



**PREDICTING SUITABLE AREAS FOR GROWING CASSAVA  
USING REMOTE SENSING AND  
MACHINE LEARNING TECHNIQUES:  
A STUDY IN NAKHON-PHANOM THAILAND**

J. Kimani Mbugua\*      Claremont Graduate University,      [Joseph.mbugua@cgu.edu](mailto:Joseph.mbugua@cgu.edu)  
Claremont, CA, USA  
Watanyoo Suksa-ngiam      Claremont Graduate University,      [Watanyoo.suksa-ngiam@cgu.edu](mailto:Watanyoo.suksa-ngiam@cgu.edu)  
Claremont, CA, USA

\* Corresponding author

**ABSTRACT**

Aim/Purpose	Although cassava is one of the crops that can be grown during the dry season in Northeastern Thailand, most farmers in the region do not know whether the crop can grow in their specific areas because the available agriculture planning guideline provides only a generic list of dry-season crops that can be grown in the whole region. The purpose of this research is to develop a predictive model that can be used to predict suitable areas for growing cassava in Northeastern Thailand during the dry season.
Background	This paper develops a decision support system that can be used by farmers to assist them determine if cassava can be successfully grown in their specific areas.
Methodology	This study uses satellite imagery and data on land characteristics to develop a machine learning model for predicting suitable areas for growing cassava in Thailand's Nakhon-Phanom province.
Contribution	This research contributes to the body of knowledge by developing a novel model for predicting suitable areas for growing cassava.
Findings	This study identified elevation and Ferric Acrisols (Af) soil as the two most important features for predicting the best-suited areas for growing cassava in Nakhon-Phanom province, Thailand. The two-class boosted decision tree algorithm performs best when compared with other algorithms. The model achieved an accuracy of .886, and .746 F1-score.

Accepting Editor: Eli Cohen | Received: December 21, 2017 | Revised: March 8, 2018 |  
Accepted: March 26, 2018.

Cite as: Mbugua, J. K., & Suksa-ngiam, W. (2018). Predicting suitable areas for growing cassava using remote sensing and machine learning techniques: A study in Nakhon-Phanom Thailand. *Issues in Informing Science and Information Technology*, 15, 43-56. <https://doi.org/10.28945/4024>

(CC BY-NC 4.0) This article is licensed to you under a [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/). When you copy and redistribute this paper in full or in part, you need to provide proper attribution to it to ensure that others can later locate this work (and to ensure that others do not accuse you of plagiarism). You may (and we encourage you to) adapt, remix, transform, and build upon the material for any non-commercial purposes. This license does not permit you to use this material for commercial purposes.

Recommendations for Practitioners	Farmers and agricultural extension agents will use the decision support system developed in this study to identify specific areas that are suitable for growing cassava in Nakhon-Phanom province, Thailand
Recommendation for Researchers	To improve the predictive accuracy of the model developed in this study, more land and crop characteristics data should be incorporated during model development. The ground truth data for areas growing cassava should also be collected for a longer period to provide a more accurate sample of the areas that are suitable for cassava growing.
Impact on Society	The use of machine learning for the development of new farming systems will enable farmers to produce more food throughout the year to feed the world's growing population.
Future Research	Further studies should be carried out to map other suitable areas for growing dry-season crops and to develop decision support systems for those crops.
Keywords	geospatial, remote sensing, machine learning, suitability, cassava prediction

## INTRODUCTION

---

World food crises resulting from shortages, cost, or inaccessibility of enough quality amounts of food are expected to get more severe in the future. The World Economic Forum (2015), for example, projects that by 2050, we will need to produce 50% more food than we do currently to meet this global demand from an increased world population of more than 9 billion people (The World Bank, 2015). To ameliorate the food crises, there is a need for farmers, public, and private organizations to collaboratively formulate and implement strategies that will improve food security in the world.

Thailand is one of the largest food producers in the world. The country is particularly well known for rice production, a crop that is very critical for Thailand's economy and its people. Rice is grown mainly in the Northeastern region of Thailand, where it is grown in about 69 % of the region's land area (National Statistical Office, 2003).

Although rice is the major crop grown by farmers in Northeastern Thailand, the region has the lowest rice production yield compared to all other regions of Thailand (Titapiwatanakun, 2012). The low yield in this region is mainly caused by the fact that most farmers can grow rice only during the wet season (Chainuvati & Athipanan, 2001) because the region lacks an adequate irrigation system (Thanawong, Perret, & Basset-Mens, 2014). For this reason, the farmers are forced to leave their farms fallow and move to cities to search for jobs in construction sites during the dry season (Craig & Pisone, 1985).

Thailand's farmers normally do not know the specific crops that they can grow in their areas during the dry season. The current agriculture planning guideline (Chainuvati & Athipanan, 2001) provides only a list of all the crops that can be grown in the region. Given that different crops require different physical and climatic conditions to grow, the guideline is not very helpful because it does not indicate suitable locations where each crop in the list can be grown successfully. There is, thus, a need for the development of a decision support system that helps farmers determine what crops to grow in their specific areas during the dry season.

