# BAYESIAN CLASSIFIER WITH *K*-NEAREST NEIGHBOR DENSITY ESTIMATION FOR SLOPE COLLAPSE PREDICTION

Min-Yuan Cheng[1], Nhat-Duc Hoang[1*], Nai-Wen Chang[1]

[1]*Dept. of Construction Engineering, National Taiwan University of Science and Technology,* #No.43, Sec. 4, Keelung Rd., Da'an Dist., Taipei City 106, Taiwan
*(*Corresponding author: ducxd85@yahoo.com)*

## ABSTRACT

Heavy rainfall and typhoon oftentimes cause the collapse of hillslopes across mountain roads. Disastrous consequences of slope collapses necessitate the approach for predicting their occurrences. In practice, slope collapse prediction can be formulated as a deterministic classification problem with two class labels, namely "collapse" and "non-collapse". Nevertheless, due to the criticality and the uncertainty of the problem, evaluating the collapse susceptibility of an area is a challenging task. This study proposes a novel Artificial Intelligence (AI) approach, named as *K*-Nearest Neighbor Based Bayesian Classifier (*K*-NNBC), to deal with slope collapse assessment. In the proposed model, Bayesian inference is used as a framework to achieve probabilistic prediction of slope collapse. Meanwhile, *K*-Nearest Neighbor (*K*-NN) is employed as a density estimation technique. Equipped with probabilistic outputs, the *K*-NNBC is able to yield predictions with different levels of confidence and diminish misclassified cases. Experimental results point out that the proposed model is very helpful for decision-makers in slope collapse assessment and disaster prevention planning.

## KEYWORDS

Slope Collapse Prediction; Bayesian Inference; *K*-Nearest Neighbor; Probabilistic Classification

## BACKGROUND

Road networks are integral to the infrastructure system. The reason is that they can support economic growth at both regional and national level by improving transportation efficiencies as well as by facilitating commercial activities and tourisms. Therefore, guaranteeing adequately serviceable roads is crucial for the enhancement of the national economy. Taiwan was formed by the collision action of Eurasia Plate and Philippine Sea Plate; and it is relatively young in geological age (Ching & Liao 2006). The island is located in the vicinity of the Pacific Ring of fire which frequently experiences seismic activities. Taiwan's topography is characterized by mountainous areas in the east and lowland plains in the west. Additionally, the region has a subtropical climate and high levels of precipitation. Therefore, earthquakes and typhoons are absolutely not unusual natural hazards within the country.

Over the past decades, an extensive network of mountain roads has been built to catch up with population expansion and economic development (Ching et al. 2011; Yang et al. 2012). Natural hazards coupled with rugged terrain lead to the fact that slope collapses may occur in many mountain roads. These catastrophic events are often triggered by earthquakes or heavy rainfalls during typhoons or monsoon storms (Lin et al. 2009; Nefeslioglu et al. 2010). Slope collapses are very undesirable since they inflict damages to man-made structures, disruption of traffic, and loss of human lives. As a consequence, slope stability analysis is an inevitable task which should be conducted regularly by roadway maintenance authorities (Cheng & Ko 2003; Das et al. 2011). The results of analysis can be utilized for identifying collapse-prone areas as well as allocating scarce resources in order to establish an overall disaster prevention program (Cheng et al. 2012). Currently, machine learning approaches have demonstrated their feasibility as well as effectiveness in slope stability analysis (Sakellariou & Ferentinou 2005).

Machine learning approaches are established based on several AI techniques and historical databases (Bishop 2006). Using this method, the slope collapse prediction can be formulated as a binary classification problem in which prediction outputs are either "collapse" or "non-collapse". Therefore, many classification techniques, such as Artificial Neural Networks and Support Vector Machines, have

shown to be feasible when applying in the problem at hand (Cheng et al. 2012). By learning the events in the past, the AI based models reduce the dependency on human judgments and can yield predictive results based on information of the new input patterns.

