参考著作No.1

06-12-2019 | Research Article

Estimating landslide occurrence via small watershed method with relevance vector machine

Journal: Earth Science Informatics

Authors: Kuo-Wei Liao, Nhat-Duc Hoang, Shih-Chun Chang

参考著作No.1

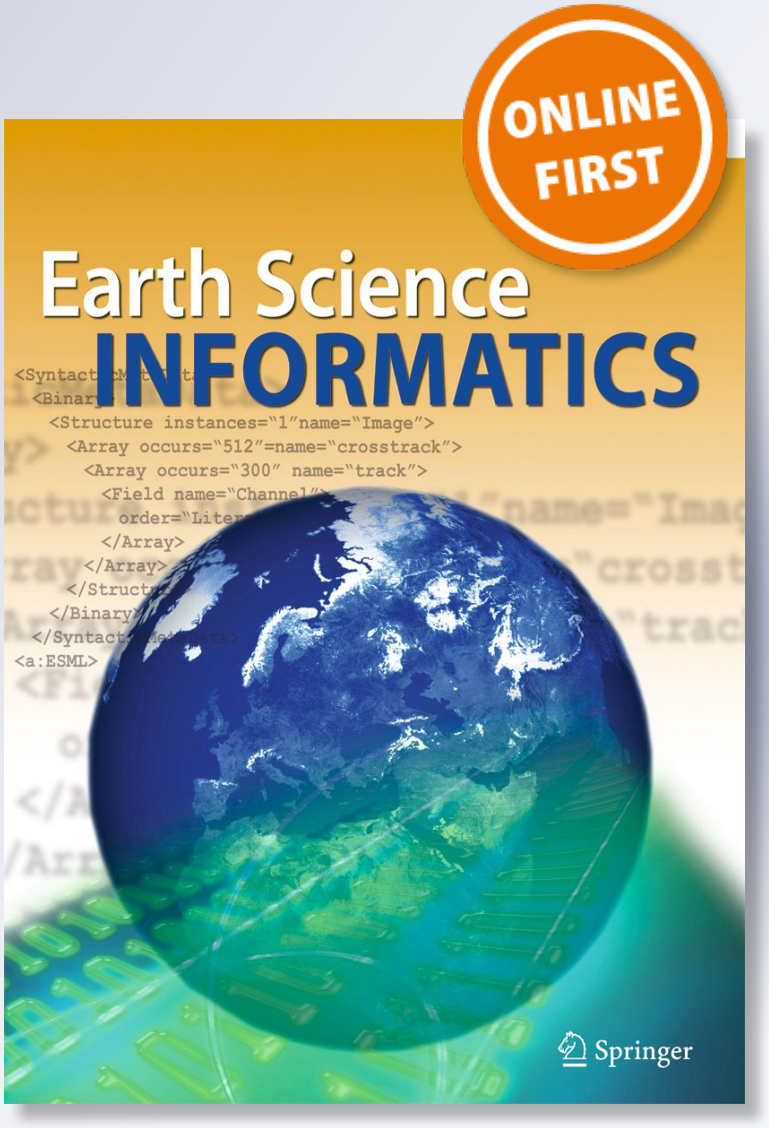
Estimating landslide occurrence via small watershed method with relevance vector machine

Kuo-Wei Liao, Nhat-Duc Hoang & Shih-Chun Chang

Earth Science Informatics

ISSN 1865-0473

Earth Sci Inform
DOI 10.1007/s12145-019-00419-7



Springer

Earth Science Informatics
https://doi.org/10.1007/s12145-019-00419-7

RESEARCH ARTICLE

Estimating landslide occurrence via small watershed method with relevance vector machine

Kuo-Wei Liao[1] · Nhat-Duc Hoang[2] · Shih-Chun Chang[3]

Received: 23 May 2019 / Accepted: 1 October 2019

Abstract

The mechanism of landslide occurrence is complicated due to the dependency and nonlinear relationship of various physical factors. A promising prediction model, which can be used to locate the high-risk regions and a corresponding prevention strategy can be prepared to reduce the slide occurrence and its consequence, is therefore desired. To perform the landslide assessment for a large-scale slope, this study proposes to use the method of small watershed that is integrated with Relevance Vector Machine (RVM) to enhance the prediction accuracy. Effect of physiographic and hydrological factors such as slope steepness, dip slope ratio, landslide ratio, and cumulative rainfall are investigated. To estimate the occurrence of landslide, RVM first maps the aforementioned factors into a feature space using Gaussian radial basis function. A linear boundary, distinguishing landslide or not, is then obtained through the search of the optimal weights. To find these weights, a Bayesian theory-based optimization problem is formulated and solved by iteratively reweighted least squares algorithm and the Laplace approximation procedure. The proposed model is validated by the data collected from Kaoping River Basin. Results indicate that the proposed RVM-based small watershed approach possesses a prediction accuracy of 87.5%, which is better than those of using Support Vector Machine (SVM), Least-Square Support Vector Machine (LS-SVM), and logistic regression, providing authorities in their hazard alert system to minimize the life or property losses caused by the landslide.

Keywords Small watershed, landslide · RVM · Physiographic factor · Hydrological factor

Introduction

Varnes (1978) divided landslides into five basic patterns: falls, topples, slides, spreads and flows. In terms of the physical mechanisms, landslides are the processes where the rocks, soils and debris slide down a slope (Cruden, 1991). In general, landslides can be classified according to the action patterns, material type and the depth of sliding surface. When classified by the depth, it can be categorized as shallow landslide or deep-seated landslide. In Taiwan, the landslide hazards often cause loss of lives and properties. Thus, the alert and prevention measures against landslides are key issues to be dealt with. When estimating the occurrence of landslide, one needs to clarify the reasons causing the landslide. Often, the landslide occurrence mechanism includes internal and external factors. Internal factors refer to the physiographic factors such as elevation difference, mean elevation, slope steepness, road ratio, dip slope ratio, landslide ratio, and distance to fault, etc.; whereas, external factors refer to the triggering mechanisms such as precipitation, earthquake or volcanic activities, etc. in which, the water content is the most important factor (Highland and Bobrowsky, 2008). The higher the landslide potential, the less the water content required for triggering the landslide. Therefore, the physiographic and hydrological factors of a hillslope area are often considered to directly affect the occurrence of landslides (Chan et al. 2015). In addition to the hydrological factor, the impacts of physiographic factors on landslides are important and described below.

Communicated by: H. Babaie

✉ Kuo-Wei Liao
kliao@ntu.edu.tw

[1] Department of Bioenvironmental Systems Engineering, National Taiwan University, Taipei, Taiwan

[2] Lecturer, Faculty of Civil Engineering, Institute of Research and Development, Duy Tan University, 254 Nguyễn Văn Linh, P, Thanh Khê, Đà Nẵng 550000, Vietnam

[3] Department of Bioenvironmental Systems Engineering, National Taiwan University, Da'an District, Taipei City, Taiwan 10617

Published online: 06 December 2019

In view of force equilibrium, stronger downward sliding force will be generated when the slope is steeper. Dai et al. (1999) pointed out that slopes with a slope of 30–40 degrees were most vulnerable to damage during the rainfall-induced landslides. For the Chi-Chi Earthquake in Taiwan, Hung et al. (2000) discovered that 90% of landslide took place on slope faces higher than 45 degrees and it indicated the steeper the slope, the higher the damage potential. Dai and Lee (2002) discovered that if the slope is less than 40 degrees, the landslide tends to increase along with the increase of slope; and most landslides are seen at the slope between 35~40 degrees.

