TOMATO DISEASE CLASSIFICATION USING ENSEMBLE LEARNING APPROACH

Ginne M James[1], S.C Punitha[2]

[1]*Research Scholar, Computer Science, PSGR Krishnammal College for Women, Coimbatore, India*
[2]*Assistant Professor, Computer Science, PSGR Krishnammal College for Women, Coimbatore, India*

Abstract

Ensemble learning methods for supervised machine learning have become trendy due to their ability to accurately predict class labels with different learner methods. It offers efficient models for good predictive capability, which tend to be large and slight imminent into the patterns or structure in the data. In this research work, the ensemble learning technique is used as a classification task to accurately predict class labels. There are six different types of tomato diseases are predicted such as Anthracnose, Bacterial canker, Bacterial spot, Bacterial speck, Early blight and Late blight. Data are collected from the local market and the database is created with 600 images and 100 images for each disease. Various techniques like contrast enhancement in preprocessing, k-means is used for segmentation of tomato disease, color, statistical color features, color co-occurrence matrix and shape features are extracted and finally ensemble learning is used to accurately predicting tomato diseases. The experimental result shows that the proposed framework outperforms well in predicting tomato diseases using ensemble learning methods. An ensemble method for multi-class classification task is compared with various boosting and bagging ensemble methods.

Keywords: *Biotic, Abiotic, Ensemble Learning, AdaBoost, LPBoost, TotalBoost and etc.*

----------***----------

1. INTRODUCTION

The edible and healthy fruit is tomato which has many vitamins and beneficial nutrients to the human body. Tomatoes are usually consumed in daily diets because they contain the natural antioxidant, also β-carotene, vitamin C, and vitamin E [1]. There is a seasonal crop and their availability is inadequate during certain seasons [2].

Tomato is an easily available and cheap fruit for analysis purpose. It is included in the very major horticulture freight [2]. Every often its demand increases at a very high level in the market due to low production or damage due to various disease attacks while supply is limited [3]. Therefore, there is a fundamental need of an advance system for detection of shelf life also provides early warning of various disease attacks [4].

Most of the diseases affect the outer area of tomato soft tissue or flesh by some color change or different spots. In general, there are two types of factors which can affect tomato fruit: biotic and abiotic agents. Different insects, bacteria, fungi, and viruses are the example of living agents [5]. Nonliving abiotic includes various atmosphere effects such as rapid temperature change, excess moisture, insufficient nutrients, poor soil pH and high humidity conditions [6].

In this work classification of tomato, the disease is carried out by using ensemble methods. Rest of this paper will explain briefly about the proposed research work. The paper is organized as follows; section II describe the proposed framework structure. Section III explains about ensemble learning algorithm. Section IV discuss the experiments and its results. Finally, section V gives the conclusion about this research work and future enrichment.

2. THE PROPOSED FRAMEWORK

The proposed framework for tomato disease classification includes four different and important phases. In previous work, the preprocessing and segmentation phases are explained in detail. The preprocessing includes image resize and contract enhancement processes. K-means clustering is used to segment out the disease portion from the image. In feature extraction, four different features are extracted like color, statistical color features, color co-occurrence matrix and shape features. Finally, ensemble learning algorithm used as a classification for classifies tomato diseases based on its feature values. The following figure (Fig1) shows the proposed framework of tomato disease classification system.

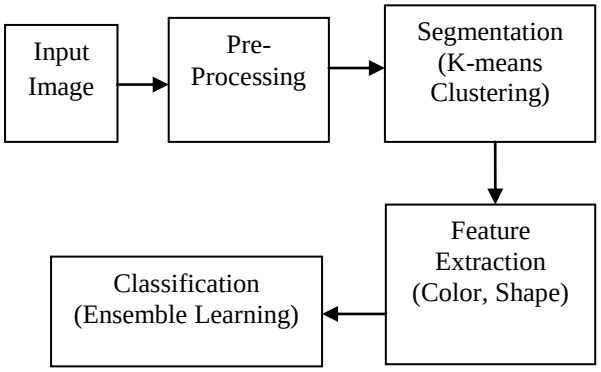


Fig-1: Proposed Framework

2.1 Data Collection

Every work the data collection is a first and most important task. In this work the data are collected the local marketplace. There are six different classes taken into consideration, 100 images for each disease are collected and stored in a database. Finally, the database contains 600 images with 100 images for each disease respectively.

