Performance Comparison of Naïve Bayes and K-NN Algorithms on Contamination Grading for Abaca Tissue Culture (In vitro)

Rhoderick D. Malangsa and Elmer A. Maravillas
College of Computer Studies
Cebu Institute of Technology - University
Cebu City, Philippines
rhojud@yahoo.com, elmer.maravillas@gmail.com

Abstract— **One of the challenges meet in the agricultural sector is the plant disease which affected food crops, causing significant losses to farmers and its spread has increased dramatically in recent years. The abaca plantation industry in the province of Southern Leyte was greatly affected by the disease and was estimated to suffer about 30 per cent in damages. There have been many studies before utilizing Naïve Bayesian algorithm and K-Nearest Neighbour on disease and contamination prediction especially for crops but the study on *in vitro* abaca specimens has not given enough focused yet. This study is created for a performance comparison contamination grading for abaca tissue culture specimen using Naïve Bayesian classification and K-Nearest Neighbour. In phase one, capturing specimen images thru masking using a camera. Phase two was feature extraction techniques of RGB mean values and binary images to obtain relevant data to be used in phase three and four where specimens were classified as either healthy or contaminated. Lastly, the performance of the two classifiers was evaluated based on the overall accuracy, precision, and recall. The average overall accuracy of the Naïve Bayes was 76%, while for K=3 were 68%, K=7 were 58% Moreover, the values generated by Naïve Bayesian in precision and recall also indicate the very good performance of the classifier. The study indicated that Naive Bayes has good potential for identifying contamination grading accurately than the K-Nearest Neighbour that mainly causes by fungi in abaca tissue culture laboratory.**

Keywords- Classification Algorithm; Contamination Grading; Expert System; Naïve Bayesian; K-Nearest Neighbour;

I. Introduction

Agriculture is one of the most ancient activities of man in which innovation and technology are usually accepted with difficultly, unless real and immediate solutions are found for specific problems or for improving production and quality. Nevertheless, a new approach, of gathering information from the environment, could represent an important step towards high quality and eco-sustainable agriculture [1].

One of the challenges meet in the agricultural sector is the plant disease. Plant disease affect food crops, causing significant losses to farmers and the spread of plant diseases has increased dramatically in recent years. Most plant [2] diseases are caused by fungi, bacteria, and viruses. Fungi are identified primarily from their morphology, with emphasis placed on their reproductive structures. Bacteria are considered more primitive than fungi and generally have simpler life cycles.

A certain disease threatened the abaca industry in the province of southern last 2003 [5]. Eighty percent of the province's abaca plantation particularly in Sogod town was greatly affected while Maasin City was estimated to suffer about 30 percent in damages.

The Philippine government is struggling to eradicate mosaic and bunchy-top diseases infesting 43 percent of abaca farms in Leyte and Southern Leyte due to a limited budget, leaving the situation out of control for years. One of the solutions to address the spread of abaca disease is through abaca tissue culture or called micro propagation. Tissue culture [3][4] is seen as an important technology for developing countries for the production of disease-free, high quality planting material and the rapid production of many uniform plants. In this way, thousands of copies of a plant can be produced in a short time. In this way, thousands of copies of a plant can be produced in a short time.

The micro propagation technique should be efficient and effective on its day to day operation [4]; it should detect early contaminants on its *in vitro* specimens. The present manual visual inspection of the laboratory technician is not enough to cover up all the specimens inside the laboratory. Ironically, the more physical contact to the specimens the more chances of contamination it brings. With the presence of diseases in the field, it is difficult to propagate disease-free planting materials using the conventional method.

This study is created for a contamination grading system for abaca tissue culture specimen using Naïve Bayes and K-Nearest Neighbour classification. The paper sets out to make comparative evaluation of the two mentioned classifiers in classifying contaminants in the tissue culture laboratory in terms of over-all accuracy, precision, and recall in vitro specimens. By utilizing Naïve Bayesian and K-NN classification algorithm in tissue culture laboratory it will increase the accuracy rate of early prediction of contaminated specimen preventing the infected specimen from infecting other specimen in the laboratory.