The dip slope means the dip direction of a natural slope is roughly identical with the true dip or apparent dip of the underlying strata. Landslide might not occur for dip slope, but a destructive dip slope usually exhibits extremely high sliding speed (potentially exceeding 100 km per hour); therefore, dip slope sliding is one of the most serious landslide events in Taiwan. The dip slope usually brings higher impact to the slope hazard. Especially, when the footing of the dip slope is disturbed or removed, there is a higher possibility of a landslide occurrence. In view of this, higher landslide potential may exist in an area with higher percentage of dip slopes in Taiwan where the mountain areas are densely developed by people.

Based on site investigation, if the overlay of a landslide area is not fully consolidated or the vegetation has not been restored, a higher landslide reoccurrence may exist. According to Chang et al. (2007), 75% of large-scale landslides during the Ji-Ji Earthquake is found in the areas that have landslides before. Fan et al. (2018) classified the surface coverage as: (1) Former landslide land. (2) Naked land. (3) Construction land. (4) Forestry land. (5) Farming land and (6) Water body for evaluating the landslides and debris occurrence. He found that a promising prediction is often obtained when the landslide ratio (i.e. number of having landslide in the past) is used to replace the surface coverage factors.

In Taiwan, the strata are mostly formed by folding structure and are easily affected by the weathering effect and lead to the occurrence of landslide. Chang (1987) divided the sedimentary rocks, igneous rocks and metamorphic rocks into soft stratum and hard stratum according to lithologic characters. The soft stratum includes inter-layer shales, alluvium, laterite accumulation, laterite gravel layer, shales or muddy rocks, porous limestone, tuff, pyroclastic rock, slate, phyllite, black schist and green schist, etc. Firm and highly consolidated stratum can resist against landslides, but weaker and highly fractured stratum are vulnerable to landslide. Because the fault is more fractured on geological side, therefore the fault distance, which is a distance from the considered site to the nearest fault, are often used as an indicator for the lithologic character.

Often, the process of landslide occurrence is complicated and involves various factors (Zhou et al. 2018). Establishing a landslide model at a regional scale is therefore, a challenging task. In light of this, AI is a reasonable choice if one has enough statistic data and in this case, no governing equations are needed. The built AI model has a potential to be utilized to investigate effects of topographic, climatic, and human-related factors on the landslide susceptibility in a systematic manner. Previous works have shown that the accuracy of conventional statistical approaches may not be sufficient due to the multivariate and nonlinear nature of the problem of interest (Pham et al. 2017, Tien Bui et al. 2012). Therefore, AI has drawn considerable attentions of researchers in modeling of landslide and other natural hazards (Kavzoglu et al. 2019).

Recently, geographic information system (GIS) is widely employed in disaster prevention research such as landslide evaluation due to its ability to incorporate with large-scale dataset with multi-layered information of spatial characteristics (Le et al. 2018). It is seen that GIS based data-driven AI approach has a great potential in landslide prediction due to its ability to provide spatial prediction for the interested measurements (i.e., the landslide occurrence in this study) which provides a chance of identifying vulnerable areas and thus is helpful for land-use planning (Jaafari et al. 2019). It is seen AI based models possess several outstanding capabilities and can be incorporated with other powerful tools such as GIS. Thus, exploring AI methods for landslide prediction is highly necessary and is one of the goals for the current study.

This study proposes a novel landslide occurrence estimation algorithm that integrates Relevance Vector Machine (RVM) and Bayesian-based optimization to implement the landslide occurrence assessment. RVM, proposed by Tipping (Tipping 2000), is a Bayesian inference approach for constructing probabilistic classification model. Compared to the Support Vector Machines (SVM) and Least Squares Support Vector Machines (LSSVM), the Bayesian-based RVM has fewer tuning parameter and therefore, is more efficient in computation. In addition, a RVM model often has a better performance in generalization due to its better sparseness property. That is, a RVM model only employs a small number of relevant vectors from the training samples to construct the classification boundary (Tipping 2004). This sparseness property is greatly useful because a sparse model is fast to establish and less susceptible to overfitting (Tzikas et al. 2006; Cheng and Hoang 2015). Recently, successful implementations of the RVM have been reported in various fields (Tien Bui et al. 2018, Samantaray et al. 2019, Abbas and Tezcan 2019, Liu et al. 2019). Nevertheless, the applications of the RVM in landslide occurrence prediction are still limited and is one of the focuses for the current study.

To estimate the occurrence of landslide, RVM first maps the physiographic and hydrological factors into a feature space using Gaussian radial basis function. A linear boundary, distinguishing landslide or not, is then obtained through the search of the optimal weights. To verify the proposed algorithm, historical records collected from Kaoping River basin

are used. Kaoping River basin is featured on its violent drop in altitude that the elevation is diminishing from northeast to southwest in constituting nearly 4000 m of height difference. Mountains with altitudes over 1000 m account for 47.8% of the basin coverage and that between 100 m to 1000 m takes 30.3% of the basin coverage. Kaohsiung and Pingtung plains are the lowest areas that are below 100 m in altitude, making up 21.9% of basin coverage. The collected slope steepness, the dip slope ratio, the landslide ratio and cumulative rainfall are integrated into the proposed RVM-based small watershed model to assess the slide occurrence. Results indicate that the proposed method possesses a prediction accuracy of 87.5%, which is better than that of using standard Support Vector Machine (SVM), Least-Square Support Vector Machine (LS-SVM) and logistic regression, providing authorities in their hazard alert system to minimize the life or property losses caused by the landslide. Details of the proposed method are provided below.

Support vector machine, Least-Square support vector machine and the employed relevance vector machine approach

Before conducting the estimation of landslide occurrence using the proposed approach (i.e., RVM), this study also trains two additional models, which are SVM (Ge et al. 2018) and LS-SVM, to evaluate/compare the performances of the proposed RVMs. These two models (SVM and LS-SVM) are briefly introduced below. SVM is a classifier that is able to solve a nonlinear problem using convex quadratic programs (QP), as shown in Eq. (1).

$$\underset{w,b,\xi}{\text{minimize}}\ \frac{1}{2}w^T w + c\sum_{k=1}^{N}\xi_k \text{ Subject to } \begin{cases} y_k\left(w^T K(x_i)+b\right)\geq 1-\xi_k \\ \xi_k \geq 0, i=1,2,...,N \end{cases} \tag{1}$$

in which y_k represents the class and $[w^T K(x_i) + b]$ indicates the classifier, w is a vector of weights that are orthogonal to the hyper-plane; c is constant number that is greater than zero; and ξ_k is the slack variable. When ξ_k is greater than one, indicating that the k-th inequality is violated. N is the number of data, and K is the kernel function, in which Gaussian radial basis function (RBF) is one of the common kernels and is adopted here, as shown in Eq. (2).

$$K(X, X_i) = e^{-\sigma\left(\|X-X_i\|\right)^2} \tag{2}$$

in which vector of X is input, σ represents kernel function parameter; and X_i are the support vectors. The least-square support vector machine, (LS-SVM, Suykens et al. 2002) does not attempt to solve the QP problem. LS-SVM actually try to solve a system of linear equations after altering the SVM via introducing the error variable (ε), as described in Eq. (3).

$$\begin{aligned} &\min\ \frac{1}{2}w^T w + \frac{\gamma}{2}\sum_{k=1}^{N}\varepsilon_k^2 \\ &\text{s.t.} y_k(w\cdot K(x_k)+b) = 1-\varepsilon_k, k=1,...,\text{n} \end{aligned} \tag{3}$$

in which γ is a constant number. It is seen that two modifications, equality constraints and a squared error variable, leading to solving a set of linear equations in LS-SVM.

