EMPIRICAL EVALUATION OF PLANT DISEASE DETECTION

Tejaswini Vempati
dept.of Computer Science(Masters)
University of Central Missouri
TXV97790@ucmo.edu

Shraddhasree Nangunoori
dept.of Computer Science(Masters)
University of Central Missouri
SXN99850@ucmo.edu

Mounika Rayapudi
dept.of Computer Science(Masters)
University of Central Missouri
MXR08230@ucmo.edu

Sravani Seelam
dept.of Computer Science(Masters)
University of Central Missouri
SXS98230@ucmo.edu

***ABSTRACT*—Crop diseases are a significant threat to food security, but because the necessary foundation is absent in many places around the world, it is still difficult to quickly identify them. Impressive results have been obtained in the field of leaf-based image classification since the development of precise approaches. In order to distinguish between healthy and diseased leaves from the generated data sets, this work uses KNN, SVM, Fuzzy KNN, and Bayesian SVM. The phases of implementation included in our proposed study are dataset construction, feature extraction, classifier training, and classification. To categorize the photos of sick and healthy leaves, the produced datasets of sick and healthy leaves are combined and trained using the algorithms mentioned and the accuracy is compared. Overall, we can clearly identify the illness present in plants on a massive scale by utilizing machine learning to train the vast data sets that are publicly available.**

Keywords— Plant disease, KNN, Fuzzy KNN, SVM, Bayesian SVM, detection, Classification.

I. INTRODUCTION

The detection of plant diseases is the key to avoiding losses in agricultural product output and quantity. The study of patterns on plants that are visible to the naked eye is referred to as "plant disease research." Plant disease detection and health monitoring are vital for sustainable agriculture. Plant disease tracking manually is quite difficult. It is feasible to recognize plant diseases with naked sight thanks to their knowledge of plant diseases and the current farming method. On a large number of plants, the procedure is time-consuming, difficult, and imprecise. Specialist consultation is expensive. In these kinds of circumstances, suggested techniques are implemented where devices are used for the quicker and less expensive automatic diagnosis of diseases. As a result, it is more accurate and beneficial. A high level of complexity is provided by visually studying the symptoms on the plant leaves, where the plant disease may be promptly detected. As a result, the classification of plant diseases is crucial to the agriculture sector.

Many initiatives have really been created to stop crop loss from diseases. Over the past ten years, integrated pest control has increasingly been used to supplement traditional ways of pesticide administration. Whatever the method, the first phase of ineffective illness management is accurate disease identification when it first manifests. Historically, institutions like local plant clinics or agricultural extension agencies have assisted in disease identification. More recently, these initiatives have also been helped by the availability of online resources for disease detection, taking advantage of the global increase in Internet usage. Even more recently, mobile phone-based tools have emerged, capitalizing on the historically unprecedented global adoption of mobile phone technology.

The scientific community has extensively investigated plant diseases, primarily concentrating on the biological aspects of diseases. A global challenge that affects food security is the issue of plant diseases. Regardless of borders, media, or technology, plant diseases have a significant negative impact on farmers' bottom lines. In the modern world, early disease detection is a difficult strategy that requires extra care. Our strategy focuses on identifying and recognizing plant-damaging diseases and pests. Over time, there has been significant growth in plant productivity. This crop is extremely vulnerable to diseases and essentially powerless against them. Additionally, viruses that affect plants have been reported, and novel viral illnesses continue to appear. Recently, a number of methods have reportedly been used to identify plant illnesses. These include using physical techniques, such as imaging and spectroscopy, to identify plant characteristics and stress-based disease detection, as well as direct methods that are directly tied to the chemical analysis of the diseased part of the plant. To seemingly identify plant diseases, the disease is currently a tough approach and needs to be addressed with unique methods. These include using physical techniques, such as imaging and spectroscopy, to identify plant characteristics and stress-based disease detection, as well as direct methods that are directly tied to the chemical analysis of the diseased part of the plant. Visibly, plant diseases come in a wide range of shapes, sizes, hues, etc. Designing more effective control measures to lessen crop loss requires an understanding of this interplay. Additionally, our method's difficult aspects include estimating how precisely we can diagnose the condition and

https://github.com/SXN99850/PlantDiseaseDetection-MLProject

the level of infection it manifests. The distinctions between the concepts of image classification and object detection must now be made clear. In contrast to a detection strategy, which deals with the class and location instances of any specific object in the image, classification assesses if an image contains any examples of an object class.

