Intelligent Traffic Accident Prediction Model for Internet of Vehicles With Deep Learning Approach

Da-Jie Lin, Mu-Yen Chen, *Member, IEEE*, Hsiu-Sen Chiang, and Pradip Kumar Sharma, *Senior Member, IEEE*

***Abstract*— In this study, a high accident risk prediction model is developed to analyze traffic accident data, and identify priority intersections for improvement. A database of the traffic accidents was organized and analyzed, and an intersection accident risk prediction model based on different mechanical learning methods was created to estimate the possible high accident risk locations for traffic management departments to use in planning countermeasures to reduce accident risk. Using Bayes' theorem to identify environmental variables at intersections that affect accident risk levels, this study found that road width, speed limit and roadside markings are the significant risk factors for traffic accidents. Meanwhile, Naïve Bayes, Decision tree C4.5, Bayesian Network, Multilayer perceptron (MLP), Deep Neural Networks (DNN), Deep Belief Network (DBN) and Convolutional Neural Network (CNN) were used to develop an accident risk prediction model. This model can also identify the key factors that affect the occurrence of high-risk intersections, and provide traffic management departments with a better basis for decision-making for intersection improvement. Using the same environmental characteristics as high-risk intersections for model inputs to estimate the degree of risk that may occur in the future, which can be used to prevent traffic accidents in the future. Moreover, it also can be used as a reference for future intersection design and environmental improvements.**

***Index Terms*— Deep neural networks, risk prediction, traffic accident intersection, deep learning.**

I. Introduction

IN THE past, traffic accident management and subsequent improvements relied mostly on the post-hoc analysis of traffic accidents and surveys of intersections that are prone to accidents. This left authorities poorly equipped to prevent accidents because their work focused on such practices as visiting accident scenes and conducting spot checks for A1 accidents (i.e., fatal accidents within the previous 24 hours). These approaches left authorities limited to enacting temporary solutions with limited ability to effectively focus resources on intersection safety improvement.

To effectively reduce accident risk, in recent years traffic accident management agencies in countries around the world have not only established standards and operating procedures for road surveys, but have also sought to develop accident risk analysis and prediction methods. The hope was that, longitudinal accident data would be used to identify and rank high risk intersections, allowing for efficient prioritization of scarce resources to minimize the frequency and severity of traffic accidents.

By analyzing the number of accidents and casualties, this study seeks to estimate the relative risk level of individual intersections, determine the risk key factors, and establish an intersection risk prediction model to predict the probability and severity of future accidents. This model marks each intersection with a different risk level and prioritization for environmental improvement, providing traffic management agencies with more an objective basis for making intersection improvement decisions.

The proposed traffic accident risk prediction model is based on the environmental factors of intersections, applying current machine learning techniques such as Naïve Bayes, Decision tree C4.5, Bayesian Network, Multilayer perceptron (MLP), Deep Neural Networks (DNN), Deep Belief Network (DBN) and Convolutional Neural Network (CNN). In practical applications, our proposed model can be used to predict the probability (risk) of accidents at different intersections by identifying similar environmental variables, thus allowing authorities to take practical steps to effectively reduce the incidence and severity of accidents along with the costs associated with such accidents. In addition, the research results identify important environmental factors that affect the occurrence of traffic accidents.

The rest of the paper is organized as follows: Section 2 reviews the relevant literature. The methods used to develop our proposed model are discussed in section

Manuscript received November 30, 2020; revised February 26, 2021; accepted April 17, 2021. This work was supported by the Ministry of Science and Technology of the Republic of China under Grant MOST109-2410-H-025-014-MY2 and Grant MOST109-2410-H-006-116-MY2. The Associate Editor for this article was T.-H. Kim. *(Corresponding author: Mu-Yen Chen.)*

Da-Jie Lin is with the Department of Transportation and Logistics, Feng Chia University, Taichung 407802, Taiwan (e-mail: dajielin@fcu.edu.tw).

Mu-Yen Chen is with the Department of Engineering Science, National Cheng Kung University, Tainan 70101, Taiwan (e-mail: mychen119@gs.ncku.edu.tw).

Hsiu-Sen Chiang is with the Department of Information Management, National Taichung University of Science and Technology, Taichung 404336, Taiwan (e-mail: hschiang@nutc.edu.tw).

Pradip Kumar Sharma is with the Department of Computing Science, University of Aberdeen, Aberdeen AB24 3FX, U.K. (e-mail: pradip.sharma@abdn.ac.uk).

Digital Object Identifier 10.1109/TITS.2021.3074987

3. Experimental results are presented in section 4, and section 5 presents findings and discussion of the results.

II. Related Works

A. Internet of Things

The rapid development and wide application of computer technologies, computer network technologies, multimedia and communication technologies, and the Internet of Things fields [1], has driven the recent development of intelligent road traffic management systems [2]. Li *et al.* The Internet of Things allows for the collection of various kinds of information through sensors [3], each of which represents an independent information source [4] from which data is collected at a certain frequency for categorization and analysis. Each independent information source would sense, measure, capture and transmit information anytime and anywhere. The development of advanced chip design and new materials have also increased the utility and longevity of such sensors [5], while also allowing for anti-interference, multi-mode, and self-adapting features [6]. These developments provide the technological basis for intelligent expressway management systems, integrating Internet of Things applications due to the introduction of mass information compatibility. High-speed wired and wireless networks have been integrated to create three-dimensional connections, ensuring the accuracy of data information, wider transmission bandwidth, higher spectrum utilization, more intelligent access, and more efficient network management [7]. The development of these advanced technologies mainly depends on NGN (Next Generation Network) communication network technologies and new wireless communication networks (3G, 4G, ZIGBEE) [8].

Expressway construction and traffic is rapidly growing around the world, and the demand for social development is growing synchronously [9]. Improving the efficiency of existing expressway traffic infrastructure requires the effective collection and analysis of usage data [10]. As cars and individual drivers are increasingly linked to wireless transmissions, drivers demand increasingly sophisticated traffic information, allowing them to assess current local traffic and driving conditions, predict future conditions, and identify optimal driving routes [11]. Expressway traffic management agencies also need to effectively monitor highway conditions and coordinate timely emergency response including police, rescue and repair units [12]. The data to drive such coordination is sourced from sensor networks that monitor traffic and environmental conditions throughout the highway network. Such monitoring data can be used to improve and simplify signal control algorithms and traffic efficiency. Wireless sensor networks can be applied to control subsystems and guidance subsystems in the execution subsystem, and to improve signal controller function to implement the bus priority function of the intelligent transportation system [13]. Besides, the position sensor can help achieve functions such as energy-saving and emission reduction.

