Machine Learning Algorithms for Soil Analysis and Crop Production Optimization: A review

Kennedy Senagi, Nicolas Jouandeau, Peter Kamoni

Machine Learning Algorithms for Predicting Land Suitability in Crop Production: A Review

Kennedy Mutange Senagi[1,2], Nicolas Jouandeau[2], Peter Kamoni[3]

[1]Dedan Kimathi University of Technology, Kenya

[2]Université Paris8, France

[3]Kenya Agricultural and Livestock Research Organization, Kenya

Abstract. Food security is an important factor to consider for a healthy nation. Agriculture is a an important sector for Kenya's economic growth and for achieving Kenya's Vision 2030. However, due to the change in farming practices from one season to the other, increased demand for land, intensive land use and failure to apply recommended farming practice results in soil degradation over time. Land evaluation is the process of assessing land performance for specified purposes *e.g.* crop production. Currently, in the Department of Soil Survey at Kenya Agricultural and Livestock Research Organization, and in many other soil survey institutions, land evaluation process is done manually and is time consuming, stressful and prone to human errors. In this paper, we review Machine Learning (ML) algorithms applied in land suitability for crop production and performance comparison of ML algorithms. We found out that parallel random forest (PRF) performed better that other supervised machine learning classification algorithms. We set up an experiment prototype, PRF scored an average RMSE of 1.03, ACC score of 0.92 and average execution time of 1532.32 milliseconds. PRF can be applied in predicting land suitability for crop production. This can reduce human errors, reduce time, and improve accuracy in land evaluation process for crop production. machine learning, parallel random forest, land evaluation, soil analysis.

1. Introduction

1.1. State of Food Security in Kenya

Kenya heavily relies on agriculture for economic growth and food security [1]. Agriculture contributes about a quarter of Kenya's Gross Domestic Product (GDP). The baseline of agriculture is expected to grow at approximately 3.7% per year during

2010-2020. The Kenyan government total public expenditure on agriculture sector has been increasing in recent years. This is to accelerate growth in the agricultural sector, increase food production, lower food prices and reduce global economic crisis [2]. In Kenya Vision 2030, agriculture is one of the sectors identified to contribute to the realization of the 10 per cent economic growth rate per annum envisaged under the economic pillar. In Vision 2030 document, Kenya desires to promote an innovative, commercially oriented and modern agriculture sector [3].

1.2. Land Evaluation for Crops Suitability

Due to changes in agricultural practices and increased demand of land, land evaluation has come to be a very important procedure before initiating land use in a particular land *e.g.,* engaging in farming activities on a piece of land or urban planning [4]. Land evaluation is the assessment of performance of Land for a defined use. The principal objective of land evaluation is to select the optimum land use for each defined land unit, taking into account both physical and socio-economic considerations as well as the conservation of environmental resources for future use [5]. It involves the analysis and interpretation of basic surveys of: soil, vegetation, climate, socio-economic (if available) and other aspects of land such as crop land use requirements. The beneficiaries of land evaluation process are the farmers who gain knowledge of what crops to grow in a particular piece of land, improve on their crop yields as well as learning new skills for sustainable land management [6]. Land evaluation, involves matching land areas called Land Mapping Units (LMU) with land uses, called Land Utilization Types (LUT), to determine the relative suitability of each land area for each land use. Land Utilization Types are specified by a set of Land-Use Requirements (LUR), which are the conditions of land necessary for the successful and sustained practice of a given LUT [6].

In practice, representative soil profiles are dug in each LMU and sampled horizon wise. The soil samples are analyzed in the laboratory for physical and chemical composition. Data from such soil profiles together with climatic, crop requirement and socio-economic data (if available) are then used to perform the land evaluation. The process involves selecting the land qualities/characteristics to be used (*e.g.*, temperature regime, soil water holding capacity, oxygen availability, rooting depth and salinity hazard) and rating them. The LMU are rated based on the selected rating scheme. The same rating scheme is used to rate the crop requirements. A matching scheme of the rated LMU qualities/characteristics with rated crop land use requirements is applied to determine the land suitability of that LMU for the specific crop. Suitability of a LMU can be categorized as either S1 (highly suitable i.e. where 75% to 100% yield is attained), S2 (moderately suitable *i.e.,* where 50% to 75% yield is attained), S3 (marginally suitable *i.e.* where 25% to 50% yield is attained) or NS (unsuitable *i.e.*, where below 25% yield attained) [7].

