A Novel Hybrid Approach Based on Instance Based Learning Classifier and Rotation Forest Ensemble for Spatial Prediction of Rainfall-Induced Shallow Landslides Using GIS

Quang-Khanh Nguyen 1, Dieu Tien Bui 2,*, Nhat-Duc Hoang 3, Phan Trong Trinh 4, Viet-Ha Nguyen 5 and Isık Yilmaz 6

1 Faculty of Information Technology, Hanoi University of Mining and Geology, Duc Thang, Bac Tu Liem, Hanoi 100000, Vietnam; nguyenquangkhanh@humg.edu.vn
2 Geographic Information System Group, Department of Business and IT, University College of Southeast Norway, Gullbringervegen 36, Bo i Telemark N-3800, Norway
3 Faculty of Civil Engineering, Institute of Research and Development, Duy Tan University, P809-K7/25 Quang Trung, Danang 556361, Vietnam; hoanghatduc@dtu.edu.vn
4 Institute of Geological Sciences, Vietnam Academy of Sciences and Technology (VASC), 84 Chua Lang Street, Dong da, Hanoi 100000, Vietnam; phantrongt@yahoo.com
5 Faculty of Geomatics and Land Administration, Hanoi University of Mining and Geology, Duc Thang, Bac Tu Liem, Hanoi 100000, Vietnam; viethatdct@gmail.com
6 Department of Geological Engineering, Faculty of Engineering, Cumhuriyet University, Sivas 58140, Turkey; isik.yilmaz@gmail.com
* Correspondence: Dieu.T.Bui@usn.no or BuiTienDieu@gmail.com; Tel.: +47-96-677-678

Abstract: This study proposes a novel hybrid machine learning approach for modeling of rainfall-induced shallow landslides. The proposed approach is a combination of an instance-based learning algorithm (k-NN) and Rotation Forest (RF), state of the art machine techniques that have seldom explored for landslide modeling. The Lang Son city area (Vietnam) is selected as a case study. For this purpose, a spatial database for the study area was constructed, and then was used to build and evaluate the hybrid model. Performance of the model was assessed using Receiver Operating Characteristic (ROC), area under the ROC curve (AUC), success rate and prediction rate, and several statistical evaluation metrics. The results showed that the model has high performance with both the training data (AUC = 0.948) and the validation data (AUC = 0.848). The results were compared with those obtained from soft computing techniques, i.e. Random Forest, J48 Decision Trees, and Multilayer Perceptron Neural Networks. Overall, the performance of the proposed model is better than those obtained from the above methods. Therefore, the proposed model is a promising tool for landslide modeling. The research result can be highly useful for land use planning and management in landslide prone areas.

Keywords: landslide; classifier ensemble; instance based learning; Rotation Forest; GIS; Vietnam

1. Introduction

Development of landslide mitigation strategy is considered to be the most effective and economical way to reduce landslide losses and minimize landslide risks [1]. Therefore, reliable landslide susceptibility and hazard maps is a key point for development, as is clearly stated by the United Nations [2]. However, producing these reliable maps is not a simple task because landslides are triggered by complex processes and relate to many causal factors. Although the recent developments
of Remote Sensing and GIS (Geographic Information Systems) have provided powerful tools for acquisition and processing of high quality data for landslide studies, the prediction power of landslide models is still a debated subject because the quality of susceptibility maps is clearly dependent on the method used [3–6]. Thus, the overall performance (the goodness of fit and the prediction power) of landslide models is not only dependent on the quality of input data but also on methods and techniques used. Therefore, various approaches have been proposed in the literature for landslide susceptibility mapping. These methods and techniques vary from simple expert knowledge to sophisticated mathematical procedures and in general, they could be divided into qualitative and quantitative groups [7]. The first group is clearly subjective because they depend heavily on expert knowledge whereas the second one is relatively objective [8].

The second group could be further categorized into three main topics: deterministic methods, statistical methods, and data mining. In general, the deterministic methods have the most accurate results due to data dependency and site-specific nature and are most suitable for site-specific locations in localized scales [9]. For large areas, application of deterministic methods is almost impossible due to the difficulty of collecting detailed geo-engineering data, therefore the use of statistical and soft computing methods has increased [10]. The statistical methods are considered to be suitable for the mapping of landslide susceptibility over large areas and using the statistical hypothesis that future landslides will occur under the same geo-environmental conditions that produce them in the past. Therefore, large amounts of data need to be collected and processed and these tasks are time consuming and involve various complex processes [11]. More importantly, because the prediction capability of statistical models is still not sufficient, data mining has been considered.

Being a branch of artificial intelligence, data mining can be defined as processes of analyzing observational data to find internal relationships and representing them in novel ways that are useful and easier understandable [12,13]. Data mining includes multiple steps, i.e., data selection, pre-processing and transformation, analysis with computational algorithms, interpretation and evaluation of the results [14]. The most common data mining methods used in landslide modeling are artificial neural networks [11,15,16], support vector machines [17–21], decision trees [10,20,22], and neuro-fuzzy [23,24]. Literature review shows that new data mining algorithms are suitable for landslide modeling for large and complex areas with good results [3,25–30], and, in general, data mining models outperform conventional methods [10,31–33]. However, recent studies on landslide modeling show that the overall performance of prediction models could be enhanced with the use of ensemble frameworks [31,34,35]. Therefore, investigation of these frameworks for landslide modeling should be carried out.

Starting in the early 1990s, ensemble-based systems have become an important research area in machine learning with various techniques have been proposed. These systems can be established through combinations of two or more methods and techniques [36–42] or ensemble frameworks such as Stacking, Bagging, AdaBoost, Random Subspace, MultiBoost, Random Forests, Diverse DECORATE (Ensemble Creation by Oppositional Relabeling of Artificial Training Examples), and Rotation Forest [43,44]. Although these ensemble-based systems often improve performances of base classifiers, the Rotation Forest outperforms the others in term of accuracy and diversity in various datasets [43,45]. In addition, exploration of the Rotation Forest for landslide analysis has seldom been carried out.

Based on such motivation, this paper explores a current state-of-the-art Rotation Forest ensemble with \( k \)-NN algorithm for landslide susceptibility mapping. The main objective of this study is therefore to create a novel methodological approach that is capable to solve the complex and high-dimensional data, for landslide susceptibility mapping. The proposed approach is a combination of an instance based learning algorithm (\( k \)-NN) and the Rotation Forest (RF) ensemble, where the Information Gain is used for feature selection. The Lang Son city area (Vietnam) is selected as a case study because it belongs to one of the most vulnerable areas with respect to landslides in the northeast region of Vietnam [46]. However, landslide studies in this area have seldom been carried out; therefore, assessment of landslide susceptibility is considered to be an urgent task. The usability of the proposed
model is assessed through comparisons with those obtained from various soft computing techniques using the same data such as Random Forest, J48 Decision Trees, and Multilayer Perceptron Neural Networks, and finally, conclusions are given.