2.2 PreProcessing

Preprocessing phase, two different processes are performed like image resize with 256×256 color index and contrast enhancement. The contrast adjustment is used as color enhancement technique and stored in the database.

2.3 Segmentation

In segmentation phase, K-means clustering algorithm is used to identify and segment out the diseased part from an original image. The cluster size is set to 5 as default for clustering and segmenting the image. The figure (Fig2) shows the segmentation result of tomato disease image.

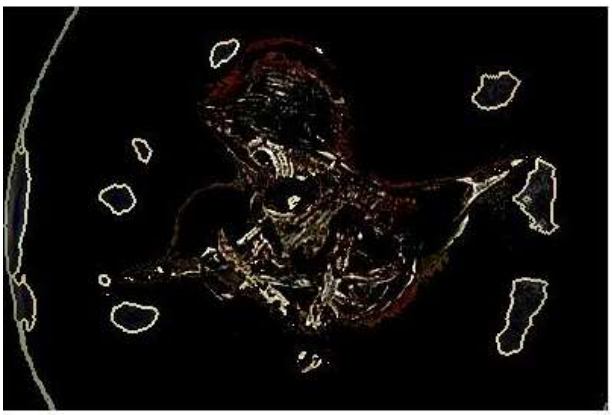


Fig-2: Segmentation Image

2.4 Feature Extraction

In feature extraction phase, there are four types of features are extracted like color, statistical color features, color co-occurrence matrix and shape features. The shape features are extracted using region properties. It includes centroid, eccentricity, orientation, diameter, solidity, extent and perimeter. The color co-occurrence matrix features like Energy, correlation, contrast, entropy, and homogeneity are extracted. The color and color statistical features like hue, saturation, variance, red, green, blue, meanR, meanG, and meanB.

2.5 Classification

The classification phase, ensemble learning algorithm is used as classification task for classifying tomato diseases more accurately and efficiently. There are three different methods are used for classification such as AdaBoost, LogitBoost, and TotalBoost approaches. The ensemble learning algorithm is explained detailed in below chapter (section III) and experimental results are explained briefly in following chapter (Section IV).

3. ENSEMBLE LEARNING

The ensemble methods are learning algorithms that create a set of classifiers and then classify new data points by taking a weighted vote of their predictions/classification. Ensemble learning is a machine learning model where multiple learners are trained to solve the same problem. In difference to ordinary machine learning methods which try to learn one hypothesis from the training dataset, ensemble methods try to construct a set of hypotheses and combine them to use for prediction.

An ensemble contains a number of learners which are generally called as base learners. The generalization ability of an ensemble is usually much stronger than base learners. In point of fact, ensemble learning is quite interesting because that it has a capable to boost weak learners which are somewhat better than random estimate to strong learners which can make very accurate predictions. So, the base learners are also referred as weak learners. Typically, an ensemble is created in two steps. The first step, a number of base learners is produced sequentially. The second step, base learners are combined to use, where among the most popular techniques are majority voting for classification and weighted averaging for regression methods.

Usually, to get a good ensemble, the base learners must be as more accurate as possible, and as more assorted as possible. This has been formally shown [7] and accentuated by many other populaces. There are many effective processes for computing the accuracy of learners, such as cross-validation, hold-out test, etc. There are many successful ensemble methods such as Boosting [8],[9], Bagging [10] and Stacking [11]. In this work, AdaBoost, LogitBoost, and TotalBoost algorithms are used in classification. Bagging is used to test the quality of ensembles.

3.1 AdaBoost Algorithm

Boosting method [12] is an ensemble classification such that each classifier has a weight which is derived from the correctness of learning. The learned models are used to predict the unknown data by the majority vote. The most popular technique is AdaBoost ensemble learning.

