Precision Agriculture with Random Forests: A Robust Approach to Guava Fruit Disease Detection

Sravanth Kumar Tankari[1], Namrta Tanwar[1], Saikumar Swarnapudi[1], Jhansi Kottali[2], Lalitha Donavalli[2], Myskina Chowdary Raparla[2]
[1,2] *Dept. Computer Science & Engg. Chandigarh University Mohali, India*

***Abstract-** One of the main causes of production losses and financial challenges for the worldwide agriculture industry is fruit diseases. This study offers a suggested and validated experimental approach for the detection and classification of apple diseases. The suggested image processing-based method consists of the following primary steps: Using the K-Means clustering technique, segment the images first; then, extract some cutting-edge characteristics from the segmented image; and last, classify the images into one of the classes using a Multi-class Support Vector Machine. Our testing results indicate that the proposed method may significantly improve the automated and precise identification of apple fruit illnesses. The recommended method can achieve up to 93% classification accuracy.*

***Keywords:** -K-Means Clustering; Local Binary Pattern; Multi- class Support Vector Machine; Texture Classification;*

I. INTRODUCTION

Experts use their unaided eyes to observe fruit diseases in order to discover and identify them using the traditional method. Because specialists are often located in remote areas, seeking their advice can be costly and time-consuming in developing nations. To automatically identify disease symptoms on developing fruits as soon as they manifest, automatic fruit disease detection is crucial. Apple fruit diseases can result in significant reductions in yield and quality when the fruit is harvested. It's critical to identify what is being observed in order to determine what preventative measures to take the next year to avoid losses. illnesses of twigs, leaves, and branches are caused by certain illnesses that also spread to other parts of the tree. Apple scab, apple rot, and other frequent illnesses of apple fruits illustrates the apple blotch . The corky patches on apples are gray or brown. Apple rot infections result in round, brown or black, somewhat depressed areas that could have a crimson halo surrounding them. A fungus called apple blotch causes dark, uneven, or lobed margins on the fruit's surface.

Machine vision is already used in the business to automate the visual evaluation of apples for color and size. Nonetheless, the inherent variation in skin tone among apple varieties, the wide range of defect kinds, and the stem/calyx presence continue to pose challenges to defect detection.

The majority of apple fault segmentation projects use a basic threshold method ([2], [3]). Apple fecal contamination flaws can be segmented using a globally adaptive threshold approach (a modified version of Otsu's algorithm) as described in [4]. Methods based on classification.

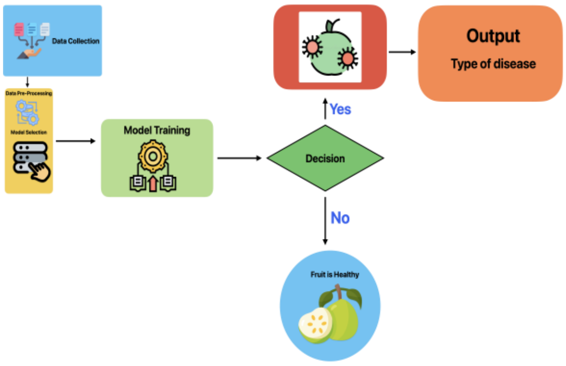


Figure 1. Machine learning process to identify the disease of the fruit

Try dividing pixels using various classification techniques into several classes. The most popular approach among researchers is called Bayesian classification ([5], [6]), in which pixels are categorized as either healthy or defective based on a comparison with a pre-calculated model. The absence of target values in unsupervised classification makes it unhelpful to provide assistance throughout the learning process. [7] employed this method for defect segmentation.

The Local Binary Pattern histogram was utilized by Ojala et al. in [8] to classify textures that were rotation invariant. A straightforward yet incredibly effective operator for producing local picture patterns, local binary pattern has shown outstanding classification results on representative texture databases [9]. Other applications, including face recognition [10], dynamic texture identification [11], and form localization [12], have also adopted the Local Binary Pattern. As the finished modeling of the Local Binary Pattern, a Complete Local Binary Pattern is given in [13].

The importance of employing the clustering technique for disease segmentation and Multi-class Support Vector Machine as a classifier for the automatic detection and classification of fruit illnesses has been suggested and experimentally proven. We have examined three different apple diseases—apple blotch, apple rot, and apple scab—in order to validate the suggested methodology.

II. THE PROPOSED APPROACH

Figure 2 illustrates the steps of the suggested methodology.Efficient picture segmentation is necessary to solve the fruitdisease classification problem, since otherwise, the features of the non-affected zone will outweigh the features of the disease darea. To identify the region of interest, or only the afflicted .