Nevertheless, the slope collapse prediction problem is not only inherently complex but also highly uncertain. One can never possess absolute confidence to argue where and when a slope will fail because there are numerous uncertainties involved. Moreover, in some circumstances, wrong classifications can be very costly if committed. Consequently, simple binary classifications are inadequate for the decision-making process. It is because they do not exhibit the uncertainties associated with the predicted outcomes. Thus, it is much more beneficial and reliable to employ a mechanism for probabilistic prediction of slope collapse in which the classification results are expressed in terms of probabilities instead of binary values.

The classifier based on Bayesian framework is an effective probabilistic approach for prediction. The Bayesian classifier is found to possess a number of advantages, such as flexibility in modeling, capability of coping with uncertainty, and resilience to noise (Langley & Sage 1994). Experimental studies in a variety of fields have revealed that the method can deliver competitive prediction performances with low computational (Domingos & Pazzani 1997). Nevertheless, very few previous studies have dedicated in investigating the capability of the Bayesian approach in predicting slope stability. Therefore, this study put forward a novel AI approach, named as *K*-Nearest Neighbor Based Bayesian Classifier (*K*-NNBC), for slope collapse assessment. In the new model, Bayesian inference is used as a framework to achieve probabilistic classifications. Meanwhile, *K*-NN algorithm is incorporated for density estimation.

## METHODOLOGY

### Bayesian Framework for Classification

In machine learning, the goal of classification is to assign an object to one of $M$ discrete classes $C_k$ where $k = 1,\ldots, M$. The input space is thereby divided into several decision regions by the decision boundaries (Bishop 2006). To classify the object based on the evidence provided by its feature vector $X$, it is requisite to obtain the conditional probability $P(C_k|X)$, which represents how likely the input $X$ belongs to the class $C_k$. Accordingly, the object will be assigned to the class with largest conditional probability. Hence, in the context of Bayesian theorem, the conditional probability $P(C_k|X)$ is computed as following (Bishop 2006; Duda et al. 2001):

$$P(C_k \mid X) = \frac{P(X \mid C_k) \times P(C_k)}{P(X)} \tag{1}$$

where $P(C_k \mid X)$ represents the posterior probability of the class $C_k$. Meanwhile, $P(X \mid C_k)$ is called the likelihood which is the class-conditional probability density function of the feature $X$. $P(C_k)$ denotes the prior probability of the class $C_k$. And, $P(X)$ represents the evidence factor. The evidence factor can be viewed as a scale factor used to ensure that the posterior probabilities sum to one (Duda et al. 2001). It can be is calculated as following:

$$P(X) = \sum_{k=1}^{M} P(X \mid C_k) \times P(C_k) \tag{2}$$

As shown in Eq. 1, the structure of Bayesian classification relies upon the prior probabilities $P(C_k)$ and the conditional densities $P(X|C_k)$ (Theodoridis & Koutroumbas 2009). The first quantity can be estimated directly from the distribution of the training samples among classes (Clark & Niblett 1989). If $N$ is the total number of available training cases and $N_k$ is the number of cases belonging to the class $C_k$, then the prior probability of this class is calculated as $P(C_k) = N_k/N$. The next step is to derive the class-conditional density $P(X|C_k)$. The $P(X|C_k)$ describes the distribution of the feature vector $X$ in each class. This conditional density is also known as the likelihood function of $C_k$ with respect to $X$.

Herein, we consider the problem in which the pattern $X$ represents a $D$-dimensional vector, and each attribute of $X$ is denoted as $X_j$ where $j = 1, \ldots, D$. Thus, to derive the likelihood function $P(X|C_k)$, the common approach is to assume that the probability distributions of attributes $X_j$, within each class, are

independent of each other. In this case, the classification approach is known as the Naïve Bayesian Classifier (Bishop 2006). Accordingly, the class-conditional density can be computed as following:

$$P(X \mid C_k) = \prod_{j=1}^{D} P(X_j \mid C_k) \tag{3}$$

where $P(X_j|C_k)$ denotes the probability distribution of the attributes $X_j$ within each class $C_k$. In addition, the density $P(X_j|C_k)$ is often assumed to be a Normal distribution.

