



Exploring Key Determinants of Traffic Accidents Fatalities in Thailand: A Hybrid Approach Machine Learning and Statistics

Nanon Sonnatthanon^{ID}, and Kasem Choocharukul*^{ID}

Department of Civil Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand

*kasem.choo@chula.ac.th

Abstract. Road traffic incidents are the 8th leading cause of mortality globally, with 90% of fatalities occurring in low- and middle-income countries. Moreover, these severe accidents are often attributed to the negligence or behavior of vulnerable road users. In Thailand, the mortality rate from road traffic accidents stand at 25 per 100,000 individuals, the highest among ASEAN nations. This study aims to investigate the various risk factors contributing to mortality from road traffic accidents, which have been collected since 2011 through the Highway Accident Information Management System (HAIMS), using statistical methodologies such as Cramer's V, Information Values (IV), which are employed to identify the most significant causes of mortality. In addition, the mortality had been investigated by deep learning model based on convolution and neural networks algorithms before being explored the significant factors influencing fatality from resulting traffic accidents from deep learning model by numerical techniques for quantifier the likelihood of associated with each significant factor. Three primary factors have been identified by Cramer's V, Information Value, and numerical analysis: using safety equipment of people, type of road user, and lighting conditions. These findings underscore the importance of targeted interventions, including stricter enforcement of safety equipment use, good road or infrastructure design for preventing accidents with different road users, and improving lighting infrastructure in some sections of a road, which could significantly reduce the road traffic mortality rate. While numerous factors contribute to road traffic fatalities in Thailand, these three factors demonstrate the strongest association with fatal outcomes.

Keywords: Road Safety, Virtual Geometry Group Model, Fatality.

1 Introduction

1.1 Problems statement

Globally, traffic accidents stands as the eighth top cause of mortality, though 90% of deaths take place within low- and middle-income countries. The rapid extension of

road transportation, in particular, has intensified safety concerns. Besides leading to major infrastructural damage and traffic backups, these collisions have also driven up the incidence of fatalities and severe harm among road users [1]. The consequences of these incidents involve economic and social detriment, both directly and indirectly [2]. Fatalities from road accidents impose a significant economic cost. Data reviewed across 31 European countries suggest that these traffic accidents are responsible for diverting 0.4% to 4.1% of the Gross Domestic Product (GDP), factoring in social and medical costs. These costs encompass medical expenses, loss of productivity, human loss, property damage, administrative costs, and other related expenses [3]. Approximately 70% of severe accidents are attributed to the negligence or risky behavior of vulnerable road users [4]. Thus, traffic accidents remain a critical concern in the country and require preventive measures. The factors influencing accident risk vary by country due to contextual differences and data availability. For example, in Iran, key factors influencing crash severity on urban highways include human factors, vehicle conditions, weather conditions, peak-hour traffic time, and traffic characteristics [5]. The risk of motorway accidents in Malaysia is substantially associated with roadway geometry. Key element of this association include the length of segments with simple curve, horizontal curve on steep slopes, and unsealed shoulders[6]. Thailand faces an alarming rate of approximately 22,000 traffic fatalities annually, with a mortality rate of 25 per 100,000 individuals [7]. The primary objective of this investigation is to determine the principal elements that increase the severity of fatal highway collision in Thailand. The research employs statistical approaches on data drawn from the Highway Accident Information Management System (HAIMS) database. It subsequently analyzes multiple variables, such as driving environment, road infrastructure quality, risk driver actions, and other related aspects that impact accident severity.

1.2 Background

Diverse methodologies have been used investigate the contributors to road traffic incidents, encompassing approaches like tableau analysis, descriptive statistics, AI and deep learning, and statistical modeling. Statistical models are particularly common because they offer interpretable insights into key accident characteristics. Specifically, logistic regression, utilizing the binary logit model, and the backward regression model are used to classify crucial safety elements. One study using these models revealed that driver age, vehicle faults, reverse gear movement, multi-vehicle crashes, certain collision types, and crashed involving motorcycles and bicycles substantially elevate crash severity on Tehran's urban highway, Iran. [8]. Deep learning technique are currently applied to conduct analysis within transportation safety. Specifically, LSTM-CNN models are leveraged for risk forecasting, employing live data inputs such as traffic volume, signal phase timing, and prevailing weather condition [9]. Convolution Neural Networks are employed to detect and quantify collision risks, drawing upon roadside radar information concerning volume, speed, and sensor occupancy [10]. Furthermore, Formosa, Quddus [11] used a Deep Neural Network to forecast potential traffic conflict, basing predictions on analyzed images collected via a signalized front-facing camera. Separately, Wu and Hsu [12] developed a successful

fusion model that integrates Gated Recurrent Units (GRUs) with Convolutional Networks. This model predicted the rate of at-fault drivers in collision with greater precision than existing machine learning methods and conventional statistical models.

