ISSN 2278-3091
Volume 9 No.2, March -April 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse31922020.pdf
https://doi.org/10.30534/ijatcse/2020/31922020

Image Processing Techniques and Data Mining Algorithms for Coffee Plant's Leaves Classification

Khenilyn P. Lewis[1], Mary Ann F. Quioc[2], Juancho D. Espineli[3]
[1]AMA University, Philippines, khenilyn@yahoo.com
[2] AMA University, Philippines, maryannquioc@gmail.com
[3] AMA University, Philippines, jcespineli@gmail.com

ABSTRACT

Arabica coffee is known for its unique taste and aroma. This coffee variety contributed majority of coffee production in the world. However, arabica coffee and other coffee varieties are prone to extinction because of several reasons including climate change, drought, diseases and issues in identification of nutritional deficiencies. Nutritional deficiencies are identified and classified manually with an expert to validate the visual symptoms occurred in the coffee leaves. On the other hand, the utilization of image processing to analyze images as well as data mining is a strong combination for classification. Therefore, this study was conducted to classify the nutritional deficiencies in arabica coffee plants including Phosphorus (P) and Potassium (K) using image processing and data mining. The images of 2045 instances with 1001 features undergone image processing techniques such as image acquisition, image pre-processing and image analysis. The 70% of data was for training and 30% was for testing using Waikato Environment of Knowledge Analysis (WEKA) and Orange Visual Programming. Random Forest, Support Vector Machine (SVM), Neural Network (ANN) and K-Nearest Neighbors (KNN) served as the classifiers of two classes. Results shows that SVM has the highest AUC of 1.000 and CA, F1, Precision and Recall of 0.983. The Correctly Classified Instances (CCI) is 98.73% and Incorrectly Classified Instances (ICI) is 1.27%. Further, the Kappa statistics of 0.97 shows an almost perfect value of agreement and implies that the classifier is better in coffee plants leave classification together with image processing.

Key words: coffee plants, data mining, image processing, machine learning

1. INTRODUCTION

Coffea Arabica is the most popular coffee variety and produces the 75% of coffee production in the world because of its rich flavor and aroma [1]. Arabica plants grows in high altitudes area and the most expensive coffee variety [2]. In the Philippines, arabica coffee marked the second largest production among four types named Robusta, Excelsa and Liberica. The volume of production in coffee varieties (mt) shown that Arabica coffee has 8,717 in production. It can be found in high elevation areas usually in low air temperature [3]. Though arabica shows second largest of coffee production in the country. It is also noted that it is hard to cultivate and grow since the country has tropical climate [4]. This coffee variety has the largest plantation in the mountainous areas in the country, like in Benguet, Mountain Province and Sagada [5]. Luckily, the researchers found some arabica coffee plants in the area of Cavite, Philippines. Cavite is part of Region IV-A and a known producer of Liberika Coffee locally known as Kapeng Barako.

However, despite the production of arabica coffee in the Philippines and in the global market, a study was conducted that 60% of coffee varieties including arabica coffee will be extinct. The extinction is due to climate change, plant diseases and nutritional deficiencies, drought and deforestation [6]. The Philippines also noted a decrease of coffee production in the country [7]. Among the mentioned causes of extinction of arabica coffee variety, this study focused in classification of nutritional deficiencies in coffee plants. It is important to identify the nutritional deficiencies in plants as it is a way of providing correct remedies and measures. It is essential to boost the nutritional content of plants to survived and produce coffee beans. As such, the proper nutritional identification can save money, effort and time to coffee farmers and growers [8]. Nevertheless, identification of nutritional deficiencies is manually performed by coffee growers and sometimes experts and laboratory machine for these are expensive and unavailable. The process of identifying and classifying the nutritional deficiencies in coffee plants is expensive and time consuming too.