Land suitability is the ability of a portion of land to allow the production of crops in a sustainable manner. The analysis identifies the main limiting factors for a particular

crop production. It also enables decision makers to develop a crop management system for increasing the land productivity or choose an appropriate fertilizer to increase productivity of the land [6].

The suitability of each LMU for each LUT is done as follows [6]:

1. Determining the actual Land Characteristic (LC) values for the LMU. These can be measured, observed in the field and estimated through laboratory measurements or remote sensing
2. Combining the Land Characteristics (LC) values into Land Quality (LQ) values (*i.e.*, inferring the LQ from the set of LC's)
3. Matching the LQ values with Land Use Requirements *e.g.*, crop requirements
4. Inferring the suitability from the set of LQ's and LURs)

1.3. Computing Techniques in Land Evaluation

In recent years, computing concepts have been successfully applied in land evaluation for crop suitability. Some of the computing techniques applied are: Geographical Information Systems (GIS) and machine learning. GIS are computer based tools that captures, stores, analyses, manages and presents all types of spatial or geographical data. GIS techniques have successfully been used to evaluate land suitability [24]. Machine learning is applied in various land evaluation processes which include: predicting soil properties [8], predicting soil fertilizer requirements [9] and land use suitability [10]. In essence, machine learning is concerned with the question of how to construct computer programs that automatically improve with experience. Besides agriculture, state of the art machine learning applications have successfully been developed, for examples, data-mining programs that learn to detect fraudulent credit card transactions, information-filtering systems that learn a users' reading preferences and autonomous vehicles that learn to drive on public highways [11].

2. Machine Learning

Machine learning is a discipline in computer science that emerged from the field of artificial intelligence. Machine learning is concerned with the construction of computer programs that automatically improve with experience. A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. It is too costly and impractical to design intelligent systems by gathering experts' knowledge and then hard-wiring it into a machine. To build intelligent machines, researchers have comprehended that these machines should

learn from and adapt to the environment they learn in. Building intelligent machines involve learning from data [12]. Common types of machine learning problems are categorized as: supervised machine learning, unsupervised machine learning and reinforcement learning.

2.1. Supervised Machine Learning

In Supervised ML algorithms, a general hypothesis is built from external supplied data. From the hypothesis we can then make predictions of future instances. Actually, in Supervised ML, a concise model of distribution of class labels is built from predictor features. This classifier is then used to assign class labels (unknown) to predictor features (known) [33]. Supervised learning problems can be categorized as: classification and regression. They are discussed below:

1. Classification Problem: The goal of classification is to learn a mapping of inputs x to outputs y, where $y \in \{1, \ldots, C\}$, with C being the number of classes. If $C = 2$, then it is a binary classification problem, if $C > 2$, then it is a multi-class classification problem. In classification problems, we assume $y = f(x)$ for some unknown function f, the goal of learning is to estimate the function f given a labeled training set, and to make predictions using $\hat{y} = \hat{f}(x)$. The main goal is actually to make predictions on new inputs that we have not seen before *i.e.* generalization. Some examples of classification supervised machine learning algorithms are: decision trees, ensembles (bagging [21] random forest [20] boosting [26]), k-NN and neural networks [13].
2. Regression Problem: Regression problems are just like classification, however, the output consists of one or more continuous variables. For example, predicting tomorrow's stock market price given the current market conditions and other relevant information. Examples of regression machine learning algorithms include: naive bayes and logistic regression [14].

2.2. Unsupervised Machine Learning

In unsupervised learning problem, we are given output/labeled data, without any inputs the goal is to learn relationships and structure from such data, this is sometimes called knowledge discovery. Unlike supervised learning, the training data consists of a set of input without any corresponding output values. Unsupervised learning problems can be formalized as one of density estimation, i.e. we want to build models of the form $p(x_i|\theta)$, unlike supervised learning which is $p(y_i|x_i, \theta)$. Supervised learning is a conditional density estimation, while unsupervised learning is an unconditional density estimation. By contrast, in unsupervised learning x_i is a vector of features and we need to create multivariate probability models, whereas, in

supervised learning, y_i is single variable that we are trying to predict. Therefore, supervised learning problems can use univariate probability models. Examples of unsupervised machine learning algorithms are: k-means and Neural Networks [13].