2. Study Area and Data

2.1. Study Area

The study area is located in the Lang Son city area, near the Vietnam-China border, which belongs to the northeastern part of Vietnam (Figure 1). It covers an area of about 168 km$^2$, between longitudes 106°41’34” E and 106°48’32” E, and latitudes 21°49’43” N and 21°57’13” N. The altitude varies from 194.5 m to 800 m above sea level with the mean of the altitude is 328 m and the standard deviation is 84.7 m. Slope angles in the study area are from 0° to 84°. Approximately 23.7% of the study area has ground slopes less than 8° and about 10.2% fall in slopes from 8° to 15°. Around 21.1% of the study area falls in slopes 15°–25°, whereas areas with slope 25°–45° account 43.5% of the total study area. Only 1.5% of the study area has slopes larger than 45°.

![Figure 1. Location of the study area and landslide inventory.](image)

Forest land covers around 43.4% of the total study area, in which 35.7% of the land is productive forest and 7.7% of the land is protective forest. Settlement areas cover 6.9% of the total study area, whereas barren land and paddy land cover 20.4% and 21.5% of the total study area, respectively. The soil types are mostly ferralic acrisols, which accounts for 78.5% of the total study area, followed by dystric gleysols (6.1%), rhodic ferralsols (5.8%), eutric fluvisols (4.8%), plinthic acrisols (1.3%), and dystric fluvisols (1.2%) soils.

Geologically, Quaternary deposits cover around 16% of the total study area that consists of granule, grit, breccia, boulder, sand, and clay. The other areas are covered by six lithological formations i.e., Na Khuat, Tam Lung, Khon Lang, Lang Son, Tam Danh, and Mau Son. The main lithologies are marl, siltstone, tuffaceous conglomerate, gritstone, sandstone, basalt, and clay shale.

The study area is characterized by monsoonal climate with rainy or dry seasons. The rainy season is normally from May to September and the dry season is from October to April. The average rainfall is in the range from 1200–1600 mm annually [46].
2.2. Data Used

Historical landslide records are the first required data for the assessment of landslide susceptibility. In the study, the landslide inventory map with a total of 172 historical landslides prepared earlier by [46] was used. This map was constructed from several sources: (i) interpretation of orthorectified aerial photographs with spatial resolution of 1 m that were acquired by the Aerial Photo—Topography Company (Vietnam) in 2003; (ii) a landslide inventory map constructed by Tam, et al. [47]; (iii) a landslide inventory map compiled by Truong, et al. [48]; and (iv) landslide locations identified from field surveys in 2012.

Among the historical landslides, 86 locations are rotational slides that account for 50% of the total landslides, whereas 52 locations are translational slides and account for 30.2% of the total landslides. The remaining inventories are debris slides with 34 locations, accounting for 19.8% of the total landslides. It is noted that rock falls are very few in this study and were excluded in this analysis.

Landslides and flash floods are the main recurrent natural hazards in the Lang Son city area. An analysis of the historical landslides shows that rainfall is the main triggering factor [46]. Landslides usually occur during the torrential rainfalls especially in tropical rainstorms. For example, many landslides occurred in the study area during the tropical rainstorm Rammasun on 19 July 2014 where the daily rainfall at the Mau Son was 504 mm. Landslides also occurred in Dong Dang town during heavy rainfalls of the tropical rainstorm Kalmaegi on 17 September 2014 that caused seven deaths and six injuries.

A digital elevation model (DEM) with spatial resolution of 5 m for this study area was constructed using the National Topographic Maps. Scales of these maps are 1:5000 for the Lang Son city and 1:10,000 for the other areas. The DEM was then used to extract morphometric properties for deriving landslide influencing factor maps i.e., slope (Figure 2a), slope length, aspect (Figure 2b), curvature, elevation (Figure 2c), and toposhade. These morphometric factors are selected because slope instability is influenced by the types of terrain [49]. They are the most commonly used factors for the assessment of landslide susceptibility in Vietnam [34,46] and in literature. In addition, valley depth (Figure 2d) was included because the increasing of upslope area could provide weight of material on the slope [50], and, thus, this is considered a key factor in slope failure assessment. Detailed explanations on valley depth for landslide susceptibility could be found in [50] and [28]. Furthermore, occurrences of rainfall-induced shallow landslides are also influenced by hydrogeological conditions [51,52], therefore topographic wetness index (TWI), stream power index (SPI), sediment transport index (STI) were used included in the analysis [28]. In this analysis, TWI, SPI, and STI were extracted from the DEM. Detailed descriptions on the calculation of these indices could be found in [53]. Detailed classes for these factors (Table 1) were determined based on a frequency ratio analysis of the landslide inventory versus factor classes [54].

Data mining techniques for the assessment of landslide susceptibility at a regional scale require the use of large amounts of non-morphometric factors for reliable analysis [10,55], therefore factors in the geographical and geological domains i.e., landuse, soil type, lithology and distance to faults were used [11]. The landuse map for the study area was extracted from the Land Use Status Map of the Lang Son province at a scale of 1:50,000, a result of the Status Land Use Project of the National Land Use Survey in Vietnam in 2010. For analysis, the landuse map was constructed with nine classes (Figure 2e). These classes were generalized from 21 original types in the Land Use Status Map. The soil type map for the study area was extracted from the National Pedology Maps at scale of 1:100,000. A total of eight layers were constructed (Figure 2f).

The geological map that provides information on underlying bedrock is an important factor for landslide modeling [56]. For this research, the geological map was constructed based on four tiles of the Geological and Mineral Resources Map (GMRM) of Vietnam at 1:50,000 scale. This map is selected because no geological map with larger scales is available for the study area. These maps were constructed by Quoc, et al. [57] and then updated by Truong, Nhi, Phuc, Quyet and The [48]. Seven geologic units (Figure 2g) were distinguished for the analysis based on lithological
similarities [56]: (i) quaternary (Granule, grit, breccia, boulder, sand, clay, and silt); (ii) conglomerate (Na Duong and Khon Lang formations); (iii) basalt (Tam Danh formation); (iv) siltstone (Na Khuat và Dong Dang formations); (v) limestone (Diem He and Bac Son formations); (vi) sandstone (Lang Son, Mau Son, and Ha Coi formations); and (vii) tuff (Tam Lung formation). Distance to faults was included in this analysis because fracturing and shearing play critical roles in slope instability [58]. In this study, the distance to faults map (Figure 2h) was compiled by buffering the fault lines. Five fault buffer categories were constructed based on an analysis of the landslide inventory map: 0–100, 100–200, 200–300, 300–400, and >400 m.