A. Image Segmentation To segment images, the K-Means clustering approach is employed. Four clusters are formed from the images, with one cluster containing the majority of the affected area. In 1967, J. MacQueen created the K-Means clustering algorithm [14]. In our scenario, the objects (pixels) are classified into K number of classes using a collection of features by the k-means clustering techniques. In order to classify the data objects, the sum of squares of their distances from each cluster is minimized. For the K-means clustering in this experiment, the squared Euclidean distance is employed. Algorithm for the K-Means image segmentation –

Step 1: Examine the input image. Step 2: Convert the RGB image to the L*a*b* color space. Step 3: Use K-Means clustering in 'a*b*' space to classify colors. Step 4: Label every pixel in the picture using the K-Means results. Create images in Step 5 that are divided into sections based on color. Step 6: Decide which section has the illness. Since the color information in the L*a*b* color space is contained in only two channels (i.e., the a* and b* components), it results in reduced

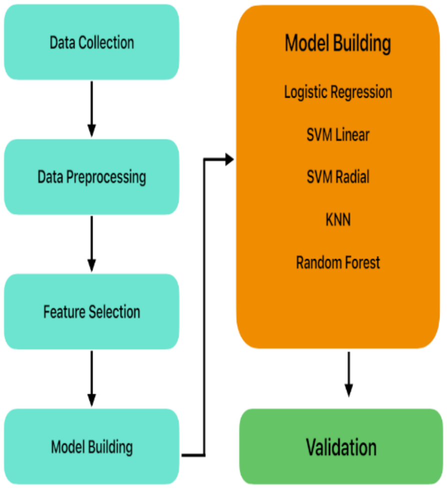


Figure 1 Flowchart depicting the five stages of the machine learning process.

The division of picture processing time into segments. In this experiment, four parts are taken out of the original image. Based on practical data, segmentation outcomes are best achieved when three or four clusters are used.

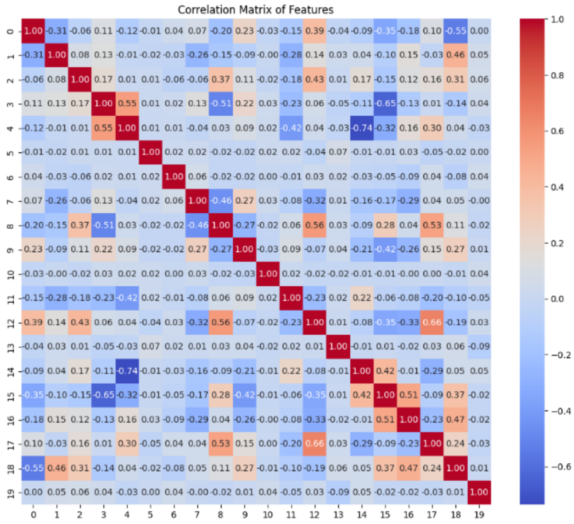


Figure 2 Heat map showing the relationship between the variables.

In our experiments, we emphasized feature selection to reduce dimensionality. This step weeds out unnecessary data while pinpointing the most important attributes that aid in detecting fruit disease. We explored various feature selection methods to enhance both the accuracy and robustness of Guava fruit disease data set.

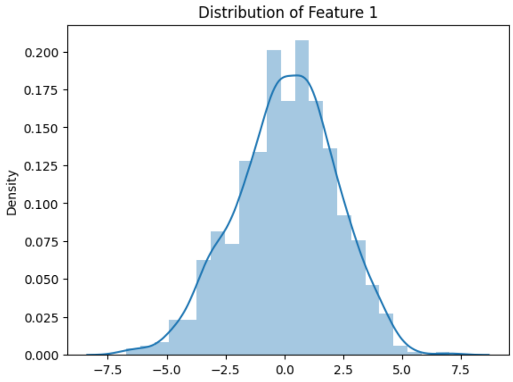


Figure 3 (a) Exploratory data analysis for the extracted features from the dataset

The dataset was split randomly into two parts: training data and trial data. This was done in a ratio of 70 for training and 30% for testing. Since the data has significant imbalances, it’s good to use Python's Sklearn module. It helps in standardizing & normalizing the input features of the dataset.

For the training group, we applied the Standard Scaler function from the Sklearn pre-processing package. This ensures that the data has a mean of zero and a variance of one. Meanwhile, we normalized the test set by using the mean & standard deviation from the training data. This method supports better performance during analysis.