B. Traffic Accident Prediction Model

Previous studies have sought to analyze and predict traffic accident risk using a wide range of analysis methods including traditional statistics, machine learning, and deep learning techniques. Predicting traffic accident risk requires the effective identification of important explanatory variables or influencing factors.

Chen *et al.*, (2019) proposed a research framework based on key feature selection of a vehicle trajectory dataset and risk prediction of lane-changing (LC) behavior on expressways [14]. This study applied fault tree analysis and K-means clustering methods, based on the Crash Potential Index (CPI) to determine the risk level of vehicle lane changes. The results of key feature selection showed that the interaction between the vehicle that changes lanes and the surrounding vehicles, along with changes in vehicle acceleration between surrounding the vehicles in the target lane, is a significant factor in the risk assessment of lane change behavior.

Wang *et al.* (2019) established a collision tendency prediction model based on the characteristics of vehicle groups collected by the floating car method based on a highway in Shanghai [15]. The binary logistic regression model and support vector machine (SVM) were used to build prediction models with an accuracy of 85%, as opposed to 60% for the binary logistic regression model.

Wang *et al.* (2019) studied real-time collision prediction issues [16] based not only on traffic and environmental forecasting factors, but also included the impact of social demographic data and trip generation parameters on immediate collision risk. They used traffic, geometric, socio-demographic and trip generation predictors to analyze the immediate collision risk of highway ramps. Two Bayesian logistic regression models were used to identify collision precursors and their impact on ramp collision risk. At the same time, four support vector machine (SVM) models were used to predict collision occurrence. The results showed that the inclusion of socio-demographic and trip generation predictors improved prediction performance.

Zheng & Sayed (2020) proposed a generalized extreme value (GEV) model based on the Bayesian hierarchical structure to predict collision risk based on three indicators: traffic volume, shock wave area and platoon ratio [17]. The proposed method was applied to four signalized intersections in Surrey, British Columbia.

Cai *et al.* (2020) sought to address extremely unbalanced traffic data in collision and non-collision cases, using a Deep Convolutional Generative Adversarial Network (DCGAN) [18], balancing the dataset through the synthetic minority oversampling technique (SMOTE) and random undersampling technique. The experimental design used logistic regression, support vector machine (SVM), artificial neural networks (ANN) and convolutional neural networks (CNN) to develop twelve models for performance evaluation. The results showed that the DCGAN provides the best prediction accuracy.

Yu *et al.* (2021) presented a novel Deep Spatio-Temporal Graph Convolutional Network (DSTGCN) to construct a traffic accident model [19]. Experimental results showed the proposed DSTGCN model outperformed the LR, LASSO, SVM, and DT methods. Meanwhile, Fang *et al.* (2021) proposed a semantic context induced attentive fusion network (SCAFNet) that first segments the RGB video frames into many individual images, which can then be used as the inputs for the graph convolution network (GCN) based on different semantic regions. Experimental results showed the proposed model is useful for prediction of driver attention.

III. Method

A. Naive Bayes

The Naive Bayes (NB) algorithm is based on Bayes' theorem. Chiang (1995) proposed a complete road traffic safety system covering data storage to analysis, for which the analysis method is mainly based on Bayes' theorem [18]. NB assumes the prior probability of a known target variable, which is often known from training samples. In addition, given any target variable or dependent variable, the participating attribute values are assumed to be mutually independent. Assuming that training materials have a set of attributes $\mathrm{X} = \{X_1, X_2, \ldots, X_\mathrm{n}\}$, X does not contain the target variable attribute, and C is the attribute value set of the target variable, $C = \{C_1, C_2, \ldots, C_\mathrm{m}\}$. $P(C|X)$ is the probability that the target category C appears under a certain set of X attributes. $P(X|C)$ is the probability of cases appearing in a set of attributes X under a target category C, as show in Eq. 1. $P(C)$ is the probability of being the target category. $P(X)$ is the probability of an event occurring under a set of X attributes.

$$P\,(C/X) = \frac{P\,(X/C) * P(C)}{P(X)} \tag{1}$$

Assuming that each feature is independent of each other based on Naïve Bayes theory, then eq.(1) becomes:

$$P\,(C/X) = \frac{\prod_{i=1}^{n} P(X_i/C) * P(C)}{\prod_{j=1}^{n} P(X_j)} \tag{2}$$

where $\mathrm{P(X_i/C)}$ is the likelihood that feature X_i occurs in a class C_m, C_m C. $\mathrm{P(c_m)}$ is the prior probability of the class C_m, $C_\mathrm{m} \in$ C in the entire data set. The output of the classifier is the highest probability class for a given set of features. Because the denominator does not depend on C and the value of the features X_i is given, the denominator can be considered a constant. The maximum class is obtained over all classes of the probability of each class $C_\mathrm{m}, C_\mathrm{m} \in$ C, computed in eq.(2) to get

$$argmax\ c = P(C = c) \prod_{i=1}^{n} P\,(X = X_i \mid C = c) \tag{3}$$

where argmax c is used to denote the function that returns the class with the highest probability. In other words, given a sequence of features $\mathrm{X_i}$, the most likely class to be classified can be obtained by eq.(3).

B. Bayesian Network

The Bayesian network is a probability model. A set of random variables $\{X_1, X_2, \ldots, X_\mathrm{n}\}$ and n sets of conditional probability distributions are obtained through a directed acyclic graph [19], [20]. The nodes in the Bayesian network represent random variables that can be observable variables, latent variables, unknown parameters, etc. The arrow connecting two nodes represents whether the two random variables have a causal relationship or are not conditionally independent. If no arrow connects the two nodes together, the random variables are called conditionally independent.

Each node has a conditional probabilistic table that describes the probabilistic distribution of the states for the corresponding variable given from the states of its parent nodes. Each row of this condition probability table lists all possible probabilities, and each row lists all possible states. The total probability of any column must be 1.

The joint probability table for a single variable X is estimated as follows:

$$P(X = i) = \frac{N_{X=i}}{N} \tag{4}$$

N is the total number of observations, $N_{X=i}$ is the number of observations when X is in state i, and P denotes the probability that X is in state i.