**Table 1. Landslide influencing factors and their classes used in this study.**

<table>
<thead>
<tr>
<th>No</th>
<th>Influencing Factors</th>
<th>Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Slope (°)</td>
<td>(1) 0–8; (2) 8–15; (3) 15–25; (4) 25–35; (5) 35–45; (6) &gt;45</td>
</tr>
<tr>
<td>2</td>
<td>Slope length (m)</td>
<td>(1) 0–10; (2) 10–30; (3) 30–50; (4) 50–80; (5) 80–120; (6) &gt;120</td>
</tr>
<tr>
<td>3</td>
<td>Aspect</td>
<td>(1) Flat; (2) North; (3) Northeast; (4) East; (5) Southeast; (6) South; (7) Southwest; (8) West; (9) Northwest</td>
</tr>
<tr>
<td>4</td>
<td>Curvature</td>
<td>(1) &lt;2; (2) –2 to –0.01; (3) –0.01 to 0.01; (4) 0.01 to 2; (5) &gt;2</td>
</tr>
<tr>
<td>5</td>
<td>Elevation (m)</td>
<td>(1) &lt;260; (2) 230–300; (3) 300–350; (4) 350–430; (5) 450–550; (6) &gt;550</td>
</tr>
<tr>
<td>6</td>
<td>Valley depth (m)</td>
<td>(1) &lt;10; (2) 10–30; (3) 30–50; (4) 50–70; (5) 70–100; (6) &gt;100</td>
</tr>
<tr>
<td>7</td>
<td>Toposhape</td>
<td>(1) Ridge; (2) Saddle; (3) Flat; (4) Ravine; (5) Convex hillside; (6) Saddle hillside; (7) Slope hillside; (8) Concave hillside; (9) Inflection hillside; (10) Unknown hillside</td>
</tr>
<tr>
<td>8</td>
<td>TWI</td>
<td>(1) &lt;5; (2) 5–6; (3) 6–7; (4) 7–8; (5) 8–9; (6) &gt;9</td>
</tr>
<tr>
<td>9</td>
<td>SPI</td>
<td>(1) &lt;30; (2) 30–100; (3) 100–200; (4) 200–300; (5) &gt;300</td>
</tr>
<tr>
<td>10</td>
<td>STI</td>
<td>(1) &lt;10; (2) 10–30; (3) 30–50; (4) 50–70; (5) &gt;70</td>
</tr>
<tr>
<td>11</td>
<td>Landuse</td>
<td>(1) Annual crop land; (2) Populated area; (3) Protective forest land; (4) Productive forest land; (5) Paddy land; (6) Barren land; (7) Perennial crop land; (8) Water surface land; (9) Grassland</td>
</tr>
<tr>
<td>12</td>
<td>Soil type</td>
<td>(1) Ferralic acrisols; (2) Dystric gleysoils; (3) Plinthic acrisols; (4) Water area; (5) Dystric fluvisols; (6) Eutric fluvisols; (7) Rhodic ferralsols; (8) Rocky mountain</td>
</tr>
<tr>
<td>13</td>
<td>Lithology</td>
<td>(1) Conglomerate; (2) Basalt; (3) Quaternary deposit; (4) Siltstone; (5) Limestone; (6) Sandstone; (7) Tuff</td>
</tr>
<tr>
<td>14</td>
<td>Distance to faults (m)</td>
<td>(1) 0–100; (2) 100–200; (3) 200–300; (4) 300–400; (5) &gt;400</td>
</tr>
</tbody>
</table>
Figure 2. Cont.
3. Theoretical Background of the Methods Used

3.1. Instance Based Learning Algorithm

The $k$-nearest neighbor ($k$-NN) is an instance-based learning algorithm that uses the nearest distance as a threshold to determine whether pixels will be added to existing clusters or a new cluster is created [59]. Despite the simplicity of its theoretical properties, this algorithm belongs to top ten methods in data mining and has been considered to be one of the most useful and effective algorithms for classification [60].

Consider a training dataset $(X, Y)$ with $X = (X_1, X_2, \ldots, X_n)$ and $Y \in [1,0]$. In the current context of landslide susceptibility analysis, $X_i$ is an input vector that represents the 14 influencing factors (slope, slope length, aspect, curvature, elevation, valley depth, toposhade, TWI, SPI, STI, landuse, soil type, lithology, and distance to faults), and $Y_i$ is the two classes, landslide and non-landslide. In the training phase, the input dataset is mapped into feature space and then the feature space is partitioned into multiple regions where decision boundaries are based on the similarity in the content of the dataset [59]. In the prediction phase, distances between pixels in the new dataset and all the training pixels are calculated. Based on $k$ thresholds, the determination of nearest neighbors is carried out by sorting these distances. Then landslide and non-landslide classes for each of the nearest neighbors are determined. Finally, the prediction value for each pixel is obtained using simple majority of the class of nearest neighbors.

The decision rule of the $k$-NN model could be written as

$$f(\text{newdata}, Y_i) = \sum_{X_i \in kNN} \text{sim(\text{newdata}, X_i)}Z(X_i, Y_i)$$  \hspace{1cm} (1)

where $\text{sim(\text{newdata}, X_i)}$ is the similarity between new data and the training data $X_i$; and $Z(X_i, Y_i)$ is the category value of the training data $X_i$. 

Figure 2. Landslide influencing factors: (a) Slope map; (b) aspect; (c) elevation; (d) valley depth; (e) landuse; (f) soil type; (g) lithology; and (h) distance to faults. ACL: Annual Crop Land; PA: Populated Area; PTL: Protective Forest Land; PDL: Productive Forest Land; PL: Paddy Land; BL: Barren Land; PCL: Perennial Crop Land; WSL: Water Surface Land; GL: Grass Land; FA: Ferralic Acrisols; DG: Dystric Gleysols; PA: Plinthic Acrisols; WA: Water Area; DF: Dystric Fluvisols; EF: Eutric Fluvisols; RF: Rhodic Ferralsols; RM: Rocky mountain; CO: COnglomerate; and QD: Quaternary Deposit.
3.2. Rotation Forest Ensemble

An ensemble-based system can be constructed by combining individual classifiers in which the individual classifiers could be trained using different: (i) subsets of features; (ii) training data sets; (ii) parameters of a given classifier; or (iv) classifier models [61]. The Rotation forest ensemble refers to the first case and is a technique formally introduced by Rodriguez, Kuncheva and Alonso [43]. This ensemble framework is a combination of the Random Subspace and Bagging techniques with Principal Component Analysis (PCA) to construct an ensemble classifier [44].