The decision support system thus developed, could, for example, be used to help the farmers identify the dry season crops that are suitable for their specific areas and to also provide information on how to grow and manage those crops. One of the crops that can be grown during the dry season is cassava (Chainuvati & Athipanan, 2001). Suitable areas for growing cassava and other dry-season crops can be identified using a suitability analysis predictive model. A decision support system for dry-season farming can subsequently be built based on the model.

The purpose of this research is to develop a predictive model that can be used to predict suitable areas for growing cassava in the Northeastern region of Thailand. The results can then be used to guide farmers and policy makers identify suitable areas for growing cassava in Thailand and other countries with similar environmental conditions. This research seeks to answer the following questions:

- What variables can be used to predict suitable areas for growing cassava in Nakhon-Phanom province of Northeastern Thailand?
- How can the variables be used to build a model for predicting suitable areas for growing cassava?

## LITERATURE REVIEW

---

It is possible to grow cassava in higher elevation areas of Northeastern Thailand following the end of the rice cultivation season (Craig, 1985). Crop health indicators can potentially be used to determine a specific crop's health conditions during the growth cycle. Remote sensing researchers use Normalized Difference Vegetation Index (NDVI), Vegetation Index (VI), and Enhanced Vegetation Index (EVI) to investigate different crop health conditions (Dusseux, Vertès, Corpetti, Corgne, & Hubert-moy, 2014; Prasad, Chai, Singh, & Kafatos, 2006; Tang, Li, Chen, Zhu, & Liu, 2012). For example, NDVI, which represents the greenness and health of crops (Prasad et al., 2006), can be used as a variable to determine crop yield.

Climate conditions influence the production of crops. Several climatic indicators that we might use to determine where cassava can grow include temperature, drought, precipitation (rain), and land surface water. In using temperature as a possible predictor of suitable areas for a given crop, we can rely on the established temperature requirements for that crop. It is known, for example, that different crops do well in different ranges of temperature. In the case of cassava, the crop performs best in areas where soil temperatures range from 25°C to 29°C (Food and Agriculture Organization, 2000) with an annual average of about 20°C (Ratanawaraha, Senanarong, & Suriypan, 2000). Further, researchers also use land surface temperature (LST) to represent suitable temperature for each crop (Suksa-ngiam, Mbugua, & Chatterjee, 2016). LST is calculated using various bands from the Landsat 8 satellite images. Precipitation or rainfall is mostly required for predicting crops like rice but is not as crucial for cassava. Precipitation estimates can be made by combining information from Microwave (MW) and Infrared (IR) satellite sensors (AghaKouchak et al., 2015). Water is also a crucial resource for growing crops. Various indicators are used to represent the availability of water for crop production. One of the water indicators that can be used to predict suitable areas for growing cassava is surface water. In remote sensing research, Normalized Difference Water Index (NDWI) represents surface water. Although cassava requires moderately- to excessively-drained soils, it can also be grown in poorly-drained soils (Caldwell, Sukchan, & Ogura, 2005). Soil moisture is another useful predictor because it affects the growth of plants and their productivity (AghaKouchak et al., 2015). Cassava performs well in areas where soils have a high soil moisture content (Food and Agriculture Organization, 2013). Scaled Draft Condition Index (SDCI) is also used as a climate condition indicator. SDCI is a composite predictor that consists of NDVI, precipitation, and, LST (AghaKouchak et al., 2015).

Soil types and their conditions are also important factors for growing crops. Different crops require different types of soil and soil conditions. For optimal cassava production, the areas should have loam to loamy-sand texture soils that are well drained, but the crop can also do well in areas with poor soils and a subsurface hardpan (Caldwell et al., 2005). The soils could also be deeper, moist, or dry, and light-textured (Food and Agriculture Organization, 2013). Because farmers can grow cassava in relatively poor-quality soils, some people tend to view the crop as a contributor to the depletion of soil quality (Idhipong et al., 2012). Soil types are represented in this study using ESRI's soil maps taken from the FAO-UNESCO soil map of Thailand (Thipruck, 2013).

The predictive model developed in this study is a machine learning (ML) model, that relies on the remotely-sensed features. ML has become crucial in the remote sensing (RS) domain since ML models tend to be highly accuracy. Many studies have applied ML to predict crop growing areas and crop yields.

Kussul, Lavreniuk, Skakun, and Shelestov (2017) used deep learning (DL) and remote sensing (RS) for crop classifications. They used multi-layer perceptron (MLP) and random forest and compared them with convolutional NNs (CNNs). They found that an ensemble of CNNs performed better than MLP in maize and soybean classification. CNNs provided more than 85% accuracy for wheat, maize, sunflower, soybeans, and sugar beet. They sourced the images for their study from Landsat-8 and Sentinel-1A satellites. However, due to their use of DL, they could not derive feature-importance information from their study. Thus, information relating to the best conditions for growing the crops can't be fully derived from their study.

You, Li, Low, and Lobell (2017) developed a dimensionality reduction method for a CNN and Long-short Term Memory network (LSTM). They introduced a Gaussian Process (GP) component to model data and enhance accuracy in DL models. They obtained satellite data from MODIS and tested their model to predict soybean yield. When compared by Mean Absolute Percentage Error, their model outperformed the USDA yield estimates. Unlike Kussul et al 2017, their model can reveal the feature importance of the predictors.

