherwise.} \end{cases} \quad (9)$$

In this case, η is the hard threshold, where m_1 is not set, but $m_1 = 2$ in most studies. Yang and Wang [24] deduced the following process of an optimization algorithm.

(1) The number of iterations is set as $t = 1$, the iteration is constrained as T, the learning coefficient is $\alpha_0 \in (0, 1]$, the soft threshold η is a small positive, and the weight w is randomly initialized.

(2) In a case of less than T, execute step (3) to step (9).

(3) Calculate $\alpha_t = \alpha_0(1 - t/T)$; set $i = 1$ and $\sigma = 0$.

(4) For observation frequency, execute step (5) to step (8).

(5) Calculate $y = w^T x_i$, $u = yw$ and $v = w^T u$.

(6) The new weight is $w^{\text{new}} = w^{\text{old}} + \alpha_t \beta(x_i)[y(x_i - u) + (y - v)x_i]$.

(7) The new temporary count is $\delta = \delta + e_1(x_i)$.

(8) Add from 1 to i.

(9) Calculate $\eta = (\delta/n)$ and add from 1 to t.

It is almost affirmative that the new weight w approaches the principal component vector [19, 27, 28].

3. Case Study

This section is divided into three parts: collection of traffic accident data, preprocessing of the research data, and substituting the data after Feature Selection in FRPCA, BPNN, and LR. Four groups of information are obtained, including BPNN, LR, FRPCA-BPNN, and FRPCA-LR.

3.1. Data Collection. The data variables used in this study are 17 input variables, including weather, light rays, road pattern, accident site, sight distance, and separation facility; there are 2,471 observations. The output variables are accident types and patterns, including (1) vehicle-vehicle accidents; (2) person-automobile/motorcycle; and (3) automobile/motorcycle data variables. There are 2,096 observations of Category (1) vehicle-vehicle; there are 146 observations of Category (2) person-automobile/motorcycle; and there are 229 observations of Category (3) automobile/motorcycle. The aforesaid variables are coded as Y_1, Y_2, and Y_3, respectively, as listed in Table 1.

Table 1: Number of traffic accidents and output variable codes in the region in 2012.

Output variable	Number of data	Percentage	Code
People and motor vehicles	146	5.9%	Y_1
Motor vehicles	2,096	84.82%	Y_2
Single vehicle	229	9.3%	Y_3
Total	2,471	100%	

3.2. Feature Selection Result. This study uses RFE for data preprocessing. The 17 variables of the database are coded before Feature Selection, as shown in Table 2. The 17 variables are sequenced according to importance, as shown in Table 3.

4. Empirical Research Result

This study uses the first 7 variables in order of importance, as obtained by the Feature Selection of RFE, as input variables, while the person-automobile/motorcycle, vehicle-vehicle, and automobile/motorcycle are output variables, as shown in Table 4. The 7 input variables and 3 output variables are substituted in FRPCA, BPNN, and LR, respectively, in order to obtain BPNN, LR, FRPCA-BPNN, and FRPCA-LR classifier models. The experimental procedure is as follows: Step (1): 2,471 data sets are used as test data, the 17 variables are sequenced according to their importance by using the RFE data preprocessing method, and the first 7 variables are rearranged as the test data set (S_1). Step (2): classifier modeling and performance evaluation, the BPNN, LR, FRPCA-BPNN, and FRPCA-LR classifiers are tested, respectively, in order to evaluate the performance of the classifiers during classification.

4.1. Analysis of BPNN Model. The 7 input variables are substituted in FRPCA in order to obtain the scores and loadings of the principal components. The scores of the principal components can be used to classify various observation points and to integrate the scores of the principal components of each observation point in order to calculate an average weighted integral indicator. The coefficient of the correlation between new variables and old variables is called loadings, which represent the influence or importance of the original variable to the new variables, where larger loadings represent higher influence.