Using the training dataset \((X, Y)\) with \(X = (X_1, X_2, \ldots, X_n)\) and \(Y \in [1,0]\), the training phase of Rotation Forest ensemble is as follows:

**Step 1.** Setup parameters: Choose \(k\)-NN algorithm as the base classifier, the ensemble size (L), the number of feature subsets (K).

**Step 2.** Training the classifier ensemble model: For \(i = 1 \ldots L\).

(a) Split \(X\) into \(K\) subsets (each subset contains \(M\) features): \(S_{i,j}\) for \(j = 1 \ldots K\).

Generate \(S^*_{i,j}\) by eliminating randomly a subset of classes.

Generate new set \(S''_{i,j}\) by selecting a bootstrap sample with a size 75% from \(S^*_{i,j}\).

Perform Principle Component Analysis on \(S^*_{i,j}\) to obtain coefficients \(a_{i,j}^{(1)}, \ldots, a_{i,j}^{(M_k)}\) and then store in a matrix \(C_{i,j}\).

Arrange the matrix \(C_{i,j}\) in a rotation matrix \(R_i\):

\[
R_i = \begin{bmatrix}
a_{i,1}^{(1)}, \ldots, a_{i,1}^{(M_k)} & [0] & \cdots & [0] \\
[0] & a_{i,2}^{(1)}, \ldots, a_{i,2}^{(M_k)} & \cdots & [0] \\
\vdots & \vdots & \ddots & \vdots \\
[0] & [0] & \cdots & a_{i,k}^{(1)}, \ldots, a_{i,k}^{(M_k)}
\end{bmatrix}
\] (2)

Construct \(R^a_i\) by rearrange the rows of \(R_i\) to match the order of the influencing factors in the training dataset.

(b) Construct base classifier \(D_i\) using the training set \(YR^a_i\).

**Step 3.** Calculating landslide susceptibility index.

The operation of the rotation forest for new data \(X_N\) is as follows: (i) Build the transformed data \(Y_N = X_NR^a_i\) run it through the \(L\) classifiers to get degree of support for the landslide and the non-landslide classes, \(d_{ij}\) with \(i = 1, \ldots, L; j = 1, 2\) for the landslide and the non-landslide classes, respectively. (ii) Landslide susceptibility index (LSI) is then estimated for each pixel of \(X_N\) using the average combination method as follows:

\[
LSI = \frac{1}{L} \sum_{i=1}^{L} d_{i1}
\] (3)

4. Proposed Hybrid Modeling Approach Based on Instance Based Learning Algorithm and Rotation Forest Ensemble for Spatial Prediction of Rainfall-Induced Shallow Landslides

This section presents the proposed hybrid modeling approach for spatial prediction of rainfall-induced shallow landslides. The hybrid model was established based on an instance based learning algorithm (\(k\)-NN) and Rotation Forest ensemble (RF). It is noted that data preparation and processing were carried out using ArcGIS@10.2 (ESRI Inc., Redlands, CA, USA, 2016), IDRISI Selva 17.0 (Clark University, Worcester, MA, USA, 2012), and R programming [62]. The RF ensemble code is
available at Kuncheva [63], whereas the proposed hybrid model was programmed by the authors in Matlab environment. Overall concept of the proposed hybrid modeling approach is shown in Figure 3.

![Figure 3. Overall concept of the proposed hybrid modeling approach in this study.](image)

4.1. The GIS Database

First, a GIS database for the study area was constructed. The database includes: (i) a landslide inventory map with 172 landslide locations; and (ii) 14 influencing factors (slope, slope length, aspect, curvature, elevation, valley depth, toposhade, topographic wetness index (TWI), stream power index (SPI), sediment transport index (STI), landuse, soil type, lithology, and distance to faults). These influencing factors were converted into a grid format with a resolution of 5 m.

For building susceptibility models, 120 landslide locations (70%, 3973 landslide pixels) were randomly selected for training the models, while the remaining landslides (1664 landslide pixels) were used for the model validation. The same amount of non-landslide pixel cells were randomly generated in the landslide-free area of the study area, and then an extraction process was conducted to obtain values of the fourteen landslide influencing factors for the training and validation data [23]. Lastly, a coding process proposed by [11] was used to prepare the training data and validation data for the proposed hybrid model.

4.2. Feature Selection

The quality of models may be affected negatively with the use of some redundant input variables [3]. Therefore, predictive abilities of influencing factors should be assessed using feature selection. The results could be used for the determination of the best subset of influencing factors that not only have high predictive abilities to the output but are also uncorrelated with each other [3]. For this study, the Information Gain technique that has been successfully used recently for feature selection and predictive ability assessment was [64] used.

The information Gain (IG) is estimated using Equation (4):

$$IG = Info(D) - \sum_{j=1}^{m} \frac{S_j}{|S|} Info(D_j); Info(D) = \sum_{i=1}^{n} \frac{n(Y_i, S)|S|}{|S|} \log_2 \frac{n(Y_i, S)|S|}{|S|}$$

where $D$ is the landslide dataset that consists of $n$ samples and $m$ influencing factor; $n(Y_i, D)$ is the number of samples associated with the class $Y_i$, landslide or non-landslide; and $S_j$ is the class $j$ of influencing factor $S$. 

---

1. GIS database

<table>
<thead>
<tr>
<th>Factor</th>
<th>Coding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slope</td>
<td>X1</td>
</tr>
<tr>
<td>Slope length</td>
<td>X2</td>
</tr>
<tr>
<td>Aspect</td>
<td>X3</td>
</tr>
<tr>
<td>Curvature</td>
<td>X4</td>
</tr>
<tr>
<td>Elevation</td>
<td>X5</td>
</tr>
<tr>
<td>Valley depth</td>
<td>X6</td>
</tr>
<tr>
<td>Toposhaide</td>
<td>X7</td>
</tr>
<tr>
<td>Lithology</td>
<td>X10</td>
</tr>
<tr>
<td>Soil type</td>
<td>X11</td>
</tr>
<tr>
<td>Landuse</td>
<td>X12</td>
</tr>
<tr>
<td>Lithology</td>
<td>X13</td>
</tr>
</tbody>
</table>

Landslide polygon is coded as Y

- Aerial photos
- Historical landslide polygons
- Topographic maps 1:50000, 1:100000
- Geologic maps 1:50000
- Soil type map 1:100000
- Landuse map 1:50000

2. Feature selection using the Information Gain (IG)

3. The hybrid model: configuration and training

4. Performance assessment and the final trained hybrid model
4.3. The Hybrid Model: Configuration and Training

With the k-NN selected as the based classifier, the configuration of the hybrid model includes determination of: (i) \( k \) value and the distance metric; and (ii) ensemble size (L) and the number of feature subsets (K). Since no rule of thumb exists for finding the optimal value of \( k \), we use a trial and error method as suggested by Pandya, et al. [65] for finding the best value of \( k \). Accordingly, the best value of \( k \) for this study was determined using the ten folds cross-validation method [44] by varying values of \( k \) versus classification accuracy estimated on the training data and the validation data. For distance metrics, Euclidean, Chebyshev, and Minkowski distances are widely used [66], therefore a trial and error test was carried out on the three distance metrics to select the best one.

