Predictability of Naïve Bayes classifier for lahar hazard mapping by weather radar

Ratih Indri Hapsari¹, Bima Ahida Indaka Sugan², Dandung Novianto¹, Rosa Andrie Asmara², Satoru Oishi³

¹ Department of Civil Engineering, State Polytechnic of Malang, Indonesia
² Department of Information Technology, State Polytechnic of Malang, Indonesia
³ Research Center for Urban Safety and Security, Kobe University, Japan

E-mail: ratih@polinema.ac.id

Abstract. The aim of this study is to develop lahar hazard vulnerability as a warning system by introducing radar-rainfall observation to data mining technique of Naïve Bayes Classifier (NBC). NBC is used to estimate lahar occurrences based on the posterior probability of rainfall, topographic factor, soil moisture, and soil type as predictors. Rainfall intensity and working rainfall were obtained from a weather radar. The soil moisture is derived from SMAP satellite imagery. A river on Mount Merapi, a very active volcano in Indonesia, was selected as the target basin. Observed rainfall and recorded lahar events in Gendol River from October 2016 to February 2018 were divided into a training dataset and a testing dataset. Qualitative evaluation through visual assessment of the hazard map product reveals that the model could estimate the occurrences of lahar. The performance of the model in terms of accuracy, Brier score, and quantitative dichotomous quality indices showed a reasonable skill. The study suggests that the NBC technique is advantageous for estimating lahar occurrences that are displayed on hazard maps. This work is expected to contribute to debris flow hazard mitigation by the data mining approach in volcanic regions.

Keywords: Naïve Bayes Classifier, Lahar, Radar, Merapi Volcano, Gendol River.

1. Introduction
Lahar flow, or volcanic material flow from the mixing of rock and sand triggered by heavy rain [1], is noticed as the most serious secondary impact of volcanic disaster. Soil moisture is also the important contributing factor to the occurrences of debris flow [2]. Due to difficulty of access to mountain flanks, remote sensing techniques are preferable rather than ground observation. A weather radar is a promising equipment for observing rainfall remotely with high accuracy. In recent times, the information of dynamic soil water content can be widely accessed from satellite images. Both data of topography and soil deposit characteristics will provide good information to identify the potential of lahar flow occurrences.

Intuitive prediction can be achieved when records of past events are available. Data mining can be helpful in estimating upcoming event likelihoods that can be displayed as hazard map. The Naïve Bayes Classifier (NBC) method provides a probabilistic prediction system that has been applied in susceptibility assessment. The algorithm is not very complex and does not need large amounts of training data [3]. Previous studies have been limited to the application of NBC for developing landslide vulnerability analyses [2, 4] and no study utilized this approach for lahar application.
In this study, NBC is applied to identify and predict the potential of lahar occurrences for display on a hazard map. NBC is used to discriminate different prone regions based on the features of rainfall, soil moisture, slope, and soil type. This is the first attempt to utilize soil moisture from remote sensing for modelling lahar. The performance of NBC is assessed and predictability is evaluated through a testing dataset. In the last part of the section, NBC model improvement and application to a practical scheme by introducing ensemble rainfall prediction are presented.

2. Methods

2.1. Study area
The study area is the upstream part of Gendol River in the Merapi volcano, in the Province of Yogyakarta Special Region, Indonesia (7.5407°S, 110.4457°E). The radar is located a Merapi Museum (-7.616009°S, 110.424357°E). The flanks of Merapi are vulnerable areas because they are highly populated, wherein 70,000 people live in the primary danger zone. The last major eruption of Merapi was in October 2010. In the following rainy season, high amounts of volcanic material was transported as far as 20 km with 110 km/hr speed. Approximately 60% of volcanic material from the mountain top flowed in Gendol River, buried 21 houses, and caused evacuation of 200 residents. Figure 1 shows the study area.