Further, since there are several reasons for coffee extinction, and it is important to provide risk management measure to save our coffee. Image processing is a popular way of enhancing and reading images to get important information or features. Thus, these features are used to processed data and even used for pattern recognition. In addition, machine learning and data mining is being utilized to predict certain forms using different classification algorithms which can be trained and further used for Artificial Intelligence [9]. With the used of image processing and machine learning algorithms, a prediction model can be developed. Machine

learning could provide prediction and classification to which nutritional deficiency is present to coffee plants.

2. METHODOLOGY

This section discusses the classification models and image processing techniques used in the conduct of the study. Also, the data gathering procedures, prediction and validation of the classifiers implemented was presented.

2.1 Classifications

Machine learning used historical data to train algorithms for prediction. The types of machine learning are supervised, unsupervised and reinforcement [10]. Machine learning is also part of Artificial Intelligence that produces knowledge in training models and historical data as input [11]. The study utilized the most popular data mining algorithms used in image processing, these are Random Forest, Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Neural Network (NN).

A. Random Forest

Random Forest can be used for classification in machine learning. It is composed of several trees during the training process and return result or prediction values of the input data. This algorithm also is known for high accuracy in returning results and has flexible nodes.

B. Support Vector Machine (SVM)

Support Vector Machine (SVM) is an algorithm that outputs hyperplane which divides the two parts of each class. Technically, SVM separate classes and best used for two classes classifications [12].

C. K-Nearest Neighbor (KNN)

This algorithm is also used for classification and regression. It is known as easy to implement and simple [13].

D. Neural Network

Neural Network patterns the process of the brain in which neurons are used to execute programs and flow. This algorithm is popularly known for Artificial Intelligence (AI) implementation as shown in Figure 1.

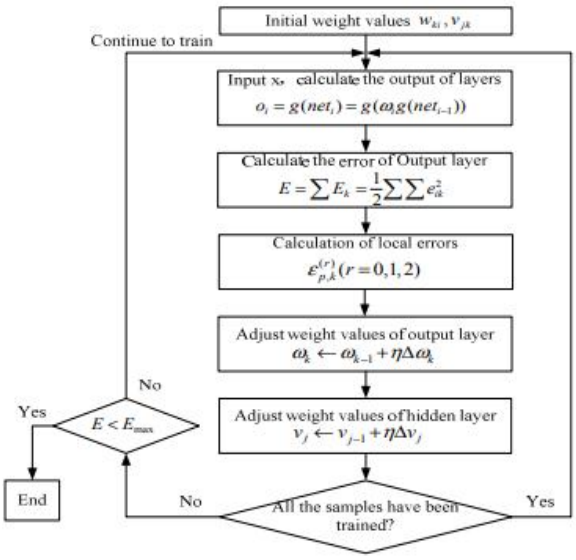


Figure 1: Flowchart of Neural Network Algorithm [14]

2.2 Image Processing

Image processing is the manipulation of images to be process and produced the desire output [15][16]. The image processing approach can be performed using image acquisition, image pre-processing and image analysis.

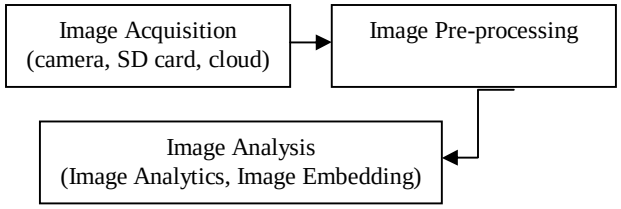


Figure 2: Image Processing Techniques

Figure 2 shows the proposed image processing techniques in classification of coffee plants nutritional deficiencies. The images of leaves were captured and save in a storage medium for retrieval and manipulation in a SD card or cloud. In image pre-processing, the images were converted from RGB to grayscale values. The images were analyzed using the input array or grayscale values. The image embedding from image analytics was utilized in image analysis.