Normally, the time-variance factors are not considered in most of the landslide occurrence assessment established by quantitative statistical analysis method, and the rainfall threshold is used for judging whether the landslide occurred or not. If the time-variance factors are considered, such assessment can be used to evaluate the landslide occurrence in an area subjecting to certain kind of hydrological and physiographic conditions. In the meantime, the research result can also be used in minimizing or preventing disasters.

Established by Tipping (2001), RVM is a powerful machine learning algorithm for constructing probabilistic model used for nonlinear pattern recognition. As a supervised learning approach, RVM employs a training data set containing N samples which is denoted as$X = \{x_n\}_{n=1}^{N}$ to make prediction of the class labels denoted as $C = \{c_n\}_{n=1}^{2}$. In the problem of interest, c_1 and c_2 represent the landslide and non-landslide categories, respectively. For the task of landslide spatial modeling, RVM is able to categorize a set of feature vectors into two decision domains (e.g., c_1 and c_2) and can be used to generalize a nonlinear classification model. Herein, for the purpose of modeling, $c_1 = 0$ means non landslide occurrence; and $c_2 = 1$ indicates a landslide occurrence. The values of the class output (either c_1 and c_2) are the ground truth results obtained from the records of landslide in the study region.

The application of RVM in physiographic and hydrological factor-based landslide modeling is appropriate since landslide is a complex phenomenon affected by various influencing factors (Hoang and Bui 2016). Moreover, the classification boundary used to quantify landslide probability is possible nonlinear. To deal with the aforementioned difficulties, two-step RVM is proposed. In the first step, RVM maps the original input data to a higher dimensional space (i.e., feature space). In the second step, RVM builds a classification model in the feature space. Figure 1 displays the model construction phase of RVM.

As mentioned earlier, the class label c_i can have two possible values: 0 for non-landslide occurrence and 1 for landslide occurrence. Accordingly, the conditional distribution of landslide occurrence given a set of the 8 landslide conditioning factors is presented as follows:

$$P(c_i|x, w) = \sigma(y) \tag{4}$$

where$\sigma(y)$is a logistic sigmoid function ($\frac{1}{1+e^{-y}}$); P

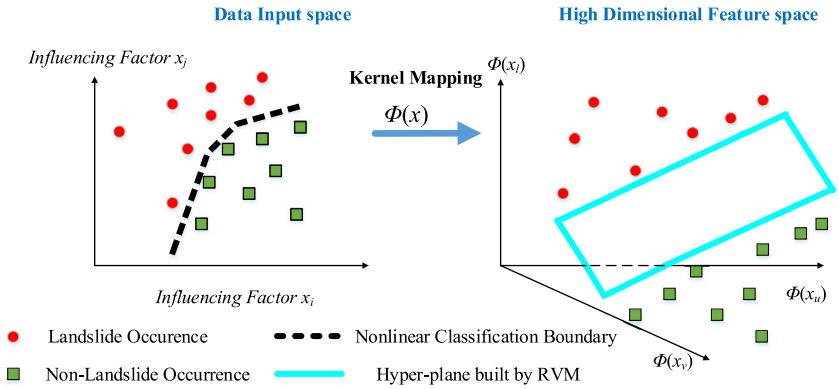


Fig. 1 Conceptual illustration of RVM learning process

denotes the posterior probability given that x and w. y is a sum of M basis functions with linear weights, as described below:

$$y(x,w) = \sum_{m=1}^{M} w_m \varphi_m(x) + w_o = w.\varphi \tag{5}$$

in which $\varphi$is a radial basis function, in which the Gaussian function is adopted as shown in Eq. (6):

$$\varphi(x_i, x_j) = \exp\left(\frac{-\|x_i - x_j\|^2}{2 \times r^2}\right) \tag{6}$$

as seen, the hyper-parameter is denoted as r which represents the width of the RBF.

In the train stage using concepts of Bayesian theory, a prior distribution with respect to the vector of model weights should be provided. In addition, due to large values of w often lead to over-fitting, a small vector of weights is often desired. A smaller weights often result in a smooth classification boundary and therefore, a better classification model with a high generalization property can be expected (Bishop and Tipping 2000; Tipping 2000). To do so, it is often to assign a Gaussian distribution with zero-mean as shown in Eq. (7).

$$p(w|\alpha) = \prod_{m=1}^{M} N\left(w_m|0, \alpha_m^{-1}\right) \tag{7}$$

where$\alpha$is a vector of independent hyper-parameters.

In addition, a different variance for the prior distribution of weights can be assumed to obtain a better prediction (Tipping 2001). In light of this, the prior distribution (w) is described as shown in Eq. (8) (Tipping 2001).

$$\begin{aligned} p(w|\alpha) &= \prod_{m=1}^{M} N\left(w_m|0, \alpha_m^{-1}\right) \\ &= (2\pi)^{-M/2} \prod_{m=1}^{M} \alpha_m^{1/2} \exp\left(-\frac{\alpha_m w_m^2}{2}\right) \end{aligned} \tag{8}$$

The most probable weight μ is found through maximizing Eq. (9) (Hoang and Bui 2016; Tipping 2001) given that initial values of the hyper-parameter α and$p(w|\, C, \alpha) \propto P(C|\, w)p(w|\, \alpha)$.

$$\begin{aligned} &\log\{P(C|w)p(w|\alpha)\} \\ &= \sum_{i=1}^{N} [c_i \log y_i + (1-c_i)\log(1-y_i)] - 0.5 w^T A w \end{aligned} \tag{9}$$

where $A = \mathrm{diag}(\alpha_1, \alpha_2, \ldots, \alpha_M)$.

To deal with the aforementioned maximization problem, the iteratively reweighted least squares algorithm and the Laplace approximation procedure is utilized. The most probable weight μ and its covariance Σ are computed in the following manners (Hoang and Bui 2016):

$$\begin{pmatrix} \mu = \Sigma.\phi^T.B.C \\ \Sigma = \left[\phi^T.B.\phi + A\right]^{-1} \end{pmatrix} \tag{10}$$

where $B = \mathrm{diag}(\beta_1, \beta_2, \ldots, \beta_N)$ with $\beta_i = \sigma\{y(x_i)\}$. $[1 - \sigma\{y(x_i)\}]$.

After the optimization problem is solved, Tipping (2001) mentioned that one may find there are many infinity among the vector α. As a result, the w having a few non-zeros are defined as the relevant vectors. When the training process terminates, the vector of model weight w can be used to construct the RVM classification model for inferring the posterior of the class label c_i of landslide occurrence status given the information of conditioning factors. This process of the RVM based prediction is illustrated in Fig. 2, in which H stands for elevation difference, Eav is mean elevation, S denotes slope steepness, RL stands for road ratio, DS is dip slope ratio, LR is landslide ratio, and DF is distance to fault.

