International Journal of Scientific Trends- (IJST)
ISSN: 2980-4299
Volume 4, Issue 4, April - 2025
Website: https://scientifictrends.org/index.php/ijst
Open Access, Peer Reviewed, Scientific Journal

Machine Learning in Agriculture Tomato Disease Classification Using Random Forest

Kilichov Najim Mirzayevich
Tashkent State Agrarian University, Tashkent, Uzbekistan
e-mail: najimqilichov@gmail.com

Abstract

This research discusses the use of the Random Forest algorithm for classifying tomato diseases based on visual characteristics of leaves. The relevance of the study is due to the need to automate the process of diagnosing plant diseases in order to increase productivity and reduce costs. The stages of data processing, feature extraction, model training and the mathematical description of the Random Forest method are presented. The results obtained show high accuracy of classification and demonstrate the potential for introducing intelligent systems into the agricultural sector.

Keywords: Machine learning, Random Forest, agriculture, tomato diseases, classification, computer vision.

Introduction

Modern agriculture faces many challenges, among which timely detection and diagnosis of plant diseases occupies a special place. Tomatoes, being one of the most widespread and economically important crops, are especially vulnerable to various diseases caused by bacteria, viruses and fungi [1]. Crop losses caused by diseases can reach critical levels, which has a negative impact on food security and agricultural economics. Traditional methods for identifying tomato diseases require the participation of experienced agronomists, as well as significant time and labor costs. In this regard, there is growing interest in the use of intelligent analysis methods, in particular in the use of machine learning algorithms. One such algorithm is Random Forest, an ensemble method that demonstrates high efficiency in classification problems based on visual features. The present study aims to develop and evaluate a tomato disease classification model using the Random Forest algorithm [2]. The work describes the stages of image processing, extracting informative features and building a model, and also presents its mathematical apparatus. The experiment results confirm the feasibility of introducing intelligent systems into agricultural practice.

International Journal of Scientific Trends- (IJST)
ISSN: 2980-4299
Volume 4, Issue 4, April - 2025
Website: https://scientifictrends.org/index.php/ijst
Open Access, Peer Reviewed, Scientific Journal

II METHODS AND MATERIALS

Random Forest is built on the basis of several decision trees that work as an ensemble [3].

The basic principles of Random Forest include:

Bagging (Bootstrap Aggregating): Creating several subsamples from the original dataset with a return (bootstrap) and training each tree on its subsample. This allows each tree to be trained on different subsets of the data, reducing the likelihood of overfitting and increasing the overall robustness of the model [4]. Random feature subset: A random subset of features is selected for each tree, reducing correlation between trees and improving overall model performance. It also helps the model to be more robust to noise in the data. Aggregation of results: For classification problems, the voting method is used (majority voting), and for regression problems, averaging the predictions of all trees is used. This means that the final prediction of the model is based on the aggregate of the predictions of all the trees, which makes the model more accurate and reliable.

For the experiment, we used a dataset of images of tomato leaves, which was taken from the https://www.kaggle.com/ platform containing labels of disease classes, including categories such as [5]:

- ✓ Tomato___Bacterial_spot;
- ✓ Tomato___Early_blight;
- ✓ Tomato___healthy;
- ✓ Tomato___Late_blight;
- ✓ Tomato___Leaf_Mold;
- ✓ Tomato___Septoria_leaf_spot;
- ✓ Tomato___Spider_mites Two-spotted_spider_mite;
- ✓ Tomato___Target_Spot;
- ✓ Tomato___Tomato_mosaic_virus;
- ✓ Tomato___Tomato_Yellow_Leaf_Curl_Virus.

The images were converted to a standard size, and features were extracted from them: color histograms, texture parameters (contrast, entropy), and histogram of oriented gradients (HOG) [6].

Feature space

Each image is represented by a vector of features [7]:

$$X_i = [x_{i1}, x_{i2}, \ldots, x_{in}], \quad i = 1, 2, \ldots, m \tag{1}$$

were, m — number of images, n — number of features.