2 Methodology

The study aims to investigate the key contributors to the lethality of road accidents in Thailand. It leverages large-scale data from the Highway Accident Information Management System (HAIMS), a resource that the Thai Department of Highways has compiled since 2011, contains detailed records on accident locations, human factors, road conditions, vehicle types, and the victim statuses. The dataset comprises 244,710 observations, with the attribute groups summarized in Table 1.

Table 1. All independent variables group and a dependent variable group from HAIMS

Feature	Group of variables	Variables
Independent variables	Road conditions	Road type, Road condition, Lane direction, Road Isle, Road traffic, Horizontal curve, Vertical curve, Intersection, U-turn at median, Access points.
	Human factors	Gender, Age, Road user type, Safety equipment usage, Alcohol consumption
	Environmental factors	Accident location, Surface pavement type, Environmental status, Lighting condition, Weather condition
	Other factors	Crash characteristics, Time of crash.
Dependent variable	Severity Level	Fatal at the scene, Fatal at hospital, Serious injuries, and Minor injuries

To identify critical elements that increase fatal accident severity, this research employs three distinct statistical approaches. The effect of these variables on accident severity is evaluated through Cramer's V and Information Values (IV). The HAIMS dataset's severity data is classified into two classifications: fatal severity (deaths occurring on-site or in the hospital) and non-fatal severity (comprising serious and minor injuries). Furthermore, a Convolutional Neural Network (CNN) deep learning model is also implemented to classify severity categories by processing pseudo-images. The pseudo-images which were input into the CNN model for classification between fatal and non-fatal events, were encoded by Weight of Evidence (WOE) technique on HAIMS data before being reshaped to square matrix format from vector row format to be as image characteristic feature, these showed how to convert HAIMS data into pseudo-images data. Additionally, the most critical factors influencing road traffic accidents are further analyzed using numerical methods derived from the deep learning model results.

2.1 The Cramer's V and Information Value technique

Cramer's V and Information Value (IV) also measure the strength of association between two categorical variables. In this study, both Cramer's V and IV are applied to evaluate the relationship between independent variables and fatality levels (fatal and non-fatal) [13, 14]. The combined application of Cramer's V and IV provides a robust framework for identifying significant factors influencing fatal accident severity. The classification power of these techniques is summarized in Table 2.

Table 2. Interpretation of Cramer's V and Information Value

Interpretation	Cramer's V value	Information Value
Powerful	≥ 0.25	≥ 0.50
Strong	≥ 0.15	≥ 0.30
Moderate	≥ 0.10	≥ 0.10
Weak	≥ 0.05	≥ 0.02
Unpredictive	≥ 0.00	≥ 0.00

2.2 Numerical analysis

The HAIMS dataset's independent variables were converted from discrete to a numeric structure before being used to train the Virtual Geometry Group deep learning framework (VGG). To maximize the classification performance concerning accident severity levels, the VGG model underwent through optimization. Techniques utilized included Bayesian optimization for parameter tuning, changes to the activation function, alterations of the learning rate and batch size, and precise setting of the number of training epochs to both minimize the loss function and ensure overfitting avoidance. A fully optimized deep learning model provides high confidence in accident severity classification, especially when addressing unseen data [15].