In addition, please note that in the training phase of RVM, a suitable r is desire because the width of RBF has significant influence on the smoothness of the classification boundary and hence affects the predictive capability of the RVM model (Hoang and Tien-Bui, 2016). Although several advance optimization algorithms have been proposed, it is often to find that

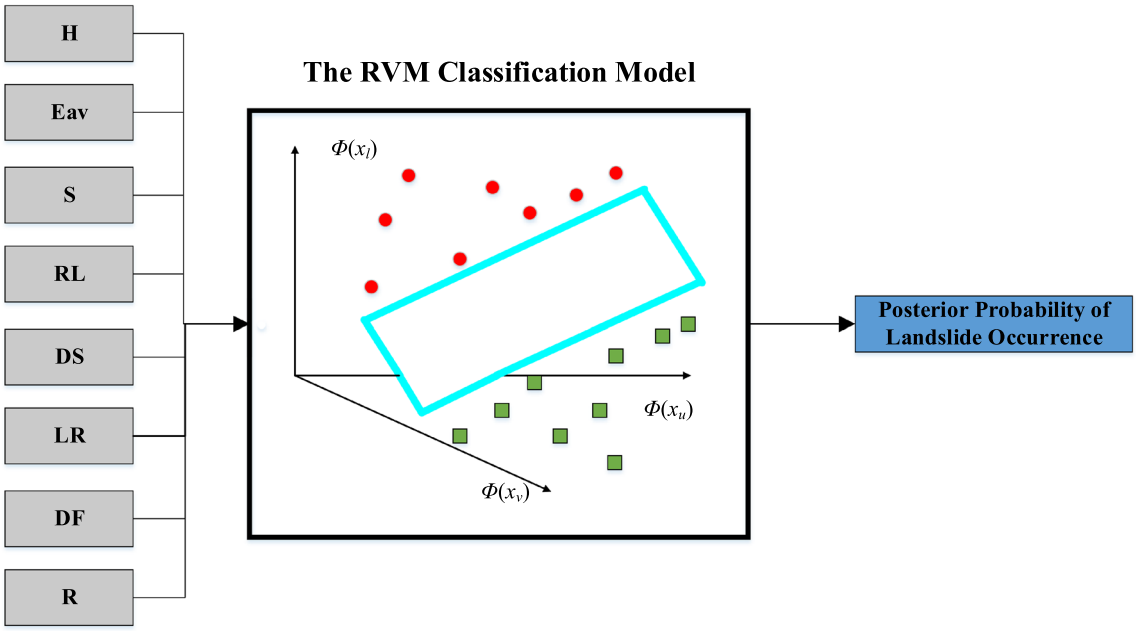


Fig. 2 The RVM-based Prediction of Landslide Posterior Probability

grid search is a very efficient method for parameter tuning and therefore, is adopted here. The value of r is searched within the range of [0, 1] with an interval of 0.1.

Comparison of SVM, LS-SVM, and RVM are summarized in Table 1:

It is seen that all three models are kernel-based algorithm and are able to perform the nonlinear classification for high dimensional problem. The formulations of three approaches are quadratic programming, system of linear equations and nonlinear programming, respectively. Only RVM can provide a probabilistic prediction with less hyper-parameter to be tuned.

Analysis results and discussions

Background

The main stream of Kaoping River is Laonung River, which has its sourced in the southern part of Xinyi Township in Nantou County. It is originated from the east peak of Yushan, running toward northeast and then turning southeast at Patonkuan where it collects the branches from the southwestern hills of Shiukuruan Mountain and the western hills of Dashuiku Mountain and then turns to south-southwest before entering the Kaohsiung City area. Here, it flows again through Meishan, Taoyuan, Baolai, Liokuei and then turning to south till to Dajing where it merges the Zuokou River from east and then turning to the southwest until it reaches to Likang and then merges Ailiau River from southeast and it is named as Erchung River. After that, it keeps running to Linkou for merging with Chishan River (Nantzeshien River) from north and it is now called as Kaoping River finally. Here, the main stream turns south and flows through Dashu, Jiuchutan, Shanliau, Liyushan and then runs into the Taiwan Strait at Dongshan.

Geographically, the Kaoping River basin is located at south of North Tropic Cancer which is covered in the sub-tropic zone. However, it presents violently different climate types because of vertical topological terrain distributed from coastal line to the mountains 3666 m high above the sea level. For this reason, the coastal plain and the medium altitude hills are classified as tropical climate and the high mountain areas, as temperate climate. In this pattern, the temperature gradually diminishes from the southwest alluvial plain to the northeast mountain areas in forming a vertical climate pattern. The

Table 1 Comparison of SVM, LS-SVM, and RVM

	Kernel-based	High-dimension	Nonlinear classification	Problem formulation	Probabilistic Estimation	No. of hyper-parameter
SVM	Y	Y	Y	Quadratic programming	Y	More
LSSVM	Y	Y	Y	System of linear equations	N	More
RVM	Y	Y	Y	Nonlinear programming	N	Less

Kaoping River basin is located in the tropical zone where distinctive line is defined between dry and humid seasons. In this area, the yearly average rainfall is about 2920 mm along the Kaoping river basin. The Majia area has the biggest yearly average rainfall, exceeding 4000 mm. The southwest plain area along the basin and the coastal area are lower in yearly average rainfall, with about 2000 mm. The yearly rainfall of the Kaoping River basin is mostly concentrated in high water period from May to September, making up 90% of the total yearly rainfall and it is mainly brought by the typhoon and the storm from southwest air current.

The stratum and regional geological condition in the Kaoping River basin area is roughly extending from north-northeast towards the south-southwest. The main fault lines are comprised of, from east to west, the following: Xiaotushan Fault, Ailiaubei River Fault, Maolin Fault, Kuangshan Fault, Kuaiku Fault, Weijingshi Fault, Meishi Fault, Shalixian River Fault, Tulungwan Fault, Kaozhong Fault, Chaozhou Fault, Liokuei Fault, Yuekuanshan Fault, Xiaolin Fault, Neiying Fault, Chishan Fault, Pingxi Fault, Jiasian Fault, Dishui Failt, and Fengshan Fault. Most of these faults are the reverse faults running from east to west in a thrust manner. Based on the Taiwan Active Fault Distribution Map announced by the Central Geological Survey under Ministry of Economic Affairs, it indicated that among the main faults in the Kaoping River basin, Chishan Fault is classified as Type-1 active fault and Chaozhou Fault (as well as the Tulungwan Fault spreading along the north area) is classified as Type-2 active fault.

