



## Research Article

# A Comparison of Support Vector Machine and Decision Tree Classifications Using Satellite Data of Langkawi Island

[H.Z.M. Shafri](#) and [F.S.H. Ramle](#)

## ABSTRACT

This study investigates a new approach in image classification. Two classifiers were used to classify SPOT 5 satellite image; Decision Tree (DT) and Support Vector Machine (SVM). The Decision Tree rules were developed manually based on Normalized Difference Vegetation Index (NDVI) and Brightness Value (BV) variables. The classification using SVM method was implemented automatically by using four kernel types; linear, polynomial, radial basis function and sigmoid. The study indicates that the classification accuracy of SVM algorithm was better than DT algorithm. The overall accuracy of the SVM using four kernel types was above 73% and the overall accuracy of the DT method was 69%.

## Services

[Related Articles in ASCI](#)

[Similar Articles in this Journal](#)

[Search in Google Scholar](#)

[View Citation](#)

[Report Citation](#)

## How to cite this article:

H.Z.M. Shafri and F.S.H. Ramle, 2009. A Comparison of Support Vector Machine and Decision Tree Classifications Using Satellite Data of Langkawi Island. *Information Technology Journal*, 8: 64-70.

**DOI:** [10.3923/itj.2009.64.70](https://doi.org/10.3923/itj.2009.64.70)

**URL:** <http://scialert.net/abstract/?doi=itj.2009.64.70>

## INTRODUCTION


Langkawi is an archipelago of 99 islands and its main island is Langkawi Island. It is located in the north-western coast of Peninsular Malaysia within the State of Kedah. The Langkawi archipelago has unspoiled rainforests, limestone and karst formations, caves and rich marine life. Langkawi definitely is geological heritage of high value based on its great geological landscape and other features such as the fossils, sedimentary structures and erosional effects. The Langkawi islands are primarily protected under the jurisdiction of the Permanent Forest Reserves, Recreational Forest or Geoforest Park that are supervised by the Forestry Department.

Statistical methods such as the minimum distance and maximum likelihood classifiers have been widely used for the classification of remotely-sensed image data. These methods have their limitations, particularly in relation to distributional assumptions and to the restrictions on data input. Numerous studies have demonstrated that **artificial intelligence** such as expert system, artificial **neural networks** and support vector machines can be alternative methodologies for classification problems to which traditional statistical approaches have long been applied ([Tseng et al., 2008](#)).

Many studies have shown that techniques such as evidential reasoning, neural networks, decision trees and Support Vector Machines (SVM) may often be able to classify a data set to a higher accuracy than conventional statistical classifiers ([Foody and Mathur, 2004](#); [Huang, 2002](#)). With these techniques, which are not based on an assumed parametric model, the standard requirement for a full and representative description of the spectral response of each class may no longer be necessary or appropriate in training the classification analysis ([Foody and Mathur, 2004](#)).

SVM is based on statistical learning theory, which has its roots in the 1960s. The learning theory developed in order to solve **pattern recognition** problems. The SVM, originally introduced by Vapnik is a classification and regression technique which is now widely used in very different fields, including in remote sensing ([Wijaya and Gloaguen, 2007](#)).

SVM separates the classes with a decision surface that maximizes the margin between the classes. The surface is often called the optimal hyperplane and the data points closest to the hyperplane are called support vectors. The support vectors are the critical elements of the training set. The ENVI SVM classifier provides four types of kernels: linear, polynomial, Radial Basis Function (RBF) and sigmoid. The default is the radial basis function kernel, which works well in most cases.

 having its origin in machine learning theory, is an efficient tool for the solution of classification and regression problems. Unlike other classification approaches that use a set of features (or bands) jointly to perform classification in a single decision step, the decision tree is based on a multistage or hierarchical decision scheme or a tree like structure. The tree is composed of a root node (containing all data), a set of internal nodes (splits) and a set of terminal nodes (leaves). Each node of the decision tree structure makes a binary decision that separates either one class or some of the classes from the remaining classes. The processing is generally carried out by moving down the tree until the leaf node is reached. This is known as a top-down approach tree ([Xu et al., 2005](#)).

ENVI can calculate special variables, such as NDVI, on-the-fly and use them in the expressions. The expressions used in ENVI's decision tree classifier are similar to those used in Band Math. They must produce single band output and have a binary result of 0 or 1. The 0 result is sent to the No branch and the 1 result is sent to the Yes branch of the decision tree. The expressions can include math functions, relational operators, boolean operators and other IDL functions.