2.3. Reinforcement Machine Learning

Reinforcement learning [15] is concerned with finding suitable actions to take in a given situation in order to maximize a reward. In contrast, with supervised learning which are given labeled data, reinforcement learning algorithms discover optimal outputs by a process of trial and error. Generally, there is a sequence of states and actions in which the learning algorithm is interacting with its environment. In many cases, the current action, not only affects the immediate reward, but also has an impact on the reward at all subsequent time steps [12].

3. Machine Learning Algorithms Applied to Land Evaluation for Suitability Assessment

In Anitha and Acharjya (2017) studies, their research objective was to predict crop to be grown in Vellore District. They came up with a hybridizing of rough set on fuzzy approximation space and ANN. The rough set on fuzzy approximation space was to get almost equivalence classes where attribute values are not qualitative. The classified information was the taken to ANN algorithm for training, prediction and testing. They used a dataset collected from Krishi Vigyan Kendra of Vellore District between 2007 to 2014. The dataset contained 2193 objects with 15 attributes. They used 26 soil attributes: soil pH, moisture, organic matter, availability of nitrogen, availability of phosphorus, availability of potassium, water pH, calcium, nitrate, magnesium, selenium rainfall, copper, zinc, manganese and iron. The dataset was divided into 55% training and 45% testing data and validated by N-folds cross-validation. The experiments were developed in R language. The average Mean Square Error (MSE) was 0.2436 and Accuracy of 93.2% [28].

Fereydoon *et al*. (2014) developed a Support Vector Machine (SVM) - Two Class Model for land suitability analysis for wheat production in Kouhin region in Iran. They used twenty two soil representative soil profiles information, each soil profile having ten features: climatic (precipitation, temperature), topographic (relief and slope) and soil-related (texture, $CaCO_3$, Organic Carbon, coarse fragment, pH, gypsum). They implemented a Two Class SVM model on a non-linear class boundaries; non-linear mapping of input vector into a high dimensional feature space. They used MATLAB 8.2 software to design and test the SVM model. They randomized the dataset and split it into training (80%) and test (20%). In performance evaluation, they got an RMSE of 3.72 and R^2 of 0.84 [4].

Land suitability classification using a large number of parameters is time consuming and costly. With this research problem, Hamzeh *et al*. (2016) presented a combination of feature selection (best search, random search and genetic search methods) and fuzzy-analytical hierarchical process(AHP) methods to improve selection of important features from a large number of parameters. On feature selection, random search performed slightly better than genetic search methods and best search. The dataset was retrieved from land classification report published by Khuzestan Soil and Water Research Institute. They found that soil texture, wetness, salinity and alkalinity were the most effective parameters for determining land suitability classification for the cultivation of barely in the Shavur Plain, southwest Iran. The report showed that soil salinity and alkalinity, soil wetness, $CaCO_3$, gypsum, pH, soil texture, soil depth and topography were the most important soil properties to consider for cultivating barley in the study area [29].

Mokarram *et al*. (2015) used AI and ML to automate the land suitability classification for growing wheat. They used a dataset with data collected from Shavur plain, northern of Khuzestan province, southwest of Iran. The dataset had the following attributes: topography (Primary slope, Secondly slope and Micro relief), salinity and alkalinity (EC and ESP), wetness (Groundwater depth and Coroma depth), soil texture, $CaCO_3$, soil depth, gypsum and pH (H_2O). Land suitability classes were classified as highly suitable (75%-100%), moderately suitable (50%-75%), marginally suitable (25%-50%) and not suitable (25%-0%). Mokarram et al. (2015 implemented Bagging, AdaBoost and RotForest algorithms and evaluated the performance of their experiments using 632+ bootstrap and 10-fold cross validation. Their results classified 26% of the land being moderately suitable, 25% being marginally suitable and 49% being not suitable. Moreover, they found that RotBoost algorithm had a better accuracy than Single Tree, Rotation Forest, AdaBoost, Bagging algorithms in predicting land suitability class. RotBoost recorded 99% and 85% accuracy score for bootstrap and cross validation respectively [30].