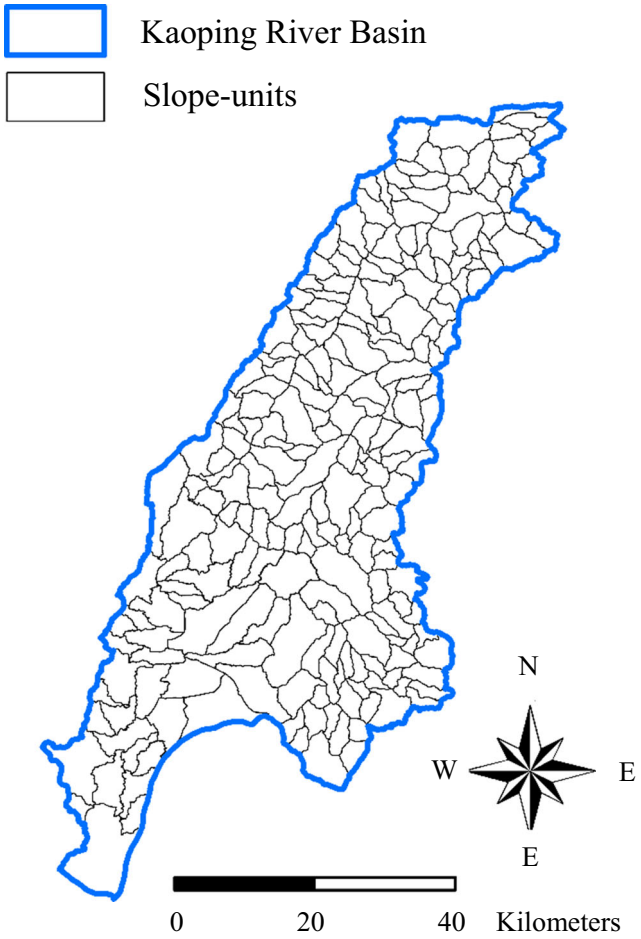


Fig. 3 Small watershed classification for Kaoping River basin

Data preparation

Classification of sub-landslide area

In order to analyze the occurrence probability of landslides at different locations in the study area, the study area was classified by the small watershed method, assuming that there is a negligible difference between the hydrological and topographic features of each watershed. In this method, the 40-m elevation model of the hydrological module software in the ArcMAP Geographic Information System is based on calculating the flow direction and the size of the catchment area for each grid. At the same time, the basin tool is also used to analyze the ridgeline and water system on the slope to classify small watersheds. Finally, a manual modification is performed according to basic data such as contour and catchment range. Shown in Fig. 3 is the classification results for each small watershed in the study area. In total, the study area is divided into 231 watershed areas, and the subsequent landslide probability assessment model will be implemented based on the 231 small watershed areas.

Selection of representing Rainfall Station

After collecting rainfall data from the Kaoping River Basin, the monitoring station needs to choose to represent the rainfall station. According to the "geographic station", "storm characteristics" and "historic disaster frequency" observed by the various meteorological stations of the Central Weather Administration, Fan et al. (2013) suggested that Dajing, Xinfa, Youyou County, Majia, Jiaxian and Yuemei rainfall stations can be the representative stations in the study area. Using the data obtained from the above-mentioned rainfall stations, the distance between each small watershed and these rainfall stations is calculated, and the rainfall station closest to the catchment area where the small watershed is located is selected as the reference rainfall station. Shown in Fig. 4 is the distribution of reference rainfall stations for each small watershed in the study area.

Classification of rain field

Fan et al. (2003) states that slopes in the mountains of Taiwan are usually steep and the flow is very fast. The runoff after the

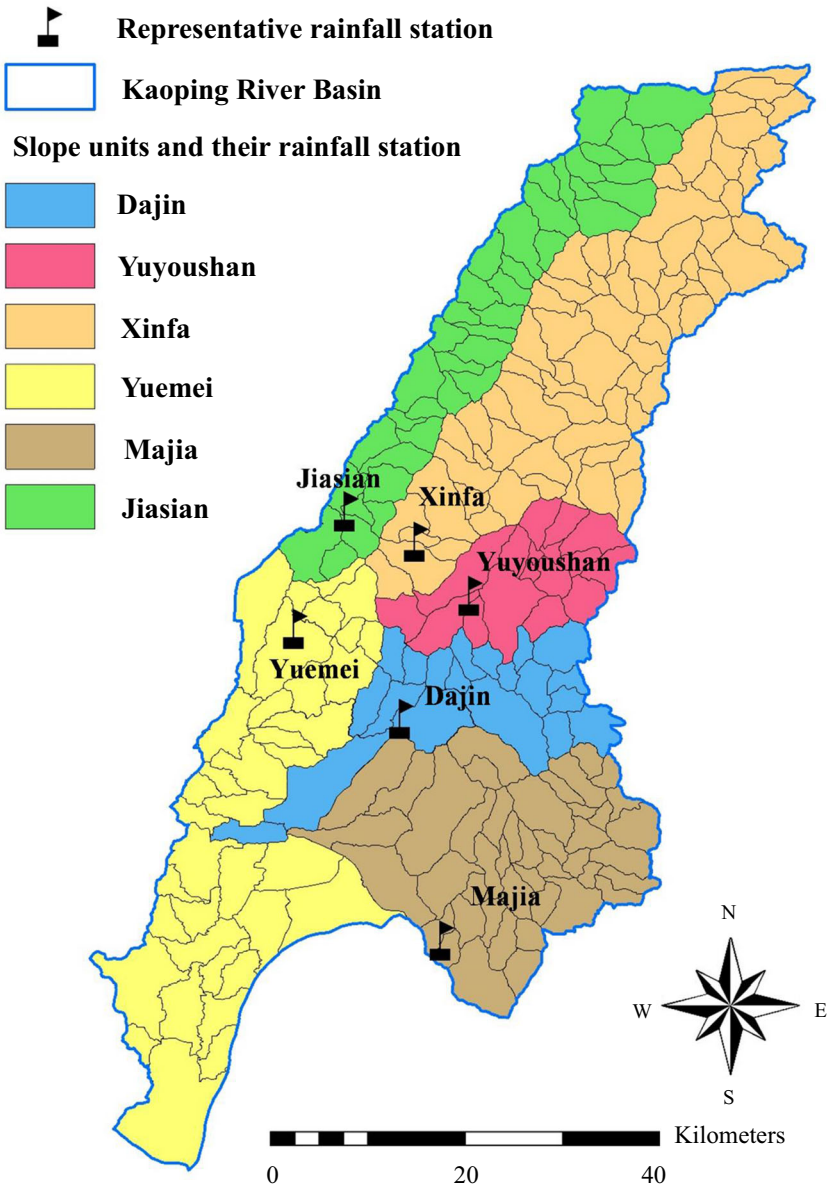


Fig. 4 Distribution of referential rainfall stations for each small watershed in Kaoping River basin

rain will soon flow into the downstream rivers. In their debris flow warning system, in addition to considering the contribution of previous rainfall, the initial amount of current rainfall is also considered. They define a rain starting time as the accumulated rainfall reached 10 mm within the first 24 h, and define a rain ending time as the accumulated rainfall within 24 h after the starting time is less than 10 mm. Based on this method, Fan et al. (2013) classify the rain field to effectively improve the accuracy of the mudslide slip warning methodology. This rain field classification method is adopted in the current study.

The graphical data in the landslide image catalogue from FORMOSAT-2 is analyzed using ArcMap 10.1, in which the difference in landslide distribution before and after the typhoon event is used to determine whether or not a landslide is occurred. In order to reduce the influence of factors other than the typhoon event between the two landslide graphic data, the precipitation during the considered period should be less than 200 mm, and earthquake magnitude should be less than 3 at that same period. Based on the aforementioned regulation, taking Typhoon Haitang in 2005, Typhoon Tower in 2005, Typhoon Kaemi in 2006, Typhoon Morak in 2009 and Typhoon Namadu in 2011 into consideration, there are 1155 cases in total, in which 826 of them are categorized as "occurrence of landslide", and 329 of them are not.

Selection of the physiographic factors

Many physiographic factors have been suggested to have significant influence on landslide occurrence (Chang 1987;

Pradhan and Lee 2010). The physiographic factors often considered are slope steepness, mean elevation, elevation difference, road ratio, dip slope ratio, landslide ratio and distance to fault. Selecting suitable representing physiographic factors are very important in the proposed approach, the factors mentioned in literatures are first selected as potential factors, followed by correlativity analysis using actual landslide records collected in this study to finalize the representing factors for the inputs of the proposed RVM. Fan et al. (2015) investigated influences of two different inputs, the original physiographic factors and their degrees of membership from fuzzy theory, to the debris occurrence. They found that the degree of membership will not only prevent the original value from having excessive discretion but will also effectively reflect its influence to the occurrence of debris.

The effect of degree of membership on the proposed RVM is therefore investigated in this study.