The AdaBoost algorithm [13] is a well-known method to build ensembles of classifiers with very excellent performance. This algorithm takes training data and defines weak classifier functions for each sample of the training dataset. Classifier function takes the sample as the argument and produces value 0 or 1 in a case of a binary classification task and weight factor for each classifier. The pseudo-code of AdaBoost is shown in below:

AdaBoost Algorithm:

Input: Sample Dataset D

Base Learning Algorithm L;

Number of Training Rounds T.

Process:

Initialize the weight distribution (D_t)

For each learning rounds

1. *Train base learner using distribution weight*
2. *Measure base learner error*
3. *Find the weight of base learners*
4. *Update the distribution, until the distribution factor reached (D_{t+1}).*

End

Output: T Classifiers

3.2 LogitBoost Algorithm

LogitBoost is a boosting algorithm that introduces a statistical elucidation to AdaBoost algorithm by using additive logistic regression model for determining classifier in each round [14]. Logistic regression is a way of describing the relationship between one or more factors. LogitBoost is a method used to minimize the logistic loss, AdaBoost technique driven by probabilities optimization. This method requires care to avoid numerical problems.

3.3 TotalBoost Algorithm

The General idea of boosting algorithm is to maintain the distribution, over a given set of samples have been optimized. A way to accomplish optimization for TotalBoost is to modify the way measuring the hypothesis edge is being constrained through iterations. AdaBoost constrains the edge with the reverse to the last hypothesis to maximum zero [15]. The totalBoost method is "totally corrective", constraining the edges of all preceding hypotheses to a maximal value that is accurately adapted. It is proven that with adaptive edge maximal value and measurement of confidence in prediction for a hypothesis weighting increases [16] in the classification process.

3.4 Bagging Algorithm

Bagging method [17] builds the models from the same learning algorithm but each algorithm learns from different instances. This method also uses majority vote for prediction of unknown data. The popular technique is Bag.

Input: Sample Dataset D

Base Learning Algorithm L;

Number of Training Rounds T.

Process:

Initialize the weight distribution

For each learning rounds

1. *Generate a bootstrap sample from dataset*
2. *Train base learner from the bootstrap sample.*

End

Output: T Classifiers

4. EXPERIMENTAL RESULT

Simulation experiments in this research are done on a PC with Intel corei3 @ 2.00 GHZ CPU and 4GB memory. The approach is designed on a platform of Matlab 2014a on the operation system Windows 8.1. The datasets used in this research were prepared from real samples for tomato at different diseases, which were collected from a local market. A dataset with 80 images and 20 images of training and testing samples respectively. There are six different classes like Anthracnose, Bacterial canker, Bacterial speck, Bacterial spot, Early blight and Late blight. Ensemble learning algorithm is used to classify the different diseases accurately based on the features, it was trained and tested. The below image (Fig3) demonstrates the accuracy output of various methods.