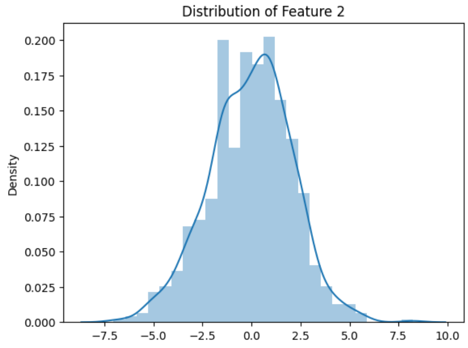


Figure 4 (b) Exploratory data analysis for the extracted features from the dataset

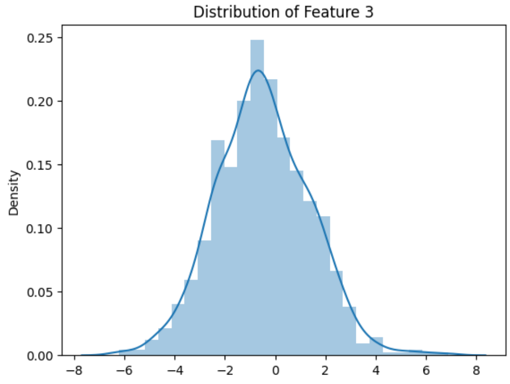


Figure 5 (c) Exploratory data analysis for the extracted features from the dataset

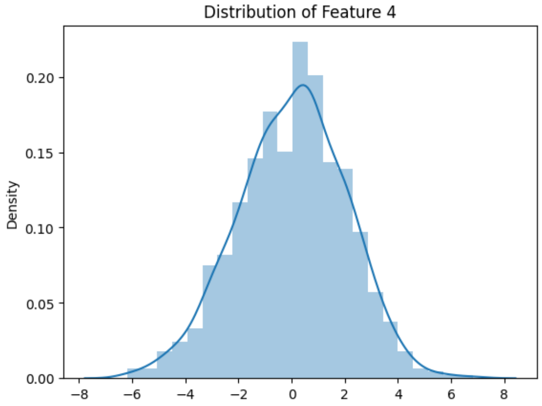


Figure 6 (d) Exploratory data analysis for the extracted features from the dataset

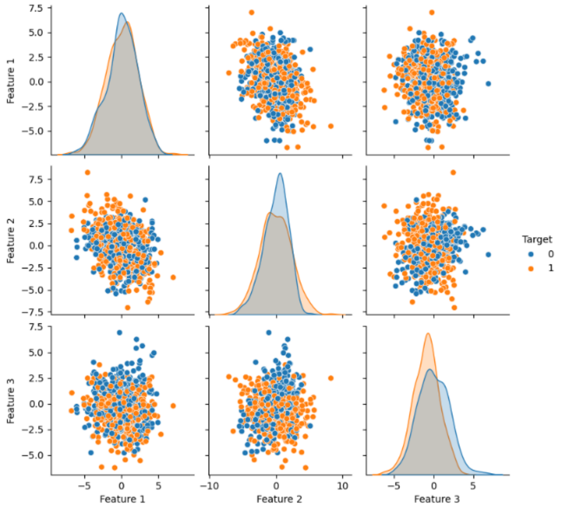


Figure 7 Exploratory data analysis for the extracted features from the dataset

III. EXPERIMENTAL SET UP

In this study, we aim to develop a model capable of detecting diseases in fruit images by leveraging deep learning techniques. The experiment is set up to explore the model’s accuracy in identifying diseases from a dataset of fruit images. The primary steps include dataset collection and preprocessing, model design, training, and evaluation

4.1. Dataset Collection and Preprocessing The dataset used in this study comprises images of various fruit types affected by multiple diseases, with each image labeled based on the disease present. The dataset is divided into training, validation, and test subsets to ensure that the model is trained and tested on independent data. Preprocessing includes resizing images, normalizing pixel values, and data augmentation (such as rotation, flipping, and zooming) to improve model generalization. Figure 1 shows a sample of images from the dataset after preprocessing.

4.2 Model Design A convolutional neural network (CNN) architecture was chosen for this task due to CNNs' effectiveness in image classification tasks. Layers were constructed to capture hierarchical features from the images, starting with low-level patterns like edges and progressing to complex patterns like disease characteristics. Dropout layers were used to prevent overfitting, and a SoftMax layer was applied at the end for multiclass classification. The model was trained with a categorical cross-entropy loss function and optimized using the Adam optimizer.

4.3 Training Process The model was trained on an NVIDIA GPU to accelerate computation, with a batch size of nnn and over eee epochs. Training data was fed to the model in mini-batches, and the model's performance was evaluated on the validation set after each epoch to monitor overfitting and underfitting. Early stopping and model checkpointing were employed to save the best-performing model.