Sun, Leinenkugel, Guo, Huang, and Kuenzer (2017) used a C5.0-based decision-tree algorithm to classify rubber plantations in China. They used multi-class classification to separate rubber plantations from other areas, such as urban and water areas. They obtained satellite data from Landsat. They improved the algorithm by adding atmospheric and topographic correction (haze and shadow removal) features. The overall accuracy of classification improved from 84.2% in 1989, 83.9% in 2000, to 86.5% in 2013. The authors concluded that bands (as features) 1-3 could not discriminate rubber plantations from other areas, while bands 4 to 7 were ideal for rubber plantation classifications.

Whereas our study employs knowledge of the required environmental conditions for growing cassava, Heumann, Walsh, and McDaniel (2011) in their study assessed the applicability of Maximum Entropy (MaxEnt), a machine learning algorithm, based on a presence-only geographic species distribution, to estimate habitat suitability for various crops in rural Northeastern Thailand. In the study, the authors used three independently obtained crop presence datasets (obtained from a household survey, remote sensing, and the field) combined with environmental data (soil type, elevation, and solar radiation). The study found that land suitability for cassava and paddy rice can be successfully estimated using the presence-only modelling.

Garzón et al. (2006) used machine learning models to develop a framework for predicting suitable areas for *Pinus sylvestris* forests in the Iberian Peninsula. The authors used environment predictor variables to compare the accuracy of three predictive ML models. The Breiman's random forest model was found to be the most accurate for this study. The framework they developed can be applied in other studies to establish potential forest areas by generating predictive maps for any other forest species.

Our study differs from previous studies in that it adds new interpretable features as predictors. For example, instead of using each Landsat's bands directly, we transformed and combined them to become interpretable maps such as temperature and water. Although we do not use DL algorithms, our machine learning algorithms can yield feature importance, which enables researchers and policy makers to interpret and understand how these features (maps) can contribute to the success of classifications.

## METHODOLOGY

---

We selected Nakhon-Phanom province, one of the provinces in Northeastern Thailand where cassava is grown, as the site for this study. The area of Nakhon-Phanom province is about 2,128 square miles. The whole province is covered by a single Landsat scene, thus images taken at different time periods do not need to be mosaiced.

### *PREDICTOR DEVELOPMENT*

In this research, we employed a data science approach to predictive analysis. Three archived cloud-free Landsat 8 satellite images covering the study area were downloaded from the U. S. Geological Survey, EarthExplorer, web-portal (<http://earthexplorer.usgs.gov/>). Landsat 8 imagery data consists of 11 bands. Most of the Landsat imagery data have a resolution of 30 meters (panchromatic has resolution of 15 meters) (United States Geological Survey, 2016). We used Arcpy, Dnnpy, Scipy, and Numpy Python libraries for this study. Arcpy was used for geoprocessing, while Dnnpy was used for top-of-atmosphere correction.

Using the Landsat 8 imagery data, the following predictors were then calculated: Normalized Difference Vegetation Index (NDVI) was used to measure crop health and the density of plants. It is calculated using equation (1) as shown in the appendix (National Aeronautics and Space Administration, 2000). Normalized Difference Water Index (NDWI) was used to assess the water content. NDWI is calculated as shown in equation (2) in the appendix (Gao, 1996). Enhanced Vegetation Index (EVI) is an alternative to NDVI. EVI is sensitive to high bio-mass areas and reduces the effects of soil and atmosphere (Jiang, Huete, Kamel, & Miura, 2008). The calculation of EVI is shown in equation (3) in appendix (Wardlow, Egbert, & Kastens, 2007). Land Surface Temperature (LST) is a measurement of the temperature on the ground surface and is calculated using equations (1) and (4) to (9) (Artis & Carnahan, 1982; Buhari, 2015; Sobrino, Jiménez-Muñoz, & Paolini, 2004; United States Geological Survey, 2015; Weng, Lu, & Schubring, 2004; Yu, Guo, & Wu, 2014). The Normalized Difference Tillage Index (NDTI) is the normalization of the difference between short wave infra-red wavelength band 6 (SWIR1) and short wave infra-red wavelength band 7 (SWIR2) of Landsat 8 data (Li, Ti, Zhao, & Yan, 2016). NDTI is used to detect crop residues and tillage practices (van Deventer, Ward, Growda, & Lyon, 1997). The Normalized Burn Ratio (NBR) is the ratio between NIR (band 5) and SWIR2 (Band 7) (United States Geological Survey, 2017). NBR is used to detect burned areas (Cocke, Fulé, & Crouse, 2005).

Besides Landsat's data, we also used elevation data (in meters) downloaded from the Diva-gis portal. Additionally, we used soil data from Food and Agriculture Organization (FAO). Based on the FAO soil types classification, there are only four soil types (Af, Ao, Ag, and Gd) in Nakhon Phanom province.

### *CASSAVA GROUND TRUTH*

The land use data for Nakhon Phanom province were obtained from the Land Development Department under the Ministry of Agriculture and Cooperatives, Thailand. We used this data to train and evaluate the ML models. Table 1 shows all initial predictors that were fed into Microsoft's Azure Machine Learning.