Moreover, this study will augment the abaca rehabilitation effort of Department of Agriculture thru the combined initiatives of LGU and SLSU.

There have been many studies before utilizing Naïve Bayesian and K-NN algorithm on disease prediction especially for crops but the study on in vitro specimens has not given enough focused yet.

The study was conducted at the Tissue Culture Laboratory located at Southern Leyte State University – Main Campus, Sogod Southern Leyte. The specimen in this study is focus on In vitro preparation. This project involves the mass production of recommended and disease-free abaca planting materials through tissue culture using the shoot tip technique

Figure 1. Map of Southern Leyte

II. Related Studies

In fruit grading study, features are extracted from the fruit to classify fruit disease.[6] Extracted features like color, texture and shape of the fruit integrated with the Naïve Bayesian classifier built on histogram matching and region based approaches.

Bayesian can be applied in expert system for human disease like liver disease classification [12] and diabetes prediction [7] while there is a possibility of existing attribute values and new samples and can be measures in terms of probability class labels. Instead of classifying the traditional testing data with training data, the initial training data to the optimal process is forwarded, to extract the optimal data set; on that optimal dataset Naive Bayesian classifier is applied.

Naive Bayes Classifier was able to classify the agricultural land soil data resulted to 100% classification of instances [8]. Others [9], [10] have demonstrated the value of image processing in inspecting and grading the quality of agricultural and food products.

While an automated system for the disease detection and grading in pomegranate plant was proposed [10]. The techniques used includes color segmentation based on linear discriminant analysis, contour curvature analysis and a thinning process, which involves iterating until the stem becomes a skeleton.

Feature extraction methods and classification techniques are applied systematically in the attempt to solve the problem in plant disease classification [11]. The classification algorithm has feasibly identified the two diseases in the banana.

On the other hand, KNN algorithm is utilized in banking sector, because of it's a very competitive environment. A study tried to tackle the question of default prediction of short term loans for a Tunisian commercial bank. A database of 924 credit records of Tunisian firms granted by a Tunisian commercial bank was used. The K-Nearest Neighbour classifier algorithm was conducted and the results indicate that the best information set is relating to accrual and cash-flow and the good classification rate is in order of 88.63 % (for k=3) [13].

The versatility of the KNN algorithm can also be manifested by the study in the field of criminal investigation. This paper applies KNN to help criminological investigators in identifying the glass type and checks if integrating KNN with another classifier using voting can enhance its accuracy in identifying the glass type [14].

The K Nearest Neighbour (KNN) is applied to determine the emotion based on this features. The research shows that various features namely prosodic and spectral have been used for emotion recognition from speech. The database used for recognition purpose was developed on Marathi language using 100 speakers. [15]The extracted features pitch and formants. Angry, stress, admiration, teasing and shocking have been recognized on the basis of features energy and formants. The result for formants was about 100% which is comparatively better than that of energy which was 80% of accuracy.

A combination of KNN and genetic algorithm for effective classification for heart disease [16]. Experimental results carried out on 7 data sets show that the approach is a competitive method for classification. The prediction model could augment the doctors in efficient heart disease diagnosis process with fewer attributes.

III. Methodology and Procedure

Prototyping methodology is a software development process which allows developers to create portions of the solution to demonstrate functionality and make needed refinements before developing the final product.

The developed system was made possible with the utilization of raspberry pi microcontroller, OpenCV, Zigbee wireless data transmission, and 12 megapixel camera.