The loadings can be obtained from

$$l_{ij} = \frac{w_{ij}}{\hat{s}_j} \cdot \sqrt{\lambda_i}, \tag{10}$$

where l_{ij} is the loadings of No. j variable on No. i principal component, w_{ij} is the weight of No. j variable on No. i principal component, λ_i is the eigenvalue of No. i principal component (i.e., variance), and $\hat{s}_j$ is the standard deviation of No. j variable.

Experimental combination 1: the experimental parameters of the BPNN forecasting model are set as follows: epochs = 1000, learning rate lr is 0.1, 0.3, and 0.5, respectively, and each experiment is repeated 5 times, with the results of the three experiments as shown in Table 5. Experimental combination 2: the FRPAC-BPNN forecasting model is different from experimental combination 1, where the principal component scores are converted by executing FRPCA before building BPNN, and then the BPNN forecasting network is built. The experimental results show the learning rate of the lr = empirical results of BPNN versus FRPCA-BPNN forecasting models.

Table 2: Codes of 17 variables.

Variable code	Variable name
X_1	Weather
X_2	Light rays
X_3	Road type
X_4	Speed limit
X_5	Road pattern
X_6	Accident site
X_7	Pavement
X_8	Pavement state
X_9	Pavement defect
X_{10}	Barrier
X_{11}	Sight distance
X_{12}	Signal type
X_{13}	Signal action
X_{14}	Separation facility
X_{15}	In fast or general lane
X_{16}	In fast and slow lanes
X_{17}	Edge of pavement

4.2. LR Analysis. Experimental combination 3: the LR classifier is constructed, and the data are imported into LR for classification forecasting according to the aforesaid test data set S_1. Experimental combination 4: the FRPCA-LR classifier is constructed as experimental combination 2, the S_1 data set is converted into principal component scores by FRPCA, and then the LR classifier is constructed. Each experiment is conducted five times; the experimental results are as shown in Table 6. The experimental results show that the average accuracy rate and standard deviation of the FRPCA-LR model are 0.8506 ± 0.0021, which is better than the 0.8514 ± 0.0031 of the LR classification model.

The LR model investigates the impacts on the pattern of traffic accidents according to the types and patterns of the traffic accidents. The optimal model is shown in Table 7. This section discusses the correlation between the vehicles and the environment, based on 2,325 pieces of data for analysis. After the deletion of 146 pieces of data involving human and vehicles, the dependent variables are divided into the two categories of "vehicle to vehicle" and "vehicle in itself" according to the types and patterns of traffic accidents. Odd ratio is adopted to represent the relevant influences of Event A to the occurrence of Event B. In Table 7, the odd ratio of crossroads among the road types is 3.01, meaning that the risk of traffic accidents on "crossroads" is higher than

Table 3: Sequence of feature attributes of environmental factors.

Rank	Code	Variable name	Importance
1	X_{16}	In fast and slow lanes	0.069791
2	X_{15}	In fast or general lane	0.045356
3	X_{17}	Edge of pavement	0.042181
4	X_5	Road pattern	0.041695
5	X_6	Accident site	0.037548
6	X_{14}	Separation facility	0.025303
7	X_{11}	Sight distance	0.018811
8	X_2	Light rays	0.017205
9	X_3	Road type	0.017063
10	X_1	Weather	0.015428
11	X_8	Pavement state	0.015322
12	X_4	Speed limit	0.012274
13	X_{13}	Signal action	0.009509
14	X_{10}	Barrier	0.008938
15	X_{12}	Signal type	0.002745
16	X_9	Pavement defect	0.002059
17	X_7	Pavement	0.000180

Table 4: Input (x) and output (y) parameters.

Code (x)	Input variable	Code (y)	Output variable
X_5	Road pattern	Y_1	People and motor vehicles
X_6	Accident site	Y_2	Motor vehicles
X_{11}	Sight distance	Y_3	Single vehicle
X_{14}	Separation facility		
X_{15}	In fast or general lane		
X_{16}	In fast and slow lanes		
X_{17}	Edge of pavement		

Table 5: Empirical results of BPNN versus FRPCA-BPNN.