200,000 random points were generated in ArcMap and the data from all the predictors were then extracted to these points. The ArcMap's point shapefile was then transformed and saved as a csv file. We used the csv file in Microsoft Azure Machine Learning.

**Table 1. Predictors**

SOURCE	ABBR	TYPE	DEFINITION
Landsat 8 Imagery	NDVI	Numeric	Normalized Difference Vegetation Index
	NDWI	Numeric	Normalized Difference Water Index
	EVI	Numeric	Enhanced Vegetation Index
	LST	Numeric	Land Surface Temperature
	NDTI	Numeric	Normalized Difference Tillage Index
	NBR	Numeric	Normalized Burn Ratio
Diva GIS	Elevation	Numeric	Elevation
Food and Agriculture Organization	Af	Discrete	Ferric Acrisols
	Ao	Discrete	Orthic Acrisols
	Ag	Discrete	Gleyic Acrisols
	Gd	Discrete	Dystric Gleysols
Land Development Department, Thailand	Cassava	Discrete	Cassava field area

### ***MACHINE LEARNING PROCESS***

The data were first cleaned by removing rows that had any missing values. We ended up with 197,989 data points (4,389 cassava cases and 193,600 non-cassava cases). We then normalized all the predictor’s data into the same scale (0 to 1) using min-max normalization as shown in equation (12) in the appendix (Microsoft Corporation, 2017). Due to class imbalance in the dataset, we partitioned the majority cases (non-cassava area) at 20 % of the samples. At this rate, we still get the margin of error less than 1% and the confidence level more than 99 % (SurveyMonkey, 2017). In addition, we applied the SMOTE algorithm to cater for the imbalance in the dataset (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). The cassava data points were increased by 200 %. After these adjustments, we ended up with 51,887 data points in total. Next, the predictors were examined for multi-collinearity. Any two predictors that shared a correlation of more than 0.9 (Pearson Correlation) were combined to avoid multi-collinearity (Hair, Black, Babin, & Anderson, 2010). EVI was highly and positively correlated with NDVI, while NDWI was highly and positively correlated with NBR. Using principle component analysis (PCA) separately, we merged EVI and NDVI to become NDVI\_EVI, and NDWI was merged with NBR to become NDWI\_NBR. The final predictors that remained were LST, NDTI, Af, Ao, Ag, Gd, NDVI\_EVI, NDWI\_NBR, and, Elevation. Finally, we split the data into 60:20:20 ratio. The first 60 % rows were used to train the machine learning algorithm, 20 % rows were used to validate the tuned model (hyperparameter tuning), and 20 % rows were used to evaluate the model.

### **RESULTS**

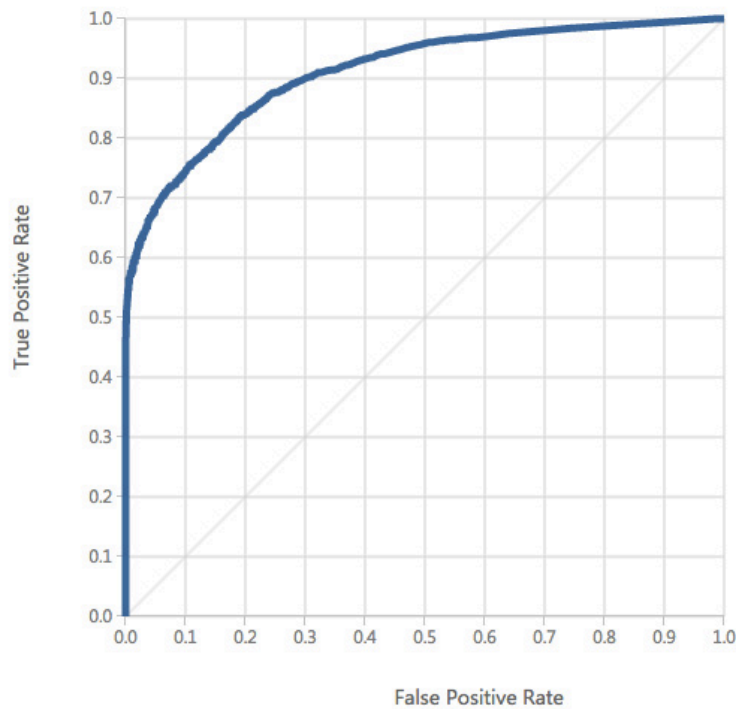
Due to the non-linear relationship nature of the predictors, we compared only algorithms used to predict non-linear relationships. These algorithms are two-class decision forest, two-class boosted decision tree, two-class decision jungle, two-class neural network, and two-class locally deep support vector machine (Ericson, Franks, & Rohrer, 2017). The performance of the algorithms, with random grid hyperparameter turning and 10-fold cross validation, are shown in Table 2. Based on this comparison, the two-class boosted decision tree algorithm was chosen for our prediction model because

it had the highest F1 score. The F1 score metric was chosen because of the class imbalance in the dataset (de Ruiter, 2015) and since both type I and II errors were important for this study.

