Fruits and Vegetables Disease Detection System Based on Indications Using Machine Learning Approach: A Systematic Review

Tajbia Karim[1], Mariam Chowdhury[2], Saima Murtuza[2], Afrida Israt Jahan[2], Afrina Khatun[3]

Abstract

In agriculture science, automated and computerized methods increase the country's growth, economy and productivity as it is highly dependent on the export of fruits and vegetables. Nowadays, it is impossible to check the quality of fruits and vegetables with bare hands as they are exported in a batch. In this world of technology, artificial intelligence plays an essential role by introducing many algorithms to detect diseases that hamper quality. This paper presents a detailed review of which algorithm best detects diseases in fruits and vegetables. The paper also includes details about pre-processing, segmentation, different algorithms for detection, and image enhancement. An analysis of different algorithms proposed by researchers for disease detection within fruits and vegetables was conducted. From contemporary research works, we have come to know that there is not one perfect method for detecting diseases of all fruits and vegetables. By careful analysis, we have recommended which machine learning method might be suitable for specific types of fruits and vegetables.

Keywords: Automated, Pre-processing, Segmentation, Fruits, Vegetables

1. Introduction

In this paper, we reviewed many papers that included diverse machine-learning approaches to detect diseases in fruits and vegetables depending on the indications. Bangladesh is an agricultural country, and this agriculture sector is crucial to the

[1] Department of Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh, Email: tajbia.karim@bup.edu.bd
[2] Department of Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh, Email: mariamchowdhury.200@gmail.com
[2] Department of Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh, Email: saima.murtuza13@gmail.com
[2] Department of Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh, Email: afridaisratjahan613@gmail.com
[3] Department of Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh, Email: afrina.khatun@bup.edu.bd

country's economy. The farmers of this country are highly dependent on producing fruits and vegetables. However, sometimes this production is hindered due to diseases caused by many pathogens. This is almost the same scenario for all fruits and vegetables. Researchers worldwide tried to give solutions to this specific problem described in their paper. By reviewing their works, we tried to analyze the best way to detect diseases among fruits and vegetables using different machine-learning approaches. As fruits and vegetables play a vital role in a country's economy as they are exported abroad, our analysis will be helpful for the researchers to find out which algorithm will be best for disease detection in fruits and vegetables with the highest accuracy rate, which will be beneficial for everyone, especially those who work in the farming sector and the industrial sector as well. The motivation behind this paper was to reduce the loss in the agricultural and food business due to fruit and vegetable diseases. As the weather in Bangladesh is changing every day, it is affecting agricultural products greatly. So, to reduce this loss, the researchers and other business personnel can follow this paper to eliminate fruit and vegetable diseases to a certain level.

By analyzing thoroughly, we found out that in many fruit businesses (export & import, juice companies, restaurants) these advanced techniques can be used to distinguish between fresh and rotten fruits and vegetables. In future research, if the researchers want to detect the maturity state of a fruit, it will become easy to find out what disease can infect it with the help of a practical machine learning approach. We compared datasets available in different papers and the approaches they followed. To the best of our knowledge, we could not find any papers that showed any comparison among the datasets and approaches. The comparison chart in our review paper reflects which algorithm is better for specific cases.