II. MOTIVATION

The current method that farmers employ to find plant diseases allows them to be seen with the unaided eye and by applying their knowledge of plant ailments. The process of doing so on a large number of plants is time-consuming, challenging, and inaccurate. The expense of consulting specialists is high. In these types of situations, proposed strategies are put into practice where gadgets are employed for the automatic identification of diseases that make the procedure cheaper and easier. This increases the accuracy rate and makes it more helpful.

In addition, Farmers with less expertise may use drugs throughout the identification process without thinking about the repercussions of their conduct and render incorrect conclusions. Additionally, it's likely that increased output and quality can lead to environmental contamination, which will bring about unwarranted financial losses. independent document. Please do not revise any of the current designations.

Producers could now detect diseases using their unassisted eyesight and their knowledge of plant maladies thanks to the existing technology. On a lot of plants, the procedure is moment, difficult, and imprecise. Specialized consultation is expensive. In such kinds of circumstances, suggested techniques are implemented where devices are used for the quicker and less expensive robotic diagnosis of ailments. As a result, it is more accurate and beneficial.

Additionally, Individuals with less experience could utilize medicines during the assignment without considering the consequences and draw false conclusions. Furthermore, it's possible that higher productivity and quality will result in radioactive pollution.

III. MAIN CONTRIBUTIONS & OBJECTIVES

The goal of this study is to categorize plant illnesses using KNN, Fuzzy-KNN, and Bayesian SVM Machine Learning classification methods to analyze leaf image data.

The execution plan is as follows.

- Data Gathering
- Data cleansing
- Building the model
- Implementation
- Comparing the models

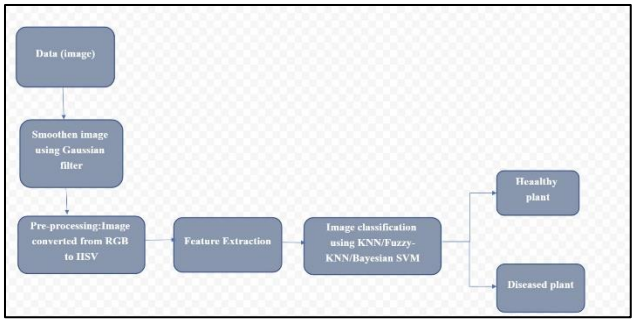


Fig: Project Flow Chart

IV. RELATED WORK

The pre-processing or purification of data is a crucial stage in the job of a machine learning engineer, and the great majority of them invested a lot of time and effort before creating a model from scratch. Data pre-processing methods include removing unwanted or noisy data, treating missing values, and detecting outliers. Identification of plant pathogens is an important subject that has been explored throughout the years and is driven by the need to create nutritious food. Expense, consumer, sensitivity, and accuracy are some desired factors to consider, nevertheless. Several works have suggested several non - destructive methods to get around those facts over the past ten years. Methods for multispectral local sensing were employed in to assess how stressed plants were by their surroundings.

For example, in, thermal and fluorescence imaging methods were introduced for evaluating plant stress caused mostly by elevated gases, radiation, water status, and insect assault, among other things. Optical technologies are useful tools taken into consideration for monitoring plant health. The investigation of plant defense mechanisms in response to pathogen presence is another crucial topic. To gauge the strength of the leaves' disease defenses, chemical components were applied to the leaves. Plants were grown in the presence of various nutritional components to assess their impacts on the crop in order to research the robustness of plants against nutritional facts. As was already indicated, various media outlets have covered the topic of plant anomaly detection. Although earlier techniques function exceptionally well in the examined circumstances, they do not yet offer a highly precise strategy for assessing illnesses and pests in real-time. Instead, they typically conduct their studies in laboratories or with pricey equipment. As a result, our strategy focuses on a low-cost method that leverages in-place photos as our source of data and includes variations of the scenario in place. Before Deep Learning gained popularity in the computer vision Field, a number of manually created feature-based techniques had been widely used just for picture recognition. The reason a method is labeled "handcrafted" is because of all the human