	BPNN			FRPCA-BPNN		
Learning rate	0.1	0.3	0.5	0.1	0.3	0.5
BPNN-1	0.8232	0.8067	0.8089	0.8607	0.8594	0.8583
BPNN-2	0.8523	0.8412	0.8335	0.8615	0.8572	0.8591
BPNN-3	0.8508	0.8543	0.8560	0.8557	0.8623	0.8599
BPNN-4	0.8550	0.8530	0.8530	0.8563	0.8553	0.8572
BPNN-5	0.8574	0.8567	0.8542	0.8584	0.8607	0.8609
Mean	0.8477	0.8424	0.8411	0.8585	0.8590	0.8591
S.D	0.0139	0.0208	0.2019	0.0026	0.0028	0.0010

the 1.38 of "one-way roads", $e^{3.01-1.38} = e^{1.23} = 3.42$. In other words, the ratio of traffic accidents at intersections is 3.24 times that of one-way roads. Likewise, the ratio of traffic accidents on "intersections" under the category of "traffic locations" is also the highest, at 7.24 times that of general roads; $e^{5.72-3.74} = e^{1.98}$. Table 7 shows the importance of traffic accident environments and road design to the ratio of traffic accidents.

Table 6: Empirical results of LR versus FRPCA-LR.

	Logistic	FRPCA-Logistic
Logistic-1	0.8480	0.8503
Logistic-2	0.8523	0.8528
Logistic-3	0.8491	0.8497
Logistic-4	0.8483	0.8497
Logistic-5	0.8551	0.8543
Mean	0.8506	0.8514
S.D	0.0031	0.0021

5. Conclusions

Traffic safety depends on road design, road configuration, vehicle performance, traffic regulations, and the effectiveness of implementation. The main means of transport in middle and low income countries include walking, bicycle, motorcycle, and bus, while that of high income countries is automobiles. Therefore, the traffic safety control measures of high income countries are not completely applicable to middle and low income countries and thus should be imported and improved to fit local transportation and road usage conditions [29].

The report of the World Health Organization (WHO) indicated that about 1.2 million people die from traffic accidents in the world annually; about 3,400 people die from traffic accidents per day; approximately 1,000 people are injured or disabled; children, pedestrians, cyclers, and the elderly are the most vulnerable road users; and 85% of fatalities and 90% of the disabled live in middle and low income countries. The scientific analysis of accident data, as well as the implementation of relevant safety measures, can prevent the occurrence of traffic accidents, thus, reducing the severity of injuries.

At present, with the rapid development of cities, it is necessary to make efficient forecasting in order that decision makers can make preventions and decisions in advance in order to reduce the death rate. This study uses RFE, FRPCA, BPNN, and LR to analyze the classification accuracy rate of the traffic accident data of the region. According to the experimental results, the classification accuracy rate of BPNN, LR, FRPCA-BPNN, and FRPCA-LR is higher than 80%; thus, forecast performance is significant. Further analysis shows that the network performance of the FRPCA-BPNN and FRPCA-LR classifiers, combined with FRPCA, is better than BPNN and LR. According to Tables 5 and 6, the accuracy rate of classifiers FRPCA-BPNN (85.89%) and FRPCA-LR (85.14%), combined with FRPCA, is higher than BPNN (84.37%) and LR (85.06%) by 1.52% and 0.08%, respectively, meaning the FRPCA-BPNN and FRPCA-LR have better classification forecast ability.

In traffic accident analysis or verification results, the human factor is mostly regarded as the first cause of traffic accidents. However, the road environment has certain

Table 7: The impacts on the types and patterns of traffic accidents, as imposed by the various factors and represented in the LR model.