2. Literature Review

Researchers have started to apply Machine Learning to detect diseases in fruits and vegetables approximately since 2015 (Behera et al., 2018). The information about the work of researchers is shown in Table 1. The first model that we reviewed here was done in 2009 by (Qin et al., 2009). They proposed a model in which citrus canker was detected on citrus using hyperspectral reflectance imaging with spectral information divergence. Spectral Information Divergence (SID) was used for detection, with an accuracy rate of 96.2%. Behera et al. (2018) proposed a model using K-means clustering and Support Vector Machine (SVM) to detect diseases of mandarin and oranges. They noticed the symptoms of the disease in those plants by observing leaves, stems, and fruits. Also, they used image pre-processing techniques such as image enhancement and lab color transformation. They were successful in detecting diseases with an accuracy of 90%. Then, Singh et al. (2019) proposed a system for detecting diseases in Mangifera indica (mango) by using Multilayer Convolutional Neural Network (MCNN). They used leaf for detection and for pre-processing, contrast enhancement and image reshaping were used. Their accuracy score was 97.13%. In the research of JJMCOE and J (2017), diseases were discovered in Malus domestica (apple). They have used the K-means clustering segmentation algorithm and Learning Vector Quantization Neural Network (LVQNN) for detection. For pre-processing, they have used image enhancement. They used 24 apple images that were real-time in nature, and they achieved an accuracy of 95%. Ullah et al. (2019) proposed a model to detect fruit disease by analyzing the leaves. They used contrast enhancement and image re-scaling for pre-processing and Convolutional Neural Network (CNN) as a detection algorithm. Their work had an accuracy rate of 97.33%. Ouyang et al. (2012) made a model to detect strawberry plant disease. The Top-hat transformation was used for pre-processing, which removes noise interference by median filtering. Their detection algorithm was an image segmentation algorithm with an accuracy rate of 99%. In the research of Pujari et al. (2013) grape, mango, and pomegranate fruits were used to grade and classify the disease of those fruits based on their statistical texture. Segmentation and Artificial Neural Networks (ANN) were used for detection purposes, and the accuracy rate was 84.65% for regular fruits and 76.6% for affected fruits. In 2019, (Doh et al.,2019) proposed a model using both SVM and K-Nearest Neighbour (KNN) for disease detection in citrus fruits and found an accuracy of 93.12% and 88.96% respectively.

In recent years, Yang et al. (2022) have achieved 99.70% accuracy in detecting disease in apples using a sorting image recognition system. Before that, Zaki et al. (2021) detected the disease in onion plants by observing young onion leaves. They have used CNN for this detection, and for pre-processing techniques, they have used image pre-processing. They achieved an accuracy of 77.05% for a data set of batch size 8 and 85.47% for a dataset of 16 images. Khattak et al. (2021) made a model which can detect infection in the citrus-based plant using their fruits and leaves and found an accuracy score of 94.55%. They have used CNN for this model. Another model where CNN is used as a detecting algorithm was proposed by Chowdhury et al. (2021). They did their research on tomato plants and found an accuracy score of 99.89%. Lastly, Mostafa et al. (2021) did their research on the guava plant using Deep Convolutional Neural Network (DCNN) structure and got an accuracy score of 97.74%.

3. Fruits and Vegetable Disease Detection

Disease detection of fruits and vegetables using Machine learning algorithms mainly involves five steps: Dataset collection, Data pre-processing, Segmentation, Feature selection, and Classifier.

3.1. Dataset

The primary data in fruit and vegetable disease detection applications are image data. Some of the main sources of collected data are google and various GitHub repositories. Some researchers also collected images from the fruits and vegetables available in the supermarket. Images acquired by different researchers are shown in Table 2.

3.2. Pre-processing

Images are acquired by various techniques consisting of multiple noises that worsen the aspect of an image. For this reason, the image without any pre-processing cannot contribute the appropriate data. Pre-processing enhances the image data, which overcomes reluctant misshaping and enlarges the image features essential for the processing and building an appropriate image (which is degraded form) than the original image for a specific application. The techniques used for image pre-processing for machine learning algorithms are Contrast enhancement and Rescaling.

3.2.1. Contrast Enhancement

The following equation is used to filter the contrast of any given image by assigning a constant intensity value to the pixel of that image using the histogram of an image. It is done with the help of the histogram equalization method.

$$H(P_{(x,y)}) = round\left(\frac{f_{(cdf)}(P_{(x,y)}) - f_{(cdf)min}}{(R \times C) - f_{(cdf)}} \times L - 1\right) \quad (1)$$

Here, $f_{(cdf)}$ = cumulative frequency of the gray level

$f_{(cdf)min}$ = minimum value of cumulative distribution function

$f_{(cdf)}(P_{(x,\ y)})$ = intensity of the current pixel,

R and C = product of several pixels in rows and columns

L = number of intensities

3.2.2. Rescaling

The images have been collected from different sources and captured with different devices. So, the images are not of the same size, therefore the images need to be resized.

3.3. Segmentation

Image segmentation is required to separate a digital image into distinct areas (Bhargava and Bansal 2021).