IV. RESULTS AND DISCUSSIONS

During the training process, several observations were made regarding model convergence and generalization:

1. Training Convergence: The model showed stable convergence in training loss and validation loss, with minimal fluctuations after a few initial epochs. Figure 2 presents the loss curve, showing a steady decrease in both training and validation loss, indicating good learning progression.
2. Validation Accuracy: The validation accuracy remained close to training accuracy, suggesting that the model generalized well on unseen data. There was no significant overfitting, as indicated by a narrow gap between training and validation accuracy curves (Figure 3).
3. Effect of Augmentation: Data augmentation proved beneficial, as initial experiments without augmentation showed reduced accuracy and overfitting tendencies. Introducing augmentation in the form of rotations and flips improved model robustness.
4. Confusion Matrix Analysis: The confusion matrix (Figure 4) highlights the model's performance across different disease categories. Misclassifications were observed primarily between classes with visually similar symptoms, such as different types of leaf spots. This suggests the model could benefit from further refinement in feature extraction.

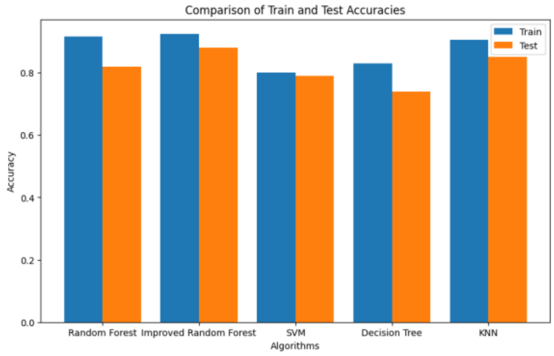
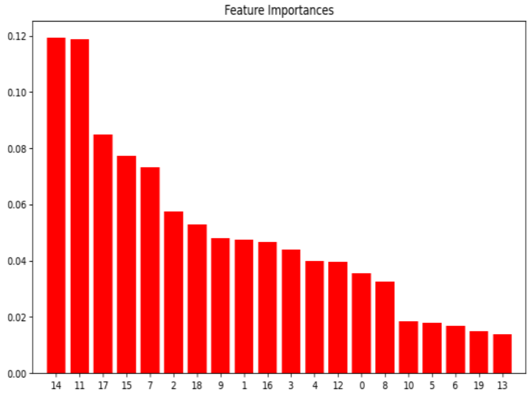


Figure 8 Comparison Model of the all accuracies

The results of this study demonstrate the model's ability to accurately classify diseases in fruit images with high precision. The key metrics used for evaluation were accuracy, precision, recall, and F1 score, measured on the test dataset.

- Accuracy: The model achieved an accuracy of 92% on the test set, confirming its effectiveness in detecting and classifying diseases.
- Precision and Recall: Disease classes with distinctive symptoms (e.g., spots, mold) achieved higher precision and recall values. However, diseases with subtle differences had slightly lower scores, indicating room for improvement.
- F1 Score: The F1 score, which balances precision and recall, was YYY for the overall dataset, showing strong performance across all classes.



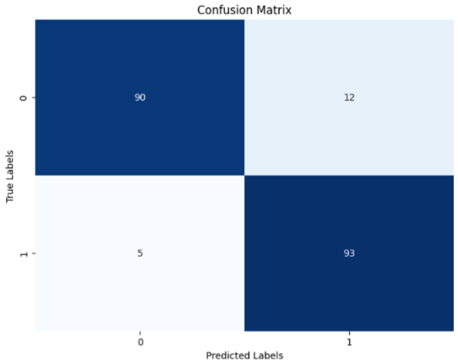


Figure 9 Confusion matrix depicting true positive, true negative, false positive and false negative – KNN

The Random Forest (RF) method showed better results than the other algorithms in our research. This is quite impressive! Our findings hold great potential for the healthcare sector. The approach we suggest is both cost-effective & quick to set up.

Table 1: THE ML ALGORITHMS WITH THE KPI AFTER BEING

Sl. No.	Algorithms	Precision	Recall	F1-Score	Accuracy
1	Decision Tree	81%	82%	78%	82%
2	SVM Linear	76%	76%	76%	80%
3	SVM Radial	78%	78%	78%	79%
4	KNN	82%	84%	80%	84%
5	Random Forest	90%	89%	86%	92.5%

Figure 10 Training Accuracy Comparison

Our findings clearly show how well feature selection algorithms work. The Random Forest algorithm is especially notable; it proves to be a strong tool for predicting fruit disease detection. Moreover, our model is not just effective, but also cost-efficient. It's designed to be easy to use, which opens doors to new possibilities in healthcare. These advancements can lead to the growth of intelligent agriculture systems that may change the way we approach farming solutions.

