Abstract—Landslide hazard mapping using decision tree on Penang Island, Malaysia is proposed in this paper. Decision tree is one of popular classification algorithm of data mining. Decision tree was constructed using Quinlan’s algorithm C4.5 with 12 landslide-causative factors to produce landslide hazard maps. Hazard map that obtained using frequency ratio is improved by decision tree in term of the risk level where non-hazardous areas are reduced and others increase.

Keywords—Decision tree, landslide, hazard map.

I. INTRODUCTION

Landslide is a common disaster happening in the world. This geological phenomenon is natural hazards that often cause damages to society. Landslides occur when the stability of a slope changes from a stable state to an unstable state. Natural and human activities are the causes of landslides. In Malaysia, landslides are common especially during monsoon seasons from May to September and November to March. Damages due to landslide have been particularly high in the recent years from 2000 to 2009 (Lim et. al., 2011). Several attempts had been performed in order to reduce the damage caused by landslide by predicting the risky areas. Studies have been conducted to detect landslides and analyse the landslide hazard using the GIS and remote sensing [1], [2]. Various techniques have been implemented in the studies to obtain the landslide analysis.

In this paper, decision tree (DT) algorithm is used to produce the landslide hazard maps. DT is one of data mining approach which uses graphical model to describe decisions and its possible outcomes. DT is chosen because it does not require statistical assumptions and able to handle data that are represented on different measurement scales. Besides that, DT model could produce a result which is simple to understand and interpret due to its white box modelling nature.

II. STUDY AREA

The study area for this research is island of Penang state, Malaysia as shown in Figure 1. Penang is one of the 13 states of Malaysia located on North West of Peninsular Malaysia. The island is located within latitudes 5°15’N to 5°30’N and longitudes 100°10’E to 100°20’E. The average rainfall amount of Penang Island ranges from 2254mm to 2903mm annually. Elevation of the terrain is from 0 to 820 meters height above the sea level. The slope ranges from 0° to 87°. The vegetation that covers in Penang Island is mainly of forest and fruit plantations and the land use is consisted of forest, urban, grassland, plantation and lakes and rivers. There are fault lines that run from north to south in the centre of the island.

III. DATA SET

Topographical, geological, soil map and various data on Penang Island were obtained from Meteorological Department, Penang Geographic Information System Center, Department of Survey and Mapping Malaysia, Department of Drainage and Irrigation Malaysia, Department of Agriculture and Mineral and Geoscience Department. From topographical database, digital elevation model with resolution of 5 meters is constructed in order to extract an elevation map. The curvature, slope angle and slope aspect are obtained from the elevation map. Distance from drainage, distance from road and distance from fault lines map obtained from their own digital map respectively. Land use map in Penang Island consists of 17 types of land usage such as transport,
settlement, industry and so on whereas vegetation cover map has 14 classes. Geology database of Penang Island contains 6 types of granites and rocks and soil texture database included sand, clay and urban land. Precipitation map was obtained based on interpolation method of the average precipitation for 29 years of historical rainfall data from 1980 to 2008. Landslides mostly happened at mountainous terrain due to rock falls, debris flow and shallow rotational debris slides.

IV. METHODOLOGY

Probabilistic method is a remarkable technique for solving many of problems in discrete mathematics. This method had been implemented into landslide hazard analysis due to its high efficiency, low cost and easy implementation [3]. This method calculates each factor’s data weights based on class distribution and its landslide density [4]. These weights are used for producing a landslide hazard index. One type of the probabilistic method is frequency ratio model and it is used in this research to generate a landslide hazard map of the study area.

DT is used after extracting landslide hazard map from frequency ratio. DT learning is one of the suitable approaches in classification because it is fast and produces models with good performance. Tree growth begins from a root node which is then split by the selection and each of the attribute selection measure uses the concept of entropy, which is defined as the degree of disorder. Entropy is calculated at a node $N$ by (1) [5]:

$$\text{Entropy}(n) = - \sum_j P(C_j|N) \log_2 P(C_j|N)$$  

(1)

where $P(C_j|N)$ is the relative frequency of $N$. Of the $k$ attributes of $N$, the Entropy for selecting attribute $A$ is given by (2)

$$\text{Entropy}_A(N) = \sum_{j=1}^{k} \frac{|N_j|}{|N|} \times \text{Entropy}(N_j)$$  