Needless to say, the assumption that the probability distributions for attributes are independent of each other can be not realistic. The reason is that correlations among attributes are not unusual. Additionally, a Normal distribution may not be the most appropriate approximation. It is because the true probability density function can possibly be multi-modal and it can take an arbitrary form. When the aforementioned assumptions are violated, the performance of the Bayesian Classifier can be unquestionably degraded.

### *K*-NN Approach for Density Estimation

Density estimation is the task of modeling a probability density function given a finite number of data instances. Among the methods for approximating density function, the *K*-NN is attractive because of its flexibility as well as simplicity. The technique does not require any assumption of the functional form for the modeled density.

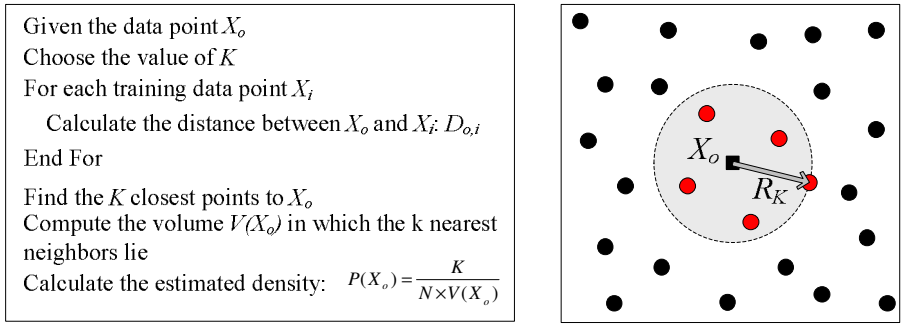


Fig. 1 *K*-NN for Density Estimation

This section of the paper reviews the *K*-NN algorithm for density estimation (Theodoridis & Koutroumbas 2009). Consider a set of *N* data points, $X_1, X_2, \ldots, X_N \in R^D$ generated from an unknown statistical distribution, the goal is to estimate the value of the unknown density function at a given point $X_o$. The procedure of *K*-NN approach for approximating probability density function is illustrated in Fig. 1. In this method, the volume of surrounding the estimation point $X_o$ is enlarged until it encloses *K* neighbors. The value of *K* can be selected by a general rule: $K \approx \sqrt{D}$ where *D* is the dimension of the data. After the *K* neighbors have been identified, we can compute the volume of the hyper-sphere surrounding $X_o$ as follows:

$$V(X_o) = C \times R_K \tag{4}$$

where $R_K$ is the distance between the estimation point and its $K^{th}$ closet neighbor (see Fig. 1). The quantity *C* is the volume of the unit sphere in *D*-dimensions, which is calculated as following:

$$C = \begin{cases} \dfrac{\pi^{D/2}}{(\frac{D}{2})!} & \text{if } D \text{ is even} \\ \dfrac{2^{(D+1)/2}\pi^{(D-1)/2}}{D!!} & \text{if } D \text{ is odd} \end{cases} \tag{5}$$

Accordingly, the estimated density $P(X_o)$ is derived as follows:

$$P(X_o) = \frac{K}{N \times V(X_o)} \quad (6)$$

where $K$ is the number of neighbors; $N$ is the total number of data point in the training set; $V(X_o)$ denotes the volume of the hyper-sphere enclosing $X_o$ and its neighbors.

## *K*-NN BASED BAYESIAN CLASSIFIER FOR SLOPE COLLAPSE PREDICTION

Fig. 2 provides the overall picture of the proposed model *K*-NNBC which is divided into 7 steps. It is noted that the model is established based on the Bayesian inference and the *K*-NN approach.