Dahikar and Rode (2014) demonstrated the use of ANN in predicting crop suitability from soil attributes. The attributes include: type of soil, pH, nitrogen, phosphate, potassium, organic carbon, sulphur, manganese, copper, calcium, magnesium, iron, depth, temperature, rainfall, humidity. They set up the experiments in MatLab [31]. However, performance results were not given in the papers. The researchers acknowledge the potential of using ANN in predicting crop suitability from soil data collected from rural district.

Elsheikh *et al*. (2013) presented ALSE, which was an intelligent system for assessing land suitability for different crops (*e.g.,* mango, banana, papaya, citrus, and guava) in tropical and subtropical regions based on geo-environmental factors. ALSE supported GIS capabilities on the digital map of an area with FAO-SYS framework model. It also had some necessary modifications to suit the local environmental conditions for land evaluation, and the support of expert knowledge. ALSE had the capability to identify crop-specific conditions and systematically computes the spatial and temporal data with maximum potential. This would help land planners to make

complex decisions within a short period taking into account sustainability of a crop. Test dataset was collected from agricultural land in Terengganu, West Malaysia. Some of the attributes they used were: rainfall, soil attributes (nutrient availability, rooting conditions, nutrient retention, soil workability and oxygen soil drainage class) and topology [32].

From the above literature, we see diverse applications of statistical and ML techniques in land suitability for better crop production: ANN, AdaBoost, SVM, fuzzy logic and PCA. Moreover, algorithms perform differently on different datasets. Experiments set up comparing different ML algorithms on standardized dataset can be better analysis.

4. Machine Learning Algorithms Comparison and Analysis

Besides ML being applied in agriculture, we found that researches who analyzed ML algorithms and measured their performance on standardized experiments. These analysis can guide to select an algorithm that is robust and has better performance. For example, Rich *et al*. (2008) did an empirical evaluation of supervised learning of high dimension data and evaluated performance on three metrics, namely: accuracy (ACC), Area Under the ROC Curve (AUC), and squared error (RMS). They also studied the effect of increasing dimensionality, ranging from 761 to 685569, the performances of the algorithms: SVM (LaSVM kernel and RBF kernels using stochastic gradient descent), ANN, Logistics Regression (LR) and Naïve Bayes (NB). Across all dimensions, RF was the best in performance, followed by ANN, then boosted trees, and lastly SVMs [18]. These results were similar to Ogutu *et al*. (2011) that showed RF outperforming SVM [23]. However, if results of each metric are examined, in terms of accuracy, boosted decision trees are the best performing models followed by RF. It's worth noting that boosted decision trees performed better than RF in relatively low dimensionality, but as dimensionality increases RF tends to perform much better than boosted decision trees. In terms of RMS performance metric, RF is marginally better than boosted trees. While in AUC, RF is a clear winner, followed by k-NN. It's important to note that, although ANN performs second best overall, it's neither first or second in either performance metrics. This is because, ANN consistently produce very good results, though perhaps not outstanding on all the performance metrics.

Kim *et al*. (2012) experiments on image classification showed that SVM performance was much better than k-NN [16]. Similarly, Sanghamitra *et al*. (2011) compared the performance of SVM and k-NN on Oriya Character Recognition and found SVM with an accuracy rate of 98.9% while k-NN had 96.47%. Thus showing that SVM had a better accuracy rate, though k-NN classifier consumed lesser storage space and had less computation than SVM [17]. SVMs performed generally better and had more accurate results than Artificial Neural Networks (ANNs) though difficult to understand the learned function. Moreover, Colas and Brazdil (2006) experiment on

text classification showed that k Nearest Neighbour (k-NN) could scale up very well and even achieve better results than SVM [19].

Similarly, Manuel *et al*. (2014) evaluated 179 classifiers got from 17 families namely: discriminant analysis, multiple adaptive regression splines, Bayesian, rule-based classifier, boosting, neural networks, bagging [21] stacking, random forests and other ensembles, SVM, decision trees, generalized linear models, logistic and multinomial regression, nearest neighbours, partial least squares and principal component regression and other methods. They were implemented in Weka, Matlab, R (with and without the caret package) C and other relevant classifiers. Parallel random forest, a version of RF, turned out to be the best classifier when implemented in R and accessed without caret. It achieved 94.1% accuracy. The second best was SVM with Gaussian and polynomial kernels when implemented in C using LibSVM which achieved 92.3% accuracy. Moreover, different learning algorithms portrayed different performance levels on change in dimensionality. Increase in dimensionality, affect the relative performance of the learning algorithm. On high dimensional data, RF performs very well followed by boosted trees and logistic regression whereas boosted decision trees perform exceptionally well when dimensionality is low [22].