A. Capture Specimen

A camera of 12 megapixels was used to capture both healthy and diseased images. In order to capture clear images

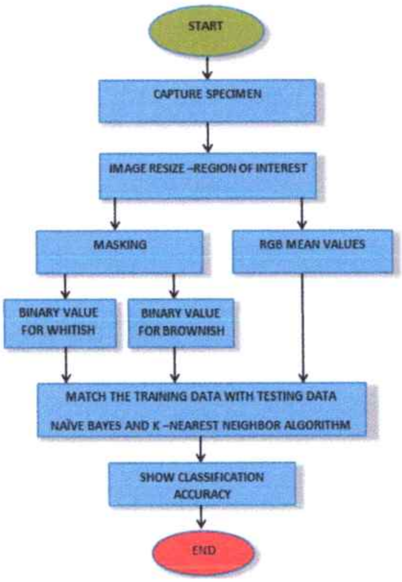

START
CAPTURE SPECIMEN
IMAGE RESIZE –REGION OF INTEREST
MASKING
RGB MEAN VALUES
BINARY VALUE FOR WHITISH
BINARY VALUE FOR BROWNISH
MATCH THE TRAINING DATA WITH TESTING DATA
NAÏVE BAYES AND K –NEAREST NEIGHBOR ALGORITHM
SHOW CLASSIFICATION ACCURACY
END

Figure 2. The architectural lay-out of the developed system

with descriptive details, the camera was kept in both horizontal and vertical resolution of 72dpi (dots per inch). The distance of the camera from the specimen is 32 cm. The elevation of the camera is 23 cm. The angle of the camera is 95^0.

B. Feature extraction

RGB color space (Red, Green, and Blue) is the combination of the primary colors of red, green, and blue, which is used by a computer monitor or television [14].The color results from a combination of three colors and each - each have a value of 8 bits of red, 8 bits of green, and 8 bits of blue. Mixture of the three primary colors balanced with porpoise will produce shades of gray. If three fully saturated colors, it will produce white. Then an RGB mean values was displayed and recorded in the database.

The captured images from video streams may contain many objects especially in the background and working with such images leads to incorrect results. These images were cropped however, cropped images then had a white background with pixel values of 255 and working with the whole image also brings inappropriate results too. To avoid this challenge a mask was applied onto the image in order to obtain the useful segment.

Masking means setting the pixel value in an image to zero or some other background value. In this step, identification of mostly whitish and brownies in the specimens. After that, based on specified threshold value that is computed for these pixels. The whitish and brownies components of the pixel intensities are set to zero if it is less than the pre-computed threshold value. Then a red, green and blue component of the pixel is assigned to a value of zero by mapping of RGB components.

C. Naïve Bayesian Classification

Naïve Bayesian Classification is commonly known as a statistical means classifier. Based on Bayes' Theorem, and uses probabilistic analysis for effective classification. [17] It give more accurate results in less computation time when applied to the large data sets consisting of hundreds of images. The formula for Naïve Bayes classifier is: P (H | E) = P (E | H) x P(H) / P(E). The basic idea of Bayes rule is that outcome of a hypothesis or an event (H) can be predicted based on some evidences (E) that can be observed.

The advantages of Naive Bayes are [18]:

• It uses a very intuitive technique. Bayes classifiers, unlike neural networks, do not have several free parameters that must be set. This greatly simplifies the design process.

• Since the classifier returns probabilities, it is simpler to apply these results to a wide variety of tasks than if an arbitrary scale was used.

• It does not require large amounts of data before learning can begin.

• Naive Bayes classifiers are computationally fast when making decisions

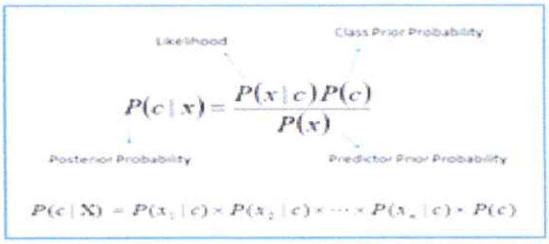

Likelihood
Class Prior Probability
$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$
Posterior Probability
Predictor Prior Probability
$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$

Figure 3. The naïve Bayesian formula

Where;

- P(c|x) is the posterior probability of class (target) given the predictor (attributes).
- P(c) is the prior probability of class
- P(x|c) is the likelihood which is the probability given class
- P(x) is the prior probability of predictor, equation 1.