Despite its strong predictive capabilities, the VGG model has an extremely complex structure, making it difficult to directly interpret its mathematical function in identifying significant factors influencing fatal accident severity. To address this, this study employs the logit model, also known as the SoftMax or Sigmoid function, to examine the main contributing factors. The logit function is expressed as:

$$F(x) = \beta_1 x_1 + \dots + \beta_n x_n = \mathbf{x}\boldsymbol{\beta} \quad (1)$$

where x represents the independent variables in the HAIMS dataset, and $\boldsymbol{\beta}$ denotes the corresponding weights, which are evaluated using the Gradient Descent Algorithm (GDA). GDA is one of the most widely used algorithms for finding optimal solutions. It iteratively adjusts these coefficients to minimize the objective function by following the steepest descent direction. Thus, the objective function for determining the optimal weight of each independent variable is given by:

$$\min_{\boldsymbol{\beta}} z = \min_{\boldsymbol{\beta}} (y - \hat{y}) = \min_{\boldsymbol{\beta}} \frac{1}{N} \sum_{i=1}^N \left(y - \frac{\exp(\mathbf{x}\boldsymbol{\beta})}{1 + \exp(\mathbf{x}\boldsymbol{\beta})} \right)^2 \quad (2)$$

Significant factor analysis

The significance of factors influencing accident severity can be evaluated by testing individual parameters. This process is essential for determining the potential value of each regressor variable in the predicted probability model. The hypothesis testing framework consists of two hypotheses: the null hypothesis (H_0) assumes that the coefficient weight is not significant, whereas the alternative hypothesis (H_1) posits that the coefficient weight is significant. Under a 95% confidence level, the test statistics given in equation (3).

$$t_o = \frac{\beta_i}{se(\beta_i)} = \frac{\beta_i}{\sqrt{c_{ii}\sigma^2}} \quad (3)$$

where σ^2 represents the variance of parameters, and c_{ii} is the diagonal element in the matrix corresponding to β_i . The null hypothesis is rejected if $|t_o| > t_{\alpha/2, n-p}$ indicating that the regressor is significant. A p-value lower than 0.05 confirms that the regressor has a significant impact on the model, making it suitable for interpretation. The variance and diagonal element can be derived from probability logit function as follows:

$$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{y}{1-y}\right)$$

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \ln\left(\frac{y}{1-y}\right)$$

The variance of $\boldsymbol{\beta}$ can be derived as:

$$V(\boldsymbol{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1} V\left(\ln\left(\frac{y}{1-y}\right)\right)$$

Since the variance of predictors is unbiased and the error variance is denoted by σ^2 , we obtain:

$$V\left(\ln\left(\frac{y}{1-y}\right)\right) = V(X\boldsymbol{\beta}) + V(\epsilon) = 0 + \sigma^2 = \sigma^2$$

Thus, the variance of parameters can be expressed in equation (4)

$$V(\beta_i) = (\mathbf{X}^T\mathbf{X})^{-1} \sigma^2 = c_{ii}\sigma^2 \quad (4)$$

The unbiased estimator of σ^2 is obtained from the residual error between actual log-odds ratio from the VGG model and predicted log-odds ratio from weights evaluated using the GDA. The denominator $n-p$ represents the residual degrees of freedom. The estimator of variance is given in equation (5)

$$\sigma^2 = \frac{\sum_{i=1}^n e_i^2}{n-p} = \frac{1}{n-p} \sum_{i=1}^n \left[\ln\left(\frac{y_i}{1-y_i}\right) - \ln\left(\frac{\hat{y}_i}{1-\hat{y}_i}\right) \right]^2 \quad (5)$$

where n is the total number of samples, p is the number of regressors, y_i is the predicted probability from the VGG model at observation i , \hat{y}_i is the predicted probability derived from GDA weight.

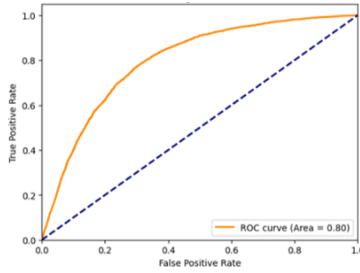


Fig. 2. AUC of the optimal VGG model on original HAIMS dataset.

Due to the complexity of deep learning models, directly interpreting the mathematic function of the VGG model is challenging. To address this, the logit function was applied to analyze the behavior of the VGG model’s classification results. Fig. 3 illustrates a scatter plot depicting the relationship between the probabilities predicted by the VGG model, the probability function, and the logits values computed using the Gradient Descent Algorithm (GDA). The plot reveals a clear pattern: lower logit values correspond to a lower probability of fatal severity, whereas higher logit values indicate an increased probability of fatal severity.