**Table 2. Evaluation of algorithms with hyperparameter with 10-fold cross-validation**

ALGORITHM	ACCURACY	PRECISION	RECALL	F1-SCORE	AUC
Boosted decision tree	.882	.843	.656	.738	.912
Locally deep support vector machine	.747	.465	.133	.198	.506
Decision forest	.753	.518	.405	.454	.765
Decision jungle	.748	.509	.229	.310	.691
Neuron network	.747	.502	.368	.423	.756

After choosing the two-class boosted decision tree algorithm, we then used the permutation- feature-importance method to eliminate features that weakened the prediction model. We then eliminated Ag and NDVI\_EVI from the model.



**Figure 1. Receiver operating characteristics (ROC) curve**

Figure 1 shows the ROC curve without a cross validation. As shown in Table 3 with a 10-fold cross validation, we achieved 0.886 accuracy, 0.860 precision, 0.660 recall, and F1 score of .746. The area under the ROC curve (AUC) is 0.910.

**Table 3. The evaluation matrixes of the two-class boosted decision tree algorithm with and without hyperparameter tuning**

ALGORITHM	ACCURACY	PRECISION	RECALL	F1-SCORE	AUC
Without a 10-fold cross validation	.881	.814	.688	.746	.911
With a 10-fold cross validation	.886	.860	.660	.746	.910

As shown in Table 4, elevation was the best predictor of the model, followed by Af in both precision and recall metrics. The worst predictor was found to be the soil type, Gd (Dystric Gleysols), but it was still useful in improving the model.

**Table 4. The precision and recall score from permutation feature importance**

FEATURE	PRECISION SCORE	RECALL SCORE
Elevation	.432	.390
Af	.067	.140
NDTI	.062	.023
NDWI_NBR	.055	.033
Ao	.018	.054
LST	.001	.096
Gd	.0001	.0001

We then analyzed the two important features: elevation and Af. Table 4.4 shows the descriptive statistics of the two important features. Farmers grow cassava on the average at the altitude of 165.320 meters and 54 percent of farms are located on Af soil.

**Table 4.4: Descriptive statistics of elevation and Af**

FEATURE	MEAN	MEAN DEVIATION
Elevation (meters)	165.320	17.360
Af (1 = yes, 0 = no)	0.540	0.496

## DISCUSSIONS

Most farmers in the Northeastern region of Thailand do not usually cultivate their farms during the dry season but instead move to urban areas to search for unskilled jobs. This happens mainly because the farmers do not have adequate information on suitable dry-season crops for their areas and how to grow or manage them. The development of a dry-season farming decision support system will enable these farmers support their families by cultivating their farms throughout the year.

In this study, we used satellite imagery data and land characteristics to develop a model for predicting suitable areas for growing cassava, which is one of the identified dry-season crops, for Thailand's Nakhon-Phanom province. We identified the elevation and Ferric Acrisols (Af) as the most im-



portant features for predicting best-suited areas for growing cassava in the province. Similar to Thailand, research has shown that Ghanaian farmers also grow cassava under the Af soil (Ezui et al., 2017).

Similar to Heumann et al. (2011), our study shows that elevation is the most important feature in classifying areas where cassava can be grown in the province. In addition, this study showed that Af is the most influential soil type in determining where farmers can grow cassava. Unlike Kussul et al. (2017), our study identifies the most important features for classifying suitable areas for cassava. This information could help researchers and policy makers in mapping other suitable areas for cassava in other regions of Thailand.

Unlike You et al. (2017), Kussul et al. (2017), and Sun et al., (2017), our study uses interpretable features to represent water, temperature, soil types, and farming characteristics. Although we can use raw imagery bands to feed into ML, having interpretable features helps us gain a deeper understanding of what the results of this study mean in farming practice.

Our study also has provided information about precision and recall, which are used to calculate F1-score as the model performance measure. Elevation helps us gain both quality (precision) and quantity (recall) of cassava classification, while Af helps us achieve the quantity, rather than the quality of classification. Elevation helps us avoid false positives (predicting the case as a cassava field when it is, in fact, a non-cassava field). Af and elevation prevent us from relying on false negatives (predicting the case as a non-cassava field when it is an actual cassava field). Because we carefully selected our features, we obtained a high accuracy rate of 88.6 %. This rate is comparable to that achieved by CNNs in the study by Kussul et al. (2017).

The model developed in this study could be improved by incorporating more land and crop characteristics such as the land slope, soil pH, soil texture, soil nutrient availability, and crop yield. The ground truth data for areas growing cassava could also be collected over several years to give a more representative sample of the areas that are suitable for cassava growing. Further studies will be carried out to predict suitable areas for other dry-season crops and to develop the decision support system.

## CONCLUSION

---

Dry-season satellite imagery for Nakhon-Phanom province in the Northeastern region of Thailand were used to generate NBR, NDVI, NDWI, EVI, NDTI, and LST features for the province. These indices were used in combination with elevation and soil type as predictors of suitable areas for growing cassava in the province. The analysis shows that elevation was the best predictor for suitable areas followed by Af. While Gd was the least accurate, it contributed to improving the overall model. The analysis also showed that the two-class boosted decision tree algorithm was the best for predicting suitable areas for growing cassava. The final prediction model achieved an accuracy score of .886 and F1 score of 0.746. The prediction model could be improved by including more ground truth data on areas growing cassava and the crop yield collected over several years.