Let the training sample be given:

$$D = \{(X_1, y_1), (X_2, y_2), \ldots, (X_m, y_m)\} \tag{2}$$

were, $X_i \in \mathbb{R}^n$ — feature vector, $y_i \in \{1, 2, \ldots, K\}$ — class label.

A random forest model consists of T trees $h_t(X), t = 1, \ldots, T$. Final prediction:

$$\hat{y} = \text{model}(h_1(X), h_2(X), \ldots, h_t(X)) \tag{3}$$

or as a probability distribution:

$$P(y = k \mid X) = \frac{1}{T}\sum_{t=1}^{T} I(h_t(X) = k) \tag{4}$$

were, I — indicator function.

The algorithm consists of four stages:

1) Random samples are created from a given data set.
2) A decision tree will be built for each sample.
3) Will conduct a vote for each prediction received.
4) Selects the prediction with the most votes as the final result.

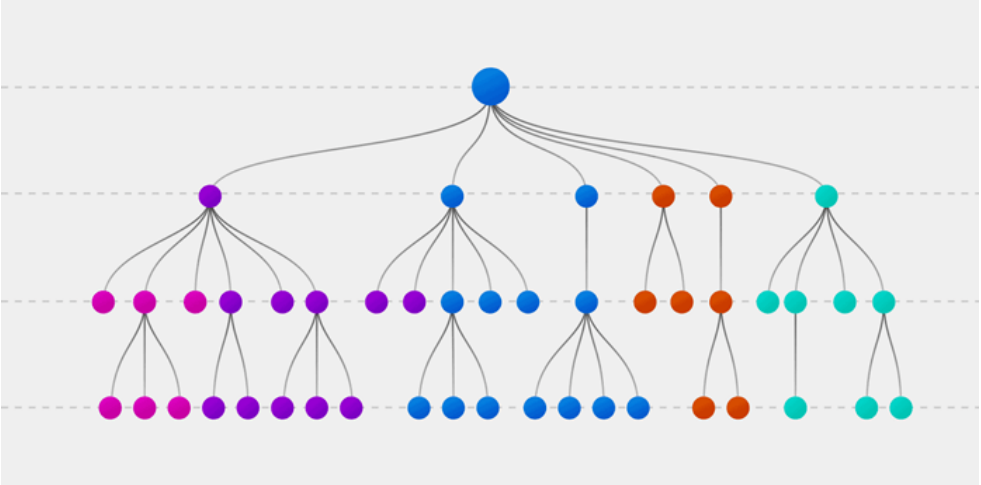

Figure 1. Random Forest

- Random forest is a collection of multiple decision trees.
- Deep decision trees can suffer from overfitting, but random forest prevents overfitting by generating trees from random samples.
- Decision Trees are Computationally Faster than Random Forests.
- Random forest is difficult to interpret, but a decision tree is easy to interpret and convert into rules.

Advantages Random Forest [8]:

➢ Random forest is considered a highly accurate and reliable method because many decision trees are involved in the forecasting process.

➢ Random forest does not suffer from overfitting problem. The main reason is that random forest uses the average of all predictions, which eliminates bias.

➢ Random Forest can be used in both types of tasks (classification and regression tasks).

➢ Random Forest can also handle missing values. There are two ways to solve this problem in Random Forest. The first uses the median to impute continuous variables, and the second calculates the weighted average of missing values [9].

➢ Random Forest also calculates the relative importance of features, which helps in selecting the most relevant features for the classifier.

Disadvantages Random Forest:

➢ Random forest is quite slow because the algorithm uses many trees to operate: each tree in the forest is given the same input data, based on which it must return its prediction. After which voting also takes place on the received forecasts. This whole process takes a lot of time [10].

International Journal of Scientific Trends- (IJST)
ISSN: 2980-4299
Volume 4, Issue 4, April - 2025
Website: https://scientifictrends.org/index.php/ijst
Open Access, Peer Reviewed, Scientific Journal

➢ The random forest model is more difficult to interpret compared to a decision tree, where you easily determine the outcome by following a path in the tree.