The joint probability table for a relation involving multiple variables is estimated as follows:

$$P(X_1 = i_1, \ldots, X_k = i_k) = \frac{N_{X_1=i_{1,\ldots,}X_k=i_k}}{N} \tag{5}$$

$N_{X1=i1}$ is the number of observations with X_1 in state $i_1, \ldots$, and $N_{Xk=ik}$ is X_k in state i_k., and P denotes the probability under all features conditions, $X_1 = i_1, \ldots, X_k = i_k$.

C. System Model

This research uses the Iterative Dichotomized (ID3) algorithm to recursively conduct top-down segmentation to construct a decision tree, and then uses the C4.5 algorithm to trim the tree branches. ID3 uses information gain [21] to select decision attributes (nodes). Information acquisition can be defined as a certain node attribute, where the "expected information entropy before being partitioned by the target variable" less the "expected information entropy before being partitioned by an attribute" obtains the degree of information clutter reduction (benefit degree), and the attribute that can obtain the maximum benefit is selected as the node.

S : A finite set of samples $\{s_1, s_2, \ldots, s_m\}$

C : Category $\{c_1, c_2, \ldots, c_m\}$

s_i : The number of samples under a certain category c_i

s_j : The number of samples under a certain attribute value (a_v) of a certain attribute (A_k)

s_{ij} : The number of samples under a certain attribute value (a_v) of a certain attribute (A_k) for a certain category (c_i)

P_i : Proportion of sample $(\frac{s_i}{S})$ under a certain category (c_i)

P_{ij} : The proportion of samples for a certain attribute value (a_v) of a certain attribute (A_k) for a certain category (c_i)

A : A collection of attributes

A_k : An attribute which contains the attribute values $\{a_1, a_2, \ldots, a_m\}$

Equation (6) calculates the expected information entropy before being partitioned by target variable $I(s_1, s_2, \ldots, s_m)$, $I(s_1, s_2, \ldots, s_m)$, representing the post-segmentation degree of entropy of the target variable of the training set (Target variable, *Dependent variable*).

$$I\left(S_1, S_2, \ldots, S_m\right) = -\sum\nolimits_{i=1}^{m} p_i\left(p_i\right) \tag{6}$$

Equation (7) calculates the sample ratio of the training sample divided by the attribute A_k. For example, the attribute "gender" has two attribute values $\{7male, 3female\}$, and the sample ratio is $\{\frac{7}{10}, \frac{3}{10}\}$.

$$E\left(A_k\right) = \sum\nolimits_{j=1}^{v} \frac{S_{1j} + \ldots + S_{mj}}{S} I \quad \left(S_1, S_2, \ldots, S_m\right) \tag{7}$$

Equation (8) calculates the information entropy for a certain attribute value (a_v) (female) for attribute (A_k), that is $I\left(s_1, s_2, \ldots, s_m\right)$, which is then substituted into Eq. (6). Multiplying the respective sample ratios, we obtain the expected information entropy before being partitioned by an attribute variable.

$$I\left(S_{1j}, S_{2j}, \ldots, S_{mj}\right) = -\sum\nolimits_{i=1}^{m} p_{ij}\left(p_{ij}\right) \tag{8}$$

By subtracting $E\left(A_k\right)$ from $I\left(s_1, s_2, \ldots, s_m\right)$, we obtain the information Gain $\left(A_k\right)$ for a certain attribute node.

$$Gain\left(A_k\right) = I\left(S_1, S_2, \ldots, S_m\right) - E\left(A\right) \tag{9}$$

By analogy, the information benefit of each attribute is obtained, and the node with the greatest information benefit is used as the node. We then follow these steps to recursively construct the ID3 decision tree. When the category status in a node is consistent or the information benefit of each attribute is consistent, there is no need for further segmentation.

C4.5 divides tree branches in two, taking the middle value of two consecutive values as the cutting point, and the information gain results are calculated and compared. The highest is the discrete cutting point of the continuous value. The following equation is used to improve the characteristics of nodes with more attributes:

$$Gain\_ratio(D, x) = \frac{IG(D, x)}{IV(x)} \tag{10}$$

$$IV\left(x\right) = \frac{D^i}{D} - (\frac{D^i}{D}) \tag{11}$$

D. Multilayer Perceptron Neural Network

Multilayer Perceptron (MLP) is a forward structured artificial neural network, including three layers: input, hidden and output. The output vector is calculated based on the input vector. MLP consists of multiple nodes, and each layer is fully connected to the next layer. Except for the input nodes, each node is a neuron (processing unit) with a nonlinear activation function. In the training process (learning) of MLP, the algorithm is used to adjust the weight and reduce the bias of the training process. MLP is a special case of deep neural network (DNN). Its main advantage lies in its ability to quickly solve complex problems [22].

E. Deep Neural Network

Deep Neural Networks (DNN) are a type of deep learning framework which can be understood as a neural network with many hidden layers (Neural Networks). A neural network uses a kind of interconnected neuron to construct a mathematical model similar to a biological neural network from artificial neurons. Normally, neurons are organized in layers, with connections only established between neurons in adjacent layers. The input low-order feature vector is added to the first layer, and converted to a high-order feature vector by gradually moving the neurons to the higher layer. The number of neurons in the output layer is equal to the number of classifications. Therefore, the output vector is a probability vector to indicate the probability that the input vector belongs to the corresponding category. The expected calculation of a single neuron and its output description are shown in Eq.(12), where a_j^i is the j th neuron in the i th layer, w^i represents the weight of the neuron's synapse, which connects the j th neuron in the i th layer with the k th neuron in the previous layer (i.e. layer i-1) Neurons.

$$a_j^i = \sigma\left(\sum\nolimits_k a_k^{i-1} w^i j_k\right) \tag{12}$$

The neural network layer can be divided into three layers: an input layer and output layer with a hidden layer in between. The original purpose of the neural network is to simulate the operation of human neurons. If the activation function is not used, then the linear combination of the above layer inputs is used as this layer's output (matrix multiplication), thus the output and input still cannot be separated from the linear relationship. The non-linear activation function is used to increase the non-linear factor of neurons, so that authentic complex models can be expressed through neural networks [23]. Common activation functions include Sigmoid, Softmax, tanh, ReLU, and ELU. The loss function maps an event (an element in a sample space) to a real number that intuitively expresses the event's economic or opportunity cost. The optimization goal is to minimize the loss function. Therefore, the performance of the neural network model and the goal of optimization are defined by the loss function.