$$P(c|x) = P(x_1|c) \text{ x } P(X_2|c) \text{ x} \ldots \text{x } P(x_n|c) \text{ x } P(c) \quad (1)$$

D. K-Nearest Neighbour Classification

The KNN algorithm is one of the most famous classification algorithms used for predicting the class of a record or (sample) with unspecified class based on the class of its neighbour records [19]. The algorithm is made of three steps as follows [20]:

1. Calculating the distance of input record from all training records

2. Arranging training records based on the distance and selection of K-nearest neighbour

3. Using the class which owns the majority among the k-nearest neighbours (this method considers the class as the class of input record which is observed more than all the other classes among the K-nearest neighbours).

The classifier assumes the distance of records from each other as a criterion for their nearness and selects the most similar records. There are numerous methods to compute the distance such as the function of Euclidean distance, Manhattan, etc., among which the function of Euclidean distance is one of the most common ones.

Given an mx-by-n data matrix X, which is treated as mx (1-by-n) row vectors x1, x2, ..., xmx, and my-by-n data matrix Y, which is treated as my(1-by-n) row vectors y1, y2,...,ymy, the various distances between the vectors xs and yt are defined as follows:

1. Euclidean Distance

The Euclidean distance is a measure to find distance between two points, defined by equation 2.

$$dst2 = (xs - \quad)(xs - yt\,)' \tag{2}$$

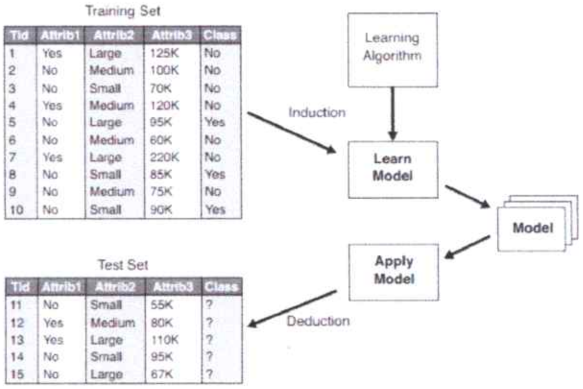
The Euclidean distance is a special case of the Minkowski metric, where p = 2.

2. Standardized Euclidean Distance

The standardized Euclidean distance is used to optimize the problem of finding the distance, defined by equation 3.

$$dst\ 2 = (xs - \quad)-1(xs - yt\,)' \tag{3}$$

where V is the n-by-n diagonal matrix whose jth diagonal element is S(j)2, S is the vector containing the inverse weights [21].



Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Test Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Figure 4. General approach for building as classification model

The choice of k also affects the performance of the k-NN algorithm. This can be determined experimentally. Starting with k=1, we use a test case to estimate the error rate of the classifier. This process is repeated each time by incrementing k to allow for one more neighbours. The K-value that gives the minimum error rate may be selected. In general, larger the number of training samples is, the larger the value of k will be [19],[21].

5. Training and Testing

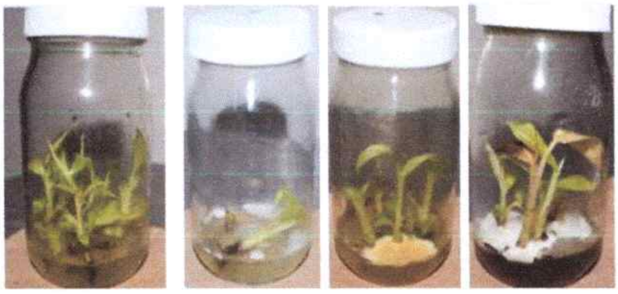
The study was conducted at the Abaca Tissue Culture, at present the laboratory technician manually inspect the in vitro specimens for contamination, by average there are 500 specimens prepared daily.

Contamination in the specimen is caused by fungi and bacteria, the visual appearance of these contamination is the whitish cottony appearance in the medium caused by fungi, while a black dots caused by bacteria. Frequently encountered bacterial and fungal contamination especially in laboratories of micro propagation posed a considerable problem. Fungus may arrive with an explant, or airborne, or enter a culture.