Moreover, Hastie *et al*. 2009 did a comparison of characteristics of ANN, SVM, trees and k-NN algorithms. Tress showed a good performance in: natural handling data mixed type, handling data that has mixed values, robustness to outliers, insensitivity to monotone transformation of inputs, computational scalability and ability to deal with irrelevant inputs. k-NN was the second best followed by SVN and ANN which performed fairly better in: ability to extract linear, combinations of features, interpretability and predictive power. This shows that Machine learning algorithms perform differently in differently in different experiment environments. Some of the factors that affect performance of machine include: number of records in a data, type of data (*e.g*., images, text, voice) and complexity of data [22]. Manuel *et al*. (2014) measured the percentage of the maximum accuracy of several algorithms including parallel RF, RF, rotational forest, SVM, k-NN and ANN. They found out that parallel random forest scored a better maximum accuracy than the others.

This research paper has discussed some of the machine learning techniques applied to land evaluation and crop suitability assessment. We have also discussed ML algorithms and applications. We see that parallel random forest performs comparatively well compared to SVM, ANN, k-NN and other machine learning classification algorithms.

5. Experiment Prototyping

5.1. Methodology

The standardized dataset [25] was tailored to predict the age of abalone tree from the attributes: sex, length, diameter, height, whole weight, shucked weight, viscera weight, shell weight and rings. Some of the reasons that led us to select the dataset to

prototype parallel random forest were: the number of attributes/features, size of dataset and the dataset is intended for classification. These are characteristics that we expect to find in soil analysis dataset [4].

In this study we implemented the experiment prototype in Python 3.5 in Ubuntu 16.04 LTS 64 bit operating system running on Intel® Core™ i3-2365M CPU @ 1.40GHz × 4 computer systems. We randomized the data and used two thirds of the 4177 records for training the model and one third for as the test data. We run each experiment 10 times and recorded the duration of each run, the ACC score, standard deviation (for y and $\hat{y}$) and RMSE. We used RMSE defined in equation (1) and ACC score defined in equation (2), where $\hat{y}$ is the predicted value, y is the actual value, n is the number or records in the test or predicted dataset and m is the number of runs. RMSE is used to measure the accuracy and validity of the predicted values. RMSE has been used to measure predictability accuracy. Standard deviation(std) of the y training set, $\hat{y}$ of the predicted and all the y labels results was captured [4]. Moreover, we used Speedup to evaluate the performance of parallelizing RF. Speedup measure how much performance gain is achieved by parallelizing RF, Speedup is defined in

$$Average\ \ RMSE_j = \frac{1}{m}\sum_{m=1}^{m=10}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i^{(j)} - \hat{y}^{(j)}\right)^2} \quad (1)$$

$$Accuracy\ \ score(y, \hat{y}) = \frac{1}{n_{samples}}\sum_{i=0}^{n_{samples}-1} 1 - (\hat{y}_i = y_i) \quad (2)$$

5.2. Results

Results were tabulated in Table **1**. We observed RF and PRF had the same RMSE, ACC and standard deviations. RF recorded a significantly high execution time compared to PRF.

Table 1. Average results for Parallel Random Forest and Random forest algorithms.

Algorithm	RMSE	ACC	Std of $\hat{y}$ (pm)	Std of y (pm)
RF	1.03	0.92	2.45	3.3
PRF	1.03	0.92	2.45	3.3

Table 2. Parallization of Random Forest algorithm.

Core(s)	Time (ms)
1	4316.0
2	3660.0
4	3159.0
8	2769.0
16	2912.0
24	3041.0
32	3135.0
40	3247.0
48	3245.0

5.3. Discussion

RMSE measures the accuracy of the predicted values and the actual value. The best RMSE value is zero; the higher the value, the lower the accuracy. The prototype experiment scored an average value of 1.03 meaning the RMSE was averagely good. ACC computes the subset accuracy. The best ACC score is 1 and the worse is 0. The results show that both PRF and RF had an ACC score of 0.92 meaning the score was significantly good. Both RF and PRF scored a standard deviation (Std) of 2.45. The best std is zero. The higher the std, the higher the deviation from the mean. Meaning the data ranges of the y labels were deviating significantly from the mean. The results of RMSE, ACC score and std could be due to the nature of the dataset(high/low biased data), complexity of the records or number of features(high/low) [22].