(2)

InfoGain is a gain obtained from the differences between the Entropy of the original node and the Entropy of the newly split nodes. The equation is as follows:

$$\text{InfoGain}(A) = \text{Entropy}(N) - \text{Entropy}_A(N)$$  

(3)

However, InfoGain has a weakness because it causes the tree grow toward continuous attributes due to tendency to select an attribute with many split points. SplitInfo is introduced to solve this problem with (4):

$$\text{SplitInfo} = - \sum_{j=1}^{k} \frac{|N_j|}{|N|} \times \log_2 \frac{|N_j|}{|N|}$$  

(4)

With the InfoGain and SplitInfo, GainRatio can be defined as follow:

$$\text{GainRatio}(A) = \frac{\text{InfoGain}(A)}{\text{SplitInfo}(A)}$$  

(5)

Gain ratio method considers different numbers of test outcomes and because of that, it is able to avoid biases towards attributes which contains larger numbers of value [6]. The C4.5 determines the best split based on gain ratio of the attributes [7]. The splitting stops when number of instances to be split is below a certain threshold or when all the instances in a subset belong to the same class. One of the advantages of C4.5 is its ability to use both continuous and discrete data. C4.5 is a popular DT algorithm with a divide-and-conquer method [8]. For this project, 10-fold cross validation was selected to evaluate the robustness of the DT model along with optimizing the DT size [8], [9].

Selection of the attributes in DT model is based on the value of the entropy and information gain. Entropy is used to measure the homogeneity or similarity of a learning set. The lower the entropy, the easier for it to be predicted and higher the information gain obtained.

Pruning is used to convert large tree into a smaller tree to avoid overfitting. Overfitting normally occurs when learning algorithm continues to develop hypotheses that reduce training set error at the cost of an increased test set error. Besides, the pruning method has possibility of increasing the prediction’s accuracy for it makes better interpretation of the DT. There are two types of pruning which are pre-pruning and post-pruning. Pre-pruning methods stop the refinement of the rules before they become too specific or overfit the data while post-pruning methods try to simplify it after finding a complete and consistent rule or rule set [10]. Post-pruning is used in C4.5 to construct trees in all training folds. The average error estimate is calculated for each number of expansions based on the temporary trees in all folds and repeated until the tree cannot be expanded. By using cross-validation, the sequence of number of expansions and their corresponding error estimates are calculated. Final number of expansion is chosen based on the number of expansions whose average error estimate is minimal.

V. LANDSLIDE HAZARD MAPPING AND ANALYSIS

Based on decision model, 1303 terminal nodes have been obtained with 192 pruning level. Experiments with pruning 60, 80, 100, 120, 140, 160, 180 are carried out to assess the performance of the landslide hazard map using DT. Comparison of percentage of total landslide between frequency ratio and each pruning level of DT are depicted in Figure 2. The result shows tremendous change in percentage of total landslide in the category of not hazardous area. Percentage of landslide at not hazardous area using frequency ratio is 18.21% while the DT shows lower percentage at this category of not hazardous area (8%- 5.83%). While the percentage of landslide at not hazardous area decreases, the percentage of landslide in other risk levels increase. Moderate area using DT has higher percentage compared with frequency ratio. Meanwhile, hazardous area shows lower percentage than frequency ratio at certain pruning level of DTs. For DT at pruning level 180, only 7.65% of the landslide falls into hazardous area. It is because the cost (average error) for pruning 180 increases due to less complexity and terminal nodes to determine the landslide hazard level risks. Hence, if pruning to maximum which is 192, the result become
irrelevant for the cost of pruning is too high.

The pruning level 140 to pruning level 100 shows the increment of percentage in hazardous level compared to frequency ratio. It is because the DT considers more attributes that increases hazardous area on the landslide occurrence. For most hazardous, DT has higher percentage of landslide than frequency ratio. Landslide hazard map using DT for pruning level 120 is shown in Figure 3.

VI. SUMMARIES

Landslide is a natural disaster that could not be avoided. This geological phenomenon occurred when a stability slope changes from a stable state to an unstable state caused by number of factors. Many studies in landslide have been conducted to detect landslides and analysed the landslide hazard using Geographic Information Systems and remote sensing with various methods.