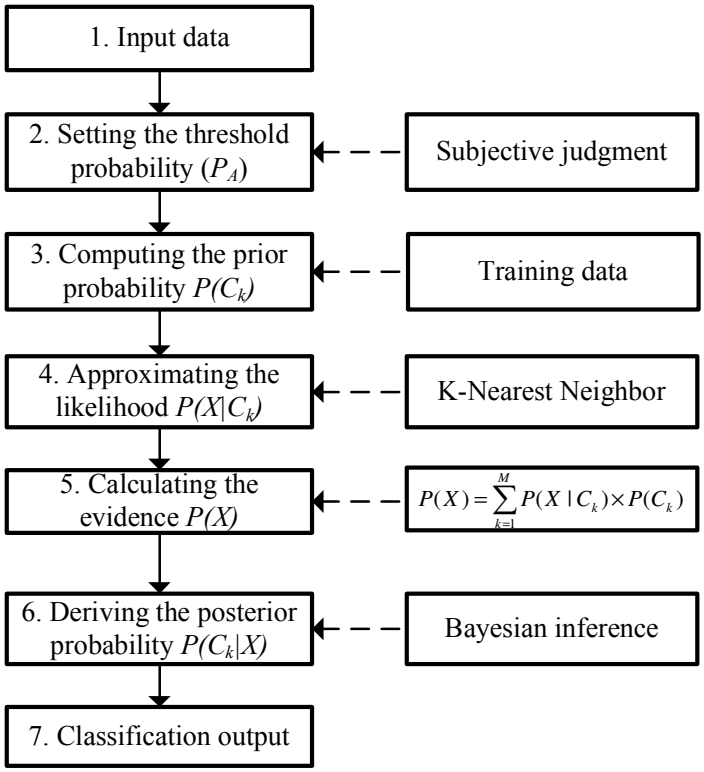


Fig. 2 *K*-Nearest Neighbor Based Bayesian Classifier

**(1) Input data**: The input data provides the pattern of an area that is under investigation of slope stability. The data attributes consist of influencing factors that impose significant impacts on the slope collapse events.

**(2) Setting the threshold probability**: In this study, we employ a threshold probability, denoted as $P_A$, to determine whether an input pattern is accepted to be in one class. Specifically, if the posterior probability of the class $P(C_k|X)$ is greater than or equal to the threshold probability, the input pattern $X$ is accepted to be classified into the class. The value of $P_A$ is selected based on the criticality of the problem and the subjective judgment of the analyst (Bishop 2006). In some critical cases, one may set a high value (e.g. 0.9) to avoid wrong classifications made by the machine.

**(3) Computing the prior probability**: The prior probability can be estimated from the relative frequency of each class in the training data set.

**(4) Approximating the likelihood**: This step aims to compute the class-conditional density $P(X|C_k)$. To do so, all of the data instances associated with the class label $C_k$ is extracted; and the *K*-NN is employed to estimate the $P(X|C_k)$. As mentioned earlier, the technique does not require any prior information of the

estimated probability density function. The estimation is determined entirely by the characteristic of the training data.

**(5) Calculating the evidence**: The evidence factor is used to scale the posterior probability into the range of 0 and 1. Because the slope collapse prediction is a two-class classification problem, the evidence is calculated as following:

$$P(X) = P(X \mid C_1)P(C_1) + P(X \mid C_2)P(C_2) \tag{7}$$

where $P(C_1)$ and $P(C_2)$ are the prior probabilities of the class 1 and class 2, respectively. Those prior probabilities are computed at the step 3 of the model. Meanwhile, $P(X|C_1)$ and $P(X|C_2)$ are the two the class-conditional densities that are estimated by the *K*-NN.

**(6) Deriving the posterior probability**: In this step, the posterior probability of each class given an input pattern *X* is given as follows:

$$P(C_k \mid X) = \frac{P(X \mid C_k)P(C_k)}{P(X)} \tag{8}$$

**(7) Classification output**: Given the threshold probability ($P_A$) and the derived posterior probability of each class, the classification outcome can be obtained. Fig. 3 illustrates the classification decision based on $P_A$. The decision rule of the proposed model can be expressed as follows:

$$\begin{cases} X \text{ belongs to } C_k, \text{ if } P(C_k \mid X) \geq P_A \\ \text{Otherwise, } X \text{ remains unclassified.} \end{cases} \tag{9}$$