2.2. Rainfall Data
Rainfall data as a predictor of lahar was observed by the X-band polarimetric Doppler radar, a multi-parameter radar (X-MP radar). It was installed in January 2015 at the Merapi Museum (755 AMSL). This radar could provide sufficient detail of the precipitation system, which allows small intense rain to be detected. The radar reflectivity is processed using a composite algorithm [5] that uses horizontal reflectivity and specific differential phases that can estimate the rainfall intensity reasonably with the two methods. Past rainfall observation in upper Gendol River from October 2016 to February 2018 were utilized in this study and divided into a training dataset and a validation dataset.

Two rainfall parameters are used as predictors; these are rainfall intensity (mm) and working rainfall (mm). Working rainfall is an antecedent rainfall calculated by accumulating the rainfall in the seven days prior to the hourly rainfall [6]. This predictor is classified by threshold lines that were developed previously (see [7]). Eq. (3) is the vertical line showing the judgment of “warning” and Eq. (4) is the diagonal line for lahar emergency judgment:
\[ x = 71.6 \]  
\[ y = -0.37 \times 94.893 \]  

where \( x \) is working rainfall (mm) and \( y \) is hourly rainfall intensity (mm/h). Figure 2, left and right, show the example of rainfall rate and working rainfall spatial distribution taken from radar observations on February 17, 2016, 16.30 LST. Here, the rainfall observations from point rain gauge is also shown (Figure 3). The temporal distribution of working rainfall is shown with a snake-line (right figure). The number of rainfall data points is 41082 for both calibration and validation, comprised of 167 mesh points and 246 time-series data points. Figure 4 shows the distribution of the whole dataset.

2.3. Soil Moisture

Soil water content is identified as a contributing factor for debris mobilization. However, difficulty arises in using the measurement of dynamic soil moisture due to unavailability of ground sensors and the communication network. Therefore, the soil data was drawn from NASA Alaska Satellite Facility and the National Snow and Ice Data Centre, as the Soil Moisture Active Passive (SMAP). SMAP Level 3 products, which are composites of surface soil moisture in fraction or m^3/m^3 unit, are used in this study. The product uses high-resolution radar backscatter gridded at 3 km. Figure 5 illustrates the example of one sheet of SMAP product soil map of the Province of Central Java. Generally, saturated water content ranges from 0.2 to 0.5 vol/vol. Therefore, the categorization of soil moisture follows these five classes: 0.0-0.1, 0.1-0.2, 0.2-0.3, 0.3-0.4, 0.4-0.5 and over 0.5. The soil water content of upper Gendol varies from 0.09751 to 0.57370 m^3/m^3.

2.4. Topography Data

Topography contributes to lahar occurrences because debris flow tends to occur on steep slope angles with high material deposits [8]. Slope steepness was generated from the Digital Elevation Model from SRTM with 90 m spatial resolution (Fig. 2a). Near the summit of the mountain, the slope varies from 0.042 to 0.648 (Fig. 2b). The slope gradient is classified according to Niu et al. [9], where rank 1 is given to the slope of 0°-3°, rank 2 to 3°-6°, rank 3 to 6°-10°, rank 4 to 10°-15°, and rank 5 for 15° and greater. After the classification was made, the 167 data points are displayed in descriptive statistics, as illustrated in Figure 8, which comes from different values of different grids.

2.5. Soil Map

Lahar formation is closely related to the soil structure of deposit materials. Due to the difficulty of observing the spatiotemporal variation of soil properties, the soil type from global soil maps is used to study the dynamics of lahars. The soil data was obtained from the Harmonized World Soil Database from the UNESCO Digital Soil Map of the World with a 30 arc-second mesh. Figure 7 is the soil map, showing that the soil type of the Gendol basin is Andosol and Arenosol. The dominant soil component is sand and silt. It is believed that both soil types can potentially experience soil instability. The number of grids with Andosols and Arenosol type soils is 19 and 148 respectively.