Regarding the ensemble size, the size of 10 was used due to ability to obtain high prediction performance of classifier ensembles as suggested in Kuncheva and Rodríguez [67]. Thus, the training dataset were separated into 10 subsets, and each subset was used to build a k-NN classifier. Finally, a committee was established with 10 k-NN classifier members. The number of feature subsets (K) influences also to performance of the hybrid model. In this study, K is selected based on a trial and error method and K = 8 is the best for the data at hand.

Finally, the model was trained and validated using the training dataset and the validation dataset using statistical criteria in Section 5.3.

4.4. Performance Assessment and the Final Trained Hybrid Model

Modeling of landslide susceptibility can be considered a two-class problem where the outputs are labeled as landslide (LS) and non-landslide (NLS) classes. Therefore, four possible outcomes, true positive (TP), false positive (FP), true negative (TN), and false negative (FN), are used to estimate performance evaluation metric such as sensitivity, specificity, positive and negative predictive values [3,44]. Accordingly, the performance of landslide susceptibility models was evaluated using classification accuracy, area under the Receiver Operating Characteristic (AUC), Kappa statistic, and several statistics evaluation measures [28,68–70].

Classification accuracy is considered a primary statistical metric that gives a proxy measure of overall performance of susceptibility models and is defined as the percentage of landslide and non-landslide pixels that are correctly classified. Goodness of fits and prediction capability of landslide models can be summarized with the use of AUC that is calculated from the area under the Receiver Operating Characteristic (ROC) curve. Interpretation of AUC values were defined as poor (<0.7), fair (0.7–0.8), good (0.8–0.9), and excellent (0.9–1.0) [71].

For the case of Kappa statistic, since this metric is a percent reduction in estimation measure that takes the cost of error into account, Kappa statistic is a good statistical measure for the inspection of landslide models. Kappa statistic value of 0 means that the agreement between the landslide models and input data is the same as one found by chance, whereas Kappa statistic value larger than 0.9 indicates that it is more than 90% better than random.

5. Results and Analysis

5.1. Determination of the Best Distance Metric and k Value

Figure 4 describes the change of the classification accuracies and AUC when the value of \( k \) is varied. The classification accuracy on the training data is generally decreased when the value of \( k \) is increased. The highest accuracy is 86.7% with \( k \) equal 1, decreased to 83.4% with \( k \) equal 21 and then the accuracy is generally stabilized. In contrast, the classification accuracy on the validation data is increased when we increase the \( k \) value. The accuracy increases from the lowest one (69.4%) with \( k \) equal 1 to the highest one (75.9%) with \( k \) is 21. The AUC of the validation data is also increased with the increasing of \( k \), from the lowest value 0.698 with \( k \) equal 1 to 0.832 with \( k \) is 21, and at this point AUC is generally stabilized. Therefore, the nearest neighbor \( k \) equal 21 is selected for this analysis.
The analysis result in this study shows that no multicollinearity existed between any of the 14 factors (VIF, VIF = 1/TOL) indices [74–76] that are widely used to measures of the degree of multicollinearity. If VIF exceeds 10 or TOL is less than 0.1 indicates multicollinearity [77].

5.2. Feature Selection and Predictive Ability of Landslide Influencing Factors

Table 2 showed the test result on four distance metrics for this study. We see that the landslide model with Manhattan distance has the highest performance. The classification accuracy is 83.2% and 75.9% for the training dataset and the validation dataset, respectively; therefore, Manhattan distance is selected for this study. This finding agrees with Bours [72,73], who concluded that Manhattan distance yielded the best performance in various studies.

**Table 2.** Classification accuracy of the k-NN model with different distance metrics.

<table>
<thead>
<tr>
<th>No.</th>
<th>Distance Metrics</th>
<th>Classification Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Training Data</td>
</tr>
<tr>
<td>1</td>
<td>Euclidean</td>
<td>83.3</td>
</tr>
<tr>
<td>2</td>
<td>Manhattan</td>
<td>83.4</td>
</tr>
<tr>
<td>3</td>
<td>Chebyshev</td>
<td>79.6</td>
</tr>
<tr>
<td>4</td>
<td>Minkowski</td>
<td>83.3</td>
</tr>
</tbody>
</table>

5.2. Feature Selection and Predictive Ability of Landslide Influencing Factors

To detect if the influencing factors are correlated, the Tolerance (TOL) and Variance Inflation Factors (VIF, VIF = 1/TOL) indices [74–76] that are widely used to measures of the degree of multicollinearity was used. If VIF exceeds 10 or TOL is less than 0.1 indicates multicollinearity [77]. The analysis result in this study shows that no multicollinearity existed between any of the 14 influencing factors (Table 3).

**Table 3.** Correlation assessment and Information Gain (IG) of influencing factors.

<table>
<thead>
<tr>
<th>No.</th>
<th>Influencing Factor</th>
<th>Tolerance</th>
<th>VIF</th>
<th>IG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Aspect</td>
<td>0.88</td>
<td>1.14</td>
<td>0.20</td>
</tr>
<tr>
<td>2</td>
<td>Slope</td>
<td>0.38</td>
<td>2.63</td>
<td>0.19</td>
</tr>
<tr>
<td>3</td>
<td>Sediment transport index</td>
<td>0.16</td>
<td>6.15</td>
<td>0.11</td>
</tr>
<tr>
<td>4</td>
<td>Stream power index</td>
<td>0.18</td>
<td>5.68</td>
<td>0.06</td>
</tr>
<tr>
<td>5</td>
<td>Distance to faults</td>
<td>0.90</td>
<td>1.11</td>
<td>0.05</td>
</tr>
<tr>
<td>6</td>
<td>Toposhade</td>
<td>0.68</td>
<td>1.46</td>
<td>0.05</td>
</tr>
<tr>
<td>7</td>
<td>Topographic wetness index</td>
<td>0.59</td>
<td>1.69</td>
<td>0.05</td>
</tr>
<tr>
<td>8</td>
<td>Curvature</td>
<td>0.68</td>
<td>1.47</td>
<td>0.05</td>
</tr>
<tr>
<td>9</td>
<td>Lithology</td>
<td>0.88</td>
<td>1.14</td>
<td>0.04</td>
</tr>
<tr>
<td>10</td>
<td>Landuse</td>
<td>0.91</td>
<td>1.10</td>
<td>0.03</td>
</tr>
<tr>
<td>11</td>
<td>Slope length</td>
<td>0.46</td>
<td>2.19</td>
<td>0.03</td>
</tr>
<tr>
<td>12</td>
<td>Soil type</td>
<td>0.94</td>
<td>1.07</td>
<td>0.03</td>
</tr>
<tr>
<td>13</td>
<td>Valley depth</td>
<td>0.91</td>
<td>1.10</td>
<td>0.02</td>
</tr>
<tr>
<td>14</td>
<td>Elevation</td>
<td>0.91</td>
<td>1.11</td>
<td>0.01</td>
</tr>
</tbody>
</table>
The result of the feature selection analysis using the Information Gain techniques is shown in Table 3. We observe that the aspect (IG = 0.2) and the slope (IG = 0.19) have the highest predictive ability values. It is closely followed by Sediment transport index (IG = 0.11) and the stream power index (IG = 0.06). It is reasonable because the slope is considered as the most important factors in landslide modeling [78–80]. The aspect reveals a high predictive ability because in this study 82.8% of the landslide pixels are occurred in south, southeast, and southwest facing slopes [46]. These slopes are the main facing directions of tropical rainstorms in the northeast of Vietnam [81,82].