3.3.1. K-Means Clustering

K-means clustering is a type of vector quantization that divides n numbers of data into k number of clusters. Every observation is assigned to the cluster with the closest mean which acts as the prototype for the cluster. As a result, Voronoi cells are created in the data space (Vertica, 2022). Within-cluster variances are minimized using K-means clustering (squared Euclidean distances). Regular Euclidean distances, on the other hand, are not the more difficult Weber problem: the mean minimizes squared errors, but only the geometric median reduces Euclidean distances. Better Euclidean solutions can be produced, for example, by employing k-medians and k-medoids. The supervised KNN classifier, which is commonly confused with k-means due to its name, is connected to the unsupervised k-means algorithm. Using the cluster generated by k-means, the 1-nearest neighbor classifier is used to categorize incoming data into existing clusters.

3.3.2. Thresholding

Thresholding is an image segmentation technique that applies to the change of the pixels of an image to make it easier to analyze. It turns into a binary image from a color or grayscale image using thresholding. While the user can or should pick the threshold T manually in some circumstances, the user often prefers the threshold to be established automatically by an algorithm. The threshold should be the "best" threshold in certain circumstances, separating the brighter foreground items from the darker background objects. The threshold for picture intensity (image brightness) can be preset manually or automatically by some applications. Pixels whose bit values are zero turn black, whereas those with a bit value larger than zero become white (a bit value of one).

3.4. Feature Selection

Feature selection is the process of selecting the most important features to input in machine learning algorithms. Feature selection techniques are implemented to reduce the number of input variables by eliminating redundant or irrelevant features and narrowing down the set of features to those most relevant to the machine learning model. Some of the most common feature selection methods are:

3.4.1. Global Color Histogram (GCH)

GCH is a common term in machine learning. Color is regarded as one of the most important forms of visual representation. Changes to an image's size, rotation, or translation into different forms do not affect the image's color. Color space, color reduction, and the color feature extraction procedure are all factors to consider when extracting color characteristics from a picture (Han et al., 2022).

3.4.2. Color Coherence Vector (CCV)

CCV works by ranging each pixel as either coherent or incoherent. For each color, the number of coherent versus incoherent pixels is stored in a CCV. By separating pixels from incoherent pixels, it is possible to make finer distinctions than color histograms. They're all partially connected hidden layers, with the output layer being the fully connected layer at the end. The size of the input image is mirrored in the output shape (Pass et al., 1997).

3.4.3. SVM

SVM is a collection of supervised learning methods for classification, regression, and detecting outliers. The goal of the SVM algorithm is to find the best line or decision boundary for categorizing n-dimensional space into classes in the future so that new data points can be easily placed in the appropriate category. The best decision boundary is known as a hyperplane. Making a straight line between two classes is how a simple SVM works (McGregor and M, 2020). There are two types of SVM. Linear SVM is used for linear regression and classification problems. Kernel SVM is used for non-linear data. It can fit a hyperplane instead of a two-dimensional space with more features which has more flexibility for non-linear data.

3.4.4. Local Binary Pattern (LBP)

LBP is a robust method for image feature extraction in image processing. LBP reflects the correlation among pixels within a local area (usually a 3×3 area) which represents the local information (Bingham et al., 2015). By applying LBP, texture pattern probability can be precise in a histogram. LBP values need to be determined for all the image pixels (Prakasa, 2016).

3.4.5. Gray level co-occurrence matrix (GLCM)

It is a typical texture-based feature extraction model. The GLCM determines the relationship between pixels by performing an operation according to the second-order statistics in the images (Öztürk and Akdemir, 2018).

3.4.6. LVQNN

LVQNN is a superior classification technique for digital pictures. In general, it's a good idea to prepare data for LVQNN the same way it is done for KNN.

Table-1: Comparison of various classification methods for disease detection of vegetables and fruits

<table>
<tr><th>Work</th><th>Fruit/Vegetable</th><th>Affected Part</th><th>Pre-processing</th><th>Algorithm</th><th>Accuracy</th></tr>
<tr><td rowspan="3">Behera et al. (2018)</td><td>Mandarin</td><td>Leaf</td><td>Image enhancement</td><td>K-means clustering</td><td rowspan="3">90%</td></tr>
<tr><td rowspan="2">Orange</td><td>Stem</td><td rowspan="2">Lab color transformation</td><td rowspan="2">SVM</td></tr>
<tr><td>Fruit</td></tr>
<tr><td>Zaki et al. (2021)</td><td>Onion</td><td>Young leaf</td><td>Image pre-processing</td><td>CNN</td><td>77.05% and 85.47% for a batch size of 8 and 16 dataset</td></tr>
<tr><td rowspan="2">Singh et al. (2019)</td><td rowspan="2">Mango</td><td rowspan="2">Leaf</td><td>Contrast enhancement</td><td rowspan="2">MCNN</td><td rowspan="2">97.13%</td></tr>
<tr><td>Rescaling of the image data</td></tr>
</table>

