***Abstract –* Food production plays a prominent role in the economy of any country. Farmers face many problems while cultivating the plants and they must take precautions at every stage of cultivation. Plants get diseases for various reasons like bacteria, insects, and fungus. Different diseases will show different effects, and some are similar. Crops will be damaged more if the farmer doesn’t take precautions during the early stage of the disease. Farmers who don’t have the proper knowledge to diagnose the plants will lose their yield. Due to the heavy population and busy work culture, it is hard for agriculture professionals to examine every farmer’s field and predict the disease. This process requires time and money too. This problem can be solved cost-effectively by using modern technologies. By using image processing and machine learning algorithms this issue can be resolved. In this project, the diseases in tomato leaves will be examined using images. The data from the images are extracted by using different vectorization methods and performing classification algorithms on each vector data to find the best model for this problem.**

**Keywords – Diseases, Machine learning, Image processing, Classification**

**Introduction**

All humans need food, and this is supplied by the agriculture field. Agriculture is a major sector and backbone of any country. Different countries implement different techniques and cultivate different crops according to their environmental and geographical conditions. Some major crops are produced by a plethora of countries like Tomato, Potato, and Corn. Policymakers are also concerned about the agriculture field because it will directly affect the people and the country's dependency on other countries. Reduction in agriculture production will rise the imports and this will change many factors in the country. Advancements in the agriculture field are much needed and some poor nations still working with old methods. Lack of knowledge in cultivation and diseases is a major issue for many farmers. Most of them implement traditional farming methods and they are unable to detect the diseases in the early stage. Modern methods are capable of detecting the diseases at early stages and help the farmers to take precautions against the disease. In this research, the Tomato plant is selected to show how these technologies can detect diseases using leaf images. The idea behind this research is to build a model which can predict the disease on the tomato leaves in the early stages. So, any farmer with minimum facilities and operatable knowledge on the computer can use this model irrespective of location, region, and time.

Tomatoes are one of the major crops in the world and it is cultivated in more than 150 countries. Since it is a nutrient-dense plant and gives good income after production, many farmers are interested to cultivate these plants. Every crop is vulnerable to certain diseases and tomato plants also get many diseases. If the farmer notices a disease, the farmer must wait for an agricultural professional to test the sample and confirm the result. This process will take so much time and the farmer have to wait so much time for the result. In poorer countries, there is no availability of agricultural professionals, and the farmers in remote places have to wait so long for the results. This waiting time will increase the spread of the disease as well as increase the risk of production reduction. These issues can be reduced using modern cutting-edge technologies. For prediction of the diseases in the tomato leaves, nine major diseases are selected. They are Bacterial spot, Early blight, Late blight, Leaf Mold, Septoria leaf spot, Target Spot, Tomato Mosaic virus, Tomato Yellow Leaf Curl Virus, and Two-Spotted Spider Mite. These diseases are very harmful to tomato plants and reduce their production. These diseases can be predicted by observing the plant leaves. For this research, the images of these diseases are collected from a trusted contributor(Huang and Chang 2020). Healthy Leaves pictures are also included along with these nine categories for reducing the bias and detecting the healthy leaves. So, the final model can be capable to find these nine diseases and healthy plants. The sample images of all the diseases and the healthy leaf are shown in figure (1).

A green leaf on a white surface

Description automatically generated with medium confidence A picture containing custard apple

Description automatically generated A picture containing ground

Description automatically generated A picture containing lettuce, mildew, vegetable

Description automatically generated A picture containing vegetable

Description automatically generated

Bacterial Spot Early Blight Late Blight Leaf Mold Septoria Leaf Spot

A green leaf on a grey surface

Description automatically generated with low confidence A green leaf on a grey surface

Description automatically generated with low confidence A green leaf on a grey surface

Description automatically generated with medium confidence A green leaf on a grey surface

Description automatically generated with medium confidence A green leaf on a grey surface

Description automatically generated with medium confidence

Target Spot Mosaic Virus Yellow Leaf Curl Two-spotted Healthy

Virus Spider Mite leaf

Figure .1 Tomato Leaves with different diseases

The computer cannot understand the image format for performing machine learning techniques. So, the images need to be converted to vector formats, and the converted data need to be saved in an optimized format. These processes are implemented using python programming language and different python libraries. Majorly, the Pillow library is used for the conversion of image data to vectors and the Scikit-learn library is used to apply machine learning algorithms. In this study, the images are converted to five different vector formats. They are Grayscale, RGB vector, Red channel, Blue channel, and Green Channel vector formats, and these are saved into a NumPy array. The RGB vector data extracted from the images are in three-dimensional array format and the remaining all are in the two-dimensional array format. These vector formats are flattened to a one-dimensional array for the analysis. The disease name and mean of each vector form are saved into different CSV files. The major problem with this data is bias in the count and high dimensions. So, to reduce these issues the analysis will be performed on the four different types of data. Those are original data, data with a maximum count of 600 in each category, 32×32-dimension image data, and 64×64-dimension image data. These four different types of analysis are explained using Table 1.