IV. Results

The purpose of the accident risk analysis is to predict the risk of accidents at various intersections. From the number of accidents and the number of casualties in the past, the risk level of each intersection is estimated. By identifying the key environmental factors that affect the occurrence of accidents at intersections, a risk prediction model for intersections is established, was used to predict the degree of accident risk at intersections that have not yet occurred.

A. Traffic Accident Intersection Selection

Traffic accident data were obtained from Taiwan's National Police Agency (NPA). Data for this accident risk prediction and analysis was taken from annual traffic accident data for 2018. The total number of traffic accidents was 320,315, of which 32,110 occurred at 19,115 provincial highway intersections. With Fig. 1 showing incident locations, analyzed

TABLE I
INTERSECTION ENVIRONMENT VARIABLES

Feature	Description
Speed limit	Low <=50 km/h
	Middle 50~70 km/h
	High >70 km/h
Road width	Narrow <=10m
	Middle 10~26m
	Width >26m
Types of signs	Traffic control sign
	Traffic control signs (with pedestrian signs)
	Flash sign
	No sign
Pavement edge line	Yes
	No
Road pattern	Underpass
	Slope
	Straight road
	Non-single road
	Alley
	Elevated road
	Culvert
	Bridge
	Tunnel
	Curved road and nearby
	Other
Crossroads	Three-way bifurcation road
	Four-way bifurcation road
	Multiple-way bifurcation road
	Not bifurcation road

based on geographic information system (GIS) data to identify geographic spatial data and spatial elements of intersections at which accidents occurred (within a radius of 50 meters), with detailed location ranges marked with red circles in Fig. 2. Problematic intersections were excluded, along with the discrete number of accidents at each intersection.

Environmental variables for each intersection were collected from the traffic accident data from NPA (the traffic accident data set has recorded the environmental factors around the location of each traffic accident) for subsequent analysis, including the continuous values of "Speed Limit" and "Road Width" (see Table I).

1) Accident Risk Calculation: In this study, all traffic accidents occurring within a radius of 50 meters around the traffic accident occurrence site were consolidated into a single accident occurrence site. The annual totals of accident occurrences, deaths, injuries and other data for the site were then used to calculate the site's Symptom Ratio Index (SRI) and Symptom Severity Index (SSI), which were then added up to generate a Combined Index (CBI), for which a larger value indicates a higher likelihood of accidents.

$$SRI_i = \frac{N_i}{MAX(N_j)}; \quad j = 1, 2, 3 \ldots n \tag{13}$$

$$SSI_i = \frac{ETA_i}{MAX(ETA_j)}; \quad j = 1, 2, 3 \ldots n \tag{14}$$

$$CBI_i = SRI_i + SSI_i \tag{15}$$

B. Comparison of Environmental Variables at Intersections

To differentiate between low, medium and high risk levels and different environment variables, the following comparison was made through the conditional probability at different risk levels to identify trends and environmental variables that affect risk level, and to determine the significant environmental variables.

TABLE II
SPEED LIMIT CONDITION PROBABILITY

Speed limit	Low risk	Middle risk	High risk
Low	71.28%	80.61%	90.96%
Medium	28.53%	19.34%	8.93%
High	0.19%	0.05%	0.11%

TABLE III
ROAD WIDTH CONDITION PROBABILITY

Road width	Low risk	Middle risk	High risk
Low	3.52%	2.57%	1.74%
Medium	62.75%	55.72%	45.97%
Width	33.74%	41.71%	52.29%

TABLE IV
SIGNAGE CONDITION PROBABILITY

Types of signs	Low risk	Middle risk	High risk
Traffic control sign	40.41%	57.15%	72.66%
Traffic control signs (with pedestrian signs)	11.66%	14.26%	16.34%
Flashing sign	9.13%	3.95%	2.29%
No sign	38.81%	24.64%	8.71%

TABLE V
ROAD WIDTH CROSSROADS CONDITION PROBABILITY

Crossroads	Low risk	Middle risk	High risk
Three-way bifurcation road	31.47%	27.78%	17.43%
Four-way bifurcation road	30.87%	46.73%	66.78%
Multiple-way bifurcation road	6.09%	7.12%	10.24%
Not bifurcation road	31.57%	18.37%	5.56%

Table II shows that, at lower speed limits, risk is positively correlated with the probability of accident occurrence, i.e., significant injuries and fatalities are more likely to occur when the speed limit is below 40 km (90.96%), and most accidents occur at lower speed limits (regardless of the level of risk).

Table III shows that narrower roads (i.e., less than 10 m in width) are more prone to low-risk accidents (as the lower number of accidents overall leads to lower casualties), while wider roads (> 26 m in width) are more prone to high-risk accidents (as the higher number of accidents leads to higher casualties), and there is a trend of increasing risk from narrower to wider roads.

Table IV shows that intersections with traffic control signals are prone to medium to serious accidents, while those with pedestrian signals are prone to serious accidents (high risk). Intersections with flashing signals are also prone to low-risk accidents, but there is a higher probability (38.81%) of low-risk accidents at intersections without signals.

TABLE VI
ROAD WIDTH PAVEMENT EDGE LINE CONDITION PROBABILITY

Pavement edge line	Low risk	Middle risk	High risk
Yes	60.74%	50.82%	30.83%
No	39.26%	49.18%	69.17%

Fig. 1. Selection of Accident Occurrence Intersections.

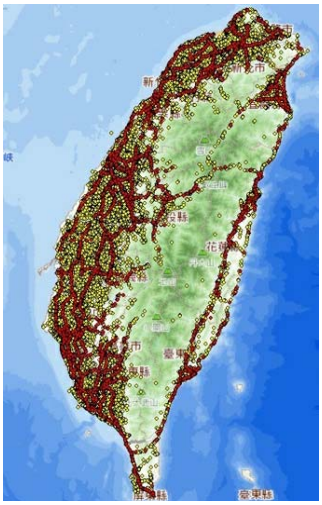

Fig. 2. Location Data of Accident Occurrence Intersections.

Table V shows that most accidents occur at four-way intersections, especially the medium and high risk accidents, and the higher the accident risk is, the higher the probability of its occurrence. However, accidents are also prone to happening at three-way intersections, while accidents at non-intersections are usually low risk. In addition, three-way intersections and non-intersections are less prone to high risk accidents.