2.6. Naïve Bayes Classifier

The Naïve Bayes approach gives the probability of an event by calculating the event likelihood and summarizing frequency with a combination of the given dataset value [10]. The basic formula of Naïve Bayes Classification is:

\[ P(\text{class} | \text{data}) = \frac{P(\text{data} | \text{class}) P(\text{class})}{P(\text{data})} \]  

(3)
Figure 2. Spatial distribution of rainfall intensity (mm/h) (left) and working rainfall (mm) (right) from X-MP radar observation for February 17, 2016, 16.30 LST

Figure 3. Temporal distribution of rainfall intensity (mm/h) (left) and working rainfall (mm) (right) from automatic rainfall gauge for 12 – 22 December 2017

where \( P(\text{class}|\text{data}) \) is the probability of a class given an event after seeing the event or a posteriori probability, \( P(\text{data}|\text{class}) \) is the probability of an event belonging to a particular class, \( P(\text{class}) \) is past event occurrence probability or a priori probability, and \( P(\text{data}) \) is the probability of that event in a whole dataset (usually neglected). All values of predictors are assumed to be conditionally independent of each other.

2.7. Training Data
The NBC approach requires a set of training data from the four attribute datasets described previously to develop a debris flow prediction system. In the rainy season in 2016 and 2017, there were two lahar incidents, on February 17, 2016 and December 20, 2017, which are regarded as “occurring” data. The number of “occurring” data points for the rainfall attribute is 1002, which consists of 501 data points for calibration and 501 data for validation. The number of “not occurring” data points is 40080, which are divided into 39579 and 501 points for calibration and validation respectively. According to Eq. (1) and Eq. (2), rainfall factor is categorized based on the warning status as well as the occurrence status.
as shown in Figure 9. In the same manner, the classification of soil moisture, slope, and soil type are shown accordingly in this figure.

![Figure 4](image-url)  
**Figure 4.** Distribution of rainfall data in terms of hourly rainfall and working rainfall

![Figure 5](image-url)  
**Figure 5.** Spatial distribution of soil moisture (m$^3$/m$^3$)

2.8. **Model Testing**

After the data training or calibration stage, the predicted event was compared with the actual incident in the disaster database. The event predicted by NBC was then compared with the actual occurrence/no-occurrence as recorded in the lahar event inventory. Success estimation was achieved when the specific event in the specific grid is assigned to the correct class, either occurring or not occurring. The most common quantitative performance index to assess the predictability of probabilistic prediction is the Brier Score:

$$BS = \frac{1}{N} \sum_{t=1}^{N} (f_t - o_t)^2$$  

(4)

where $BS$ is Brier Score, $N$ is number of data, $f_t$ is the probability that was forecasted, and $o_t$ is the actual outcome of the event. When the event does not happen, “0” is given for $o_t$. Conversely, when the event does happen, “1” is assigned to $o_t$. Perfect performance is achieved when the $BS$ is 1.
Other measures to evaluate the prediction accuracy are the critical success index (CSI), the probability of detection (POD), and the false alarm rate (FAR), which come with contingency tables [11]:

\[
CSI = \frac{N_{hit}}{N_{hit} + N_{miss} + N_{false}}
\]  

(5)

\[
POD = \frac{N_{hit}}{N_{hit} + N_{miss}}
\]  

(6)

\[
FAR = \frac{N_{false}}{N_{hit} + N_{false}}
\]  

(7)
where $N_{hit}$ is the number of hit events (lahar present and model predicts), $N_{miss}$ is the number of miss events (lahar present but model does not predict), and $N_{false}$ is the number of false events (lahar not present but model predicts). After a satisfying performance was obtained, the lahar prediction model and disaster warning system based on the NBC approach by introducing short-term rainfall prediction was developed. Figure 10 presents the procedure of lahar warning system development by the NBC approach.