knowledge that went into creating the algorithm and the intricate parameters that are used in the procedure. These methods' high computing costs and time requirements as a result of the intricate preprocessing, feature extraction, and classification are some of its drawbacks. Some of the most well-known manual feature extraction techniques, which are typically paired with classifiers include Support Vector Machines (SVM). Unlike when employing handcrafted-based approaches that use distinct processes, the capabilities of Machine Learning have allowed researchers to develop systems that can be trained and evaluated end-to-end (all incorporated in the same process). Due to machine learning's (ML) exceptional performance as a feature extractor in picture identification tasks, the concept has been applied to a variety of fields, including robotics, agriculture, and automation. Some agricultural applications use computer vision and machine learning to handle challenging problems like plant recognition. In order to identify the species among the 57 different types of trees, we employed an ML for leaf segmentation, image processing techniques to extract hand-crafted features, feature vectors to train an SVM, and an SVM and KNN. The Plant Village Dataset, however, only includes pictures of leaves that have already been clipped in the Sensors. This contrasts with the images in our plant Diseases and Pest Dataset, which were captured on-site by various cameras at various resolutions and show not only leaves at various stages of infection but also other infected parts of the plant, such as fruits and stems, as well as leaves infected by specific pathogens. Additionally, dealing with background fluctuations in our dataset that are mostly brought on by the environment or the location itself is difficult (greenhouse). The problems, such as pattern variation, infection state, various diseases or pests and their placement in the image, and surrounding objects, among others, are still difficult to overcome even though the works listed above provide remarkable performance on leaf disease recognition.

V. PROPOSED FRAMEWORK

In order to distinguish between healthy and diseased leaves from the generated data sets, we proposed the below algorithms for classification.

KNN:

The k Nearest Neighbors (kNN) approach uses a database in which the data points are divided into numerous unique classes to predict the categorization of a new sample point. By choosing the k data points that are closest to the new observation and choosing the class with the highest frequency among them, the approach identifies which points from the training set are sufficiently similar to be taken into account when choosing the class to forecast for a new observation. Consequently, it is called the k Nearest Neighbors algorithm.

1. A new sample and a positive integer k are supplied.
2. We choose the k database records that are most similar to the new sample.
3. We discover how these entries are most frequently categorized.
4. We categorize the new sample in this manner.

FUZZY K-NEAREST NEIGHBOR:

One of the best techniques for tackling supervised learning problems is the k-Nearest Neighbors (kNN) classifier. The Fuzzy-kNN algorithm determines the fuzzy degree of each instance's membership in the problem's classes.

Based on how closely the previously encountered examples resemble the training data, it classifies them. But when it comes to categorization, it gives every labeled sample the same weight. It is possible to increase precision in many ways, with the Fuzzy k-Nearest Neighbors (FuzzykNN) classifier being one of the most effective. You may determine the fuzzy degree of membership of each instance to the problem's classes using FuzzykNN. As a result, it produces more seamless boundaries between classes.

There is no fuzzy alternative for handling such a big amount of data besides the traditional kNN technique for handling massive datasets. Nevertheless, because of the added processing cost and lengthy latency when dealing with large datasets, computing class membership is even less scalable.

The fuzzy-kNN algorithm offers a significant improvement over the standard kNN approach. It has been established that it is quite competitive in terms of accuracy when compared to other fuzzy approaches. To perform this technique effectively, it is crucial to precalculate the class memberships while using the training set. Then, using the knowledge, it calculates the kNN for each sample in the test set.

The following formal notation can be used to represent the fuzzy-kNN algorithm:

Let T R stand for a training dataset and T S for a test set; both are the result of n and t samples, respectively, which are predetermined numbers. Each sample, xi, is represented by a vector, (xi1, xi2, xi3,.., xij), where xij represents the value of the j-th feature of the i-th sample, and xi represents the value of the i-th sample.

There is a known class to which every sample of T R belongs, but T S has no discernible class. The fuzzy-kNN algorithm consists of two sections. While continuing to use a leave-one-out strategy for the remainder of the process, the first stage determines the kmemb nearest neighbors of the T R against itself.

This is done by calculating the distances between each sample of T R and x train, and then looking for the samples that are kmemb closest to those. After computing the neighbors, the class membership is created, as shown in

Equation 1. As a result, the T R now includes a class membership vector in place of the original class label.

BAYESIAN SVM:

This "Support Vector Machine" algorithm is a supervised machine learning algorithm that can be used to address classification and regression issues. It was created using the "support vectors" theory. It is typically used in practice to solve different kinds of categorization challenges. We represent each data point as a point in an n-dimensional space (where n is the number of features you have), with the value of each feature being the value of a particular coordinate. This method is referred to as the SVM algorithm. Then, classification is performed by locating the hyperplane that most clearly distinguishes between the two classes.