The distance to faults, the toposhade, the topographic wetness index, the curvature and the lithology have almost equal predictive ability. The lowest predictive ability is for the elevation factor where IC is of 0.01. Although IG value is varies among factors, none of them reveals null value; therefore all the factors were used for building the hybrid model.

5.3. Model Training and Assessment

The training result of the proposed hybrid model is shown in Table 4. We see that the hybrid model has a high degree of fit with the training data where the classification accuracy is 85.8% and AUC is of 0.948. The classification accuracy of the hybrid model is higher than 2.4% those obtained by the base classifier. The positive predictive value is 94.4% indicating that the probability the hybrid model classifies pixels correctly in the landslide class is 94.4%. The negative predictive value is 77.3%, which means that the probability the hybrid model classifies pixels to the non-landslide class is 77.3%. The sensitivity is 80.6% indicating that 80.6% of the landslide pixels in this study are classified to the landslide class correctly. The specificity is 93.2% indicating 93.2% of non-landslide pixels are classified to the non-landslide class correctly. Kappa statistic is 0.716 demonstrating that it is 71.6% better than random, a substantial agreement between the models and the training data.

<table>
<thead>
<tr>
<th>No</th>
<th>Parameter</th>
<th>Proposed Hybrid Model</th>
<th>Random Forest Model</th>
<th>J48 Decision Trees Model</th>
<th>Neural Nets Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>True positive</td>
<td>3579</td>
<td>3637</td>
<td>3531</td>
<td>3528</td>
</tr>
<tr>
<td>2</td>
<td>True negative</td>
<td>2931</td>
<td>3385</td>
<td>3296</td>
<td>2781</td>
</tr>
<tr>
<td>3</td>
<td>False positive</td>
<td>214</td>
<td>156</td>
<td>262</td>
<td>265</td>
</tr>
<tr>
<td>4</td>
<td>False negative</td>
<td>862</td>
<td>408</td>
<td>497</td>
<td>1012</td>
</tr>
<tr>
<td>5</td>
<td>PPV (%)</td>
<td>94.4</td>
<td>95.9</td>
<td>93.1</td>
<td>93.0</td>
</tr>
<tr>
<td>6</td>
<td>NPV (%)</td>
<td>77.3</td>
<td>89.2</td>
<td>86.9</td>
<td>73.3</td>
</tr>
<tr>
<td>7</td>
<td>Sensitivity (%)</td>
<td>80.6</td>
<td>89.9</td>
<td>87.7</td>
<td>77.7</td>
</tr>
<tr>
<td>8</td>
<td>Specificity (%)</td>
<td>93.2</td>
<td>95.6</td>
<td>92.6</td>
<td>91.3</td>
</tr>
<tr>
<td>9</td>
<td>Accuracy (%)</td>
<td>85.8</td>
<td>92.6</td>
<td>90.0</td>
<td>83.2</td>
</tr>
<tr>
<td>10</td>
<td>Kappa index</td>
<td>0.716</td>
<td>0.851</td>
<td>0.799</td>
<td>0.663</td>
</tr>
<tr>
<td>11</td>
<td>AUC</td>
<td>0.948</td>
<td>0.981</td>
<td>0.942</td>
<td>0.905</td>
</tr>
</tbody>
</table>

The prediction performances of the hybrid model are assessed using the validation data that were not used during the training phase. The detailed result is shown in Table 5. The result shows that the hybrid model performs well where the classification accuracy is 76.1% and AUC is of 0.848. The positive predictive value of 75.5% indicates that the probability the ensemble model classifies pixels correctly in the landslide class is 75.5%. The negative predictive value is 76.8%, indicating that the probability the hybrid model classifies pixels to the non-landslide class is 76.8%. The sensitivity of 76.5% indicates that 76.5% of the landslide pixels are classified correctly to the landslide class. The specificity is 76.1%, indicating 76.1% of non-landslide pixels are classified to the non-landslide class correctly. Kappa statistic is 0.523 indicating a moderate agreement between the models and the validation data.
Table 5. Model validation using the validation data (PPV: Positive predictive value; NPV: Negative predictive value).

<table>
<thead>
<tr>
<th>No</th>
<th>Parameter</th>
<th>Proposed Hybrid Model</th>
<th>Random Forest Model</th>
<th>J48 Decision Trees Model</th>
<th>Neural Nets Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>True positive</td>
<td>1256</td>
<td>762</td>
<td>1017</td>
<td>1227</td>
</tr>
<tr>
<td>2</td>
<td>True negative</td>
<td>1278</td>
<td>1528</td>
<td>1421</td>
<td>1176</td>
</tr>
<tr>
<td>3</td>
<td>False positive</td>
<td>408</td>
<td>902</td>
<td>647</td>
<td>437</td>
</tr>
<tr>
<td>4</td>
<td>False negative</td>
<td>386</td>
<td>135</td>
<td>242</td>
<td>488</td>
</tr>
<tr>
<td>5</td>
<td>PPV (%)</td>
<td>75.5</td>
<td>45.8</td>
<td>61.1</td>
<td>72.2</td>
</tr>
<tr>
<td>6</td>
<td>NPV (%)</td>
<td>76.8</td>
<td>91.9</td>
<td>85.5</td>
<td>70.7</td>
</tr>
<tr>
<td>7</td>
<td>Sensitivity (%)</td>
<td>76.5</td>
<td>85.0</td>
<td>80.78</td>
<td>71.6</td>
</tr>
<tr>
<td>8</td>
<td>Specificity (%)</td>
<td>75.8</td>
<td>62.9</td>
<td>68.71</td>
<td>72.9</td>
</tr>
<tr>
<td>9</td>
<td>Accuracy (%)</td>
<td>76.1</td>
<td>68.8</td>
<td>73.3</td>
<td>72.2</td>
</tr>
<tr>
<td>10</td>
<td>Kappa index</td>
<td>0.523</td>
<td>0.376</td>
<td>0.466</td>
<td>0.444</td>
</tr>
</tbody>
</table>