The statistics and histograms of the historical dataset are described in Tables 2 and 3, and Figs. 5 and 6. It is seen that although the ranges with and without landslide are similar, the mean, standard deviation and median are apparently different for most input factors, resulting to different histograms.

Results and discussions

In this research, the landslide occurrence is established according to the data obtained from each small watershed and rainfall station as the physiographic and hydrological factors for machine learning. It is known that the approach of machine learning often lacks the basics of mechanics, therefore, the selection of input factors plays an important role. In light of this, this study exclusively investigates the influence of physiographic and hydrological factors on the landslide occurrence.

As mentioned, 7 physiographic factors, that are often believed to have significant effect on landslide, are considered in the current study. That is, these 7 factors are the input candidates for the proposed RVM. Fan et al. (2018) stated that among these 7 factors, mean elevation and elevation difference can be excluded after independent test. Road ratio and distance to fault can be excluded after correlation test. That is, after independent and correlation tests, only 3 physiographic factors that are dip slope ratio, landslide ratio, and slope steepness are remained as the machine learning inputs. Together with the cumulative rainfall in each typhoon event (i.e., the hydrological factor), the total input factors are 4 in their logistic model (Fan et al. 2018). Based on Fan et al. (2018), the raw data of physiographic and hydrological factors are converted into the degree of membership using the following functions.

Table 2 Statistics of the input factors (landslide occurred)

Input Factor (IF)	Min	Mean	Std	Median	Max
H	0.000	7.889	1.722	15.255	110.256
Eav	0.000	0.033	0.014	0.048	0.241
S	0.531	26.450	30.965	10.996	40.158
RL	9.621	1208.492	1168.586	813.352	3206.798
DS	16.000	1251.372	1280.000	627.111	3919.000
LR	0.023	0.130	0.051	0.226	1.000
DF	0.000	0.024	0.004	0.052	0.730
R	159.5	1015	741	608	2965

Table 3 Statistics of the input factors (landslide not occurred)

Input Factor (IF)	Min	Mean	Std	Median	Max
H	0.000	21.11	23.14	14.97	110.256
Eav	0.000	0.022	0.04	~0	0.241
S	0.531	14.7	12.11	12.54	38.44
RL	9.621	607.98	827.57	195.25	3206.798
DS	16.000	653.17	561.78	542	2173
LR	0.023	0.24	0.32	0.08	1.000
DF	0.000	0.004	0.013	0.0	0.11
R	159.5	713.65	495.75	577.5	2158

The degree of membership S_N for the slope steepness factor is given by

$$S_N = \begin{cases} 0.00023 \times S^{2.283} \\ 1, \text{if } S \geq 39.23^\circ \end{cases} \quad (11)$$

where S means the slope steepness (°).

The degree of membership DS_N for the dip slope ratio factor is given by

$$DS_N = 1 - \exp(-21.01 \times DS) \quad (12)$$

where DS means the dip slope ratio.

The degree of membership LR_N for the landslide ratio factor is given by

$$LR_N = 1 - \exp(-36.4 \times LR) \quad (13)$$

where LR means the landslide ratio.

The degree of membership R_N for the cumulative rainfall factor is given by

$$R_N = 0.0006 \times R - \left(8.9 \times 10^{-8}\right) R^2 \quad (14)$$

where R is the cumulative rainfall (mm).

Based on the above discussion considering effects of input factors and modelling, 8 different models are built/used to examine their corresponding performances. To be specific, these 8 models are described in Table 4.

The collected data consisting of 1155 historical cases are divided into 2 sets: training set (90%) and testing set (10%). The former set is employed for model construction. The later set is reserved for evaluating the model testing. Moreover, to

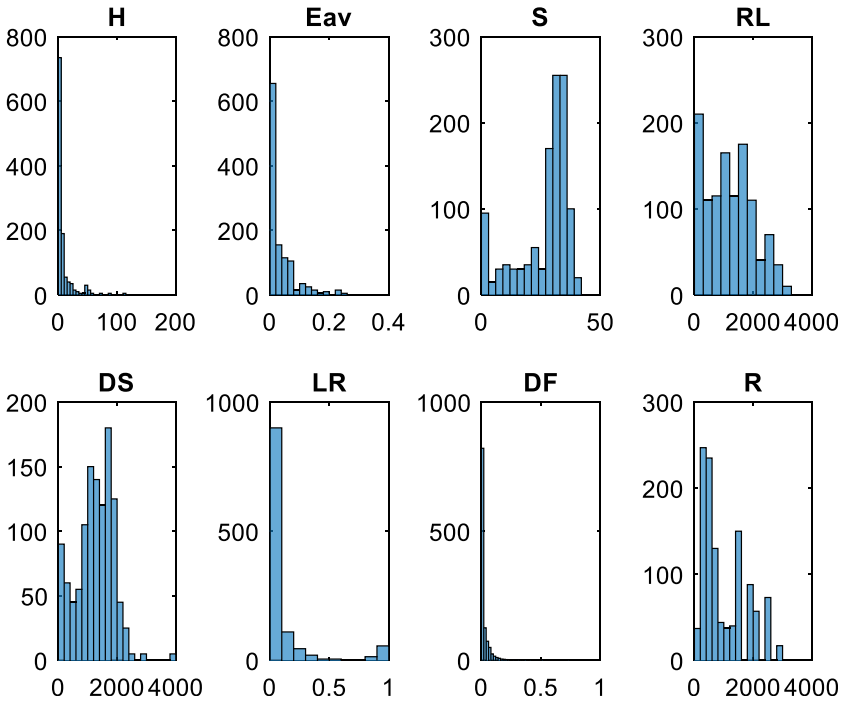


Fig. 5 Histogram of the input factors (landslide occurred)

avoid overfitting, a 10-fold cross-validation is utilized in this study. The model performance is assessed according to the sensitivity, specificity and overall accuracy of the confusion matrix. Sensitivity, also called the true positive rate, measures the proportion of actual positives that are correctly identified by the estimation model. Specificity, also called the true negative rate, measures the proportion of actual negatives that are correctly identified by the estimation model. To avoid overfitting, 10-fold cross-validation is adopted.

The measurements of these 8 models are displayed in Table 5. As shown, the performances of three different machine learning models are all better than that of the previous study (i.e., logistic regression). As seen, the SVM enhances the estimation accuracy from 81.3% to 82.44%. the LS-SVM further increases the accuracy to 84.7%. The proposed RVM with any kind of normalization is superior to those of the logistic model, SVM and LS-SVM. As Table 5 indicates that the overall accuracies of the RVM models are stably greater