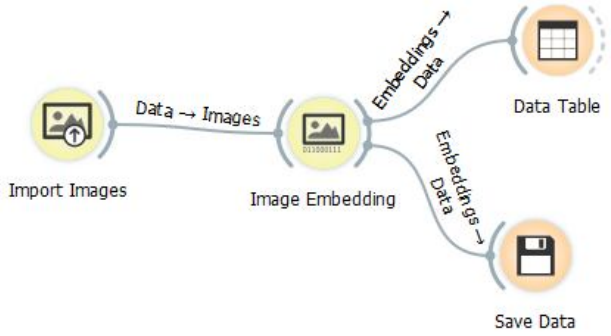


Figure 3: Image Processing Analytical Framework

The imported images composed of coffee leaves will undergone image embedding. In image embedding, the images were connected to the server. The embedders are SqueezeNet (local), Inception v3, CGG-16, VGG-19, Painters, Deeploc and openface [17].

2.3 Data Preparation

Two nutritional deficiencies named Phosphorus (P) and Potassium (K) of Arabica coffee were found during the farm visit in Cavite, Philippines. The leaves were manually identified together with an agriculturist.

Step 1: The leaves were manually identified by two agriculturists during the coffee farm visits.

Step 2: The leaves were captured using a Nikon Digital SLR Camera D5300 with single lens reflex digital camera.

Step 3: The images were saved in SD card and cloud as storage.

(a) (b)

Figure 4: Phosphorus deficiency (a) and Potassium deficiency (b) [18]

Figure 4 (a) and (b) shows the nutrient deficiencies in Phosphorus (P) and Potassium (K). Phosphorus deficiency has symptoms in plant growth and produced mottled appearance while Potassium (K) deficiency has scorch tip and necrosis within the leaves.

2.3 Proposed Method

The proposed method is presented using the analytical framework. The converted values of images into vector array with 1001 features each were trained using the classifiers, SVM, Random Forest, KNN and ANN. The evaluation of results is shown using the confusion matrix, ROC Analysis, Scatter Plot and Distributions.

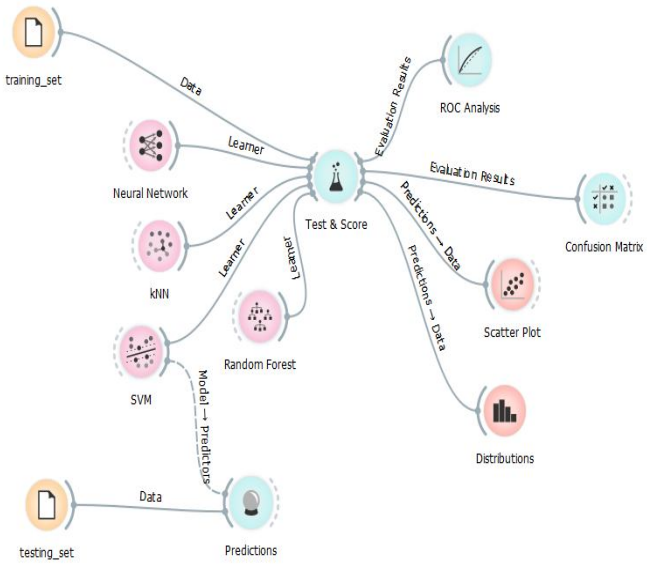


Figure 5: Analytical Framework of the Proposed Method

Receiver Operating Characteristics (ROC) is a plot used to present trade off among classifiers [19]. Scatter plot is used to present data points within x and y axis to show how variables affect each other [20]. In addition, the testing set utilized the SVM as best fit classifier to perform the prediction of the classifier.

2.4 Training of the Prediction Model

The models were trained using the Waikato Environment for Knowledge Analysis (WEKA) and Orange Visual Programming. The training data set is composed of 2405 instances with 1001 features and the testing data set is composed of 101 instances with 1001 features using 10-fold cross validation to avoid overfitting.