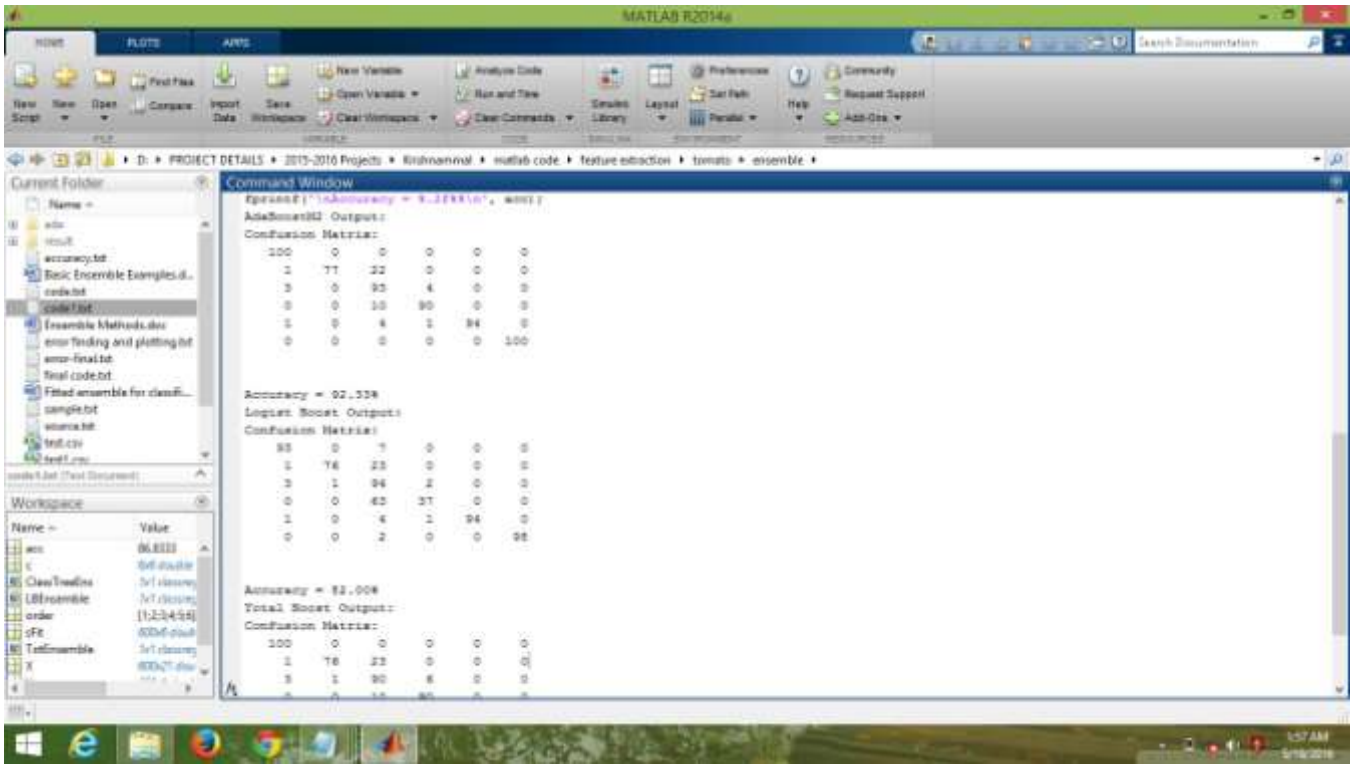


Fig-3:Accuracy Output

The multi-class ensemble methods like AdaBoost, LogitBoost and TotalBoost are used to evaluate the accuracy. The AdaBoost algorithms outperform well in predicting tomato diseases with the high accuracy rate of about 92% when to weigh against to other learning methods. The next table (Table1) has shown the accuracy of various methods and its accuracy rate, ensemble size and accuracy loss rate. AdaBoost learner methods yield better accuracy with low accuracy loss rate of0.0233 and other two methods are 0.0733 and 0.1033 respectively. The chart representation of accuracy comparison is given in the following figure (Chart-1).

Table1. Accuracy Comparison

Algorithm / Accuracy	Accuracy	Ensemble Size	Accuracy Loss
AdaBoost	92.33%	1626526	0.0233
LogitBoost	82.00%	28018	0.1033
TotalBoost	86.83%	34516	0.0733

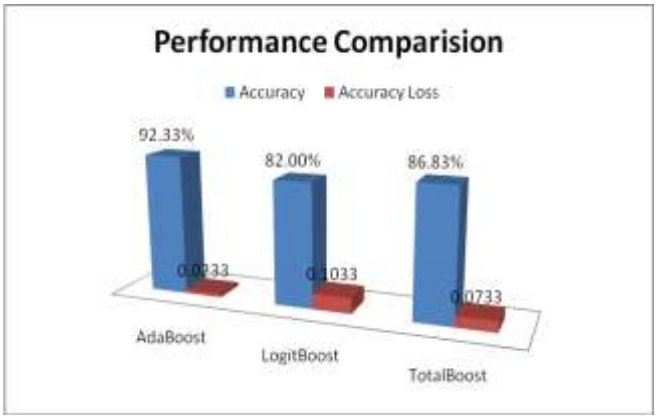


Chart-1: Performance Comparison chart

The natural performance measures for classification error are the error rate. In this work various types of error rates are shown in the following figures. Error rates are compared with three different methods like AdaBoost, LogitBoost, and TotalBoost with various error rates such as Resubstitution Loss Error, Generalization Error, Cross-Validation Error, Training Error, Test Classification Error and Cross-Validation for Test Classification.