III RESULTS

Removing highly correlated features ($C > 0.9$):

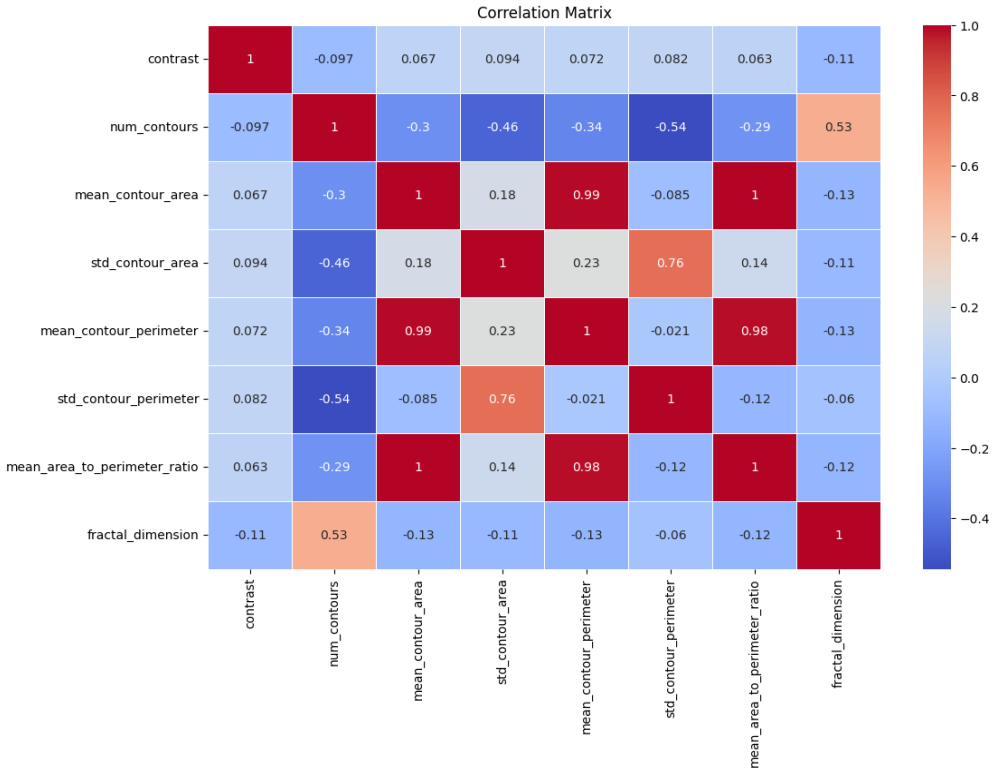

Figure 2. Correlation Matrix

Selection of hyperparameters Random Forest:

✓ Number of trees,
✓ Tree depth,
✓ Minimum number of samples for node splitting,
✓ Minimum number of samples per sheet.

The model was trained on 80% of the sample, 20% were used for testing.

Parameters Random Forest:

- Number of trees: 100
- Tree depth: 10
- Partition criterion: 'gini'

Training the RandomForest visualization model on the error matrix.

✓ 0 = Tomato___Bacterial_spot;
✓ 1 = Tomato___Early_blight;
✓ 2 = Tomato___healthy;
✓ 3 = Tomato___Late_blight;

International Journal of Scientific Trends- (IJST)
ISSN: 2980-4299
Volume 4, Issue 4, April - 2025
Website: https://scientifictrends.org/index.php/ijst
Open Access, Peer Reviewed, Scientific Journal

- ✓ 4 = Tomato___Leaf_Mold;
- ✓ 5 = Tomato___Septoria_leaf_spot;
- ✓ 6 = Tomato___Spider_mites Two-spotted_spider_mite;
- ✓ 7 = Tomato___Target_Spot;
- ✓ 8 = Tomato___Tomato_mosaic_virus;
- ✓ 9 = Tomato___Tomato_Yellow_Leaf_Curl_Virus