Table VI shows that accidents on roads with roadside markings are mostly low risk accidents (as the lower number of accidents leads to lower casualties), and the higher the accident risk is, the less likely it is that an accident will occur at an intersection with roadside markings. Therefore, serious accidents (with greater rates of incidence and higher casualties), are prone to happen on roads without roadside markings, and the higher the accident risk is, the more likely it is that an accident will occur at an intersection without roadside markings.

C. Traffic Intersection Forecast Risk Model

Mechanical learning and deep learning methods were used to model the intersection risk prediction, and a two-stage

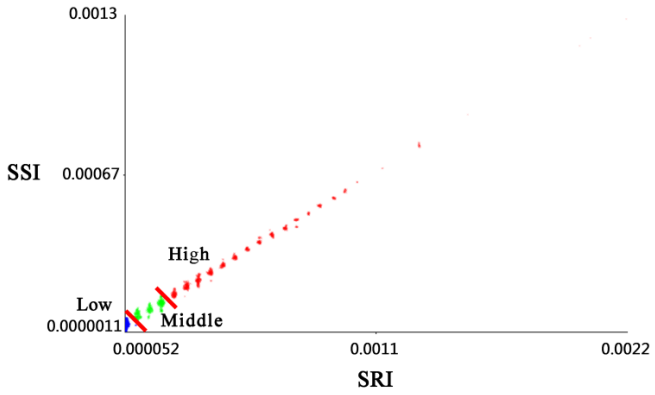

Fig. 3. Risk grouping (First level).

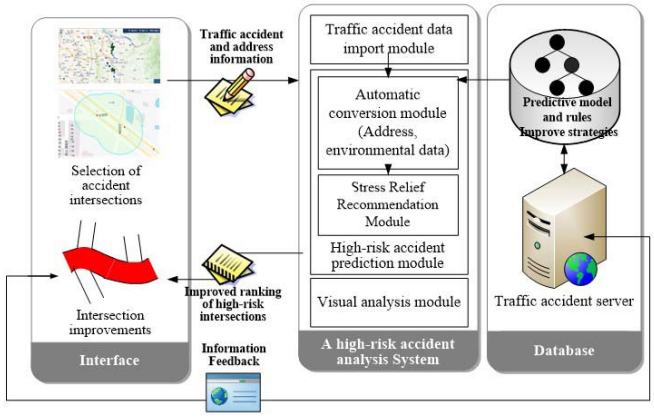

Fig. 4. Architecture of a high-risk accident prediction and analysis platform.

analysis mode was used to implement the prediction model of intersections with higher probability of high-risk accidents. Based on the CBI of 19,115 intersections, the intersection clusters with low, medium, and high-risk levels were clustered through the K-means cluster algorithm to calculate the number of intersections and cluster centers for each cluster, as shown in Fig. 3.

In the first stage, all 19,115 accident-prone intersections were clustered into low, medium and high-risk groups, while in the second stage, the "high-risk accident intersections" obtained from the first clustering stage were again subjected to K-means clustering into the three clusters of high-low, high-medium and high-high to identify key environmental factors associated with high-risk intersections.

D. Risk Intersections Analysis

The prior probability of low, medium and high-risk intersection clusters is calculated in Table VII, while Table VIII shows each section's prior probability for environmental variables as pertains to the conditions for accident occurrence risk.

E. Evaluation

The OS environment of this research is Windows 10, the hardware specification is I5 CPU, NVIDIA 1060 IT GPU,

TABLE VII

INTERSECTION GROUPING WITH DIFFERENT RISK LEVELS PRIOR PROBABILITY FIRST STAGE GROUPING

Grouping	Number of intersections	Probability
Low	13,881	72.58%
Middle	4,324	22.62%
High	910	4.8%

TABLE VIII

PRIOR PROBABILITY OF EACH FEATURE FIRST STAGE GROUPING

Speed limit			**Pavement edge line**		
Attributes	**Probability**	**Frequency**	**Attributes**	**Probability**	**Frequency**
Low	74.33%	14,217	Yes	57.06%	10,913
Middle	25.51%	4,879	No	42.94%	8,213
Width	0.16%	30	**Road pattern**		
Crossroads			**Attributes**	**Probability**	**Frequency**
Attributes	**Probability**	**Frequency**	Tunnel	0.06%	11
Three-way bifurcation road	29.96%	5,731	Underpass	0.06%	11
Four-way bifurcation road	36.18%	6,920	Bridge	0.44%	85
Mulitple-way bifurcation road	6.52%	1,247	Culvert	0.05%	9
Not bifurcation road	27.33%	5,228	Elevated road	0.07%	13
Types of signs			Curved road and nearby	1.29%	246
Attributes	**Probability**	**Frequency**	Slope	0.12%	22
Traffic control sign	45.74%	8,749	Alley	0.12%	22
Traffic control signs (with pedestrian signs)	12.47%	2,385	Straight road	24.14%	4,617
Flash sign	7.63%	1,459	Other	1.06%	202
No sign	34.16%	6,533	Non-single road	72.61%	13,888
Road width					
Attributes	**Probability**		**Frequency**		
Narrow	3.22%		615		
Middle	60.35%		11,543		
Width	36.43%		6,968		

with 20G memory for program execution. Programming was done in Python 3.6 and the deep learning framework used was Tensorflow 1.14.0. The mechanical learning methods Decision tree (C4.5), Bayes Net (BN), Naïve Bayes (NB), Multilayer perceptron (MLP), Deep Neural Networks (DNN), Deep Belief Network (DBN) and Convolutional Neural Network (CNN) were used to build the traffic intersection risk prediction models. Data for 19115 intersection traffic accidents were used for training and testing in a 70:30 ratio.

In the C4.5 parameter configuration, this research used a batch size of 100, a confidence factor of 0.25, and minimum number of instances per leaf of 2. In the BN parameter configuration, the batch size is 100, the estimator alpha is 0.5, and the search algorithm uses the hill climbing method. In the Naïve Bayes parameters configuration, the batch size is 100, and precision is 0.1. In the Multilayer Perceptron parameter configuration, the hidden layer is 1, the learning rate is 0.3, momentum is 0.2, and the number of training epochs is 500.