Following figure (Fig5) shows the error rate of Resubstitution Loss error, Generalization Error, Cross-Validation Error, and Training Error. The Adaboost algorithm performs well and reaches maximum iteration with low error rate weigh against to other two algorithms. The error rates are computed for training dataset. Adaboost error rates are lies between logit boost and total boost algorithm.

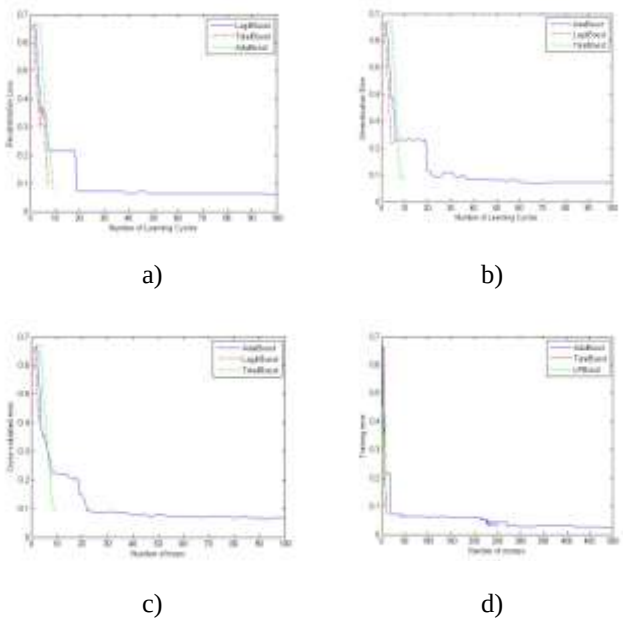


Fig 5. Error rate evaluation: a) Resubstitution Loss Error, b) Generalization Error, c) Cross-Validation Error, and d) Training Error.

The ensemble weights for training dataset are displayed in following figure (Fig6). Weights are calculated for each iteration; Adaboost algorithm performs learning process up to 500 rounds and the lpboost and total boost algorithms are stopped at 8 and 10 iterations respectively. It clearly tells that Adaboost algorithm performs well for a large database with minimum loss.

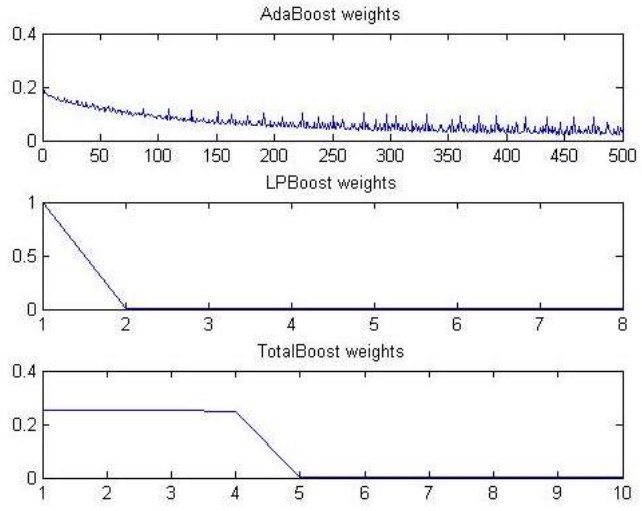
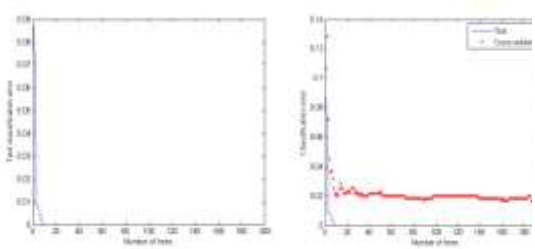