## ACKNOWLEDGEMENT

---

We would like to thank the Land Development Department, the Ministry of Agriculture and Cooperatives, Thailand for providing us with the ground-truth data. Without this dataset, this project would not have been possible.

## REFERENCES

---

- AghaKouchak, A., Farahmand, A., Melton, F. S., Teixeira, J., Anderson, M. C., Wardlow, B. D., & Hain, C. R. (2015). Remote sensing of drought: Progress, challenges and opportunities. *Reviews of Geophysics*, 53(2), 2014RG000456. <https://doi.org/10.1002/2014RG000456>
- Artis, D. A., & Carnahan, W. H. (1982). Survey of emissivity variability in thermography of urban areas. *Remote Sensing of Environment*, 12(4), 313–329. [https://doi.org/10.1016/0034-4257\(82\)90043-8](https://doi.org/10.1016/0034-4257(82)90043-8)
- Buhari, U. (2015). *Landsat 8: Estimating land surface temperature using ArcGIS*. Retrieved June 7, 2015, from <https://www.youtube.com/watch?v=uDQo2a5e7dM>
- Caldwell, J. S., Sukchan, S., & Ogura, C. (2005). Management of tropical sandy soils for sustainable agriculture. Retrieved December 13, 2015, from <http://www.fao.org/docrep/010/ag125e/AG125E10.htm>
- Chainuvati, C., & Athipanan, W. (2001). *Crop diversification in Thailand*. Bangkok, Thailand: Food and Agriculture Organization of The United Nations Regional Office for Asia And the Pacific.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321-357.
- Cocke, A. E., Fulé, P. Z., & Crouse, J. E. (2005). Comparison of burn severity assessments using differenced normalized burn ratio and ground data. *International Journal of Wildland Fire*, 14, 189-198. <https://doi.org/10.1071/WF04010>
- Craig, I. A. (1985). *Cropping systems trial results from the northeast rainfed agricultural development project (NERAD)*. In Soil, Water and Crop Management Systems for Rainfed Agriculture in Northeast Thailand. Khon Kaen, Thailand: Khon Kaen University.
- Craig, I. A., & Pisone, U. (1985). *Overview of rainfed agriculture in northeast Thailand*. In Soil, Water and Crop Management Systems for Rainfed Agriculture in Northeast Thailand. Khon Kaen, Thailand: Khon Kaen University.
- de Ruiter, A. (2015). Performance measures in Azure ML: Accuracy, precision, recall and F1 score. Retrieved December 20, 2017, from <https://blogs.msdn.microsoft.com/andreasderuiter/2015/02/09/performance-measures-in-azure-ml-accuracy-precision-recall-and-f1-score/>
- Dusseux, P., Vertès, F., Corpetti, T., Corgne, S., & Hubert-moy, L. (2014). Agricultural practices in grasslands detected by spatial remote sensing. *Environmental Monitoring and Assessment*, 186(12), 8249-8265. <https://doi.org/10.1007/s10661-014-4001-5>
- Ericson, G., Franks, L., & Rohrer, B. (2017). How to choose machine learning algorithms. Retrieved June 12, 2017, from <https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-algorithm-choice>
- Ezui, K. S., Franke, A. C., Ahiabor, B. D. K., Tetteh, F. M., Sogbedji, J., Janssen, B. H., ... Giller, K. E. (2017). Understanding cassava yield response to soil and fertilizer nutrient supply in West Africa. *Plant and Soil*, 420(1-2), 331-347. <https://doi.org/10.1007/s11104-017-3387-6>
- Food and Agriculture Organization. (2000). *Impact of cassava production on the environment*. Retrieved December 12, 2015, from <http://www.fao.org/docrep/007/y2413e/y2413e07.htm>
- Food and Agriculture Organization. (2013). *Save and grow: Cassava*. Retrieved December 13, 2015, from <http://www.fao.org/ag/save-and-grow/cassava/en/4/index.html>
- Garzón, M. B., Blazek, R., Neteler, M., Dios, R. S. de, Ollero, H. S., & Furlanello, C. (2006). Predicting habitat suitability with machine learning models: The potential area of *Pinus Sylvestris* L. in the Iberian Peninsula. *Ecological Modelling*, 197(3), 383-393. <https://doi.org/10.1016/j.ecolmodel.2006.03.015>
- Gao, B.-C. (1996). NDWI - A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment*, 58(3), 257-266. [https://doi.org/10.1016/S0034-4257\(96\)00067-3](https://doi.org/10.1016/S0034-4257(96)00067-3)
- Heumann, B. W., Walsh, S. J., & McDaniel, P. M. (2011). Assessing the application of a geographic presence-only model for land suitability mapping. *Ecological Informatics*, 6(5), 257-269. <https://doi.org/10.1016/j.ecoinf.2011.04.004>

- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis: A global perspective* (7th ed.). Upper Saddle River, New Jersey: Pearson Prentice Hall.
- Idhipong, S., Pong-sed, A., Maolanont, T., Wani, S. P., Rego, T. J., & Pathak, P. (2012). *Improved crops and cropping systems for rainfed northeast Thailand*. International Crops Research Institute for the Semi-Arid Tropics.
- Jiang, Z., Huete, A. R., Kamel, D., & Miura, T. (2008). Development of a two-band enhanced vegetation index without a blue band. *Remote Sensing of Environment*, 112, 3833-3845. <https://doi.org/10.1016/j.rse.2008.06.006>
- Kussul, N., Lavreniuk, M., Skakun, S., & Shelestov, A. (2017). Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing Letters*, 14(5), 778-782. <https://doi.org/10.1109/LGRS.2017.2681128>
- Li, B., Ti, C., Zhao, Y., & Yan, X. (2016). Estimating soil moisture with Landsat data and its application in extracting the spatial distribution of winter flooded paddies. *Remote Sensing*, 8(1), 38. <https://doi.org/10.3390/rs8010038>
- Microsoft Corporation. (2017). *Normalize data*. Retrieved June 8, 2017, from <https://msdn.microsoft.com/en-us/library/azure/dn905838.aspx>
- National Aeronautics and Space Administration. (2000). *Measuring vegetation (NDVI & EVI): Feature articles*. Retrieved June 8, 2017, from [https://earthobservatory.nasa.gov/Features/MeasuringVegetation/measuring\\_vegetation\\_2.php](https://earthobservatory.nasa.gov/Features/MeasuringVegetation/measuring_vegetation_2.php)
- National Statistical Office. (2003). *2003 Agricultural census northeastern region*. Bangkok, Thailand: Ministry of Information and Communication Technology.
- Prasad, A. K., Chai, L., Singh, R. P., & Kafatos, M. (2006). Crop yield estimation model for Iowa using remote sensing and surface parameters. *International Journal of Applied Earth Observation and Geoinformation*, 8(1), 26-33. <https://doi.org/10.1016/j.jag.2005.06.002>
- Ratanawaraha, C., Senanarong, N., & Suriypan, P. (2000). *A review of cassava in Asia with country case studies on Thailand and Viet Nam*. Retrieved December 13, 2015, from <http://www.fao.org/docrep/009/y1177e/y1177e04.htm>
- Sobrino, J. A., Jiménez-Muñoz, J. C., & Paolini, L. (2004). Land surface temperature retrieval from LANDSAT TM 5. *Remote Sensing of Environment*, 90(4), 434-440. <https://doi.org/10.1016/j.rse.2004.02.003>
- Suksa-ngiam, W., Mbugua, J., & Chatterjee, S. (2016). *A GIS decision support system for crop cultivation*. Paper presented at Twenty-second Americas Conference on Information Systems, San Diego. Retrieved from <http://aisel.aisnet.org/amcis2016/Decision/Presentations/5>
- Sun, Z., Leinenkugel, P., Guo, H., Huang, C., & Kuenzer, C. (2017). Extracting distribution and expansion of rubber plantations from Landsat imagery using the C5.0 decision tree method. *Journal of Applied Remote Sensing*, 11(2), 26011. <https://doi.org/10.1117/1.JRS.11.026011>
- SurveyMonkey. (2017). *Sample size calculator*. Retrieved December 20, 2017, from <https://www.surveymonkey.com/mp/sample-size-calculator/>
- Tang, R., Li, Z.-L., Chen, K.-S., Zhu, Y., & Liu, W. (2012). Verification of land surface evapotranspiration estimation from remote sensing spatial contextual information. *Hydrological Processes*, 26(15), 2283-2293. <https://doi.org/10.1002/hyp.8341>
- Thanawong, k, Perret, S. R., & Basset-Mens, C. (2014). Eco-efficiency of paddy rice production in Northeastern Thailand: A comparison of rain-fed and irrigated cropping systems. *Journal of Cleaner Production*, 73, 204-217. <https://doi.org/10.1016/j.jclepro.2013.12.067>
- Thipruck, P. (2013). Thai soil survey. Retrieved May 25, 2017, from <http://www.arcgis.com/home/item.html?id=d6457372db2e42938db0503a18235830>
- Titapiwatanakun, B. (2012). *The rice situation in Thailand (technical assistance consultant's report)*. Asian Development Bank.
- United States Geological Survey (2015). *Using the USGS Landsat 8 product*. Retrieved December 15, 2015, from [http://landsat.usgs.gov/Landsat8\\_Using\\_Product.php](http://landsat.usgs.gov/Landsat8_Using_Product.php)