Moreover, RF recorded a higher execution time than PRF. Meaning PRF was faster than RF. This could be because PRF utilized multiple CPU cores compared to RF that used a single core. Logically, we expected to get more work done by N processors to be N times faster, which was not the case as seen in Table **2**. This could be due to a certain amount of overhead incurred in keeping the parallelized system working. Some of the overheads include: time spent in interprocess interaction and idling time as a result of load imbalance and/or synchronization [27].

6. Conclusion

Kenya is keen in food security in order to feed her population. Land evaluation for crop suitability assessment is an important procedure for identifying crops that grow optimally in a specific soil sample. In Kenya Soil Survey Department, Kenya Agricultural and Livestock Research Organization and many other institutions, land evaluation is done manually, time consuming, stressful and prone to human errors. Making the process tedious and prone to errors. Computer science techniques have been proposed to leverage the challenges including geographical information systems

and machine learning. We reviewed literature of applications of ML algorithms in land suitability evaluation for crop production. From the review, we found that PRF performed comparatively well compared to other algorithms like k-NN, ANN and SVM. PRF had a good handling of data of mixed type and values, was robust to handle outliers in an input space, had a good predicative accuracy and could be scaled to large datasets. We set up a simple experiment prototype to implement PRF. PRF had a significant performance in terms of RMSE, Accuracy and Time of Training. Among other machine learning preparation procedures, we need to: have sufficient number of records for training the model, take note of data complexity have enough number of features in the dataset. In essence, this prototype can be scaled up to land suitability dataset, as we intend to utilize machine learning algorithms in soil analysis to optimize crop production.

References

1. Ngeno V, Mengist C, Langat B. (2011). Technical efficiency of maize farmers in Uasin Gishu district, Kenya. 10th African Crop Science Conference Proceedings, Maputo, Mozambique, pp. 41-47 ref.25.
2. Athur M, Karl P, Samuel B. (2012). Regional Strategic Analysis and Knowledge Support System (ReSAKSS). ReSAKSS Africa, Washington.
3. Government of Kenya. (2007). A Globally competitive and prosperous Kenya. www.opendata.go.ke.
4. Fereydoon S, Ali K, Azin R, Ghavamuddin Z, Hossein J, Munawar I. (2014). Support vector machines based-modeling of land suitability analysis for rainfed agriculture. Journal of Geosciences and Geomatics, Vol. 2, Issue No. 4, pp. 165-171.
5. FAO. (1985). Guidelines for land evaluation for rainfed agriculture. FAO, Rome.
6. FAO. (1976). FAO And Agriculture Organization Of The United Nations. FAO, Rome.
7. Abbas T, Fereydoon S. (2015). Qualitative land suitability evaluation for maize (Zea mais L.) in Abyek, Iran using FAO method. Global Journal of Research and Review (GJRR), Vol. 2, Issue No. 1, pp. 37-44.
8. Jay G, Anurag I, Jayesh G, Shailesh G, Vahida A. (2012). Soil data analysis using classification techniques and soil attribute prediction. International Journal of Computer Science Issues, Volume 9, Issue No. 3.
9. Gholap J. (2012). Performance tuning of J48 algorithm for prediction of soil fertility. Asian Journal of Computer Science and Information Technology, Vol 2, Issue No. 8.
10. Shattri BM, Saied P, Ahmad RBM, Saied P. (2012). Optimization of land use suitability for agriculture using integrated geospatial model and genetic algorithms. ISPRS Ann Photogramm Remote Sens Spat Inf Sci I-2, pp. 234-299.
11. Welling M. (2011). A First Encounter with Machine Learning. Irvine, CA.: University of California.