Fig-6: Ensemble Weights for various learning methods

The following figure (Fig7) shows the test error rate with cross-validation result. Test error shows the quality of ensembles also the accuracy efficiency. Compare to training error the test error rate is quite optimistic.



a) b)

Fig-7: Test error evaluation a) Test Classification Error and b) Cross-Validation for Test Classification

5. CONCLUSION

In this research work, tomato disease classification system was developed. The system has four main stages; preprocessing, segmentation, feature extraction and disease classification. Contrast enhancement technique is applied to each image then k-means clustering algorithm is used to segment the disease portion of tomato image, RGB, HSV histogram values, statistical color moments and color co-occurrence matrix are obtained as feature vectors for each image. Finally, ensemble models are developed for tomato disease classification. Three types of ensemble methods such as AdaBoost, LogitBoost and TotalBoost are used to create models and test tomato diseases. Based on the obtained results, the tomato disease classification accuracy is 92.33% for AdaBoost learning method with a minimum error of 0.0233 while comparing to other two methods. In regards to future enhancement of this work is to increase database size and the algorithm implementation limited to three methods. So in future various algorithms are used for classification.

REFERENCES

[1]. Mazza, G.Velioglu, Y.S.Gao, L.Oomah, B.D. Antioxidant activity and total phenolics in selected fruits, vegetables, and grain products. J. Agric. Food Chem. 1998, 46, 4113–4117.

[2]. R.Kalaivani, Dr.S. Muruganand, Dr.Azha.Periasamy, "Identifying The Quality Of Tomatoes In Image Processing Using Matlab", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 2, Issue 8, August 2013, Pp: 3525-3531.

[3]. Z. B. Husin, A. Y. B. Md Shakaff, A. H. B. Abdul Aziz, and R. B. S. Mohamed Farook, "Feasibility study on plant chili disease detection using image processing techniques," in Proceedings of the 3rd International Conference on Intelligent Systems Modelling and Simulation (ISMS '12), pp. 291–296, Kota Kinabalu, Malaysia, February 2012.

[4]. T. Brosnan and D.-W. Sun, "Improving quality inspection of food products by computer vision—a review," Journal of Food Engineering, vol. 61, no. 1, pp. 3–16, 2004.

[5]. J. A. Kodagali and S. Balaji, "Computer vision and image analysis based techniques for automatic characterization of fruits—a review," International Journal of Computer Applications, vol. 50, no. 6, 2012.

[6]. S. R. Dubey, Automatic recognition of fruits and vegetables and detection of fruit diseases [thesis], 2012.

[7]. Krogh, A., Vedelsby, J.: Neural network ensembles, cross validation, and active learning. In Tesauro, G., Touretzky, D.S., Leen, T.K., eds.: Advances in Neural Information Processing Systems 7. MIT Press, Cambridge, MA (1995) 231–238

[8]. Schapire, R.E.: The strength of weak learnability. Machine Learning 5(2) (1990) 197–227.

[9]. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to Boosting. Journal of Computer and System Sciences 55(1) (1997) 119–139.

[10]. Breiman, L.: Bagging predictors. Machine Learning 24(2) (1996) 123–140

[11]. Wolpert, D.H.: Stacked generalization. Neural Networks 5(2) (1992) 241–260.

[12] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning, Data Mining, Inference, and Prediction. Stanford, California.2nd ed. Springer. 2001.

[13]. Yoav Freund, and Robert E. Schapire, "Experiments with a new boosting algorithm", Proceedings 13th International Conference on Machine Learning, Vol. 96, pp. 148-156, 1996

[14]. Sateesh B. Boosting techniques on rarity mining. IJARCSSE. 2012,2-10.

[15]. Sateesh B. Boosting techniques on rarity mining. IJARCSSE. 2012,2-10.

[16]. Ogutu JO, Piepho HP, Streeck TS. A comparison of random forests, boosting and support vector machines for genomic selection. BMC Proc. 2011,5 Suppl 3-S11.

[17]. Leo Breiman: "Bagging predictors", Machine learning, Vol. 24, No. 2, pp. 123-140, 1996