|  |  |
| --- | --- |
| **Data Type** | **Description** |
| Original Images from contributor | Data with 227×227-dimension images |
| Maximum count 600 | Data with a maximum count of 600 in each category |
| 32×32 | Data with 32×32-dimension images |
| 64×64 | Data with 64×64-dimension images |

Table .1 Data used in this study

To find the best algorithm for this project, the contents in the NumPy array are used as the features and the disease in the CSV file are used as the targeted variables. Different classification algorithms like the Logistic classification algorithm and KNN algorithm will be implemented on the data. The algorithms will be selected based on the related works. The accuracies will be noted for all the vector forms and the different algorithms. For further analysis, the dimensions of the data will be reduced using dimension reduction techniques, and then apply the machine learning algorithm to the reduced data and find the best model based on the accuracy and different metrics. In the final stage, the vector method which gives good accuracy and shows reasonable metrics will be picked for predicting the disease in new images and for the practical implementation. The major advantages of this project are economic and extendibility factors. Once the model is built, it can be used as many times as needed without investing more money. This project can be extended to other crops also and the only problem is to collect the data of each crop for performing analysis.

Helping farmers and motivating them to use the new technologies is the aim of this project. I believe the breakthrough in machine learning will help in the detection of disease using images. Image processing is the major field for this project and this project needs many tools and libraries for the analysis.

**Research Objectives:**

This study is intended to explore and find answers to the below research objectives.

* Explore different vectorization methods for image processing and check which method is more suitable for tomato leaves disease detection.
* Perform different scaling and modeling techniques and find which method is more suitable for this problem.
* Perform different classification algorithms and evaluate the metrics to check which algorithm is good for disease detection.
* Evaluate the results and find which features are giving good results in disease detection.

This paper is structured as follows: 2. Literature Review, 3. Methodology, 4. Data Processing, 5. Ethical Considerations and SWOT Analysis, 6. Evaluation and Future Plan, and 7. References.

**3. Methodology**

The methodology is the route map or step-by-step process of any project. There are many methodologies available for the practical implementation of data analytical research. This study follows the structure and process of the CRISP-DM methodology. CRISP-DM is an acronym for the Cross-industry standard process for data mining. CRISP-DM is a powerful and robust method for implementing the research. This is an open-source methodology and is widely used by data practitioners for research. It consists of six major phases and the approach follows a step-by-step implementation of each phase. The phases in the CRISP-DM methodology are Business understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Other advantages of this methodology are flexibility and dependencies. The researcher can easily move back to the sequence or restart the phase based on the research and the dependency of the result in each phase can be observed clearly. All these processes start after the data collection process.

Data

Collection

Figure 2. Methodology of the project

**3.1 Data Collection**

In the data collection phase, the primary data needed for this research is the images. It is not possible to take the image data from a particular region or a particular season because this will reduce the efficiency of the model. So, the best way to collect data is by taking data from the internet. There are some ethical and economic considerations for collecting data from the internet. The data should be from an open-source with no restriction on usage, data should be relevant, and data should be hygiene. The data which follows all these principles should be taken for further analysis. Data relevance and data hygiene is also the most important factors because irrelevant or blurry images are not good for the analysis. After searching and exploring different related works and datasets related to the tomato leaves images, there is a website named Mendeley which gives information about the tomato leaves. It consists of images about nine different diseases infecting the tomato leaves along with the healthy tomato leave images. The data is open-source, hygiene, and relevant to the tomato leaves.

**3.2 Context Understanding**

The next phase of the research is to understand the context of the research. The context of the research is to help farmers by implementing a model which is capable of detecting diseases infecting tomato leaves using image processing and machine learning techniques. To achieve this context, the research should be conducted on different vector types and images of different sizes. A programming language is needed to start the research and perform an analysis of the images. There is a need for different python libraries for converting the images to different vectorization models and performing the machine learning algorithms. Python programming language is used for the overall analysis of the data because it is one of the popular and robust languages for data analysis. For handling the data different python libraries are needed and the major libraries are Pandas and NumPy. And Pillow library can be used for the conversion of the images to vector forms. Finally, the sci-kit learn library is needed for analyzing the data using machine learning techniques. There is no special equipment or permissions needed for proceeding with the analysis because the data is taken from an ethically significant source and the output of the models is shown on the computer. There is no personal information or human-related data used in this model.