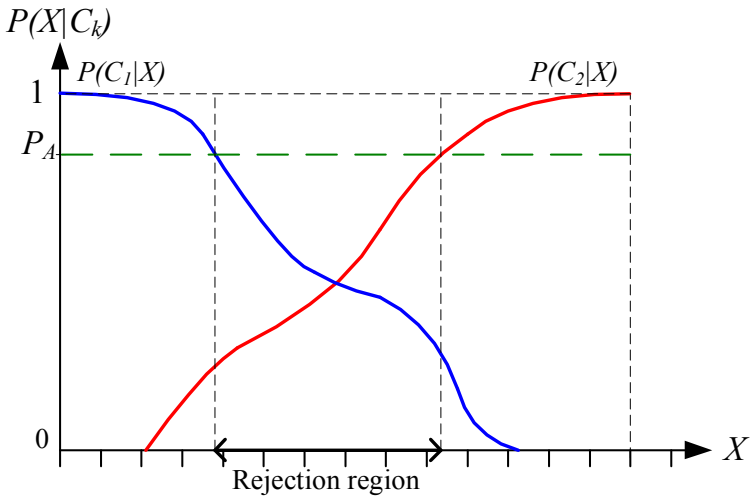


Fig. 3 Classification decision based on $P_A$

## MODEL APPICATION

### Historical data

The historical database utilized in this article contains 211 slope evaluation samples collected in the Provincial Highway No. 18 and No. 21. Within the database, there are 105 failure cases and 106 non-failure cases. For slope collapse prediction, this study employs 16 influencing factors divided into 9 groups: landforms, geological structure, stratigraphy, rock properties, vegetation coverage, water

condition, road properties, earthquake, and rainfall. Table 1 provides the information of the influencing factors and their statistical descriptions.

Table 1 Influencing factors and statistical description

| Group | Note | Description | Min | Average | Std. Dev. | Max |
|---|---|---|---|---|---|---|
| Landforms | IF1 | Slope aspect | 0.0 | 154.1 | 99.8 | 355.0 |
| | IF2 | Slope gradient | 10.0 | 60.7 | 14.2 | 90.0 |
| | IF3 | Slope height | 5.0 | 20.7 | 16.8 | 150.0 |
| | IF4 | Slope form | -55.4 | 1.0 | 18.7 | 50.0 |
| Geological Structure | IF5 | Formation type | 1.0 | 4.8 | 0.9 | 6.0 |
| Stratigraphy | IF6 | Angle between slope aspect and trend | 0.0 | 89.7 | 32.1 | 180.0 |
| | IF7 | Angle between gradient and inclination | -20.0 | 46.2 | 23.8 | 85.0 |
| Rock Properties | IF8 | Rock mass size | 0.1 | 0.5 | 0.4 | 2.5 |
| | IF9 | Rock mass volume | 20.0 | 68.0 | 17.2 | 100.0 |
| Vegetation coverage | IF10 | Vegetation coverage percentage | 5.0 | 59.8 | 27.6 | 98.0 |
| | IF11 | Vegetation coverage thickness | 0.1 | 1.4 | 1.1 | 4.5 |
| Water condition | IF12 | Catchment area | 197.0 | 23553.5 | 69365.2 | 888751.0 |
| Road Properties | IF13 | Excavation height at slope toe | 0.0 | 4.7 | 5.4 | 50.0 |
| | IF14 | Change of slope gradient due to toe cutting | 0.0 | 9.8 | 11.6 | 45.0 |
| Earthquake | IF15 | Maximum ground acceleration | 0.0 | 249.1 | 84.1 | 391.9 |
| Rainfall | IF16 | Maximum accumulated rainfall | 384.6 | 1238.0 | 525.8 | 1947.3 |

## Experimental Results

Among 211 samples in the database, 171 samples are used to train the prediction model and 40 samples are reserved for the testing process. After the training process, the proposed classification model (*K*-NNBC) can be utilized to forecast new input patterns from the testing set. Detailed prediction results of the *K*-NNBC for testing are illustrated in Table 2. In this table, $P(C_1|X)$ and $P(C_2|X)$ represent the posterior probability of collapse and non-collapse, respectively.