2.5 Prediction and Validation

Validation is used to determine the accuracy of the proposed model. To validate a classifier, precision, recall, f-measure and interrater reliability can be used [9]. In addition, to measure the performance evaluation of a classifier, confusion matrix can be utilized [21]. As confusion matrix measures classification in machine learning with two or more classes [22]. It is also a table that shows the performance of the classifiers [23]. Precision is the ratio of relevant instances in the retrieved instances that are referred to as a positive value where tp is truly positive and fp is a false-positive as shown in (1).

$$Precision = tp/(tp/fp) \quad (1)$$

Recall is defined as the true positive rate where p is true positive and fn is false-negative as shown in (2).

$$Recall = tp/(tp/fn) \quad (2)$$

The weighted average of Precision and Recall is called F-Measure as shown in (3).

$$F\ Score = 2*(Recall * Precision) / (Recall + Precision) \quad (3)$$

Cohen's Kappa statistic is one among the list of Interrater Reliability within raters. Po is the relative observed agreement among raters, P_e is the hypothetical probability of chance agreement and K is the Kappa value[24].

$$K=(P_o-Pe)/1-Pe \quad (4)$$

Table 1: Kappa Value and Level of Agreement

Value of Kappa	Level of Agreement
0.00-0.20	None
0.21-0.39	Weak
0.40-0.59	Minimal
0.60-0.79	Moderate
0.80-0.90	Strong
Above 0.90	Almost Perfect

Table 1 shows the Kappa value and level of agreement. The value of kappa from 0.00-0.20 is none, 0.21-0.39 is weak, 0.40-0.59 is minimal, 0.60-0.79 is moderate, 0.80-0.90 is strong and above 0.90 is almost perfect.

3. RESULTS AND DISCUSSION

This section discusses the results of the study conducted. Two classes were analyzed in four different classification models. Table 2 shows the result of evaluation in the classification models.

Table 2: Evaluation Results of the Algorithms

Algorithm	AUC	CA	F1	Precision	Recall
KNN	0.996	0.975	0.975	0.975	0.975
SVM	1.000	0.983	0.983	0.983	0.983
Random Forest	0.990	0.954	0.954	0.954	0.954
NN	0.988	0.983	0.983	0.983	0.983

Support Vector Machine (SVM) has the highest AUC (1.000), CA (0.983), F1 (0.983), Precision (0.983) and Recall (0.983). Second is KNN with AUC (0.996), CA (0.975), F1 (0.975), Precision (0.975) and Recall (0.975). Third is Random Forest with AUC (0.990), CA (0.954), F1 (0.954), Precision (0.954), and Recall (0.954). Neural Network also return high AUC (0.988), CA (0.983), F1 (0.983), Precision (0.983) and Recall (0.983).

Table 3: Detailed Accuracy by Class

Class	TP Rate	FP Rate	Precision	Recall	F measure
Phosphorus (P)	0.985	0.010	0.992	0.985	0.989
Potassium (K)	0.990	0.015	0.981	0.990	0.986
	0.987	0.012	0.987	0.987	0.987

Table 3 shows the detailed accuracy by class. Two classes were determined as Phosphorus (P) and Potassium (K). Phosphorus has TP Rate (0.985), FP Rate (0.010), Precision (0.992), Recall (0.985) and F-Measure (0.989). Potassium has TP Rate (0.990), FP Rate (0.015), Precision (0.981), Recall (0.990) and F-Measure (0.986).

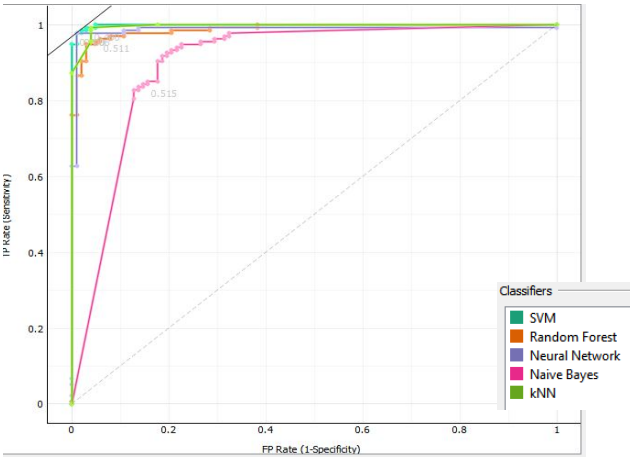


Figure 6: ROC Analysis in Phosphorus (P)

ROC analysis was used to interpret the results in Phosphorus and Potassium as shown in figures 6 and 7 respectively. Five classifiers were presented to show the FP Rate (1-Specificity).