JJMCOE and J (2017)	Apple	Fruit	Image enhancement	K-Means clustering segmentation algorithm LVQNN	More than 95%
Ullah et al. (2019)	Apple Bell paper Blueberry Cherry(sour) Corn(maize) Grape Orange Peach Potato Raspberry Soybean Strawberry Squash Tomato	Leaf	Contrast enhancement Image resizing	CNN	97.33%
Ouyang et al. (2012)	Strawberry	Fruit	Top hat transform Median filtering for removing noise interference	Image segmentation algorithm	99%
Pujari et al. (2013)	Grape Mango Pomegranate	Fruit	None	ANN classifier	84.65% and 76.6% for healthy and diseased fruits
Qin et al. (2009)	Citrus	Fruit	None	SID method for classification	96.2%

Khattak et al. (2021)	Citrus	Fruit and leaves	Keras image pre-processing using data generator class and API	CNN	94.55%
Yang et al. (2022)	Apple	Fruit	Grayscale processing Binarization Enhancement Processing Feature extraction	Sorting image recognition system	99.70%
Doh et al. (2019)	Citrus	Fruit	Image segmentation	ANN	SVM (93.12%)
			Feature extraction	SVM Phenotyping K-means clustering	ANN (88.96%)
Chowdhury et al. (2021)	Tomato	Leaves	Resizing and Normalizing Leaf segmentation Augmentation	CNN	99.89%
Mostafa et al. (2021)	Guava	Plant	Image acquisition Illumination	DCNN structures (AlexNet, SqueezeNet, GoogLeNet, ResNet-50, and ResNet-101)	97.74%

Table-2: Features of the datasets obtained by different researchers

Work	Fruit/Vegetable	Total Number of Image Data	Image Source
Ullah et al. (2019)	Apple	2,17,204	Internet
	Bell paper		Smartphone
	Blueberry		
	Cherry(sour)		
	Corn(maize)		
	Grape		
	Orange		
	Peach		
	Potato		
	Raspberry		
	Soybean		
	Strawberry		
	Squash		
	Tomato		
Singh et al. (2019)	Mango	2200	Real-time image
			GitHub repository
Zaki et al. (2021)	Onion	1000	Real-time image
JJMCOE and J (2017)	Apple	24	Real-time image
Khattak et al. (2021)	Citrus	2293	Different dataset
Chowdhury et al. (2021)	Tomato leaves	18,161	PlantVillage tomato leaf images

4. Comparative Analysis

After reviewing the works of multiple researchers, we can summarize that different algorithms work best on different fruits and vegetables. In Figure 1, we can get a brief idea about the accuracy rate for detecting disease in mango plants by using different algorithms.