![Figure 8. Distribution of datasets for slope and soil moisture parameters](image)

Figure 8. Distribution of datasets for slope and soil moisture parameters

![Figure 9. Factors of rainfall, soil moisture, slope, and soil type](image)

Figure 9. Factors of rainfall, soil moisture, slope, and soil type

### 3. Results and Discussion

#### 3.1. Likelihood of Predictors Class
The first step of the NBC algorithm is analysing the likelihood of “occurrence” or “no-occurrence” or $P(data|class)$ for each classification of rainfall factor, soil moisture factor, slope factor, and soil type. From 40080 cases of rainfall for calibration, 501 of them are lahar cases. Therefore, the likelihood ratio of lahar occurrences to total events $P(B|A)$ are 0.0125 and 0.9875 for occurrence and no-occurrence, respectively. Through this step, a model that generalizes how lahar attributes relate to the disaster occurrence status was obtained.
The prediction outcome was obtained through this procedure and shown in spatial distribution. The example results are shown in Figure 11 representing “occurrence” and “no-occurrence” events on the lahar hazard map for visual evaluation. The selected events were February 17, 2016, 14.00 LST for a non-lahar case and February 17, 2016, 16.30 LST for a lahar case. Also, the previously unseen data from December 16, 2017, 13.00 LST as a non-lahar case and December 20, 2016, 16.30 LST as a lahar case are shown in this figure. From this figure, it can be seen that both “no-occurrence” events were estimated correctly. In the case of December 17, 2016, 16.30 LST, the bottom of the catchment is estimated to have a risk of occurrence. This is because the hourly rainfall and working rainfall were high in this part.

3.2. NBC Model Predictability

The accuracy of the model was evaluated through accuracy, Brier Score, CSI, POD, and FAR, which are shown in Table 1. Overall, the reproducibility of the NBC model was quite good in predicting lahar occurrences, though the CSI and POD are somewhat low. This fact is attributable to...
the unavailability of spatially distributed lahar measurement along the upper Gendol River. In this study, the lahar events were mainly determined from CCTV cameras and resident reports because there was no sediment measurement on the rivers during the rainfall observation period. Consequently, the exact location of lahar occurrences was not recognized precisely. Moreover, this experiment was based upon the lahar database since 2015 that consists of only two lahar events, which are insufficient to allow the model to learn the features. There are 13 rivers that drain the slopes of Mount Merapi and are vulnerable to debris flow. During the observation period from October 2015 to February 2018, the lahar database recorded at least 16 lahar flow events on the Putih River, Pabelan River, Krasak River, and BLongkeng, in addition to Gendol River, which have the potential to construct the training data. A future study will involve all river basins in Merapi at once as training data.

| Table 1 Performance of the NBC model |
| No. | Indices | Value | Perfect Skill |
|-----|---------|-------|---------------|
| 1   | Accuracy | 82.28% | 100%          |
| 2   | Brier Score | 0.68 | 1             |
| 3   | CSI     | 0.56  | 1             |
| 4   | POD     | 0.56  | 1             |
| 5   | FAR     | 0.00  | 0             |

Developing a reliable spatial prediction of lahar susceptibility is challenging due to the complex nature of volcanic debris flow. The use of NBC is expected to discover hidden patterns from existing geomorphological and hydrological data that lead to lahar flow. NBC often works better in complex real situations than expected ones. Moreover, this easily and quickly provides predictions, which is essential in emergency situations.

Estimation of rain event characteristics in a short duration and scale is indispensable for mitigating lahar disasters. In Figure 10, the comprehensive model for lahar susceptibility mapping involves short-term rainfall predictions. In future investigations, high-resolution nowcasting products from the X-MP radar echo extrapolation model, the advection model, will be introduced to the NBC model with lead times of 1-4 hours. The use of a hydro-meteorological prediction model for debris flow hazard estimation in the Merapi area has been reported by Hapsari et al. [12]. However, it was found that the use of rain-radar prediction in sediment-related disasters involves a high amount of uncertainties. Some of the issues emerging from this finding relate specifically to the use of ensemble prediction instead of single deterministic forecast. The ensemble is generated by perturbing the rain advection vector with its eigenvalues, producing five members (see [13] for details). This model is used to predict the rain echo motion observed by the X-MP radar. By evaluating five ensemble rainfall spatial distributions at once, the river sections that are most likely to be hit by lahar flow are analysed through hazard zoning. Accordingly, the appropriate policy can be chosen based on what hazard is most likely.