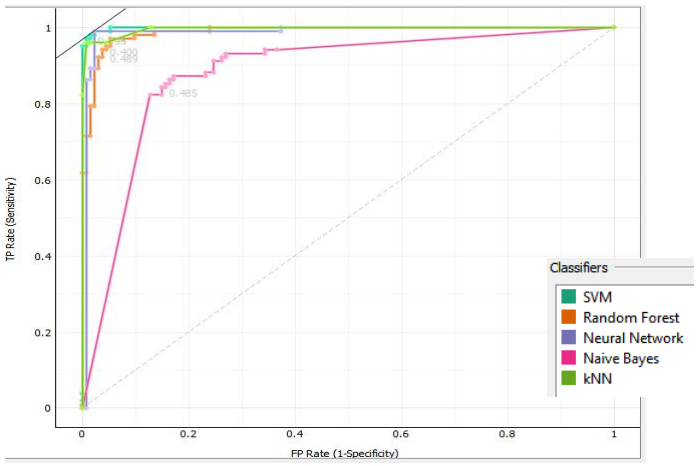


Figure 7: ROC Analysis in Potassium (K)

These classifiers were SVM, Random Forest, Neural Network, Naïve Bayes and KNN. Both ROC analysis shows high FP Rate for SVM since two classes were used for comparison.

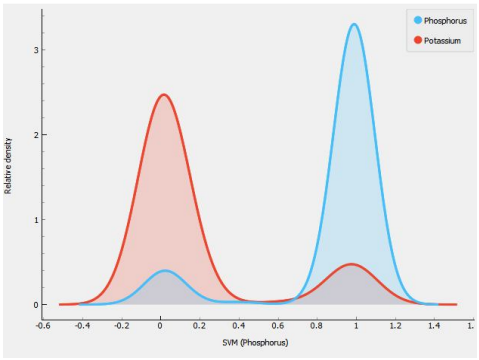


Figure 8: Distribution of SVM for Phosphorus

Figure 8 shows the distribution of SVM for Phosphorus. Phosphorus has larger number compared to Potassium. Likewise, the relative density also shows the higher values.

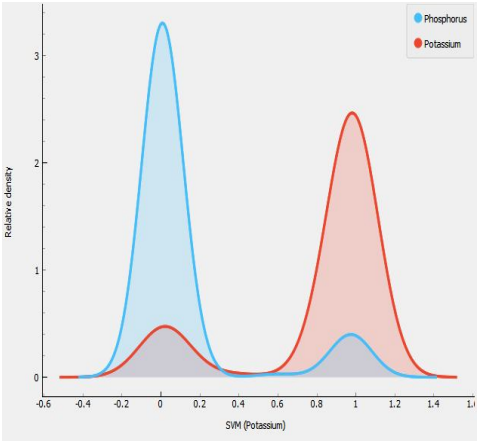


Figure 9: Distribution of SVM for Potassium

On the other hand, the distribution of SVM for Potassium also presents higher number and relative density of Phosphorus.

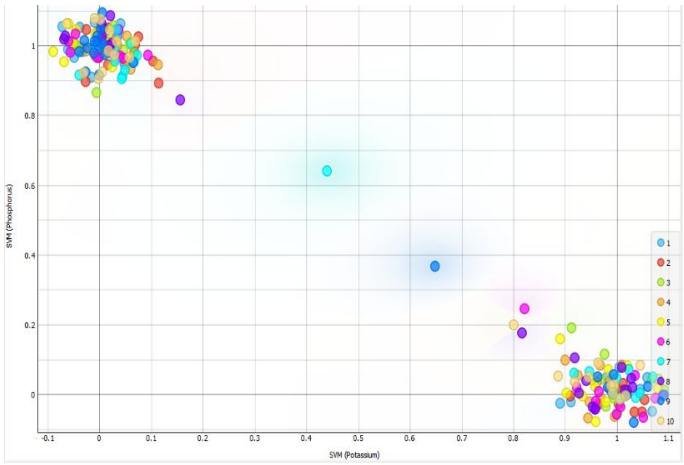


Figure 10: Scatter Plot of SVM for P and K classes

The scatter plot was used to present the SVM data for P and K classes. In which, a 10-fold cross validation was applied to avoid over fitting.