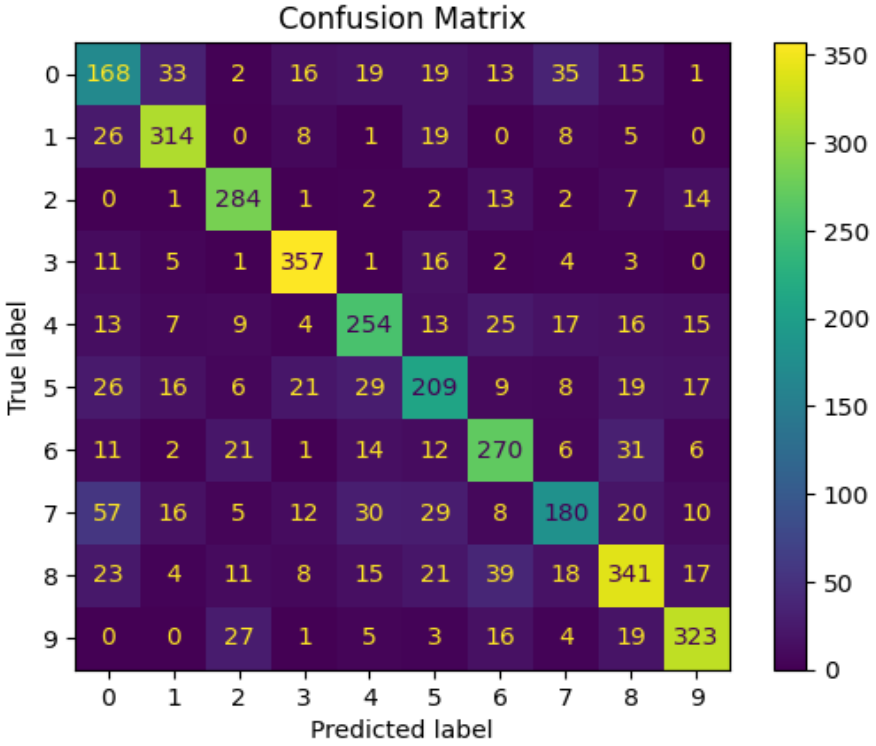

Figure 3. Confusion Matrix

Metrics:

- ➢ Accuracy: 83%
- ➢ Recall, Precision, and f1-score for each class are presented in the classification report.

Table 1. Classification report

	precision	**recall**	**f1-score**
Tomato___Bacterial_spot	79	81	82
Tomato___Early_blight	81	82	81
Tomato___healthy	82	87	82
Tomato___Late_blight	83	89	86
Tomato___Leaf_Mold	83	82	83
Tomato___Septoria_leaf_spot	87	88	87
Tomato___Spider_mites Two-spotted_spider_mite	81	85	80
Tomato___Target_Spot	80	80	81
Tomato___Tomato_mosaic_virus	85	84	86
Tomato___Tomato_Yellow_Leaf_Curl_Virus	80	81	82
Accuracy			**83**

IV DISCUSSION

Random Forest showed high stability and accuracy in the classification task. Through the use of an ensemble of trees, resistance to overfitting and noise is achieved. Moreover, the model is easy to interpret and requires minimal setup compared to neural network methods.

V CONCLUSION

This paper discusses the use of the Random Forest algorithm for automated classification of tomato diseases based on visual features extracted from leaf images. The results obtained demonstrated the high accuracy and stability of the model, which confirms the effectiveness of this method in agricultural problems. The advantages of the Random Forest algorithm are its ability to handle high-dimensional data, resistance to overfitting, and ease of implementation and interpretation. These qualities make it particularly attractive for use in agricultural production environments where there may be noise, incomplete data and various sources of variability (lighting, camera angle, etc.). The implementation of such systems can significantly reduce diagnostic time, reduce dependence on the human factor and increase the efficiency of crop monitoring. This is especially important for small and medium-sized farms with limited resources to attract specialists. However, despite the positive results obtained, tasks remain open to improve the generalization ability of the model and expand its applicability in real conditions. Possible directions for future research include:

- using deep neural networks (for example, Convolutional Neural Networks) for automatic feature extraction;
- development of mobile apps and embedded systems based on trained models;
- collection and replenishment of the training sample, taking into account regional and climatic features;
- integration of machine learning algorithms into the predictive agricultural technology and agronomic recommendations system.