Variables	Saliency	Odd ratio
Constant terms	0.067	0.540
Relevant environmental factors		
Road types		
One-way road	0.044	1.38
Crossroad	0.028	3.01
Location of the accident		
Intersection	0.012	**5.72**
Ordinary road sections	0.032	**3.74**
Visibility range		
Satisfactory	—	—
Poor	0.024	1.59
Relevant road design		
Directional facilities		
Two-way no-passing markings	0.084	1.67
Directional markings	0.074	1.45
No directional facilities	0.040	2.26
Traffic separation facilities		
Fast lanes and slow lanes		
No lane-change markings (without signs)	0.17	1.19
Lane markings (with signs)	0.024	1.89
Lane markings (without signs)	0.049	3.02
No lane markings	0.005	5.71
Separation lines		
Separations lines between the fast lanes and slow lanes	0.142	1.91
No separations lines between the fast lanes and slow lanes	0.028	4.25
Sidelines		
Yes	—	—
No	0.032	3.99
Total sample size	2,471	

correlation, and improper intersection design or planning is likely to cause traffic accidents. In comparison to other traffic accident sites, a forked road intersection is the most probable accident site. This study used RFE to select 7 input variables from 17 input variables. Based on the 7 input variables, the environmental factor and road design are found to be the causes of road traffic accidents in the region. According to the statistical data of the Taichung Police Station, the 4 main causes among the 67 causes of accidents are as follows: (1) not allowing other vehicles to pass as per regulations, (2) not aware of the situation ahead, (3) violating specific sign (line) bans, and (4) not maintaining a safe driving distance, accounting for 23.72%, 13.54%, 7.51%, and 8.22%, respectively, of the total number of traffic accidents, and the total proportion is as high as 52.99%. The road authorities may refer to the 7 variables of traffic accidents and road design, as proposed in this study, regarding future road designs and plans. As for the 4 main causes of accidents on road sections involving the above seven variables of traffic accidents and road design, they should be the priorities in the future elimination actions of the police force. If improvements and preventive measures are made, road safety can be substantially increased, thereby reducing traffic accidents and fatalities. The findings can serve as reference for the police force and management authorities to improve roads, as well as the assessment and management models for the elimination of traffic offences.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] D. D. Clarke, P. Ward, C. Bartle, and W. Truman, "Killer crashes: fatal road traffic accidents in the UK," *Accident Analysis & Prevention*, vol. 42, no. 2, pp. 764–770, 2010.

[2] K.-V. Hung and L.-T. Huyen, "Education influence in traffic safety: a case study in Vietnam," *IATSS Research*, vol. 34, no. 2, pp. 87–93, 2011.

[3] H. Gjerde, A. S. Christophersen, P. T. Normann, and J. Mørland, "Toxicological investigations of drivers killed in road traffic

accidents in Norway during 2006–2008," *Forensic Science International*, vol. 212, no. 1–3, pp. 102–109, 2011.

[4] Y. Komada, S. Asaoka, T. Abe, and Y. Inoue, "Short sleep duration, sleep disorders, and traffic accidents," *IATSS Research*, vol. 37, no. 1, pp. 1–7, 2013.

[5] S. A. Shappell and D. A. Wiegmann, "Human factors investigation and analysis of accidents and incidents," in *Encyclopedia of Forensic Sciences*, pp. 440–449, Academic Press, 2nd edition, 2013.

[6] L.-Y. Chang, D.-J. Lin, C.-H. Huang, and K.-K. Chang, "Analysis of contributory factors for driving under the influence of alcohol: a stated choice approach," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 18, pp. 11–20, 2013.

[7] H. Fu and Y. Zhou, "The traffic accident prediction based on neural network," in *Proceedings of the 2nd International Conference on Digital Manufacturing and Automation (ICDMA '11)*, pp. 1349–1350, IEEE, Zhangjiajie, Hunan, August 2011.

[8] C.-Y. Kong and J.-K. Yang, "Logistic regression analysis of pedestrian casualty risk in passenger vehicle collisions in China," *Accident Analysis and Prevention*, vol. 42, no. 4, pp. 987–993, 2010.

[9] Y.-K. Ou, Y.-C. Liu, and F.-Y. Shih, "Risk prediction model for drivers' in-vehicle activities—application of task analysis and back-propagation neural network," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 18, pp. 83–93, 2013.

[10] R. Gang and Z. Zhuping, "Traffic safety forecasting method by particle swarm optimization and support vector machine," *Expert Systems with Applications*, vol. 38, no. 8, pp. 10420–10424, 2011.