There are two objectives in this project; to classify plant types using **artificial intelligence** method which consists of Support Vector Machine (SVM) and Decision Tree (DT) classifiers and to compare the classification accuracy between the SVM and DT classifier.

## MATERIALS AND METHODS

**Study area:** The study area is located between latitude 6° 29' 33.20" to 6°23' 6.24" and between longitude 99° 48' 0.34" to 99° 55' 30.86" at the northeast of Langkawi Island within the State of Kedah, Malaysia ([Fig. 1](#)). The area is mainly covered by forest, mangroves, agricultural land and sand beaches.

The topography varies from flat coastal plains, hilly areas to rugged mountains. The area comprises the three river basins of Kilim, Air Hangat, Kisap and the neighbouring island of Langgun and Tanjung Dandang. This study was conducted between July 2006 and July 2007.

**Data:** Several data used consist of the Systeme Probatoire pour l'Observation de la Terre (SPOT) 5 satellite imagery, topography map with 1:50 000 scale and ground truth data. The SPOT 5 image of Langkawi Island (Path 264 Row 337) acquired on March 7, 2005 was used for image classification. The image were already geometrically corrected and registered to WGS 84 datum and UTM Zone 47 projection. The digital processing of the SPOT 5 imagery was carried out using the ENVI 4.3 software.

In this study, two methods were used to classify the SPOT 5 image. The classifiers are Decision Tree (DT) ([Waheed et al., 2006](#)) and Support Vector Machine (SVM) ([Pal and Mather, 2004](#)).

**Decision tree classifier:** NDVI is an index calculated from reflectance measured in the red visible and near infrared channels. The chlorophyll (green pigment) absorbs incoming radiation in the visible band, while the leaf structure and water content is responsible for a very high reflectance in the near-infrared region of the spectrum. Firstly, NDVI map was produced. From the NDVI map, NDVI range of each land cover types was generated ([Fig. 2](#)). The equation of NDVI is shown as below:

$$NDVI = (NIR - Red) / (NIR + Red) \quad (1)$$

The purpose of image classification is to group together pixels that have similar patterns of brightness values across a series of image bands or information channels. Therefore, brightness value (BV) for each land cover type was obtained ([Table 1](#)).

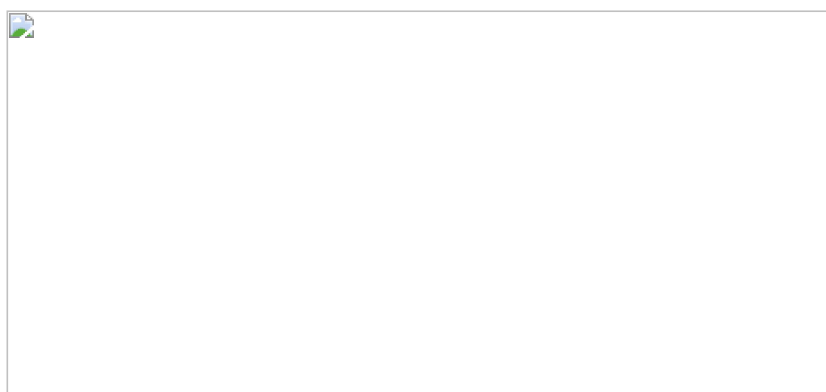


Fig. 1: Study area at the northeast of Langkawi Island within the state of Kedah, Malaysia

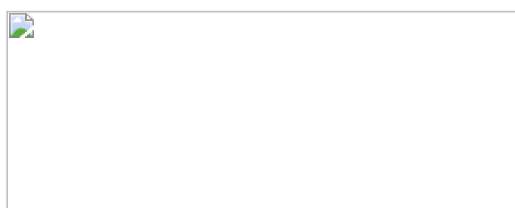
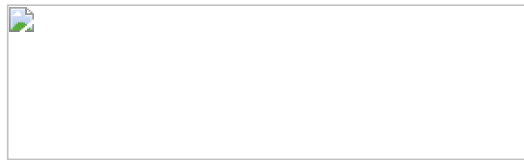


Fig. 2: Range of NDVI values for each land cover

Table 1: Range of brightness value for each land cover



The DT classification rules were developed manually based on the Brightness Value (BV) and Normalized Difference Vegetation Index (NDVI) of each land use type. The decision tree classifier performs multistage classifications by using a series of binary decisions to place pixels into classes ([Fig. 3](#)).