Thus, machine learning, and in particular the Random Forest algorithm, is a powerful tool in modern agriculture that can increase the productivity and sustainability of agricultural systems by intellectualizing the diagnostic and decision-making processes.

REFERENCES

[1] Lokendra Nath Yogi, Tara Thalal, Sarada Bhandari, "The role of agriculture in Nepal's economic development: Challenges, opportunities, and pathways for modernization," Heliyon, pp. ISSN 2405-8440, https://doi.org/10.1016/j.heliyon.2025.e41860, 2025.

[2] Fengzhi Wu, Yong Wang, Maihong Zheng, Jinliang Wang, Jiya Pan, Lanfang Liu, "Prediction and quality zoning of potentially suitable areas for Panax notoginseng cultivation using MaxEnt and random forest algorithms in Yunnan Province," China, Industrial Crops and Products, vol. Volume 229, no. ISSN 0926-6690, https://doi.org/10.1016/j.indcrop.2025.120960, 2025.

[3] Yutong Lai, Ci Peng, Weipeng Hu, Dejun Ning, Luhaibo Zhao, Zhiyong Tang, "Adaptive optimization random forest for pressure prediction in industrial gas-solid fluidized beds,"

Powder Technology, vol. Volume 453, no. ISSN 0032-5910, https://doi.org/10.1016/j.powtec.2025.120607, 2025.

[4] Eleftherios Kouloumpris, Konstantinos Moutsianas, Ioannis Vlahavas, "SABER: Stochastic-Aware Bootstrap Ensemble Ranking for portfolio management," Expert Systems with Applications, Vols. Volume 249, Part B, no. ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2024.123637, 2024.

[5] Kaggle.com, "https://www.kaggle.com/," [Online]. Available: https://www.kaggle.com/.

[6] Diego Legarda, Karen Pérez, Daniel M. Muñoz, "A comparative hardware implementation of histogram of oriented gradients as a descriptor in embedded tracking of swarm robots," Journal of Parallel and Distributed Computing, vol. Volume 198, no. ISSN 0743-7315, https://doi.org/10.1016/j.jpdc.2024.105026, 2025.

[7] Yi Zhong, Jie Jiang, Weize Quan, Mingyang Zhao, Dong-ming Yan, "Distinctive learning of latent space feature for occlusion-aware facade parsing," Building and Environment, vol. Volume 279, no. ISSN 0360-1323, https://doi.org/10.1016/j.buildenv.2025.112955, 2025.

[8] Yihan Ding, Xuanpei He, Rui Zhang, Haotian Wu, Yingaridi Bu, "Random forest-assisted Raman spectroscopy and rapid detection of sweeteners," Infrared Physics & Technology, vol. Volume 148, no. ISSN 1350-4495, https://doi.org/10.1016/j.infrared.2025.105871, 2025.

[9] Tao Li, Jie-Xue Jia, Jian-Yu Li, Xian-Wei Xin, Jiu-Cheng Xu, "A novel random fast multi-label deep forest classification algorithm," Neurocomputing, vol. Volume 615, no. ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2024.128903, 2025.

[10] Wenxiang Li, Shengqun Chen, Lijin Lin, Li Chen, "Random-forest-based task pricing model and task-accomplished model for crowdsourced emergency information acquisition," Systems and Soft Computing, vol. Volume 7, no. ISSN 2772-9419, https://doi.org/10.1016/j.sasc.2025.200235, 2025.