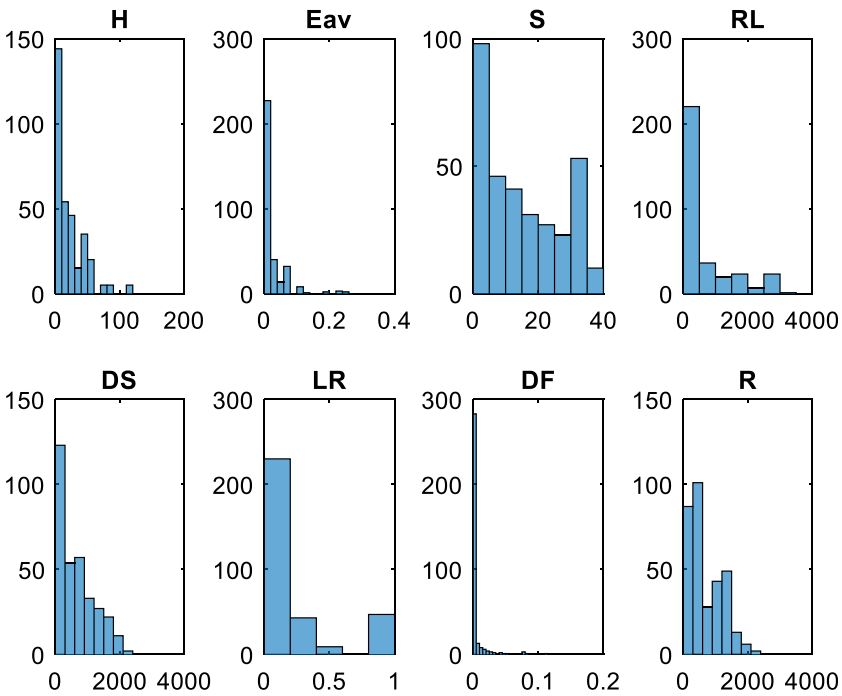


Fig. 6 Histogram of the input factors (landslide not occurred)

Table 4 Different Models investigated in the current study

Model	No. of input factor	Normalization technique	Machine learning
No. 1	8	Standardization	RVM
No. 2	8	–	RVM
No. 3	4	Standardization	RVM
No. 4	4	Membership function	RVM
No. 5	8	Membership/ Standardization	RVM
No. 6	4	Standardization	SVM
No. 7	4	Standardization	LSSVM
No. 8*	4	Membership function	Logistic

*Fan et al. (2018)

than 87.50%. The poor performance of No. 2 model indicates that normalization is necessary for estimating landslide. In addition, the techniques of normalization do not have significant impact on the estimation performance by comparing results between No. 1 and No.5 models, and between No.3 and No.4 models. Reducing input factors from 8 to 4 is not necessary to low the estimation accuracy, indicating that the independent and correlation analysis can be utilized to simplify the machine learning network. As shown, the estimation performances of No. 1, No.3, No.4 and No.5 models are all very promising. However, it is observed that logistic regression possesses the best estimation in specificity among all the investigated models. In addition, it is known that the accuracy/ sensitivity/specificity of a binary estimation model depends on the values of threshold. In order to further compare these models, the ROC (receiver operating characteristic curve) for each model is displayed in Fig. 7, the corresponding AUC (area under curve) is displayed in Table 5. Similarly, the performances of models No.3 and No.4 are similar to those of models using more input factors.

Conclusions

The mechanism of landslide occurrence is often too complicated to estimate via a physical model. The machine learning model provides an alternative way to assess such task. In order to analyze the landslide occurrence at different locations, the Kaoping River Basin is divided into 231 small watersheds that have minimal variations in their hydrological and physiographic characteristics. Hysterical records indicate that in these small watershed area, there are 826 cases with slides and 329 cases without slides, which are used in the construction of the assessment model via the proposed machine learning algorithm. Together with the concept of small watershed method, this study proposes a Relevance Vector Machine (RVM) to enhance the prediction accuracy. Effect of physiographic and hydrological factors such as slope steepness, dip slope ratio, and cumulative rainfall are investigated. RVM first maps the aforementioned factors into a feature space and a Bayesian theory-based optimization problem is formulated to find the weights of a linear boundary for binary classification. Results of the proposed approach are compared to that of literature study, indicating the proposed RVM-based small watershed approach is superior to the logistic regression. Several important issues are observed and they are described below.

1. According to correlation test, independent test and results of models No. 1, No.3, No.4 and No.5, mean elevation,

Table 5 Mean measurements* for different models in the testing stage

Model	RVM					SVM	LS-SVM	Logistic
	No.1	No.2	No.3	No.4	No.5	No.6	No.7	No.8**
Accuracy	88.23%	28.44%	89.31%	87.58%	87.58%	82.44%	84.70%	81.3%
Sensitivity	95.43%	0.00%	95.42%	92.29%	95.66%	87.76%	95.32%	79.50%
Specificity	70.31%	100.00%	73.94%	75.76%	67.27%	70.12%	63.64%	81.8%
AUC	0.94	–	0.92	0.91	0.89	–	–	–

*mean of 10 runs, **Fan et al. (2018)

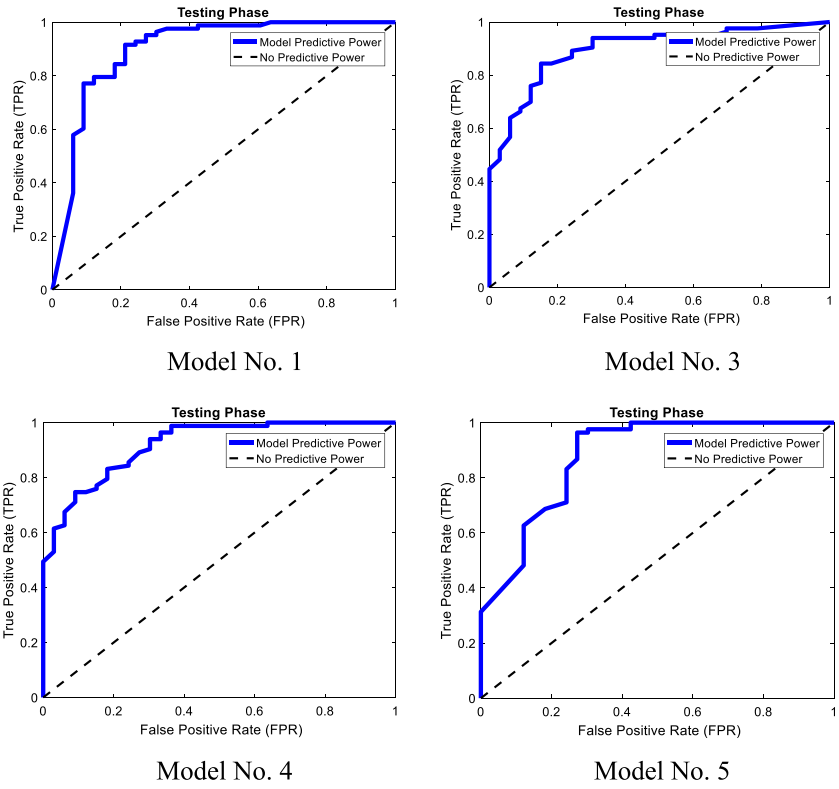


Fig. 7 AUC of different RVM models

elevation difference, road ratio and distance to fault only have a moderate influence on the estimation of landslide occurrence. That is, to simplify a machine learning network, one can establish a learning machine only with inputs of dip slope ratio, landslide ratio, slope steepness and cumulative rainfall.

2. Techniques of normalization does not have significant effect on the performance of the built RVM. Both standard normalization and degree of membership can deliver a promising estimation.
3. Compared to the logistic regression model, the standard SVM and LS-SVM have better estimation accuracies. The performance of the proposed RVM is able to continuously improve the estimation accuracy and is a superior tool.
4. The relative large number of AUC are observed for all RVM models, indicating that the proposed RVM models can consistently provide a promising estimation.
5. The proposed method can provide authorities in their hazard alert system to minimize the life or property losses caused by the landslide if physiographic and hydrological factors are given with a prescribed risk level.
6. It is worth of investigating other more sophisticated metaheuristic methods for selecting the hyper-parameter in establishing an RVM.