In the parameter configuration of DNN is 500, the learning optimizer uses Adagrad, and the activation function is the Relu function. In the DBN parameter settings, the learning rate of RBM is set to 0.05, the learning rate of ANN is 0.05, the number of training epochs is 200, the batch size is 100, and the Relu activation function is also used. The CNN uses the Adam learning optimizer, the loss function is Cross Entropy, the number of training epochs is 20, and the batch size is 32. All classification methods are tested using 10-fold cross-validation.

Different traffic intersection forecast risk models were evaluated in terms of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). Precision is defined by TP / (TP + FP), recall is defined by TP / (TP + FN), F1-score = 2 * Precision * Recall / (Precision + Recall) and overall accuracy is defined by (TP + TN) / (TP + FP + FN + TN).

1) First Stage Risk Clustering: Different mechanical learning methods are used for analysis to identify the environmental factors and rules that affect intersection risk of accident occurrence at different levels (low, medium, and high). Table IX shows that, with the exception of NB, most methods have a higher precision rate for low-risk intersections, but low accuracy for medium and high-risk intersections. This may be due to the heavily skewed distribution of data of the three clusters of risk intersections, resulting in a relative lack of training data for medium and high-risk intersections with fewer accidents, thus the model cannot determine relevant environmental factors or rules for medium and high-risk intersections. Overall, MLP and DNN outperform the other methods for prediction accuracy.

2) Second Stage Risk Clustering: Due to the imbalanced data clusters based on the first stage analysis results, none of the analysis methods are able to determine effective models in the original risk clusters (especially in the middle-risk and high-risk clusters). Therefore, the 910 high risk intersections were subjected to the K-means clustering to further cluster the advanced high-low, high-moderate and high-high risk clusters, with 498, 313 and 99 intersections respectively.

For second stage clustering of high-risk intersections, the different mechanical learning methods were used again to determine rules for the environment variables in terms of accident-prone intersections. Table X shows the result of the confusion matrix for decision tree analysis, indicating that the

TABLE IX
METHOD PERFORMANCE EVALUATION-FIRST STAGE CLUSTERING

Method	Attributes	TP	FP	FN	TN	Precision	Recall	F1-score	Accuracy
C4.5	Low risk	13881	5234	0	0	72.62%	100%	84.14%	72.62%
	Middle risk	0	0	4324	14791	0%	0%	0%	
	High risk	0	0	910	18205	0%	0%	0%	
BN	Low risk	13143	4345	738	889	75.15%	94.68%	83.80%	71.81%
	Middle risk	583	1044	3741	13747	35.83%	13.48%	19.59%	
	High risk	0	0	910	18205	0%	0%	0%	
NB	Low risk	13187	4393	694	841	75.01%	95.00%	83.83%	71.84%
	Middle risk	546	989	3778	13802	35.57%	12.63%	18.64%	
	High risk	0	0	910	18205	0%	0%	0%	
MLP	Low risk	13515	4893	366	341	73.42%	97.36%	83.71%	71.94%
	Middle risk	236	471	4088	14320	33.38%	5.46%	9.38%	
	High risk	0	0	910	18205	0%	0%	0%	
DNN	Low risk	13881	5234	0	0	72.62%	100%	84.14%	72.62%
	Middle risk	0	0	4324	14791	0%	0%	0%	
	High risk	0	0	910	18205	0%	0%	0%	
DBN	Low risk	13881	5234	0	0	72.62%	100%	84.14%	72.62%
	Middle risk	0	0	4324	14791	0%	0%	0%	
	High risk	0	0	910	18205	0%	0%	0%	
CNN	Low risk	13881	5234	0	0	72.62%	100%	84.14%	72.62%
	Middle risk	0	0	4324	14791	0%	0%	0%	
	High risk	0	0	910	18205	0%	0%	0%	

TABLE X
DECISION TREE CONFUSION MATRIX-SECOND STAGE CLUSTERING

Low risk	Middle risk	High risk	
416	82	0	**Low risk**
224	89	0	**Middle risk**
57	42	0	**High risk**

number and proportion of high-high risk clusters is still too small for effective classification, resulting in low accuracy. However, the high-low and high-medium risk clusters can be more effectively classified than in the first stage analysis result.

From the evaluation results of the first and second clustering (Tables X and XI), the neural network model outperforms the other methods. Among the two methods based on probability theory, NB outperforms BN. Among all methods, MLP has the best detection effect and is best suited for prediction of risky intersections. In the case of unbalanced data training, NB performs best. However, DNNs perform well for low-risk intersection detection due to over-learning (Precision = 100%) and other intersections are almost impossible to detect, probably due to overfitting caused by deep learning neural networks.

TABLE XI
METHODS PERFORMANCE EVALUATION-SECOND STAGE CLUSTERING

Method	Attributes	TP	FP	FN	TN	Precision	Recall	F1-score	Accuracy
C4.5	Low risk	416	281	82	131	59.68%	83.53%	69.62%	55.49%
	Middle risk	89	124	224	473	41.78%	28.43%	33.84%	
	High risk	0	0	99	811	0%	0%	0%	
BN	Low risk	426	273	72	139	60.94%	85.54%	71.18%	56.92%
	Middle risk	92	119	221	478	43.60%	29.39%	35.11%	
	High risk	0	0	91	819	0%	0%	0%	
NB	Low risk	431	291	67	121	59.70%	86.55%	70.66%	56.15%
	Middle risk	80	108	233	489	42.55%	25.56%	31.94%	
	High risk	0	0	99	811	0%	0%	0%	
MLP	Low risk	408	274	88	140	59.82%	82.26%	69.27%	55.28%
	Middle risk	95	130	217	468	42.22%	30.45%	35.38%	
	High risk	0	3	99	808	0%	0%	0%	
DNN	Low risk	424	257	74	156	62.35%	85.14%	71.93%	57.80%
	Middle risk	101	127	212	470	44.30%	32.27%	37.34%	
	High risk	1	1	98	810	50%	1.01%	1.98%	
DBN	Low risk	418	243	80	169	63.24%	83.94%	72.13%	58.35%
	Middle risk	113	136	200	461	45.38%	36.10%	40.21%	
	High risk	0	0	99	811	0%	0%	0%	
CNN	Low risk	498	412	0	0	54.72%	100%	70.73%	54.73%
	Middle risk	0	0	313	597	0%	0%	0%	
	High risk	0	0	99	811	0%	0%	0%	