The principal microbial contaminants frequently reported in plant in vitro cultures are bacteria and fungi [22],[23]

TABLE 1. Characterization and identification of fungal contaminants of tissue-cultured abaca (musa textiles nee)

Contaminants	Cultural characteristics	Morphological characteristics
Aspergillus sp	Colonies are flat, circle, Filamentous, velvety, wolly or cottony texture. Colony color is gray to green at center with a white border. The reverse is yellow to Pale yellow.	Conidiophores bear heads, long and hyaline that terminates in bulbous heads while conidia are globose to subglobose and usually yellowish green and dark brown
Chrysosporium sp.	Colonies are semi elevated, circle, fairly rapid grower, smooth. Colony color is white to off-white. The reverse is white to off-white color	Produced septate, hyaline hyphae. Conidia often appeared to be minimally differentiated from the hyphae and may appear to form directly on the hyphae. Conidia more often formed at the ends of simple or branched conidiophores of varying lengths. Conidiophores were ramified, forming tree-like structures.



a.) b.) c.) d.)

a.) a healthy specimen, b). moderately contaminated, c) critically contaminated, and d.) critically contaminated.

During the training phase, there were 100 specimen of which divided into. 40 for healthy specimens, 30 moderately contaminated, and 30 for heavily contaminated specimens.

For testing phase, there were new 50 specimens tested, in this phase the decision of the classifier and the decision of the laboratory technician were evaluated in actual contamination grading.

Figure 5. Preparation of abaca tissue culture specimen (in vitro)

Figure 6. The specimen inside the tissue culture laboratory

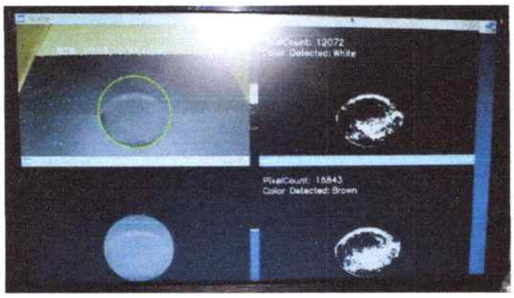


Figure 7. The monitor showing the feature extraction during the training phase

Figure 8. The system inside the abaca tissue culture laboratory

IV. Results and Discussion

A confusion matrix [24] contains information about actual and predicted classifications done by a classification system. It shows the number of correct and incorrect predictions made by the classification model compared to the actual outcomes (target value) in the data.

Performance of such systems is commonly evaluated using the data in the matrix.

Accuracy is calculated as the sum of correct classification divided by the total number of classification (the diagonal) in equation 4.

$$AC = \frac{a+d}{a+b+c+d} \tag{4}$$

Precision is the proportion of the predicted positive cases that were correct, as calculated using the equation in equation 5;

$$\text{Precision} = TP_A / (TP_A + FP_A) \tag{5}$$

Recall or true positive rate (TP) is the proportion of positive cases that were correctly identified, as calculated using the equation in equation 6;

$$\text{Recall} = TP_A / (TP_A + FN_A \tag{6}$$

Result for Binary Classification

In this study, the binary classification consisted of; a.) Healthy Specimen (Contaminated Free) and b.) Contaminated specimen. The Naïve Bayes Algorithm, KNN (K=3), and KNN (K=7) were used and tested.

TABLE 2. The Result of the Performance of the Classifier in Binary Class

	Naive Bayes	K=3	K=7
Over-all Accuracy	92%	72%	62%
Precision Healthy	92%	54%	33%
Precision Contaminated	91%	76%	71%
Recall Healthy	80%	40%	26%
Recall Contaminated	97%	85%	77%

Result for Multi Class Classification

In this study, the multi-class classification consisted of ; a.) Healthy Specimen (Contaminated Free) , b.) Moderately Contaminated Specimen, and c.) Heavily Contaminated Specimen. The Naïve Bayes Algorithm, KNN (K=3), and KNN(K=7) were used and tested.