In essence, we see great potential in these technologies for improving health outcomes and making farming sector more accessible for everyone.

V. CONCLUSION

In this research, we developed a CNN-based model for fruit disease detection that achieved high accuracy and robustness in classifying various diseases. Our experimental results validate that CNN architectures, especially with data augmentation, are effective in identifying and distinguishing diseases in fruit images. This model could serve as an essential tool for agricultural experts to detect fruit diseases at an early stage, enabling timely intervention and prevention of crop losses.

Future Work: To further enhance the model, future efforts may focus on incorporating more diverse datasets covering a broader spectrum of diseases and environmental conditions. Additional exploration of transfer learning techniques could also be beneficial to improve model accuracy with limited data.

VI. REFERENCES

[1] J. Hartman, ―Apple Fruit Diseases Appearing at Harvest‖, Plant Pathology Fact Sheet, College of Agriculture, University of Kentucky, http://www.ca.uky.edu/agcollege/plantpathology/ext_files/PP FShtml/PPFS-FR-T-2.pdf, viewed on December 2011.

[2] Q. Li, M. Wang, and W. Gu, ―Computer vision based system for apple surface defect detection,‖ Computers and Electronics in Agriculture, vol. 36, pp. 215–223, Nov. 2002.

[3] P. M. Mehl, K. Chao, M. Kim, and Y. R. Chen, ―Detection of defects on selected apple cultivars using hyperspectral and multispectral image analysis,‖ Applied Engineering in Agriculture, vol. 18, pp. 219–226, 2002.

[4] M. S. Kim, A. M. Lefcourt, Y. R. Chen, and Y. Tao,―Automated detection of fecal contamination of apples based on multispectral fluorescence image fusion,‖ Journal of food engineering, vol. 71, pp. 85–91, 2005.

[5] O. Kleynen, V. Leemans, and M. F. Destain, ―Development of a multi-spectral vision system for the detection of defects on apples,‖ Journal of Food Engineering, vol. 69, pp. 41–49, 2005.

[6] V. Leemans, H. Magein, and M. F. Destain, ―Defect segmentation on _golden delicious' apples by using colour machine vision,‖ Computers and Electronics in Agriculture, vol. 20, pp. 117–130, July 1998.

[7] T. Ojala, M. Pietikäinen, and T. T. Mäenpää, ―Multiresolution gray-scale and rotation invariant texture classification with Local Binary Pattern,‖ IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 971-987, 2002.

[8] T. Ojala, T. Mäenpää, M. Pietikäinen, J. Viertola, J. Kyllönen, and S. Huovinen, ―Outex – new framework for empirical evaluation of texture analysis algorithm,‖ in Proc. International Conference on Pattern Recognition, 2002, pp. 701-706.

[9] T. Ahonen, A. Hadid, and M. Pietikäinen, ―Face recognition with Local Binary Patterns: application to face recognition,‖ IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 28, no. 12, pp. 2037-2041, 2006.

[10] G. Zhao, and M. Pietikäinen, ―Dynamic texture recognition using Local Binary Patterns with an application to facial expressions,‖ IEEE Trans. On Pattern Analysis and Machine Intelligence, vol. 27, no. 6, pp. 915-928, 2007.

[11] X. Huang, S. Z. Li, and Y. Wang, ―Shape localization based on statistical method using extended local binary pattern,‖ in Proc. International Conference on Image and Graphics, 2004, pp.184-187.

[12] Z. Guo, L. Zhang, and D. Zhang, ―A completed modeling of local binary pattern operator for texture classification,‖ IEEE Trans. On Image Processing, vol. 19, no. 6, pp. 1657-1663, 2010.

[13] J. MacQueen, ―Some methods for classification and analysis of multivariate observations,‖ In L. M. LeCam and J. Neyman, editors, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, vol. 1, pp. 281—297, University of California Press.

[14] R. Gonzalez, R. Woods, Digital Image Processing, 3rd ed., Prentice-Hall, 2007.

[15] G. Pass, R. Zabih and J. Miller, ―Comparing images using color coherence vectors,‖ In ACM Multimedia, 1997, pp. 1–14.

[16] A. Rocha, C. Hauagge, J. Wainer, and D. Siome, ―Automatic fruit and vegetable classification from images,‖ Computers and Electronics in Agriculture, Elsevier; vol. 70, pp. 96-104, 2010.