Table 2 *K*-NNBC prediction result for testing cases

| Case | $P(C_1|X)$ | $P(C_2|X)$ | Actual output | Predicted output | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | $P_A$= 0.5 | $P_A$= 0.6 | $P_A$=0.7 | $P_A$=0.8 | $P_A$=0.9 |
| 1 | 0.011 | 0.989 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0.002 | 0.998 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0.647 | 0.353 | 1 | 1 | 1 | x | x | x |
| 4 | 0.532 | 0.468 | 1 | 1 | x | x | x | x |
| 5 | 0.004 | 0.996 | 0 | 0 | 0 | 0 | 0 | 0 |
| … | … | … | … | … | … | … | … | … |
| 38 | 0.005 | 0.995 | 0 | 0 | 0 | 0 | 0 | 0 |
| 39 | 0.004 | 0.996 | 0 | 0 | 0 | 0 | 0 | 0 |
| 40 | 0.993 | 0.007 | 1 | 1 | 1 | 1 | 1 | 1 |

Moreover, the classification outputs corresponding to different level of the threshold probability $P_A$ are also provided. As mentioned earlier, the value of $P_A$ is used to derive the final classification result of the input pattern. It is also worth reminding that if the posterior probability of the class $P(C_k|X)$ is greater than or equal to $P_A$, the input pattern $X$ is classified into the class $C_k$. Therefore, if the posterior probabilities of both classes cannot surpass the threshold value, then the input $X$ remains unclassified. In Table 2, outputs of unclassified patterns are denoted with the symbol 'x'.

Furthermore, to better verify the capability of the new approach, its performance is compared to results obtained from the Naïve Bayesian Classifier (NBC). In the NBC, the probability distributions of attributes within each class are considered to be independent of each other. Moreover, in this method, the class-conditional densities are assumed to be Normal distributions which parameters (mean and variance) are estimated directly from the training samples. The detail of result comparison is provided in Table 3.

Table 3 Result comparison

| Case | $P_A$ | 0.5 | | 0.7 | | 0.9 | |
|---|---|---|---|---|---|---|---|
| | Model | NBC | *K*-NNBC | NBC | *K*-NNBC | NBC | *K*-NNBC |
| Training | Classified cases | 171 | 171 | 169 | 166 | 159 | 154 |
| | Unclassified cases | 0 | 0 | 2 | 5 | 12 | 17 |
| | Misclassified cases | 6 | 4 | 4 | 2 | 2 | 1 |
| | Accuracy Rate (%) | 96.5 | 97.7 | 97.6 | 98.8 | 98.7 | 99.4 |
| Testing | Classified cases | 40 | 40 | 37 | 34 | 36 | 33 |
| | Unclassified cases | 0 | 0 | 3 | 6 | 4 | 7 |
| | Misclassified cases | 4 | 2 | 1 | 0 | 1 | 0 |
| | Accuracy Rate (%) | 90.0 | 95.0 | 97.3 | 100.0 | 97.2 | 100.0 |

Observable in Table 3, if the threshold $P_A$ is set to be 0.5, the NBC misclassifies 4 testing cases and thus, its accuracy rate is 90%. Meanwhile, the number of misclassifications and accuracy rate of K-NNBC for testing process is 2 and 95%, respectively. It is noted that if the $P_A$ is set to be 0.7 or larger values, the proposed *K*-NNBC achieves 100% of prediction accuracy in the testing process. On the other hand, even with high values of threshold probability, the NBC still committed misclassified cases.

**Discussions**

Essentially, with different values of the threshold probability, the classifier is allowed to deliver predictions with different level of confidence. The capability of exhibiting and adjusting confidence can be pivotal in slope collapse prediction. The reason is that, for critical areas, it is reliable as well as beneficial to set a high value of threshold probability to minimize the risk of erroneous classification. Typically, when $P_A$ = 0.7, the proposed method can predict accurately 85% of the cases, and leaves 6 sensitive cases for expert decision. In this way, it can be seen that the method can help reduce the effort of human expert and thus boost the productivity of the analyzing process.

Furthermore, the prediction outputs expressed in terms of probabilities can be inferred as an index for prioritizing different areas in mountainous regions. This function can be helpful for the decision-making process since there can be a large amount of man-made structures that are susceptible to damages caused by slope collapse. Moreover, the workforces as well as financial resources for constructing retaining structures are definitely limited. Therefore, Government agencies can utilize the proposed model as a tool for resource allocation to set up an optimal disaster prevention program.