All coefficients in the logit function were evaluated using the Gradient Descent Algorithm, which aimed to minimize the Mean Square Error (MSE) between class prediction and probability derived from the logit function. Following this, individual parameters of the predicted probability model were tested to determine their statistical significance. The hypothesis testing results identified eight significant variables that correlate with fatal accident severity under a 95% confidence level, as summarized in Table 3

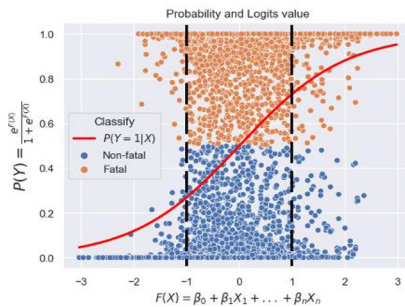


Fig. 3. Probability and Logits’ function for investigating the black box VGG model.

Table 3. Result of individual significant regressor for fatal and non-fatal.

Name	Coefficient	Std. Err.	t-statistic	p-value
Safety equipment	1.0043	0.1892	5.3079	1.12E-07
Road user type	1.0080	0.2001	5.0387	4.74E-07
Light environment	1.3080	0.2770	4.7222	2.35E-06

Name	Coefficient	Std. Err.	t-statistic	p-value
Road type	1.0323	0.2786	3.7055	2.12E-04
Month in a year	1.5694	0.4783	3.2812	1.04E-03
Number of road lane	0.5399	0.2414	2.2364	2.53E-02
Gender	0.8376	0.3799	2.2051	2.75E-02
Time	0.7946	0.3902	2.0362	4.17E-02

3.3 Likelihood of significant factors

Five significant factors were consistently identified across three analytical techniques – Cramer’s V, Information Value, and Numerical Analysis – as key contributors to fatal accident severity. These factors include safety equipment usage, road user type, light condition, number of road lanes, and driver gender. However, this study primarily focuses on physical factors and vehicle types to develop recommendations aimed at improving road safety in Thailand and significantly reducing the severity of traffic accidents.

For safety equipment usage, the analysis considers safety-belt usage, helmet usage and no safety equipment usage. The odds of fatal severity for individuals not using any safety equipment are 2.27 times ($\exp(\beta \Delta x) = \exp(1.004 \times (0.63529 - (-0.182425))) = 2.27$) higher than for those using a safety belt. Furthermore, the odds of fatal severity for individuals driving motorcycles or bicycles are 26% higher compared to those driving cars. These findings emphasize the critical role of safety equipment and vehicle type in influencing accident severity. Additional details are summarized in Table 4.

Table 4. Numeric value for analysis likelihood of Safety equipment factor.

Name	Numeric value
Safety	-0.182425
Helmet	0.052709
Not used	0.635290
Coefficient of safety equipment	1.004319

For road user type, there are numerous types of vehicles in Thailand; however, bikers exhibit the highest numerical values, indicating a significantly higher risk of fatal severity compared to other vehicle types. This finding aligns with the fact that bikers are vulnerable road users due to their minimal protective measures. Conversely, cars have the lowest numerical value, highlighting their relative safety in accidents. This safety advantage is primarily attributed to protective features such as seat belts, which reduce the impact of collisions and lower the likelihood of severe outcomes. The details supporting this observation are summarized in Table 5.

Table 5. Numeric value for analysis likelihood of road user type factor.

Name	Numeric value
Car	-0.291024
Pedestrian	0.000000
Tricycle (no motor)	0.236679
Motorcycle	0.308825
Tricycle	0.655437
Bike	1.034902
Coefficient of road user type	1.00804

For light environmental factors, the odds of fatal severity increase as environmental conditions shift from daytime to nighttime. When the light environment changes from daytime to nighttime with street lighting, the odds of a fatal accident increase by approximate 30%, with an odds ratio of 1.30. However, without street lighting, the risk of fatal severity increases 2.59 times compared to daytime conditions. This finding implies that in nighttime conditions without street lighting, the likelihood of fatal severity is twice as high as in nighttime conditions with street lighting, emphasizing the critical role of proper road lighting in reducing accident severity. Details are summarized in Table 6.

Table 6. Numeric value for analysis likelihood of light condition factor.