Table 4: Kappa, CCI and ICI Values of SVM

Measure	Value
Kappa	0.9743
Correctly Classified Instances (CCI)	98.7342%
Incorrectly Classified Instances (ICI)	1.2658%

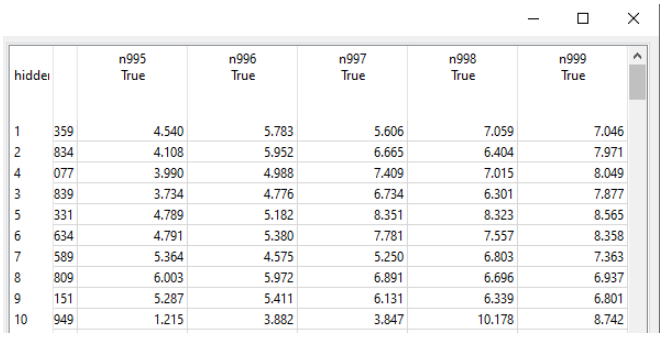
Table 4 shows the Kappa, CCI and ICI values for SVM. The Correctly Classified Instances (CCI) is 98.7342% and the Incorrectly Classified Instances (ICI) is 1.2658%. The Kappa value is 0.9743 which implies that the model is almost perfect in predicting the nutritional deficiencies in arabica coffee using two classes.

Table 5: Confusion Matrix

A	B	←Classified as
1303	2	A-Phosphorus
1	1099	B-Potassium

The confusion matrix out of instances was presented in Table 5. Phosphorus was classified in 1303 images and 2 for Potassium. Likewise, Potassium was classified 1099 and 1 for Phosphorus. Providing a high accuracy for two classes using the SVM classifier.

On the other hand, the image processing techniques used in the images were image acquisition, image pre-processing and image analysis. During the image acquisition, a camera was used to capture the images and saved in a storage medium. In the pre-processing, images were converted from RGB values to grayscale values. In image analytics, the images were connected to a server called the SqueezeNet (local) for getting the grayscale values. The converted grayscale values can be seen in a data table composed of several different number of values. Those values were saved in a .csv file and was processed using WEKA to determine the Kappa, CCI and ICI.



	hidde	n995 True	n996 True	n997 True	n998 True	n999 True
1	359	4.540	5.783	5.606	7.059	7.046
2	834	4.108	5.952	6.665	6.404	7.971
4	077	3.990	4.988	7.409	7.015	8.049
3	839	3.734	4.776	6.734	6.301	7.877
5	331	4.789	5.182	8.351	8.323	8.565
6	634	4.791	5.380	7.781	7.557	8.358
7	589	5.364	4.575	5.250	6.803	7.363
8	809	6.003	5.972	6.891	6.696	6.937
9	151	5.287	5.411	6.131	6.339	6.801
10	949	1.215	3.882	3.847	10.178	8.742

Figure 11: Sample data table of images in image embedding

Figure 11 shows the sample data table of images after image embedding. From one single image, each image has grayscale values in each pixel. In here, from one image, 1001 values were found in the image. The training and testing dataset were evaluated using the four classifiers in which SVM returned the highest accuracy.

4. CONCLUSION

This study was conducted to utilized image processing techniques and data mining algorithms in classifying coffee plants. During the data gathering, Phosphorus (P) and Potassium (K) are the nutritional deficiencies occurred and used in classifications. The classifiers used were KNN, SVM, Random Forest and ANN. The image processing techniques conducted were image acquisition, image pre-processing and image analysis. The images were captured, converted from RGB to grayscale values and image embedding was performed to get the data table or input vector. Among the four classifiers, SVM has the highest almost perfect Kappa value and implies that it is an appropriate model for coffee plants classification with two classes.