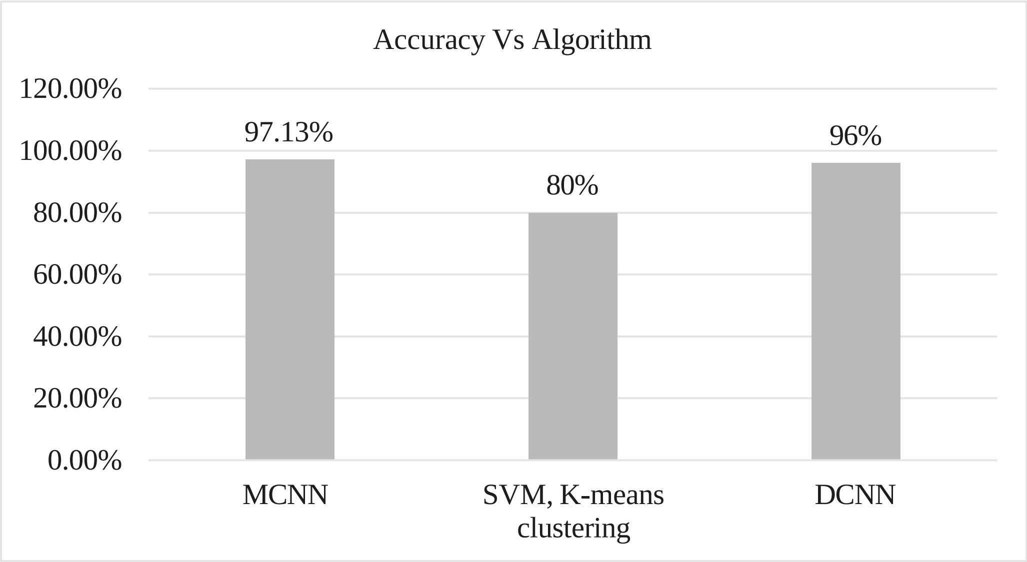

Figure 1: Accuracy vs algorithms for Mango disease detection

We have collected the data from the work of multiple researchers. Singh et al. (2019) used MCNN and found an accuracy score of 97.13%. Mia et al. (2020) proposed a model that used SVM and K-means clustering as detection algorithms and the accuracy rate was 80%. Ashok et al. (2021) used DCNN and found an accuracy of 96%. So, we can understand that MCNN has the highest score for accuracy in detecting disease in mango plants. By analyzing the comparison, we observed that MCNN shows exceptionally high accuracy in image recognition problems. It can also detect necessary features without any human guidance. On the other hand, for SVM and K-means clustering with an 80% accuracy rate, we observed if the dataset is large then SVM is not suitable as well as for K-means clustering, it is generally prone to biases.

We can conclude that MCNN is the best among the algorithms we analyzed due to its high accuracy rate and other features.

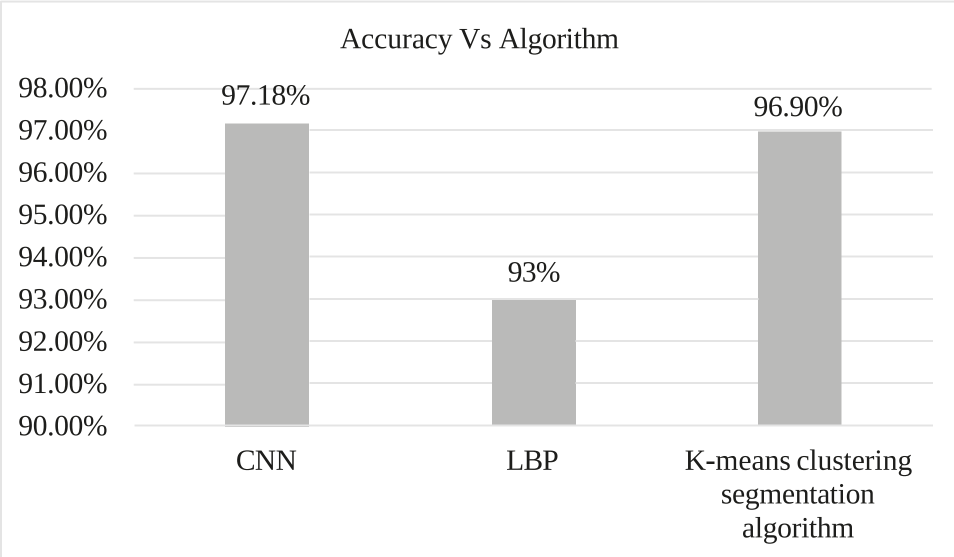

Figure 2: Accuracy vs algorithms for Apple disease detection

Figure 2 shows a bar chart of accuracy Vs different algorithms for detecting diseases in Apple. Khan et al. (2021) used CNN and found an accuracy score of 97.18%. 18%. Dubey et al. (2012) used LBP to complete the detection and found an accuracy of 93%. In the research of JJMCOE and J (2017), they used the K-means clustering segmentation algorithm and LVQNN and found an accuracy of 96.6%. So, we can say that the CNN algorithm has the best accuracy rate for detecting apple diseases. By analyzing the comparison of figure 2, we observed an accuracy rate of 97.18% for CNN. From the earlier comparison, we know that it gives a high accuracy rate in image recognition which is one of the notable features of CNN. In the figure, the LBP shows an accuracy rate of 93% which is the lowest among the other two algorithms. One of the disadvantages is its level of recognition is still low.

We can conclude that CNN is the best among the algorithms we analyzed due to its high accuracy rate and other features.