TABLE 3. The Result of the Performance of the Classifier in Multi Class Classification

	Naive Bayes	K=3	K=7
Over-all Accuracy	60%	64%	54%
Precision Healthy	92%	54%	33%
Precision Moderately Contaminated	33%	27%	38%
Precision Heavily Contaminated	56%	68%	66%
Recall Healthy	80%	40%	27%
Recall Moderately Contaminated	33%	45%	27%
Recall Heavily Contaminated	61%	88%	83%

It can be seen in Table 2. The Naïve Bayes has achieved the highest over-all accuracy of **92%** among the classifiers.

Followed by K=3 or KNN with 3 neighbors with the over-all accuracy of **72%** and the last is K=7 with over-all accuracy of **62%**. It can also be noted that on the same table, the percentage in precision and recall is also the highest in Naïve Bayes.

Still on Table 2, in Naïve Bayes the Healthy class, the precision is 92% and recall is 80%. This means that for precision, out of the times Healthy class was predicted, 92% of the time the system was in fact correct. And for recall, it means that out of all the times Healthy class should have been predicted by 80% of the labels were correctly predicted.

Table 3 shows the Result of the Performance of the Classifier in Multi Class Classification, this time the K=3 has the highest over-all accuracy of **64%**, followed by Naïve Bayes with **60%** and the last is K=7 with the percentage overall accuracy of **54%.** In the table also can be seen the recall of Heavily Contaminated class under K=3 which is **88%** however the precision is only **68%.**

By average, the over-all accuracy of the algorithms in binary classification and multi-class classification; Naïve Bayes is **76%**, K=3 is **68%** and K=7 is **58%**.

V. CONCLUSION

With a average over-all accuracy of 76% of Naïve Bayes classification algorithms in contamination grading, this research has proved that there is a consistent and more accurate way of detecting contamination in the tissue culture specimens rather than the naked eye inspections. Moreover, the percentage generated in precision and recall also indicate the very good performance of the three chosen classifier. The study indicated that Naive Bayes has good potential for identifying contamination grading accurately that mainly causes by fungi in abaca tissue culture laboratory over the K-Nearest Neighbors (K=3 & K=7). Abaca tissue culture in vitro is a huge field and the factors that affects the contamination is one challenging research field. Though some factors needs to be considered during the training phase like the distance of the camera to the specimens, the amount of light, the angle of the camera, and memory of the microcontroller. The system cannot match the precision and accuracy of the human eye, but the speed and the cost at which they work can be easily be overcome.

REFERENCES

[1] Di Palma, D., Bencini, L., Collodi, G., Manes, G., Chiti, F., Fantacci, R., "Distributed monitoring systems for agriculture based on wireless sensor network technology", Department of Electronics and Telecommunications, University of Florence Via di Santa Marta 3, 50019 Florence Italy

[2] S. B. Dhaygude and N. P. Kumbhar (2013), "Agricultural plant leaf disease detection using image processing" International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, vol. 2, no. 1, pp.599-602 Available at www.ijareeie.com

[3] S. Thiart (2003), "Manipulation of growth by using tissue culture techniques" Combined Proceedings International Plant Propagators' Society, vol. 53, no. 66.

[4] G. E., Pedroso, "Module 5: Tissue Culture of Banana and Abaca", Western Mindanao State University, Zamboanga City 2009

[5] Erna S. Gorne (30 October 2006). "Bunchy top virus in Southern Leyte scales down abaca production". PIA Information Services - Philippine Information Agency. Archived from the original on 5 May 2007.