## CONCLUSION

This paper has presented and verified a new prediction model, named as *K*-NNBC, to assist decision-makers in slope collapse prediction. The proposed model is developed by the fusion of the Bayesian inference framework and the *K*-NN approach. The *K*-NNBC utilizes the Bayesian framework to compute the posterior probability of slope collapse event given an input pattern that provides features of the investigated area. Furthermore, *K*-NN is employed to approximate the class-conditional probability density without any assumption of the form of the density. The proposed model also does not require the assumption of independent attributes which can be delusive in the real-world situation. Superior prediction accuracy in the experiment process have convincingly proved the capability of the new approach for supporting decision-makers in slope collapse prediction as well as in disaster prevention planning.

## REFERENCES


Bishop, C. (2006). Pattern Recognition and Machine Learning. *Springer Science+Business Media.*

Cheng, M.-Y., A. F. V. Roy & K.-L. Chen (2012). Evolutionary risk preference inference model using fuzzy support vector machine for road slope collapse prediction. *Expert Systems with Applications*, 39 (2), 1737-1746. doi: 10.1016/j.eswa.2011.08.081

Cheng, M. Y. & C. H. Ko (2003). Automated Safety Monitoring and Diagnosis System for Unstable Slopes. Computer-Aided Civil and Infrastructure Engineering, 18 (1), 64-77. doi: 10.1111/1467-8667.t01-1-00300

Ching, J. & H.-J. Liao (2006). Predicting landslides probabilities along mountain road in Taiwan, *in Proceedings of the TAIPEI2006 International Symposium on New Generation Design Codes for Geotechnical Engineering Practice, Nov. 2~3, 2006, Taipei, Taiwan*.

Ching, J., H.-J. Liao & J.-Y. Lee (2011). Predicting rainfall-induced landslide potential along a mountain road in Taiwan. *Geotechnique 61, No. 2, 153–166*. doi: 10.1680/geot.8.P.119.3740

Clark, P. & T. Niblett (1989). The CN2 Induction Algorithm. *Machine Learning*, 3 (4), 261-283. doi: 10.1023/a:1022641700528

Das, S., R. Biswal, N. Sivakugan & B. Das (2011). Classification of slopes and prediction of factor of safety using differential evolution neural networks. *Environmental Earth Sciences*, 64 (1), 201-210. doi: 10.1007/s12665-010-0839-1

Domingos, P. & M. Pazzani (1997). On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning* , 29 103–130. doi: 10.1.1.40.1930

Duda, R. O., P. E. Hart & D. G. Stock (2001). Pattern Classification, 2nd Edition. *John Wiley & Sons*.

Langley, P. & S. Sage (1994). Induction of selective Bayesian classifiers. *In R. Lopez de Mantaras & D. Poole (Eds.), in Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence (pp. 399–406). San Francisco, CA: Morgan Kaufmann.*

Lin, H.-M., S.-K. Chang, J.-H. Wu & C. H. Juang (2009). Neural network-based model for assessing failure potential of highway slopes in the Alishan, Taiwan Area: Pre- and post-earthquake investigation. *Engineering Geology*, 104 (3-4), 280-289. doi: 10.1016/j.enggeo.2008.11.007

Nefeslioglu, H. A., E. Sezer, C. Gokceoglu, A. S. Bozkir & T. Y. Duman (2010). Assessment of Landslide Susceptibility by Decision Trees in the Metropolitan Area of Istanbul, Turkey. *Mathematical Problems in Engineering*. doi:10.1155/2010/901095

Sakellariou, M. G. & M. D. Ferentinou (2005). A study of slope stability prediction using neural networks. *Geotechnical & Geological Engineering*, 23 (4), 419-445. doi: 10.1007/s10706-004-8680-5

Theodoridis, S. & K. Koutroumbas (2009). Pattern Recognition. *Academic Press, Elsevier Inc*,

Yang, S., C. Shen, C. Huang, C. Lee, C. Cheng & C. Chen (2012). Prediction of Mountain Road Closure Due to Rainfall-Induced Landslides. *Journal of Performance of Constructed Facilities*, 26 (2), 197-202. doi:10.1061/(ASCE)CF.1943-5509.0000242