The performance and prediction power of the hybrid model is further verified using the success-rate and prediction-rate method [83] as suggested in [35]. The success-rate curve was obtained by comparing the landslide susceptibility indices with the landslide pixels in the training data (3793 landslide pixels). In the same way, the prediction-rate curve was constructed using the landslide pixels in the validation data (1164 landslide pixels). Then, the areas under the two curves (AUC) were estimated (Figure 5). It could be observed that the AUC of the success-rate is 0.944, indicating a high degree of fit of the ensemble model with the training pixels. The AUC of the prediction rate is 0.846, indicating that the prediction power of the model is high.

![Figure 5. Success rate and prediction rate curves, and their areas under the curve (AUC) for the landslide susceptibility map in this study.](image)

5.4. Cartographic Presentation of the Landslide Susceptibility Map

Once the ensemble model was successfully constructed, it was used to calculate the landslide susceptibility index for each of all the pixels of the study area and then, the result is converted to a GIS format to open in the ArcGIS 10.1 software using an application developed in C++ programming. One of critical concerns in landslide susceptibility modeling is to interpret the classes of the resulting landslide susceptibility map. For this purpose, a graphical curve was constructed based on the cumulative percentage of landslide pixels versus landslide susceptibility map (Figure 6). First,
the landslide inventory map was overlaid with the landslide susceptibility map to extract a landslide pixel value table. Then the landslide pixel values were descending sorted corresponding to landslide susceptibility indices and cumulative percentages of landslide pixels and the susceptibility maps were estimated.

According to Chung, et al. [84], the study area should be classified into five classes based on the susceptibility index values and five percent of pixels with the highest values can be classified into the “very high” susceptibility class. Therefore, the landslide susceptibility map in this study is classified as follows: (i) very low (40%); (ii) low (20%); (iii) moderate (20%); (iv) high (15%); and (iv) very high (5%). Finally, thresholds that are used to separate these five susceptibility degrees are determined. The landslide susceptibility map result is shown in Figure 7.

Landslide density analysis was carried out for these susceptibility classes by overlaying all the landslide pixels on the landslide susceptibility map, and then density values were calculated. Theoretically, these values should increase from the very low to the very high class [23]. The result is shown in Figure 8. We see that landslide density increases smoothly and gradually from the very low to the very high classes in this study area.

Figure 6. Cumulative percentage of landslide pixels versus landslide susceptibility map.
5.5. Usability Assessment of the Proposed Hybrid Model

Figure 7. Landslide susceptibility map using the proposed hybrid model for the study area.

Figure 8. Landslide density plots of susceptibility classes for the study area (VH: Very high).
5.5. Usability Assessment of the Proposed Hybrid Model

Since this study aims to propose a new approach for landslide susceptibility mapping, therefore the usability of the proposed hybrid model should be assessed. Accordingly, the performance of the hybrid model was compared with those produced by several state-of-the-art methods such as Random Forest, J48 Decision Trees, and Multi-layer Perceptron Neural Networks (Neural Nets). Random Forest is selected because this is an innovative technique and has just recently used for landslide susceptibility but proven great performance [85,86]. To build the Random Forest model for this study, 500 trees were used, as suggested in Stevens, et al. [87]. For the case of J48 Decision Trees, this method has successfully applied in many fields with high accuracy, including landslide susceptibility [31,34]. To construct the J48 Decision Trees model in this study, 10 pixels per leaf and the confident factor of 0.15 were used. These are the best parameter values that were determined based on a test in Tien Bui, Pradhan, Revhaug and Trung Tran [34]. Neural Net is considered to be one of the best methods for modeling of complex problems such as landslides [3]. For building the Neural Nets model, the logistic sigmoid is used as the activation function. Training iteration, learning rate, momentum were used as 500, 0.3, and 0.2, respectively, as suggested in [88,89]. The best structure of the Neural Nets model with 14 input layers, one hidden layer (six neurons), and an output layer was determined using the method in Tien Bui, Tuan, Klempe, Pradhan and Revhaug [3].

The results of the training landslide susceptibility models Random Forest, J48 Decision Trees, and Neural Net are shown in Table 4. The results show that all the three models have high performances with the training data. The highest degree of fit is for the Random Forest model (AUC = 0.981 and Accuracy is 92.57%). The performances of the hybrid model and J48 Decision Trees models are almost the same. In contrast the Neural Net model performed worst. The prediction performances of the three models were assessed using the validation data and the results are shown in Table 5. It could be observed that the overall prediction performances of the three susceptibility models are lower than those obtained from the proposed hybrid model, in terms of accuracy, kappa index, and PPV. Although AUC of the Random forest model (0.857) in Figure 9 is almost equal that of the proposed model (0.848), the PPV of the landslide class is only 45.8% (Table 5) indicating that the AUC of the Random forest model is strongly influenced by the non-landslide pixels. Therefore, the prediction capability of landslides and AUC of the Random Forest model did not correspond strictly. This finding is in agreement with [35,90]. In addition, the Random forest model presents an overfitting problem (Tables 4 and 5). This is because the prediction is made based on the weighted average [91,92] of the training dataset, therefore it was difficult to extrapolate values in the validation dataset that were somewhat outside its known values [93].

![Figure 9](image-url)
In order to confirm the prediction performances of the proposed hybrid model better than the three susceptibility models in this study, McNemar’s test at the 95% significant level is used. The null hypothesis is that there is no difference of the prediction performances between the classifier ensemble model and each of the three landslide susceptibility models. The Chi-square ($\chi^2$) is then calculated (using Equation (5)) and then Chi-square comparisons with the critical table values at the significant level $\alpha = 5\%$ are employed to assess the significance of differences between the susceptibility models. If the Chi-square value exceeds the critical table values of 3.841, the null hypothesis is rejected and the prediction power of the two susceptibility models is said to be significantly different [63].

$$
\chi^2 = \frac{(|PI_{ij} - PI_{ji}| - 1)^2}{PI_{ij} + PI_{ji}} \tag{5}
$$

where $PI_{ij}$ is the number of pixels misclassified by the susceptibility model $i$; and $PI_{ji}$ is the number of pixels misclassified by the susceptibility model $j$.