ACKNOWLEDGEMENT

The authors would like to acknowledge the help and assistance of the National Coffee Research, Development and Extension Center (NCRDEC) in Cavite State University, the Municipal Agricultural Office of Amadeo, Cavite, AMA University-Quezon City and Commission on Higher Education (CHED).

REFERENCES

1. Fellemedia, **Moyee Coffee _ Speciality Ethiopian Coffee in Ireland _ FairChain,** 2020. .
2. V. Aristizábal-Marulanda, Y. Chacón-Perez, and C. A. Cardona Alzate, ***The biorefinery concept for the industrial valorization of coffee processing by-products.*** Elsevier Inc., 2017. https://doi.org/10.1016/B978-0-12-811290-8.00003-7
3. Department of Agriculture Philippines, **Code of Good Agricultural Practices for Coffee**, *Philipp. Natl.*

Stand., no. 632, pp. 1–31, 2015.
4. R. Leaño, **Types of Coffee that Grow in the Philippines | Philippine Primer**. 2017.
5. **Arabica Coffee FMR _ Philippine Rural Development Project.** 2019
6. K. Lauren, **The world's most popular coffee species are going extinct. And scientists say we are to blame**, 2019.
7. Department of Trade & Industry and Department of Agriculture, **2017-2022 Philippine Coffee Industry Roadmap**, p. 58, 2017.
8. **Nutritional Problems - Welcome Coffee Growers!** . 2019
9. H. D. Gadade, **10_Machine Learning Approach towards Tomato Leaf Disease Classification,** no. 1, pp. 3–8, 2020. https://doi.org/10.30534/ijatcse/2020/67912020
10. Rajendra Akerar, **Artificial Intelligence for Business**, *Am. J. Roentgenol.*, pp. 1–30, 2019.
11. O. R. Devi, **International Journal of Advanced Trends in Computer Science and Engineering** Available Online at http://www.warse.org/ijatcse/static/pdf/file/ijatcse02422015.pdf," vol. 4, no. 2, pp. 15–21, 2015.
12. A. C. Braun, U. Weidner, and S. Hinz, **Vector Machines for Hyperspectral Classification - a Comparison,** *2011 3rd Work. Hyperspectral Image Signal Process. Evol. Remote Sens.*, vol. 2, no. 3, pp. 1–4, 2011.
13. O. Harrison, **Machine Learning Basics with the K-Nearest Neighbors Algorithm,** *Towards Data Science2*. pp. 1–16, 2018.
14. Z. Zeng, L. Zheng, and D. Ling, **Network Algorithm,** vol. 0, no. 3, pp. 1716–1718, 2008.
15. E. A. B. da Silva and G. V. Mendonca, **Digital Image Processing,** *Electr. Eng. Handb.*, pp. 891–910, 2005.
16. X. Wang, **Moving window-based double haar wavelet transform for image processing,** *IEEE Trans. Image Process.*, vol. 15, no. 9, pp. 2771–2779, 2006. https://doi.org/10.1109/TIP.2006.877316
17. **Orange Data Mining - Download.** 2019 .
18. **Coffee Nutritional Deficiencies** — Vikaspedia. 2019.
19. C. S. Specificity, **Plotting and Intrepretating an ROC Curve.** pp. 4–5, 2014.
20. MDH, **Scatter Plot What is a Scatter Plot ?** p. 2014, 2016.
21. J. M. Victoriano, M. Luis, and C. D. Santos, **Predicting Pollution Level Using Random Forest: A Case Study of Marilao River in Bulacan Province, Philippines**, vol. 3, no. 1, pp. 151–162, 2017.
22. Sarang Narkhede, **Understanding Confusion Matrix – Towards Data Science**, *Towards Data Science*. 2018.
23. M. Learning, **Simple guide to confusion matrix terminology**. pp. 1–9, 2014.
24. Stephanie, **Cohen's Kappa Statistic - Statistics How To**. 2014.