**Support vector machine classifier:** Before the classification, training sites representative of land-cover classes of interest was acquired using ROI tool. The areas selected to serve as training sites should be relatively homogeneous and extensive enough to provide good statistics.

The SPOT 5 image was classified automatically using four kernel types. The kernel types are linear, polynomial, radial basis function and sigmoid. Firstly, the classification was executed by using the default parameters. The parameters are penalty parameter, pyramid levels and classification probability threshold. The default value of penalty parameter is 100, pyramid levels is 0 and classification probability threshold is 0.

Secondly, the classification was performed by using modified parameters value. The new values of penalty parameter were 200, 300 and 400. The new values of pyramid levels were 1, 2 and 3. The new values of Classification Probability Threshold were 0.3, 0.7 and 1.

Lastly, the classification was performed by using the optimum parameters value to find the most accurate classification image. The selection of the optimum value of parameter was based on the accuracy assessment. The final classification was performed by using the following parameters value as in [Table 2](#).

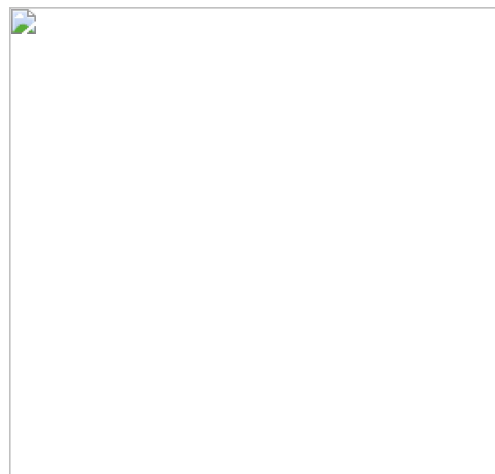


Fig. 3: Decision tree classifier

Table 2: The parameters value for the final classification of SVM

## RESULTS

**Decision tree:** There are six classes which consist of water, mangrove, limestone forest, dipterocarp forest, rubber and non-vegetated. The dipterocarp forest class was not well classified. Some of the dipterocarp forest area was misclassified as rubber and mangrove. This occurred because the NDVI ranges for dipterocarp forest, mangrove and rubber are quite similar ([Fig. 4](#)).

**Support vector machine:** After the changes of penalty parameter value, the classified images were quite similar to the default classified image, except that the classified image of sigmoid kernel is better.

After the pyramid levels value was changed, the new classified images were better than the default classified images for all the kernel types (linear, polynomial, radial basis function and sigmoid kernel).

However, after the probability threshold value was modified, the new classified images did not show any changes.

[Figure 5](#) shows the difference between the final classified images and the default parameter value of classified images. The default parameter value of classified images shows that the polynomial and radial basis function kernels can classify the non-vegetated class (sand). However, linear and sigmoid kernels have misclassified the non-vegetated. After using the optimal parameters values, the results show that the non-vegetated class (sand) was well classified.

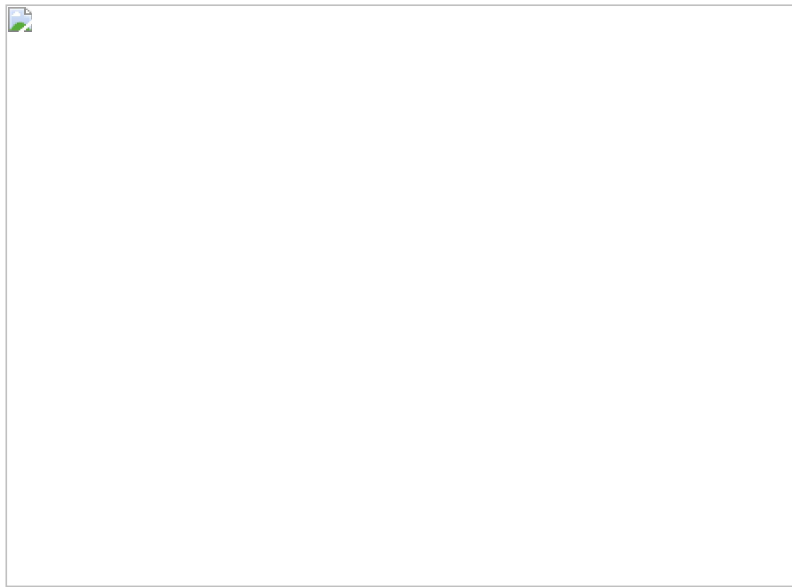


Fig. 4: Classified image using decision tree

Fig. 5: Comparison of SVM classifications using, (a) Default and (b) Optimum parameter values