## Predicting Suitable Areas for Growing Cassava

- United States Geological Survey. (2016). *What are the band designations for the Landsat satellites?* Landsat missions. Retrieved June 4, 2017, from <https://landsat.usgs.gov/what-are-band-designations-landsat-satellites>
- United States Geological Survey. (2017). *Product guide: Landsat surface reflectance-derived spectral indices*. Department of the Interior U.S. Geological Survey. Retrieved from [https://landsat.usgs.gov/sites/default/files/documents/si\\_product\\_guide.pdf](https://landsat.usgs.gov/sites/default/files/documents/si_product_guide.pdf)
- van Deventer, A. P., Ward, A. D., Growda, P. H., & Lyon, J. G. (1997). Using thematic mapper data to identify contrasting soil plains and tillage practices. *Photogrammetric Engineering & Remote Sensing*, 63(1).
- Wardlow, B. D., Egbert, S. L., & Kastens, J. (2007). Analysis of time-series MODIS 250 m vegetation index data for crop classification in the U.S. Central Great Plains. *Remote Sensing of Environment*, 108(3), 290-310. <https://doi.org/10.1016/j.rse.2006.11.021>
- Weng, Q., Lu, D., & Schubring, J. (2004). Estimation of land surface temperature-vegetation abundance relationship for urban heat island studies. *Remote Sensing of Environment*, 89(4), 467-483. <https://doi.org/10.1016/j.rse.2003.11.005>
- The World Bank. (2015). *Food security*. [https://doi.org/10.1596/978-1-4648-0484-7\\_food\\_security](https://doi.org/10.1596/978-1-4648-0484-7_food_security)
- Retrieved September 27, 2015, from <http://www.worldbank.org/en/topic/foodsecurity>
- World Economic Forum. (2015). *Global risks 2015* (10th ed.). Geneva: World Economic Forum.
- You, J., Li, X., Low, M., & Lobell, D. (2017). *Deep Gaussian process for crop yield prediction based remote sensing data*. Paper presented at the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17), San Francisco, CA.
- Yu, X., Guo, X., & Wu, Z. (2014). Land surface temperature retrieval from Landsat 8 TIRS-comparison between radiative transfer equation-based method, split window algorithm and single channel method. *Remote Sensing*, 6(10), 9829-9852. <https://doi.org/10.3390/rs6109829>

**APPENDIX: EQUATIONS USED FOR ANALYSIS**

EQUATIONS	NUMBER	WHERE
$NDVI = (RED - NIR) / (RED + NIR)$	(1)	RED = Red wavelength (band4)
$NDWI = (NIR - SWIR1) / (NIR + SWIR1)$	(2)	NIR = Near infrared wavelength (band 5) EVI = enhanced vegetation index
$EVI = (NIR - RED) / (NIR + 6.0 * RED - 7.5 BLUE + 1)$	(3)	BLUE = blue wavelength (Band 2)
$P_v = \left[ \frac{NDVI - NDVI_{min}}{NDVI_{max} - NDVI_{min}} \right]^2$	(4)	$\epsilon$ = Spectral emissivity
$\epsilon = 0.004P_v + 0.986$	(5)	ML = Band-specific multiplicative rescaling factor
$L_\lambda = M_L Q_{cal} + A_L$	(6)	AL = Band-specific additive rescaling factor from
$T_B = \frac{K_2}{\ln\left(\frac{K_1}{L_\lambda} + 1\right)}$	(7)	Qcal = Quantized and calibrated standard product pixel values (DN) $L_\lambda$ = TOA spectral radiance (Watts/ (m2 * srad * $\mu$ m))
$\rho = hxc / \sigma$	(8)	
$S_t = \frac{T_B}{1 + \left(\frac{\lambda T_B}{\rho}\right)} \ln(\epsilon)$	(9)	$K_1$ = Band-specific thermal conversion constant
$NDTI = \frac{SWIR1 - SWIR2}{SWIR1 + SWIR2}$	(10)	$K_2$ = Band-specific thermal conversion constant
$NBR = \left[ \frac{NIR - SWIR2}{NIR + SWIR2} \right]$	(11)	$T_B$ = At-satellite brightness temperature (K)
$Z = \frac{(x - min)}{(max - min)}$	(12)	$\lambda$ = Wavelength (band 10) h = Planck's constant (6.626 $\times$ 10 <sup>-34</sup> J s) $\sigma$ = Boltzmann constant (1.38 $\times$ 10 <sup>-23</sup> J/K) C = Velocity of light (2.998 $\times$ 10 <sup>8</sup> m/s) $S_t$ = Land Surface Temperature (K) SWIR1 = short wave infra-red wavelength (band 6) SWIR2 = short wave infra-red wavelength (band 7) Z = normalized score (0 to 1) x = score of the case min = the minimum value of the feature max = the maximum value of the feature

## BIOGRAPHIES

---



nell University, USA.

**Kimani Mbugua** is a Ph.D. student at Claremont Graduate University, California. His research is mainly focused in GIS, Remote Sensing, and Machine Learning. He is currently researching on the application of computer vision in aerial and satellite imagery to automate the prediction, identification, classification, and mapping of crop and plant conditions in near real-time. Kimani has also worked as a testing, design, and development engineer in various Agricultural Machinery Testing Centers in Kenya. He holds a B.Sc. in Engineering from University of Nairobi, Kenya, and a Master's in International Agriculture and Development from Cor-



**Watanyoo Suksa-ngiam** is a Ph.D. student at Claremont Graduate University, California. He received a B.E. in Control System Engineering from King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand. He also received an MBA from National Institute of Development Administration, Bangkok, Thailand and M.S. in Information Systems and Technology from Claremont Graduate University. Suksa-ngiam conducted research in business and technology before he came to the USA. His primary research is currently focused on digital economy. Besides, he often conducts research in the fields of data science and geographic information system.