V. CONCLUSION

A risk prediction model based on traffic accident data is established and validated for accident risk prediction on intersections using 2018 traffic accident data for Taiwan. The number of traffic accidents at intersections with similar environmental characteristics is used to predict the likelihood of traffic accidents in the future. Of a total of 320,315 traffic accidents reported, 32,110 occurred at 19,115 provincial highway intersections. Risk clustering in terms of CBI was performed for accident data for provincial highway intersections and various mechanical learning methods were used to establish a prediction model for high-risk intersections. Results show that environmental variables such as road width, speed limit and presence of roadside markings are significant predictors of accident incidence. On the other hand, the relatively high

numbers of accidents at low-risk and medium-risk intersections made it easier to identify their environmental characteristics. Prediction accuracy for high-risk accident intersections suffered from a relative scarcity of data, but decision tree rules and detection models were found to provide acceptable prediction accuracy for high-low and high-medium risk intersection clusters. In addition, the DBM model was found to be best suited for intersection risk prediction, while NB performs best for model training with unbalanced data. The results of this study can provide a reference for traffic management agencies to minimize accident risk at intersections.

Small-scale unmanned aerial vehicles (UAVs), or drones, have shown exceptional performance in many IoT applications, such as the Internet of Vehicles. Tian *et al.* (2019) designed an efficient privacy-preserving authentication architecture for UAVs under an Internet of Drones (IoDs) environment [27]. Gope and Sikdar (2020) presented an efficient privacy aware authenticated key agreement method for edge-computing IoD environments [28]. Both approaches provided third-party communication ability, and mobile-edge computing function integrated with UAV devices. Future work can consider information security and privacy issues in intelligent traffic prediction models.

Future work will seek to develop a high-risk accident prediction and analysis platform based on the environmental factors of intersections, as shown in Fig. 4. Based on the location of traffic accidents, data collection and analysis will be carried out to achieve the following goals:

1). This system platform can integrate and analyze traffic accident data and GIS layer information, thus understanding the overall accident location. Then, through analyzing the accident cause and the impact of environmental factors at the various intersections, we can formulate relevant improvement strategies as a reference for future intersection design and environmental improvements.

2). Use predictive models to estimate the likely locations of high-risk accidents to allow traffic management authorities to better prevent high-risk road accidents or serious casualties.

REFERENCES

[1] Z. Zhu, S. Chen, Y. Yang, A. Hu, and X. Zheng, "VISSIM simulation based expressway exit control modes research," *Proc. Eng.*, vol. 137, pp. 738–746, Feb. 2016, doi: 10.1016/j.proeng.2016.01.311.

[2] M. Li, X. Chen, and W. Ni, "An extended generalized filter algorithm for urban expressway traffic time estimation based on heterogeneous data," *J. Intell. Transp. Syst.*, vol. 20, no. 5, pp. 474–484, Sep. 2016.

[3] Y. Feng, B. Hu, H. Hao, Y. Gao, Z. Li, and J. Tan, "Design of distributed cyber–physical systems for connected and automated vehicles with implementing methodologies," *IEEE Trans. Ind. Informat.*, vol. 14, no. 9, pp. 4200–4211, Sep. 2018.

[4] J. Yang, J. Zhou, D. Fan, and H. Lv, "Design of intelligent recognition system based on gait recognition technology in smart transportation," *Multimedia Tools Appl.*, vol. 75, no. 24, pp. 17501–17514, Dec. 2016.

[5] L.-W. Chen and Y.-F. Ho, "Centimeter-grade metropolitan positioning for lane-level intelligent transportation systems based on the Internet of vehicles," *IEEE Trans. Ind. Informat.*, vol. 15, no. 3, pp. 1474–1485, Mar. 2019.

[6] S. Korjagin and P. Klachek, "Innovative development of intelligent transport systems based on biocybernetical vehicle control systems," *Transp. Res. Procedia*, vol. 20, pp. 326–333, Jan. 2017, doi: 10.1016/j.trpro.2017.01.038.

[7] D. Su, Z. Guo, Z. Li, and Y. Zhou, "Operation risk model and monitoring-warning system of expressway tunnels," *Transp. Res. Procedia*, vol. 14, pp. 1315–1324, Jun. 2016, doi: 10.1016/j.trpro.2016.05.204.

[8] L. Wang, Y. Li, and Z. Lin, "Design of intelligent power supply system for expressway tunnel," *IOP Conf. Ser., Earth Environ. Sci.*, vol. 108, Jan. 2018, Art. no. 052062.

[9] J. Sun, Z. Li, and J. Sun, "Study on traffic characteristics for a typical expressway on-ramp bottleneck considering various merging behaviors," *Phys. A, Stat. Mech. Appl.*, vol. 440, pp. 57–67, Dec. 2015.

[10] Y. Zhu, N. Gao, J. Wang, and C. Liu, "Study on traffic flow patterns identification of single intersection intelligent signal control," *Procedia Eng.*, vol. 137, pp. 452–460, Feb. 2016, doi: 10.1016/j.proeng.2016.01.280.

[11] L. Wang, M. Abdel-Aty, Q. Shi, and J. Park, "Real-time crash prediction for expressway weaving segments," *Transp. Res. C, Emerg. Technol.*, vol. 61, pp. 1–10, Dec. 2015.

[12] X. Xue, Y. Jia, and Y. Tang, "Expressway project cost estimation with a convolutional neural network model," *IEEE Access*, vol. 8, pp. 217847–217866, 2020.

[13] X. Jia, W. Zhou, T. Lei, L. Jing, and Y. Shen, "Impact analysis of expressway construction on ecological carrying capacity in the three-river headwater region," *J. Traffic Transp. Eng.*, vol. 7, no. 5, pp. 700–714, Oct. 2020.

[14] T. Chen, X. Shi, and Y. D. Wong, "Key feature selection and risk prediction for lane-changing behaviors based on vehicles' trajectory data," *Accident Anal. Prevention*, vol. 129, pp. 156–169, Aug. 2019.

[15] J. Wang, T. Luo, and T. Fu, "Crash prediction based on traffic platoon characteristics using floating car trajectory data and the machine learning approach," *Accident Anal. Prevention*, vol. 133, Dec. 2019, Art. no. 105320.

[16] L. Wang, M. Abdel-Aty, J. Lee, and Q. Shi, "Analysis of real-time crash risk for expressway ramps using traffic, geometric, trip generation, and socio-demographic predictors," *Accident Anal. Prevention*, vol. 122, pp. 378–384, Jan. 2019.