Perform data augmentation on datasets where copies of photos from classes with few examples were added to the dataset after the images in the dataset were translated, rotated, flipped, and/or brightened. Once this is finished, the masks are added to the RGB images to create the final 100 x 100 x 4 input. The images and labels are then added to the RGB images and stored in separate binary (.npy) files. To discover the characteristics of the dataset, it will be necessary to train SVM using Bayesian Optimization on the preprocessed images. The Bayesian analysis is simplified using the Gaussian Process (GP) regression model.

In order to formalize the relationship between the result (in this case, the root mean square error) and the SVM optimization parameters, this function builds a regression model.

Because it is a Bayesian model, the regression parameters also get a prior distribution that is multivariate normal. This model uses the widely accepted assumption that the residuals are normal. In order to support nonlinearity in the SVM tuning parameters, which is not achievable with the SVM model, GP regression models use a kernel basis expansion (similar to the SVM model). The covariance function of the multivariate normal prior is implemented as a radial basis function kernel, and the kernel parameters of the generalized radial basis function are estimated via maximum likelihood estimation.

VI. DATA DESCRIPTION

Plant Village is a publicly available dataset(1) - that contains 54,306 photos of 38 different healthy and diseased leaves associated with their 14 plant species. This dataset served as the training ground for all deep learning models (some of the plant diseases are shown in Figure below). The photos were reduced to 224 * 224 * 3 pixels in size, and normalization was taken into consideration by dividing the pixel values by 255 in order to make them more appropriate for the initial values of the models.

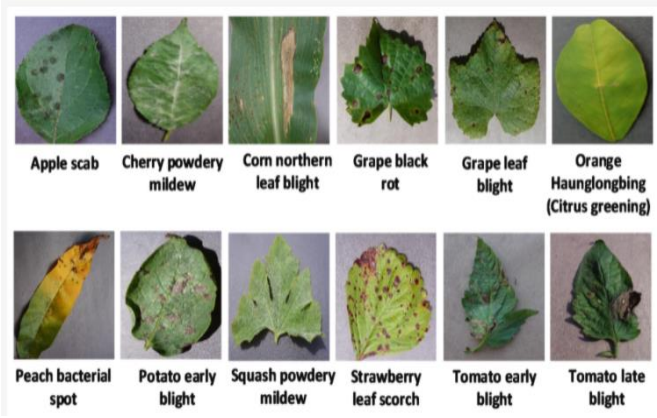


Fig1 - Sample images from the Plant Village dataset

The dataset was split into three categories, with training datasets and testing datasets being separated by 80% and 20%, respectively, to prevent overfitting. Purification of data is a crucial stage in the job of a machine learning engineer, and the great majority of them invested a lot of time and effort before creating a model from scratch. Data pre-processing methods include removing unwanted or noisy data, treating missing values, and detecting outliers. To get more accurate findings, some background noise should be eliminated before feature extraction. After the image has been transformed from RGB to grayscale(2), it is smoothed using a Gaussian filter, as demonstrated in the sample below: To gauge the amount of green color existing in the image, the image is first transformed to the HSV color space. Feature selection is a critical step in the solution of any machine-learning problem. Specifically, in this project, we are selecting features based on the correlation between various variables and the target variable. The correlation between the feature green part of the leaf (F1) and the feature green part of the leaf (F2) is extremely high, indicating that both variables are highly dependent on one another. As a result, we can eliminate F2.

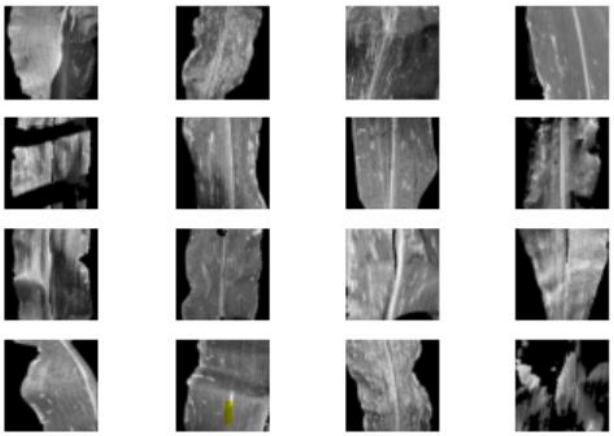


Fig2 – Images from the Plant Village dataset after Gray scaling

https://github.com/SXN99850/PlantDiseaseDetection-MLProject

Validation Method:

Accuracy: Accuracy is the most intuitive performance metric, and it is simply the ratio of correctly predicted observations to the total number of observed observations (or observations correctly predicted).