[11] S. S. Durduran, "A decision making system to automatic recognize of traffic accidents on the basis of a GIS platform," *Expert Systems with Applications*, vol. 37, no. 12, pp. 7729–7736, 2010.

[12] K. El-Basyouny and T. Sayed, "Safety performance functions using traffic conflicts," *Safety Science*, vol. 51, no. 1, pp. 160–164, 2013.

[13] T. Kanchan, V. Kulkarni, S. M. Bakkannavar, N. Kumar, and B. Unnikrishnan, "Analysis of fatal road traffic accidents in a coastal township of South India," *Journal of Forensic and Legal Medicine*, vol. 19, no. 8, pp. 448–451, 2012.

[14] M. Karacasua and A. Er, "An analysis on distribution of traffic faults in accidents, based on driver's age and gender: Eskisehir case," *Procedia–Social and Behavioral Sciences*, vol. 20, pp. 776–785, 2011.

[15] J. De Oña, G. López, R. Mujalli, and F. J. Calvo, "Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks," *Accident Analysis & Prevention*, vol. 51, pp. 1–10, 2013.

[16] A. T. Kashani, S. M. Afshin, and A. Ranjbari, "Analysis of factors associated with traffic injury severity on rural roads in Iran," *Journal of Injury and Violence Research*, vol. 4, no. 1, pp. 36–41, 2012.

[17] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1–3, pp. 389–422, 2002.

[18] C. Chu, A.-L. Hsu, K.-H. Chou, P. Bandettini, C. Lin, and Alzheimer's Disease Neuroimaging Initiative, "Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images," *NeuroImage*, vol. 60, no. 1, pp. 59–70, 2012.

[19] X.-H. Lin, F.-F. Yang, L. Zhou et al., "A support vector machine-recursive feature elimination feature selection method based on artificial contrast variables and mutual information," *Journal of Chromatography B*, vol. 910, pp. 149–155, 2012.

[20] D. Wei, S. Li, and M.-K. Tan, "Graph embedding based feature selection," *Neurocomputing*, vol. 93, pp. 115–125, 2012.

[21] M. Ahmadzadeh, A. H. Fard, B. Saranjam, and H. R. Salimi, "Prediction of residual stresses in gas arc welding by back propagation neural network," *NDT & E International*, vol. 52, pp. 136–143, 2012.

[22] P. Reed and Y.-Q. Wu, "Logistic regression for risk factor modelling in stuttering research," *Journal of Fluency Disorders*, vol. 38, no. 2, pp. 88–101, 2013.

[23] P. Reed, "The effect of traffic tickets on road traffic crashes," *Accident Analysis & Prevention*, vol. 64, pp. 86–91, 2014.

[24] T.-N. Yang and S.-D. Wang, "Robust algorithms for principal component analysis," *Pattern Recognition Letters*, vol. 20, no. 9, pp. 927–933, 1999.

[25] J. Karhunen, P. Pajunen, and E. Oja, "The nonlinear PCA criterion in blind source separation: relations with other approaches," *Neurocomputing*, vol. 22, no. 1–3, pp. 5–20, 1998.

[26] L. Xu and A. L. Yuille, "Robust principal component analysis by self-organizing rules based on statistical physics approach," *IEEE Transactions on Neural Networks*, vol. 6, no. 1, pp. 131–143, 1995.

[27] F. Gharibnezhad, L. E. Mujica, J. Rodellar, and C.-P. Fritzen, "Damage detection using robust fuzzy principal component analysis," UPCommons Conference Report, 2013.

[28] P. Luukka, "Nonlinear fuzzy robust PCA algorithms and similarity classifier in bankruptcy analysis," *Expert Systems with Applications*, vol. 37, no. 12, pp. 8296–8302, 2010.

[29] G.-G. Zhang, K. K. W. Yau, and G. Chen, "Risk factors associated with traffic violations and accident severity in China," *Accident Analysis & Prevention*, vol. 59, pp. 18–25, 2013.