4. Conclusion
In this study, lahar hazard vulnerability mapping by Naïve Bayes Classifier (NBC) was developed and evaluated in upper Gendol River Basin. The posterior probability from the predictors, involving rainfall rate, soil moisture, slope factor, and soil type were estimated from the prior probability of lahar events. The model performance is acceptable, indicated by 82.28% accuracy and 0.68 Brier Score. In terms of dichotomous indices, the CSI, POD, and FAR are 0.56, 0.56, and 0.00 respectively. The research reveals that radar echo extrapolation model with NBC technique is advantageous for estimating lahar occurrences. Future studies are suggested to use wider lahar event records from other basins in the Merapi region in order to improve the data training procedure. In future investigations, high-resolution nowcasting products from the X-MP radar echo extrapolation model, as the advection model, will be introduced to the NBC model with lead times of 1-4 hours.
Acknowledgments
This study is supported by JICA-JST Science and Technology Research Partnership for Sustainable Development (SATREPS), Integrated Study on Mitigation of Multimodal Disasters Caused by Ejection of Volcanic Products in Indonesia, Kyoto University, and Gadjah Mada University.

References
[1] Lavigne F and Thouret J C 2003 Sediment transportation and deposition by rain-triggered lahars at Merapi Volcano, Central Java, Indonesia. Geomorphology 49, 45-69
[2] Liu R, Chen Y, Wu J, Gao L, Barrett D, Xu T, Li L, Huang C, and Yu J 2015 Assessing spatial likelihood of flooding hazard using Naive Bayes and GIS: a case study in Bowen Basin, Australia. Stoch. Environ. Res. Risk Assess 6, 1575-1590.
[3] Barde N C and Patole M 2013 Classification and Forecasting of Weather using ANN, k-NN and Naive Bayes Algorithms. International Journal of Science and Research 5(2), 1740-1742.
[4] Bui D T, Pradhan B, Lofman O, and Revhaug I 2012 Landslide Susceptibility Assessment in Vietnam Using Support Vector Machines, Decision Tree, and Naive Bayes Models. Mathematical Problems in Engineering 2012, 1-26.
[5] Park S G, Maki M, Iwanami K, Bringi V N, and Chandrasekar V 2005 Correction of radar reflectivity and differential reflectivity for rain attenuation at X Band Part II: Evaluation and application. J. of Atmospheric and Oceanic Tech. 22, 1633-1655.
[6] Lavigne F, Thouret J C, Voight B, Suwa H, and Sumaryono A 2000 Lahars at Merapi volcano, Central Java: an overview. Journal of Volcanology, 100(1-4), 423-456.
[7] Hapsari R I, Oishi S, Syarifuddin M, Asmara R A, Legono D 2019 X-MP Radar for developing a lahar rainfall threshold for the Merapi Volcano using a Bayesian approach. Journal of Disaster Research 14(5) 811-828.
[8] Takahashi T 2007 Debris Flow: Mechanic, Prediction and Countermeasures. Taylor and Francis: London.
[9] Niu F, Luo J, Lin Z, Liu M, Yin G 2014 Thaw-induced slope failures and susceptibility mapping in permafrost regions of the Qinghai–Tibet Engineering Corridor, China. Bull Volcanol 59, 460–480.
[10] Bayes T 1763 An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society 53, 370–418.
[11] Roebber P J 2009 Visualizing Multiple Measures of Forecast Quality. Weather and Forecasting 24(2), 601-608.
[12] Hapsari R I, Oishi S, Syarifuddin M, Asmara R A 2017 Uncertainty in debris flow hazard estimation using X-MP radar rainfall forecast. E-proceedings of the 37th IAHR World Congress, 1-10.
[13] Hapsari R I, Oishi S, Sunada K, Nakakita E, and Sano T 2011 Singular vector method on short-term rainfall prediction using radar for hydrologic ensemble prediction. Annual Journal of Hydraulic Engineering JSCE 55, S109-S114.