Precision: The precision of a prediction is defined as the ratio of correctly predicted positive observations to the total number of correctly predicted positive observations.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall(sensitivity): The recall of positive observations is the ratio of correctly predicted positive observations to all positive observations in the actual class.

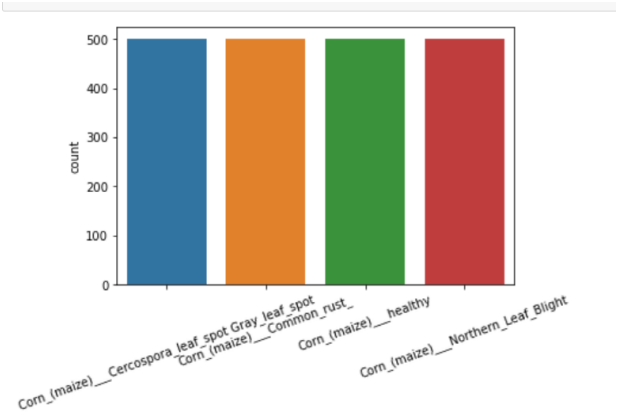
$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

F1 Score: Precision and recall are combined to form the F1 Score, which is a weighted average. As a result, both false positives and false negatives are taken into consideration when computing this score.

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

VII. RESULTS & COMPARISION/ANALYSIS

If enough data is provided for training, machine-learning approaches may identify plant leaf diseases with high accuracy. In this project, fuzzy KNN and bayesian SVM have been studied to better comprehend plant diseases for the Corn dataset.



The importance of collecting large datasets with high variability, data augmentation, and improving classification accuracy will all be covered in this document, along with the significance of detecting plant leaf disease in small samples and the significance of hyperspectral imaging for early detection of plant disease. Although there are certain advantages, there are also some drawbacks. It is suggested that deep learning-based algorithms be utilized in order to produce good detection effects on their datasets. With an accuracy rating of 73, the Model has performed admirably.

We compute descriptive statistics for the position of the proper class in the probability space. The classifier's results are arranged in descending order to achieve this. The higher you are on the list, the better your classification. With an accuracy of 73.2%, the Bayesian SVM performed brilliantly in the categorization of plant diseases.

	precision	recall	f1-score	support
0	0.48	0.62	0.54	84
1	0.86	0.91	0.88	88
2	0.65	0.53	0.58	125
3	0.98	0.92	0.95	103
accuracy			0.73	400
macro avg	0.74	0.74	0.74	400
weighted avg	0.74	0.73	0.73	400

Output: Classification report for Bayesian SVM

The accuracy is more than that of the prior method, Fuzzy KNN, which had an accuracy rate of roughly 72%.

	precision	recall	f1-score	support
0	0.63	0.65	0.64	105
1	0.63	0.91	0.75	65
2	0.64	0.60	0.62	108
3	1.00	0.80	0.89	122
accuracy			0.72	400
macro avg	0.73	0.74	0.72	400
weighted avg	0.75	0.72	0.73	400

Output: Classification report for Fuzzy KNN

We also have trained our Corn dataset with the

https://github.com/SXN99850/PlantDiseaseDetection-MLProject

KNN algorithm which has performed with accuracy of 72%

	precision	recall	f1-score	support
0	0.66	0.61	0.63	117
1	0.67	0.90	0.77	69
2	0.61	0.65	0.63	95
3	1.00	0.82	0.90	119
accuracy			0.73	400
macro avg	0.73	0.74	0.73	400
weighted avg	0.75	0.73	0.73	400

Output: Classification report for KNN

The final comparison(3) between the trained algorithms is displayed in the pie chart for the Corn dataset. All three algorithms performed similarly with 72% accuracy.

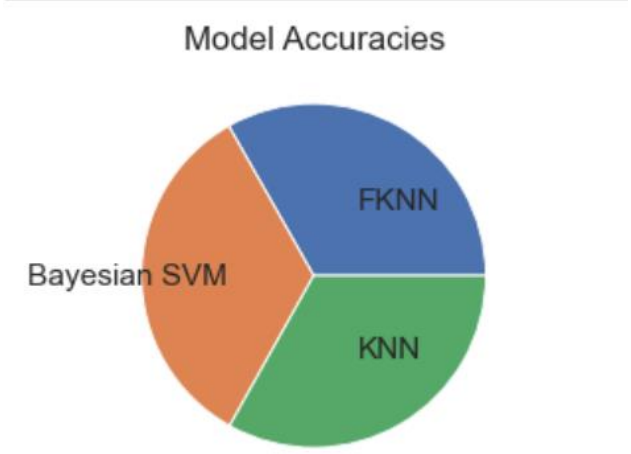


Fig3: Pie chart comparision

VIII. REFERENCES

[1] Liakos, K.G.; Busato, P.; Moshou, D.; Pearson, S.; Bochtis, D. Machine learning in agriculture: A review. *Sensors* **2018**, *18*, 2674.