[17] L. Zheng and T. Sayed, "A novel approach for real time crash prediction at signalized intersections," *Transp. Res. C, Emerg. Technol.*, vol. 117, Aug. 2020, Art. no. 102683.

[18] Q. Cai, M. Abdel-Aty, J. Yuan, J. Lee, and Y. Wu, "Real-time crash prediction on expressways using deep generative models," *Transp. Res. C, Emerg. Technol.*, vol. 117, Aug. 2020, Art. no. 102697.

[19] L. Yu, B. Du, X. Hu, L. Sun, L. Han, and W. Lv, "Deep spatio-temporal graph convolutional network for traffic accident prediction," *Neurocomputing*, vol. 423, pp. 135–147, Jan. 2021.

[20] J. Fang, D. Yan, J. Qiao, J. Xue, and H. Yu, "DADA: Driver attention prediction in driving accident scenarios," *IEEE Trans. Intell. Transp. Syst.*, early access, Jan. 1, 2021, doi: 10.1109/TITS.2020.3044678.

[21] Y. F. Chiang, "Appling the GIS techniques to develop an accident analysis system for urban streets," M.S. thesis, Dept. Transp. Logistics Manage., Nat. Chiao Tung Univ., Hsinchu, Taiwan, 1996. [Online]. Available: https://ndltd.ncl.edu.tw/cgi-bin/gs32/gsweb.cgi/ccd=qB.K2S/record?r1=1&h1=1

[22] D. Gryaznov, "Scanners of the year 2000: Heuristics," in *Proc. 5th Int. Virus Bull.*, 1999, pp. 225–234. [Online]. Available: https://vx-underground.org/archive/VxHeaven/lib/adg00.html

[23] J. Han and M. Kamber, *Data Mining Concepts and Techniques*. San Mateo, CA, USA: Morgan Kaufmann, 2001, pp. 284–287.

[24] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.

[25] R. Y. M. Li, B. Tang, and K. W. Chau, "Sustainable construction safety knowledge sharing: A partial least square-structural equation modeling and a feedforward neural network approach," *Sustainability*, vol. 11, no. 20, p. 5831, Oct. 2019.

[26] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, no. 3, pp. 807–814.

[27] Y. Tian, J. Yuan, and H. Song, "Efficient privacy-preserving authentication framework for edge-assisted Internet of drones," *J. Inf. Secur. Appl.*, vol. 48, Oct. 2019, Art. no. 102354.

[28] P. Gope and B. Sikdar, "An efficient privacy-preserving authenticated key agreement scheme for edge-assisted Internet of drones," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 13621–13630, Nov. 2020.

Da-Jie Lin received the Ph.D. degree from the Department of Civil and Environmental Engineering, University of California at Berkeley, Berkeley, USA. He currently serves as an Assistant Professor for the Department of Transportation and Logistics, Feng Chia University, Taiwan, where he also serves as the Director for the Center for Advanced Transportation Management Systems. His research interests include transportation planning and management, intelligent transportation systems, elaboration and evaluation of traffic improvement proposal, and global logistics management/E-commerce.

Mu-Yen Chen (Member, IEEE) received the Ph.D. degree in information management from National Chiao Tung University, Taiwan. Previously, he was a Professor of information management with the National Taichung University of Science and Technology, Taiwan. He is currently working as an Associate Professor with National Cheng Kung University, Taiwan. His current research interests include artificial intelligence, soft computing, bio-inspired computing, data mining, deep learning, context-awareness, and machine learning, with more than 100 publications in prestigious venues, such as IEEE TRANSACTIONS ON FUZZY SYSTEMS, IEEE INTERNET OF THINGS (IOT) JOURNAL, IEEE SENSORS JOURNAL, IEEE ACCESS, *ACM Transactions on Internet Technology*, *Applied Soft Computing*, *Soft Computing*, *Neurocomputing*, *Computer Networks*, and *FGCS*. He has served as the Editor-in-Chief on International Journal of Big Data and Analytics in Healthcare, and an Associate Editor of IEEE ACCESS, *Granular Computing*, *Human-centric Computing and Information Sciences*, *Journal of Medical and Biological Engineering*, and *Journal of Information Science and Engineering*, while he is an editorial board member on several SCI journals.

Hsiu-Sen Chiang received the Ph.D. degree from the National Yunlin University of Science and Technology, Taiwan. He is currently a Professor with the Department of Information Management, National Taichung University of Science and Technology, Taiwan. His current research interests include data mining, petri nets, deep learning, and machine learning, biomedical science, and Internet marketing, with more than 50 publications in these areas, including *Applied Soft Computing*, *Information Fusion*, *Bioinformatics*, IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, *Journal of Medical Systems*, *Journal of Medical and Biological Engineering*, and *Information Fusion*. He is also an Associate Editor of *International Journal of Big Data and Analytics in Healthcare (IJBDAH)*.

Pradip Kumar Sharma (Senior Member, IEEE) received the Ph.D. degree in computer science and engineering from the Seoul National University of Science and Technology, South Korea, in August 2019.

He was a Software Engineer at MAQ Software, India, and involved on variety of projects. He is currently an Assistant Professor in cybersecurity with the Department of Computing Science, University of Aberdeen, U.K. He also worked as a Post-Doctoral Research Fellow with the Department of Multimedia Engineering, Dongguk University, South Korea. He is proficient in building largescale complex data warehouses, OLAP models, and reporting solutions that meet business objectives and align IT with business. He has published many technical research articles in leading journals from IEEE, Elsevier, Springer, and MDPI. His current research interests include the areas of cybersecurity, blockchain, edge computing, SDN, SNS, and the IoT security. He has also been invited to serve as a Technical Programme Committee Member and the Chair for several reputed international conferences, such as IEEE ICC2019, IEEE MENACOMM'19, and 3ICT 2019. He has served as an expert reviewer for IEEE TRANSACTIONS, Elsevier, Springer, and MDPI journals and magazines. He received a Top 1% Reviewer in computer science by the Publons Peer Review awards 2018 and 2019, Clarivate Analytics. He has served as a Guest Editor for international journals of certain publishers, such as Springer, Willey, MDPI, and JIPS. He is an Associate Editor of *Human-centric Computing and Information Sciences (HCIS)* and *Journal of Information Processing Systems (JIPS)* journals.