Acknowledgments This study was supported by the Ministry of Science and Technology (MOST) of Taiwan and Soil and Water Conservation Bureau. under grant number MOST 107-2622-E-011-020 -CC2. The support is gratefully acknowledged.

References

Abbas H, Tezcan J (2019) Relevance vector machines modeling of non-stationary ground motion coherency. Soil Dyn Earthq Eng 120:262–272. https://doi.org/10.1016/j.soildyn.2019.02.002

Bishop CM, Tipping ME (2000) Variational relevance vector machines, proceedings of the. In: 16th conference on uncertainty in artificial intelligence, Morgan Kaufmann publishers Inc, San Francisco, CA, USA, pp 46–53

Chan HC, Chang CC, Chen SC, Wei YS, Wang ZB, Lee TS (2015) Investigation and analysis of the characteristics of shallow landslides in mountainous areas of Taiwan. J Chin Soil Water Conserv 46(1): 19–28

Chang KT, Chiang SH, Hsu ML (2007) Modeling typhoon-and earthquake-induced landslides in a mountainous watershed using logistic regression. Geomorphology 89(3–4):335–347

Chang SC (1987) The prediction of potential geological hazards of Slopeland and its applications in environmental impact assessment. J Chin Soil Water Conserv 18(2):41–62

Cheng MY, Hoang ND (2015) Typhoon-induced slope collapse assessment using a novel bee colony optimized support vector classifier. Nat Hazards 78(3):1961–1978. https://doi.org/10.1007/s11069-015-1813-8

Cruden DM (1991) A simple definition of a landslide. Bulletin of Engineering Geology and the Environment 43(1):27–29

Dai F, Lee CF, Wang S, Feng Y (1999) Stress–strain behaviour of a loosely compacted volcanic-derived soil and its significance to rainfall-induced fill slope failures. Eng Geol 53(3–4):359–370

Dai FC, Lee CF (2002) Landslide characteristics and slope instability modeling using GIS. Lantau Island, Hong Kong, Geomorphology 42:213–228

Fan JC, Yang CH, Chang SC, Huang HY, Guo JJ (2013) Effects of climate change on the potential of the landslides in the basin of Kaoping stream. J Chin Soil Water Conserv 44(4):335–350

Fan, JC, Liu, CH, and Wu, MF (2003) Determination of Critical Rainfall Thresholds for Debris-Flow Occurrence in Central Taiwan and Their Revision after the 1999 Chi-Chi Great Earthquake, Proceedings of 3rd International Debris Flow Hazard Mitigation Conference, Davos, Switzerland, 10–12 : 103–114

Fan JC, Huang HY, Liu CH, Yang CH, Guo JJ, Chang CH, Chang YC (2015) Effects of landslide and other physiographic factors on the occurrence probability of debris flows in Central Taiwan. Environ Earth Sci 74:1785–1801

Fan JC, Chang SC, Liao KW, Guo JJ, Liu CH, Chang YC,... Yang CH (2018) The impact of physiographic factors upon the probability of slides occurrence: a case study from the Kaoping River Basin, Taiwan. Journal of the Chinese Institute of Engineers 41(5):419–429

Ge Y, Chen H, Zhao B, Tang H, Lin Z, Xie Z, Lv L, Zhong P (2018) A comparison of five methods in landslide susceptibility assessment: a case study from the 330-kV transmission line in Gansu region, China. Environ Earth Sci 77(19):662

Highland L, Bobrowsky PT (2008) The landslide handbook: a guide to understanding landslides (p. 129). Reston: US Geological Survey

Hoang ND, Bui DT (2016) A novel relevance vector machine classifier with cuckoo search optimization for spatial prediction of landslides. J Comput Civ Eng 30:04016001

Hung JJ, Lin ML, Chen TC, Wang KL (2000) 921 chi-chi earthquake disasters characteristics of slope failure case analysis: dip slope failure. Sino-Geotechnics 81:17–32

Jaafari A, Panahi M, Pham BT, Shahabi H, Bui DT, Rezaie F, Lee S (2019) Meta optimization of an adaptive neuro-fuzzy inference system with grey wolf optimizer and biogeography-based optimization algorithms for spatial prediction of landslide susceptibility. CATENA 175:430–445. https://doi.org/10.1016/j.catena.2018.12.033

Kavzoglu T, Colkesen I, Sahin EK (2019) Machine learning techniques in landslide susceptibility mapping: a survey and a case study. In: Pradhan SP, Vishal V, Singh TN (eds) Landslides: theory. Practice and Modelling. Springer International Publishing, Cham, pp 283–301. https://doi.org/10.1007/978-3-319-77377-3_13

Le HV, Bui QT, Tien Bui D, Tran HH, Hoang ND (2018) A hybrid intelligence system based on relevance vector machines and imperialist competitive optimization for Modelling Forest fire danger using GIS. J Environ Inf. https://doi.org/10.3808/jei.201800404

Liu Y, Ye Y, Wang Q, Liu X, Wang W (2019) Predicting the loose zone of roadway surrounding rock using wavelet relevance vector machine. Appl Sci 9(10):2064

Pham BT, Tien Bui D, Prakash I (2017) Landslide susceptibility assessment using bagging ensemble based alternating decision trees, logistic regression and J48 decision trees methods: a comparative study. Geotech Geol Eng 35(6):2597–2611. https://doi.org/10.1007/s10706-017-0264-2

Pradhan B, Lee S (2010) Landslide susceptibility assessment and factor effect analysis: Backpropagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression Modelling. Environ Model Softw 25(6):747–759

Samantaray AK, Singh G, Ramadas M (2019) Application of the Relevance Vector Machine to Drought Monitoring. In, Singapore, Soft computing for problem solving. Springer Singapore, pp 891–898

Suykens JA, De Brabanter J, Lukas L, Vandewalle J (2002) Weighted least squares support vector machines: robustness and sparse approximation. Neurocomputing 48(1-4):85–105

Tien Bui D, Pradhan B, Lofman O, Revhaug I (2012) Landslide susceptibility assessment in Vietnam using support vector machines, decision tree, and Naïve Bayes models. Math Probl Eng 2012:26. https://doi.org/10.1155/2012/974638

Tien Bui D, Shahabi H, Shirzadi A, Chapi K, Hoang ND, Pham B, Bui QT, Tran CT, Panahi M, Bin Ahamd B, Saro L (2018) A novel integrated approach of relevance vector machine optimized by imperialist competitive algorithm for spatial modeling of shallow landslides. Remote Sens 10(10):1538

Tipping ME (2000) The relevance vector machine. In Advances in neural information processing systems, MIT Press 12:652–658

Tipping ME (2001) Sparse bayesian learning and the relevance vector machine. Journal Machine Learn Research 1:211–244

Tipping ME (2004) Bayesian inference: an introduction to principles and practice in machine learning. In: Advanced lectures on machine learning, vol 3176. Lecture Notes in Computer Science. Springer, Berlin Heidelberg, pp 41–62. https://doi.org/10.1007/978-3-540-28650-9_3

Tzikas DG, Wei L, Likas A, Yang Y, Galatsanos P (2006) A tutorial on relevance vector machines for regression and classification with applications. EURASIP News Letter 17(2):4–23

Varnes DJ (1978) Slope movement types and processes. Special report, 176:11–33

Zhou C, Yin K, Cao Y, Ahmed B, Li Y, Catani F, Pourghasemi HR (2018) Landslide susceptibility modeling applying machine learning methods: a case study from Longju in the three gorges reservoir area, China. Comput Geosci 112:23–37. https://doi.org/10.1016/j.cageo.2017.11.019