12. Jordan M, Kleinberg J, Schölkopf B. (2007). Information science and statistics. Springer Science+Business Media, LLC.
13. Murphy KP. (2012). Machine learning a probabilistic perspective. London, England: The MIT Press.
14. James G, Witten D, Hastie T, Tibshirani R. (2013). An introduction to statistical learning. Springer (Vol. 6). New York: Springer.
15. Sutton RS, Barto AG. (1998). Reinforcement learning: An introduction. Cambridge: MIT press, Vol. 1, Issue No. 1.
16. Kim J, Byung-Soo K, Silvio S. (2012). Comparing image classification methods: K-nearest-neighbor and support-vector-machines. American-Math' 12/CEA' 12 Proceedings of the 6th WSEAS International Conference on Computer Engineering and Applications, and Proceedings of the 2012 American conference on Applied Mathematics, ISBN: 978-1-61804-064-0, pp. 133-138.
17. Sanghamitra M, Himadri NDB. (2011). Performance comparison of SVM and k-NN for oriya character recognition. International Journal of Advanced Computer Science and Applications, Special Issue on Image Processing and Analysis, pp. 112-116.
18. Rich C, Alexandru N-M. (2006). An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on Machine learning, ACM, pp. 161-168.
19. Colas F, Brazdil P. (2006). Comparison of SVM and some older classification algorithms in text classification tasks. In IFIP International Conference on Artificial Intelligence in Theory and Practice, pp. 169-178, Springer US.
20. Breiman L. (2001). Random forests. Machine learning, Kluwer Academic Publisher, 45(1), pp. 5-32.
21. Breiman L. (1996). Bagging predictors. Machine Learning, Kluwer Academic Publishers, Boston, Vol 24, Issue No. 2, pp. 123-140.
22. Manuel FD, Eva C, Senen B. (2014). Do we need hundreds of classifiers to solve real world classification problems?. Journal of Machine Learning Research, pp. 3133-3181.
23. Ogutu JO, Piepho HP, Schulz-Streeck T. (2011). A comparison of random forests, boosting and support vector machines for genomic selection. In BMC proceedings, Vol. 5 (Suppl 3), pp. 1-5, BioMed Central.
24. Abraham M, Daniel H, Abeba N, Tigabu D, Temesgen G, Hagos G. (March, 2015). GIS based land suitability evaluation for main irrigated vegetables in Semaz Dam, Northern Ethiopia. Research Journal of Agriculture and Environmental Management, ISSN 2315-8719. Vol. 4(3), pp. 158-163.
25. UCI, "Abalone Data Set." https://archive.ics.uci.edu/ml/datasets/Abalone. Date Accessed: 9th January 2017.
26. Shapire R, Freund Y , Bartlett P, Lee W. (1998) Boosting the margin: A new explanation for the effectiveness of voting methods. Annals of Statistics, Vol. 26, No. 5, pp. 1651-1686.
27. Gagne SG. (2005). Operating Systems Concepts, 7th Edition. John Wiley and Sons, 2005.

28. Anitha A, Acharyjya DP. (2017). Crop suitability prediction in Vellore District using rough set on fuzzy approximation spapce and neural network. The Natural Computing Applications Forum 2017, Springer.
29. Hamzeh S, Mokarram M, Haratian A, Bartholomeus H, Ligtenberg A, Bregt AK. (2016). Feature selection as a time and cost-saving approach for land suitability classification (Case Study of Shavur Plain, Iran). Journal of Agriculture, Multidisciplinary Digital Publishing Institute, Vol. 6, Issue No. 4.
30. Mokarram M, Hamzeh S, Aminzadeh F, Zarei AR. (2015). Using machine learning land suitability classification. West African Journal of Applied Ecology. Vol. 23, Issue No. 1, pp. 63-73.
31. Dahikar SS, Rode SV. (2014). Agricultural crop yield prediction using artificial neural network approach. International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering, Vol. 2, Issue No. 1.
32. Elsheikh R, Shariff ARBM, Amiri F, Ahmad NB, Balasundram SK, Soom MAM. (2013). Agriculture Land Suitability Evaluator (ALSE): A decision and planning support tool for tropical and subtropical crops. Journal of Computers and Electronics in Agriculture. Elsevier Vol. 93, pp. 98-110.
33. Kotsiantis SB, Zaharakis ID, Pintelas PE. (2016). Machine learning: a review of classification and combining techniques. Artificial Intelligence Review, Vol. 26, Issue No. 3, pp. 159-190, Springer.