[6] U. Solanki, U.K. Jaliya, and D.G. Thakore (2015), " A survey on detection of disease and fruit grading', *International Journal of Innovative and Emerging Research in Engineering,* vol. 2, no. 2, pp. 109-114 Available: http://www.ijiere.com

[7] A. Ambica, S. Gandi, and A. Kothalanka (October 2013), "An efficient expert system for diabetes by naïve Bayesian classifier", *International Journal of Engineering Trends and Technology,* vol. 4, no. 10. pp. 4634-4639 Available: http://www.ijettjournal.org

[8] P. Bhargani and S. Jyotchi (August 2009),"Applying naïve Bayes data mining technique for classification of agricultural land soils". *International Journal of Computer Science and Network Security*, vol. 9, no. 8, pp. 117-122 Available: http://www.ijcsns.org

[9] N.V.G. and H.K.S. (May 2010), "Quality inspection and grading of agricultural and food products by computer vision", *International Journal of Computer Applications*, vol.2, no. 1

[10] S, Sannakki, V. Rajpurohit, V. Nargubnd, and R. Arun Kumar.(2011), "Hybrid Intelligent system for automated pomegranate disease detection and grading" , *International Journal of Machine Intelligence* , vol. 3, no. 2, , pp-36-44

[11] G. Owomugisha, J.A. Quinn, and E. Mwebaze (2014), "Automated vision-based diagnosis of banana bacterial wilt disease and black sigatoka disease', *International Proceedings of the 1st International Conference on the use of Mobile in ICT in Africa* 2014. Pp. 1-5, December 9-10, 2014, Stellenbosch, South Africa

[12] S.Dhanodharam (May 2014), " Liver disease prediction using Bayesian classification", Proceeding on 4th National Conference on Advance Computing, Applications and Technologies, special issue, pp. 1-3

[13] A. K. Abdelmoula (2015). "Bank credit risk analysis with k-nearest neighbour classifier: Case of Tunisian banks, Accounting and Management Information Systems, vol. 14, no. 1, pp. 79-106

[14] M. S. Aldayel "K-nearest neighbor classification for glass identification problem"

[15] R. P. Gadhe, R. R. Deshmukh, and V. B. Waghmare (2015), "KNN based emotion recognition system for isolated Marathi speech" International Journal of Computer Science Engineering (IJCSE), vol. 4, no.4.

[16] M.Akhil jabbar, B.L Deekshatulua, and Priti Chandra b, "Classification of heart disease using k- nearest neighbor and genetic algorithm ".International Conference on Computational Intelligence: Modeling Techniques and Applications. Available online at www.sciencedirect.com

[17] R. Rulaningtyas, A. B. Suksmono, L.R Mengko, and P. Saptawati (December 2012), "Color segmentation using Bayesian method of tuberculosis bacteria images in ziehl-neelsen sputum smear", *Conference Paper* Available: https://www.researchgate.net/publication/259636216

[18] M. K. Stern, J. E. Beck, and B. P. Woolf, "Naïve Bayes Classifiers for User Modeling," Available: http://citeseerx.ist.psu.edu

[19] AH. Wahbeh, QA. Al-Radaideh, MN. Al-Kabiand, EM. Al-Shawakfa (2011), "A comparison study between data mining tools over some classification methods", International Journal Of Advanced Computer Science And Applications. vol. 35, pp. 18-26.

[20] M. Kuhkan (2016) , "A Method to improve the accuracy of k-nearest neighbor algorithm" International Journal of Computer Engineering and Information Technology, vol. 8, no. 6, June 2016, pp. 90–95 Available online at: www.ijceit.org

[21] M. Sharma and S. K. Sharma (2013), "Generalized k-nearest neighbour algorithm- a predicting tool", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, issue 11. Available online at: www.ijarcsse.com

[22] R. A., Bajet," Epidemiology and Integrated Management of Abaca Bunchy Top in the Philippines". Department of Plant Pathology, University of the Philippines, Los Baños, Philippines.

[23] J.S. Cobrado and A.M. Fernandez (2016), "Common fungi contamination affecting tissue-cultured abaca (musa textiles nee) during initial stage of micro propagation", Asian Research Journal of Agriculture 1(2): 1-7 Available: http://www.journalrepository.org

[24] R. Kohavi and F. Provost (1998)," Special issue on applications of machine learning and the knowledge discovery process", Machine Learning, Kluwer Academic Publishers, Boston, Manufactured in The Netherlands 30, 271-274 Available: http:// robotics.stanford.edu