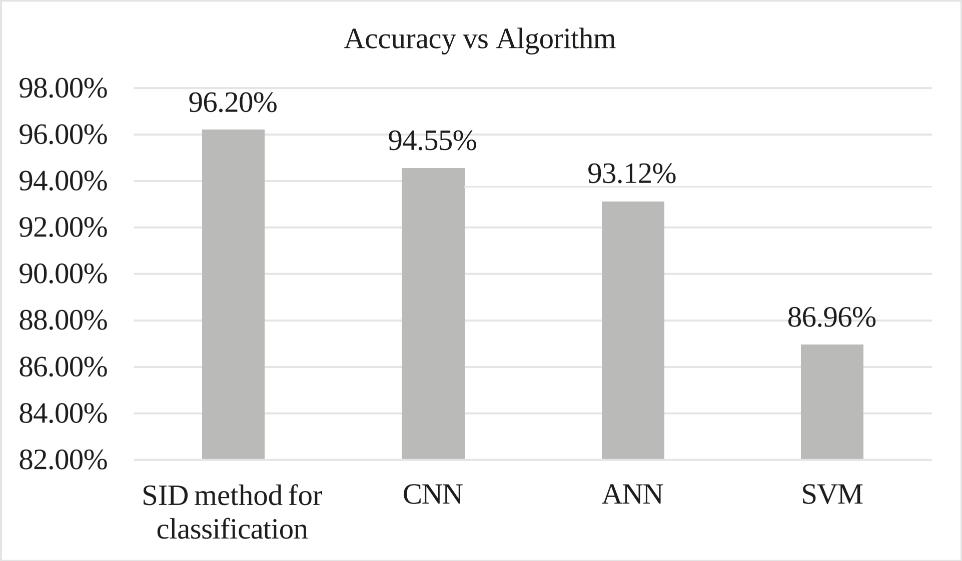

Figure 3: Accuracy vs algorithms for Citrus disease detection

Figure 3 represents a bar chart of different machine learning algorithms along with their accuracy rate to detect diseases in citrus fruit and leaves. From the above chart, we distinguished that the SID method has the highest accuracy rate. One of the major reasons for this value is that the SID method is an image-based classification procedure that is used to discriminate citrus diseases from other confounding diseases. On the other hand, SVM has the lowest accuracy rate among all other methods because it is sensitive to noisy attributes and requires a large search time.

We can conclude that for citrus disease detection SID method for classification is most suitable.

5. Conclusion

This paper highlights detailed descriptions of different algorithms for disease detection within fruits and vegetables along with applied pre-processing factors. The most significant features of agricultural by-products are their size, texture, and color. After reviewing and studying many research papers we came to a position where we can say which algorithm is best and will give high accuracy in detecting diseases in fruits and vegetables. More precise results can be obtained if researchers review papers with a large number of datasets which will allow them to acquire higher accuracy rates and will help them to understand which machine learning algorithm is appropriate for detecting diseases in any kind of agro-based products. We can conclude that this comparative study of different machine learning algorithms can be applied to early-stage detection of commonly available fruits and vegetables preventing massive economic loss with an increase in GDP rate as a result of which any country can progress in doing business with agro-based products without any complication.

References

Ashok, V., & Vinod, D. S. (2021). A novel fusion of deep learning and android application for real-time mango fruits disease detection. In Intelligent System Design. Springer, Singapore, pp. 781-79.

Behera, S. K., Jena, L., Rath, A. K., & Sethy, P. K. (2018, April). Disease classification and grading of orange using machine learning and fuzzy logic. In 2018 International Conference on Communication and Signal Processing (ICCSP). IEEE, pp. 0678-0682.

Bhargava, A., & Bansal, A. (2021). Fruits and vegetables quality evaluation using computer vision: A review. Journal of King Saud University-Computer and Information Sciences, 33(3), pp. 243-257.

Bingham, E., Kaski, S., Laaksonen, J., & Lampinen, J. (Eds.). (2015). Advances in independent component analysis and learning machines. Academic Press.

Chowdhury, M. E., Rahman, T., Khandakar, A., Ayari, M. A., Khan, A. U., Khan, M. S., ... & Ali, S. H. M. (2021). Automatic and reliable leaf disease detection using deep learning techniques. AgriEngineering, *3*(2), pp. 294-312.

Doh, B., Zhang, D., Shen, Y., Hussain, F., Doh, R. F., & Ayepah, K. *(2019, September). Automatic citrus fruit disease detection by phenotyping using machine learning*. In 2019 25th International Conference on Automation and Computing (ICAC). IEEE, pp. 1-5.