[2] Römer, C.; Bürling, K.; Hunsche, M.; Rumpf, T.; Noga, G.; Plümer, L. Robust fitting of fluorescence spectra for pre-symptomatic wheat leaf rust detection with support vector machines. *Comput. Electron. Agric.* **2011**, *79*, 180–188.

[3] Chen, T.; Zhang, J.; Chen, Y.; Wan, S.; Zhang, L. Detection of peanut leaf spots disease using canopy hyperspectral reflectance. *Comput. Electron. Agric.* **2019**, *156*, 677–683.

[4] Coops, N.; Stanford, M.; Old, K.; Dudzinski, M.; Culvenor, D.; Stone, C. Assessment of Dothistroma needle blight of Pinus radiata using airborne hyperspectral imagery. *Phytopathology* **2003**, *93*, 1524–1532.

[5] Leucker, M.; Mahlein, A.-K.; Steiner, U.; Oerke, E.-C. Improvement of lesion phenotyping in Cercospora beticola–sugar beet interaction by hyperspectral imaging. *Phytopathology* **2015**, *106*, 177–184.

[6] Saleem, M.H.; Potgieter, J.; Mahmood Arif, K. Plant Disease Detection and Classification by Deep Learning. *Plants* **2019**, *8*, 468.

[7] Xie, C.; Yang, C.; He, Y. Hyperspectral imaging for classification of healthy and gray mold diseased tomato leaves with different infection severities. *Comput. Electron. Agric.* **2017**, *135*, 154–162.

[8] Kobayashi, T.; Kanda, E.; Kitada, K.; Ishiguro, K.; Torigoe, Y. Detection of rice panicle blast with multispectral radiometer and the potential of using airborne multispectral scanners. *Phytopathology* **2001**, *91*, 316–323

[9] Brahimi, M.; Boukhalfa, K.; Moussaoui, A. Deep learning for tomato diseases: Classification and symptoms visualization. *Appl. Artif. Intel.* **2017**, *31*, 299–315.

[10] Ferentinos, K.P. Deep learning models for plant disease detection and diagnosis. *Comput. Electron. Agric.* **2018**, *145*, 311–318.

[11] Fuentes, A.; Yoon, S.; Kim, S.; Park, D. A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. *Sensors* **2017**, *17*, 2022.

[12] Mohanty, S.P.; Hughes, D.P.; Salathé, M. Using deep learning for image-based plant disease detection. *Front. Plant Sci.* **2016**, *7*, 1419.

[13] TÜRKOĞLU, M.; Hanbay, D. Plant disease and pest detection using deep learning-based features. *Turk. J. Electr. Eng. Comput. Sci.* **2019**, *27*, 1636–1651.

[14] Zhang, K.; Wu, Q.; Liu, A.; Meng, X. Can Deep Learning Identify Tomato Leaf Disease? *Adv. Multimed.* **2018**, *2018*.

[15] Oppenheim, D.; Shani, G.; Erlich, O.; Tsror, L. Using Deep Learning for Image-Based Potato Tuber Disease Detection. *Phytopathology* **2019**, *109*, 1083–1087.

[16] Too, E.C.; Yujian, L.; Njuki, S.; Yingchun, L. A comparative study of fine-tuning deep learning models for plant disease identification. *Comput. Electron. Agric.* **2019**, *161*, 272–279.

[17] Chen, J.; Liu, Q.; Gao, L. Visual Tea Leaf Disease Recognition Using a Convolutional Neural Network Model. *Symmetry* **2019**, *11*, 343.

https://github.com/SXN99850/PlantDiseaseDetection-MLProject

[18] Kamal, K.; Yin, Z.; Wu, M.; Wu, Z. Depthwise separable convolution architectures for plant disease classification. *Comput. Electron. Agric.* **2019**, *165*, 104948.

https://github.com/SXN99850/PlantDiseaseDetection-MLProject