Name	Numeric value
Day	-0.15171
Night with light bulb	0.048900
Night without light bulb	0.576114
Coefficient of light condition	1.307913

4 Conclusion and future work

Using safety equipment, such as helmets and seat belts, significantly reduces the risk of fatal traffic accidents by approximately 1.79 to 2.27 times compared to not using any safety equipment. Moreover, bicyclists face a significantly higher fatality risk—about 3.81 times greater than car users—when an accident occurs, highlighting that cyclists are the most vulnerable road users in Thailand. Additionally, inadequate or unsupportive infrastructure for cyclists further increases the likelihood of severe accidents. Furthermore, driving at night or in poorly lit conditions significantly increases the risk of mortality by 2.60 times. This underscores the critical impact of poor visibility on fatal accident rates.

The key factors contributing to road fatalities in Thailand include road user type, especially cyclists, nighttime driving—particularly in areas without lighting—and the behavior of individuals who do not use safety equipment while driving. To mitigate these risks, road infrastructure improvement such as installing adequate road lighting, constructing obstacle medians, and implementing cyclist safety policies could play an important role in reducing traffic-related fatalities. Encouraging and enforcing the use

of safety equipment is also essential for improving road safety. For future research, further analysis using advanced techniques to investigate significant group and individual factors influencing fatal accidents severity in the dataset is recommended. Approaches such as Shapley Additive Explanations analysis through artificial intelligence, deep learning models, or new machine learning models could provide deeper insights into accident risk factors and enhance predictive modeling.

References

1. Shen, Y., et al., *Inter-national benchmarking of road safety: State of the art*. Transportation research part C: Emerging technologies, 2015. **50**: p. 37-50.
2. Alrifi, S.A. and K.F. Alkahtani, *Economic and Social Costs of Traffic Crashes in Saudi Arabia*, in *Department of Civil Engineering*. 2021, King Saud University.
3. Wijnen, W., et al., *Crash cost estimates for European countries, deliverable 3.2 of the H2020 project SafetyCube*. 2017.
4. Antos, S.E., et al., *Detecting Urban Clues for Road Safety*. 2021.
5. Khalili, M. and A. Pakgohar, *Logistic regression approach in road defects impact on accident severity*. Journal of emerging technologies in web intelligence, 2013. **5**(2): p. 132-135.
6. Rusli, R., et al., *Crash severity along rural mountainous highways in Malaysia: An application of a combined decision tree and logistic regression model*. Traffic injury prevention, 2018. **19**(7): p. 741-748.
7. World Health Organization, *Global status report on road safety 2023*. 2023. p. 8.
8. Moghaddam, F.R., et al., *Crash severity modeling in urban highways using backward regression method*. International Journal of Computer and Information Engineering, 2009. **3**(12): p. 2779-2784.
9. Li, P., M. Abdel-Aty, and J. Yuan, *Real-time crash risk prediction on arterials based on LSTM-CNN*. Accident Analysis & Prevention, 2020. **135**: p. 105371.
10. Huang, T., S. Wang, and A. Sharma, *Highway crash detection and risk estimation using deep learning*. Accident Analysis & Prevention, 2020. **135**: p. 105392.
11. Formosa, N., et al., *Predicting real-time traffic conflicts using deep learning*. Accident Analysis & Prevention, 2020. **136**: p. 105429.
12. Wu, Y.-W. and T.-P. Hsu, *Mid-term prediction of at-fault crash driver frequency using fusion deep learning with city-level traffic violation data*. Accident Analysis & Prevention, 2021. **150**: p. 105910.
13. Akoglu, H., *User's guide to correlation coefficients*. Turkish Journal of Emergency Medicine, 2018. **18**(3): p. 91-93.
14. Yap, B.W., S.H. Ong, and N.H.M. Husain, *Using data mining to improve assessment of credit worthiness via credit scoring models*. Expert Systems with Applications, 2011. **38**(10): p. 13274-13283.
15. Sonnatthanon, N. and K. Choocharukul, *Crash severity prediction using a virtual geometry-group-based deep learning approach with images-based feature representation*. Results in Engineering, 2025. **27**: p. 106155.

16. Nahm, F.S., *Receiver operating characteristic curve: overview and practical use for clinicians*. Korean journal of anesthesiology, 2022. **75**(1): p. 25-36.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