Fig. 6: Classified images using SVM linear (left) and SVM polynomial (right)

Fig. 7: Classified images using SVM RBF (left) and SVM sigmoid (right)

Table 3: Accuracy assessment

The final classifications (optimum parameter value) of each kernel types are shown in [Fig. 6](#) and [7](#). The final classified images (optimum parameter value) give better classification result than the default classified images. There are six classes which consist of water, mangrove, limestone forest, dipterocarp forest, rubber and non-vegetated. The parameters used in the classification using SVM are based on the values in [Table 2](#).

**Accuracy assessment:** [Table 3](#) shows that the accuracy of optimum parameters values is better than the default parameter values. The result shows that the SVM Radial Basis function gives the highest overall accuracy which is 76.0004%. The lowest overall accuracy is Decision Tree (DT) with 68.7846%. This means that image classification using Support Vector Machine (SVM) method is better than Decision Tree (DT) in this study.

## CONCLUSION

The result shows that the SVM algorithm gives better classification image than DT algorithm. The overall accuracy of the SVM method using four kernel types are above 73% and overall accuracy of the DT method is 69%.

The study concludes that the use of the optimal parameter values for Penalty Parameter and Pyramid Levels would give better classification result. However, modifying the Probability Threshold values is not needed because the result is similar to the default value. The Radial Basis Function gives the highest overall accuracy which is 76.0004% and the sigmoid is the weakest kernel of SVM.

The overall accuracy of the DT is low because the main variable used is NDVI, however NDVI range for dipterocarp forest, mangrove and rubber class are quite similar. Therefore, the brightness value variable was also used to separate the land cover. The study also shows that NDVI and brightness value variables are suitable to be used in DT classifier.

## RECOMMENDATION

There are a few recommendations that are suggested to improve the classification results of satellite image. Firstly, to improve the classification result using Decision Tree classifier, the elevation variable can be used. In this study, mangrove, rubber and forest dipterocarp have been misclassified because they have approximately similar NDVI. We could improve the separation of the classes by using elevation variable as the features exist at different altitudes. Secondly, SVM also includes other parameters such as degree of kernel polynomial, bias in kernel function, gamma in kernel function and pyramid classification threshold that can be assessed in future studies. The parameters could possibly give different image classification results.

## ACKNOWLEDGMENTS

The authors would like to thank the Malaysian Centre for Remote Sensing (MACRES) for providing the satellite imagery and Dr. Nurul Salmi Abdul Latip and her team from the School of Biological Science, Universiti Sains Malaysia (USM) for providing the ground truth data. This research is partially funded by the Remote Sensing and GIS programme at the Faculty of Engineering, UPM (Funding No. 66604).



Foody, G.M. and A. Mathur, 2004. Toward intelligent training of supervised image classifications: Directing training data acquisition for SVM classification. *Remote Sens. Environ.*, 93: 107-117.

[CrossRef](#) |

Huang, K.Y., 2002. The use of a newly developed algorithm of divisive hierarchical clustering for remote sensing image analysis. *Int. J. Remote Sens.*, 23: 3149-3168.

[CrossRef](#) |

Pal, M. and P. Mather, 2004. Assessment of the effectiveness of support vector machines for hyperspectral data. *Future Generat. Comput. Syst.*, 20: 1215-1225.

[CrossRef](#) | [Direct Link](#) |

Tseng, M.H., S.J. Chen, G.H. Hwang and M.Y. Shen, 2008. A genetic algorithm rule-based approach for land-cover classification. *ISPRS J. Photogram. Remote Sens.*, 63: 202-212.

[CrossRef](#) |

Waheed, T., R.B. Bonnell, S.O. Prasher and E. Paulet, 2006. Measuring performance in precision agriculture: CART-A decision tree approach. *Agric. Water Manage.*, 84: 173-185.

[CrossRef](#) |

Wijaya, A. and R. Gloaguen, 2007. Comparison of multisource data support vector machine classification for mapping of forest cover. *Proceedings of the International Geoscience and Remote Sensing Symposium*, July 23-28, 2007, IEEE Xplore, London, pp: 1275-1278.

Xu, M., P. Watanachaturaporn, P.K. Varshney and M.K Arora, 2005. Decision tree regression for soft classification of remote sensing data. *Remote Sens. Environ.*, 97: 322-336.

[CrossRef](#) |