The result is shown in Table 6. We see that the lowest Chi-square (10.081) is for the proposed hybrid model vs. the Neural Net model that exceeds the critical table values of 3.841, whereas the $p$-value (0.0015) is less than 0.05. The other Chi-square values are far larger than the critical table values and the $p$-values are also far smaller than 0.05, therefore we conclude that the prediction performance of the proposed hybrid model is significantly higher than the other landslide models in this study.

<table>
<thead>
<tr>
<th>No</th>
<th>Pairwise Comparison</th>
<th>Chi-Square ($\chi^2$)</th>
<th>$p$-value</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The hybrid model vs. Random Forest</td>
<td>687.077</td>
<td>&lt;0.0001</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>The hybrid model vs. J48 Decision Trees</td>
<td>181.845</td>
<td>&lt;0.0001</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>The hybrid model vs. Neural Net</td>
<td>10.081</td>
<td>0.0015</td>
<td>Yes</td>
</tr>
</tbody>
</table>

6. Discussion and Conclusion

The most effective way to prevent casualties and economic losses due to landslides is to avoid constructions in the vicinity of steep terrains [94]. However, it is not possible in many areas due to limitation of land and the rapid growth of human population [95], therefore high quality of landslide susceptibility and hazards maps is an important tool for reducing landslide risk through landuse
planning and management. However, the prediction performances of landslide susceptibility models is still one of the most debated subjects in recent decades [96]. Literature review shows that a perfect landslide model that makes no error is almost impossible; therefore, a highly accurate model for particular area requires assessment studies to find an algorithm with the highest overall performances. For this purpose, classifier ensemble approaches have been considered to be important strategies to enhance model performances [63]. Only several of the increment of percentage of the prediction accuracy could influence the resulting landslide susceptibility [3,97]. We address this issue in this study by proposing a novel hybrid machine learning approach for mapping of rainfall-induced shallow landslides using GIS.

The proposed model is a combination of an instance base learning algorithm ($k$-NN) and the Rotation Forest ensemble that has seldom been used for modeling of landslide. The $k$-NN is one of the most well-known nonparametric algorithm and belong to the top 10 algorithms in data mining [60]. Although the $k$-NN algorithm is considered to be a lazy learner due to its simplicity, the algorithm has been demonstrated to be one of the most useful and effective algorithms in data mining applications [98]. The result of this study shows that the base $k$-NN model has a high performance (classification is 83.4%). For the case of the Rotation Forest, this is current state-of-the art ensemble that outperforms other frameworks i.e., Bagging, AdaBoost, and Random Forest [99]. The performance of the based classifier ($k$-NN, classification is 83.4%, Figure 4) was increased 2.4% with the used of the Rotation Forest ensemble (classification is 85.8%, Table 4). The result in this study confirmed that the proposed model performs well in both the training and validation data in terms of classification accuracy, AUC, and other statistical evaluation metrics (Table 3). This result agrees with Althuwaynee, et al. [100] and Tien Bui et al. [31,34] who conclude that the ensemble frameworks increase accuracy of based classifiers significantly.

The overall performance of the proposed model was further compared with those produced by J48 Decision Trees, and Neural Net. They are state-of-the art methods that are widely used in data mining [63]. Although the models in this study fit the training data well, their prediction capabilities are clearly lower than the proposed model (Tables 3–5). To confirm the difference of the prediction performances of the classifier ensemble model and the other susceptibility assessment models, McNamar’s test should further used. The test result shows that, statistically, the prediction performance of the proposed model is significantly higher (Table 6).

The determination of landslide influencing factors is a crucial point and has been discussed [101] and 14 factors have been selected for this analysis based on analysis of the landslide types as well as the failure mechanisms, however, the influence of specific factor on the classification performance should be quantified with the use of feature selection procedures [3]. Redundant factors where predictive ability values are null or negative should be removed from the original dataset. This will help to improve overall performances of resulting models [3,64]. In this study, the predictive abilities of the fourteen influencing factors are quantified with the use of the Information Gain technique. The result shows that aspect and slope have the highest predictive ability value whereas the lowest value is for elevation (Table 2). The result is reasonable because most of the landslides in this study occurred in south, southeast, and southwest facing slopes [46] and they are the main facing directions of tropical rainstorms in the northeast of Vietnam [81]. For slope, this factor is considered to be the most important factor that influences occurrences of landslides in many areas (i.e., [79]). For the case of elevation factor, the elevation in this study area varies from 194.5 m to 800 m and the distribution of landslide pixels are quite even with regard to altitude.

Overall, the results of this study have demonstrated the effectiveness of a classifier ensemble strategy with the use of the $k$-NN algorithm and the Rotation Forest framework for the assessment of landslide susceptibility. The classifier ensemble model outperforms the three susceptibility models in this study; therefore, the proposed model is promising, and could considered as an alternative for the susceptibility mapping of rainfall-induced shallow landslides. Finally, the results in this study may useful for land use planning and management in landslide prone areas.
Acknowledgments: This research was supported by the Geographic Information System group, Department of Business and IT, School of Business, University College of Southeast Norway, Bø i Telemark, Norway. The authors would like to thank four anonymous reviewers for their valuable and constructive comments on the earlier version of the manuscript.

Author Contributions: Quang-Khanh Nguyen and Dieu Tien Bui designed the experiment, processed the data, and wrote the manuscript. Nhat-Duc Hoang, Phan Trong Trinh, Viet-Ha Nguyen, and Isık Yılmaz performed the analysis and checked the results.

Conflicts of Interest: The authors declare that there is no conflict of interest.

References

4. Hong, H.; Pradhan, B.; Xu, C.; Tien Bui, D. Spatial prediction of landslide hazard at the yihuang area (China) using two-class kernel logistic regression, alternating decision tree and support vector machines. Catena 2015, 133, 266–281. [CrossRef]
7. Chung, C.-J.; Fabbri, A.G. Predicting landslides for risk analysis—Spatial models tested by a cross-validation technique. Geomorphology 2008, 94, 438–452. [CrossRef]


45. Pham, B.T.; Tien Bui, D.; Dholakia, M.; Prakash, I.; Pham, H.V.; Mehmood, K.; Le, H.Q. A novel ensemble classifier of rotation forest and naive bayer for landslide susceptibility assessment at the luc yen district, yen bai province (viet nam) using GIS. *Geomat. Nat. Hazard Risk.* 2016, 1–23. [CrossRef]


55. Mancini, F.; Ceppi, C.; Rittrovato, G. GIS and statistical analysis for landslide susceptibility mapping in the Daunia area (Italy). *Nat. Hazards Earth Syst. Sci.* 2010, 10, 1851–1864. [CrossRef]
58. Brideau, M.-A.; Yan, M.; Stead, D. The role of tectonic damage and brittle rock fracture in the development of large rock slope failures. Geomorphology 2009, 103, 30–49. [CrossRef]