In this paper, 12 landslide-causative factors have been used. The main objective of this research is to implement DT in landslide hazard map in order to improve the prediction of landslide hazard risk levels. Frequency ratio is used to extract the correlation between landslides occurrence with landslide causative factors. After that, DT model is calculated and constructed using DT algorithm. Finally, pruning is implemented to observe the performance of each pruning level with the percentage of past landslide event. The use of DT method improves the landslide hazard map where percentage of past landslide event increases in three risk level, i.e most hazardous, hazardous and moderate while it reduces in non hazardous level.

ACKNOWLEDGMENT

The authors thank Universiti Sains Malaysia for the financial support through research short term grant.

REFERENCES

Study of Improving Classification Accuracy between Groups of Depressive and Remitted Persons based on Filtered Speech Signal using FIR

Patiyuth Pramkeaw and Thaweesak Yingthawornsuk

Abstract—This study is to present the results of pairwise classification based on using the Mel-Scale Frequency Cepstral Coefficients (MFCC) extracted from speech samples between two subject groups which are depressive (DPR) and remitted (RMT) speakers. Only voiced segments detected from the pre-processed speech signal with FIR Low-Pass Filter at different cutoff frequencies of 0.75 KHz and 1 KHz was used before feature extraction and classification. ML classifier can classify 70% correctly, when first filtering all speech samples with FIR-LPF at cutoff-frequency of 1KHz, compared to accuracy of 53% in case without pre-filtering to speech. Further NN classifier also agrees on the higher correct score of classification with 69% in accuracy when speech signals were filtered first before classifying, as compared to 60% the lower accuracy obtained from without-filtering case. The improvement of classification performance can be effectively increased by signal filtering technique.

Keywords-- Speech Signal, FIR Filter, PCA, ML, NN.

I. INTRODUCTION

Speech classification is the process of automatic identification and separation between the speech samples. Two different speech samples comprise of the depressed and remitted speech samples used in this proposed study. Based on information in speech signal, classification which is a part in the recognition technique makes it possible to the speaker’s voice to be used in verifying their identity and control access to services such as voice dialing, banking by telephone, telephone shopping, database access services, information service, voice mail, security control for the confidential information areas, and remote access to computers. The acoustical parameters of speech signal used in recognition tasks have been popularly studied and investigated, and being able to be categorized into two types of processing domain: First group is spectral-based parameters and another is dynamic time-series. The most popular spectral-based parameter used in recognition approach is the Mel Frequency Cepstral Coefficients called MFCC [1,5,8].

Due to its advantage of less complexity in implementation of feature extraction algorithm, only sixteen coefficients of MFCC corresponding to the mel-scale frequencies of the speech cepstrum are extracted. Signal filter is most required for many tasks, which can be considered into analog and digital filters. Most analog filters are designed and all extracted MFCC features are then statistically analyzed for principal components, at least two dimensions minimally required in further recognition performance evaluation. Acoustical parameters extracted by several speech processing techniques such as statistical properties of fundamental frequency, speech jitters and shimmers, formants, pitch contour, Glottal Spectral Tilt, and frequency distribution of PSD energies [3,5,6,7] have suggested to be the effective indicators for monitoring the symptom of major depression in speakers[4] through their sound outcomes. This study focused on the effectively discriminative parameter the Mel-Scale Frequency Cepstral Coefficients (MFCC). The extracted features are used as input to ML and NN classifiers in attempt to identify the suicidal patients among depressed patients and remitted subjects.

The following sections in this paper describe methods of signal filtering, voiced/unvoiced detection, acoustical feature extraction, principal component analysis, feature classification. Experimental results, discussion and conclusion are provided.

II. METHODOLOGY

2.1 Database Information

Database consists of female two groups of speech samples recorded in an environmentally controlled recording room to have all possibly less acoustical interferes to the quality of sound sample during the recording time. The first group comprises of two depressed “DPR” females and another is a group of two remitted “RMT” females. All sound signals are recorded under most similar condition such as the same length of recording time, and the level of sound amplitude. All subjects admitted in the program are between the youngest are of 23 years old and the oldest one of 65 years old. Individual subject is advised to complete the Beck Depression Inventory-II (BDI-II) before the participating in interviewing session with a psychiatrist. The BDI-II inventory used in this program is a mood weighting measure in a form of standard, brief and self score question are having detail of clinical