Dubey, S. R., & Jalal, A. S. (2012, November). Detection and classification of apple fruit diseases using complete local binary patterns. In 2012 Third International Conference on Computer and Communication Technology. IEEE, pp. 346-351.

Han, Ju & Ma, Kai-Kuang. (2002). Fuzzy color histogram and its use in color image retrieval. IEEE transactions on image processing: a publication of the IEEE Signal Processing Society, 11, pp. 944-52.

JJMCOE, J. (2017). Detection and Classification of Apple Fruit Diseases using K-means clustering and Learning Vector Quantization Neural Network.

Khan, A. I., Quadri, S. M. K., & Banday, S. (2020). Deep learning for apple diseases: classification and identification. arXiv preprint arXiv:2007.02980.

Khattak, A., Asghar, M. U., Batool, U., Asghar, M. Z., Ullah, H., Al-Rakhami, M., & Gumaei, A. (2021). Automatic detection of citrus fruit and leaves diseases using deep neural network model. IEEE Access, *9*, pp. 112942-112954.

McGregor, M. (2020, July 02). SVM machine learning tutorial – what is the support vector machine algorithm, explained with code examples. Retrieved June 23, 2022, from https://www.freecodecamp.org/news/svm-machine-learning-tutorial-what-is-the-support-vector-machine-algorithm-explained-with-code-examples/ (Last accessed 21 July 2022)

Mia, M., Roy, S., Das, S. K., & Rahman, M. (2020). Mango leaf disease recognition using neural network and support vector machine. Iran Journal of Computer Science, 3(3), pp.185-193.

Mostafa, A. M., Kumar, S. A., Meraj, T., Rauf, H. T., Alnuaim, A. A., & Alkhayyal, M. A. (2021). Guava Disease Detection Using Deep Convolutional Neural Networks: A Case Study of Guava Plants. Applied Sciences, *12*(1), 239, pp. 1-19.

Ouyang, C., Li, D., Wang, J., Wang, S., & Han, Y. (2012, October). The research of strawberry disease identification based on image processing and pattern recognition. In International Conference on Computer and Computing Technologies in Agriculture. Springer, Berlin, Heidelberg. pp. 69-77.

Öztürk, Ş., & Akdemir, B. (2018). Application of feature extraction and classification methods for histopathological image using GLCM, LBP, LBGLCM, GLRLM and SFTA. Procedia computer science, *132*, pp. 40-46.

Pass, G., Zabih, R., & Miller, J. (1997, February). Comparing images using color coherence vectors. In Proceedings of the fourth ACM international conference on Multimedia, pp. 65-73.

Prakasa, E. (2016). Texture feature extraction by using local binary pattern. *INKOM Journal*, *9*(2), pp. 45-48.

Pujari, J. D., Yakkundimath, R., & Byadgi, A. S. (2013). Grading and classification of anthracnose fungal disease of fruits based on statistical texture features. International Journal of Advanced Science and Technology, 52(1), pp. 121-132.

Qin, J., Burks, T. F., Ritenour, M. A., & Bonn, W. G. (2009). Detection of citrus canker using hyperspectral reflectance imaging with spectral information divergence. Journal of food engineering, 93(2), pp.183-191.

Singh, U. P., Chouhan, S. S., Jain, S., & Jain, S. (2019). Multilayer convolution neural network for the classification of mango leaves infected by anthracnose disease. IEEE Access, 7, pp. 43721-43729.

Ullah, M. R., Dola, N. A., Sattar, A., & Hasnat, A. (2019, November). Plant Diseases Recognition Using Machine Learning. In 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART). IEEE, pp. 67-73.

VerticaPy. (2022). Retrieved 22 June 2022, from https://www.vertica.com/python/documentation_last/learn/KMeans/ (Last accessed 21 July 2022)

Yang, M., Kumar, P., Bhola, J., & Shabaz, M. (2022). Development of image recognition software based on artificial intelligence algorithm for the efficient sorting of apple fruit. *International Journal of System* Assurance Engineering and Management, *13*(1), pp. 322-330.

Zaki, M. A., Narejo, S., Ahsan, M., Zai, S., Anjum, M. R., & u Din, N. (2021). Image-based Onion Disease (Purple Blotch) Detection using Deep Convolutional Neural Network. Image, *12*(5), pp. 448-458.