**3.3 Data Understanding**

In the data understanding stage, the raw data downloaded from the source is examined under different factors. Firstly, the data is already separated according to the diseases and placed in different folders with the disease name as the folder name. All the images are in jpg format and 227×227 size. All the images are in colour format and there are no grayscale images. There is a total of ten folders among them, nine are related to diseases infecting tomato leaves and one folder contains images of healthy tomato leaves. The nine diseases are Bacterial spot, Early blight, Late blight, Leaf mold, Septoria leaf spot, Target spot, Tomato mosaic virus, Tomato yellow leaf curl virus, and Two-spotted spider mite. The count of the images is not equal and there are uneven counts in the images on all folders. This may cause bias in the model so there is a need to take another dataset with a maximum count of 600 for each disease. All images are taken by placing the leaf on a surface and some images have shadows on the back. So, there may be chances of miscalculations of the images in the analysis.

**3.4 Data preparation**

Data preparation is the major step in the analysis, and it needs to be carried out very carefully. Image data needs to be changed to different vector forms for analysis. Image is made up of a combination of pixels and each pixel has an RGB value. The colour of the pixel is displayed based on the RGB value. For grayscale images, there is no RGB value but there is a single colour value for each pixel. In the preparation phase, these vector values of all pixels in each image are extracted and saved in the NumPy array. The array of vectors is saved into a NumPy file and the other features like mean and disease name are saved in the CSV file. These vector values can be extracted using functions from the pillow library. The images are opened in the python library and the RGB vector of the image is taken into a list. Later, the image is converted to grayscale, and its vector value is saved to an array. The RGB vector is a three-dimensional array and the grayscale vector is a two-dimensional array. The grayscale vector is flattened to a one-dimensional array and saved to a NumPy file in the local system. Coming to the RGB vector, the red channel vector data, the green channel vector data, and the blue channel vector data are extracted from the RGB vector and saved into different lists. The name of the disease and the mean of the vector are saved in a CSV file for all the formats. Finally, there will be two files each for the five vector methods. The same process will be repeated by resizing the images to 32×32 size and 64×64 size. The same process will be repeated by taking a maximum of 600 images of each disease in all three formats. The preparation of the data is explained more clearly in the data pre-processing section of this paper.

**3.5 Modeling**

In the modeling phase, The converted data from the preparation phase is used to analyze the prediction by implementing different machine learning algorithms. The idea is to analyze the different data types separately. The array from the Numpy file is used as the features and the disease data in the CSV files will be used as the target variable. Since this is a classification problem, algorithms that are popular and used in related works like Logistic, SVM, and KNN classifiers will be performed on the data. From the available data, 80 percent of the data will be used for training and 20 percent of the data will be used for the testing. In the first stage, the data will be analyzed without doing the feature reduction techniques and note the accuracy and other metrics of the different models. In the later stages, the data after doing feature scaling will be analyzed. Different scaling methods will be also performed on the data and observe which method will be giving good prediction.

**3.6 Evaluation**

Evaluation of the models is not an easy task. Looking at the accuracy alone will not give the best analysis model. The models will be analyzed using different metrics and the model which gives reasonable outputs in all the metrics will be moved to the deployment phase. The evaluation metrics that are considered in this research are accuracy, precision, confusion matrix, F1 score, recall, and classification report. In the first stage, the models with more than 70 percent accuracy will be taken. Later, the metrics of all the models will be examined to find the best model for the analysis. Apart from accuracy, the remaining evaluation methods help in choosing the best methods among the high accuracy models. If there is any problem with the results then the process will be started again from the data preparation stage. That’s why it is important to evaluate the models carefully throughout the examination process.

**3.7 Deployment**

Finally, the model which shows good metrics will be taken for the deployment. It is better to pick the best methods and train the method for deployment to ensure the performance of the code. The major goal of the deployment is to evaluate the context of the model. After the deployment, we can examine how the model is working and how it is predicting the diseases of the tomato leaves. Building the model again with the best methods will increase the efficiency of the model. Once the model is deployed we can use it for the analysis as many times as possible. If the model fails in the deployment phase in terms of context then the entire process should be re-run again.

**4. Data Pre-processing**

Images are understandable to humans and normally people will get more information from an image. This is not the case for analyzing the images for the machines. Machines cannot understand the image directly so the data need to be converted to different vector forms for the analysis. Many Python libraries are available to extract the features from the image data. In this study, Pillow Library is used to convert the image data to vector form because it is a free and one of the widely used libraries to extract the features from images. Pillow library can do many functionalities on the images from basic operations like rotation and resizing to colour space conversions. It supports manipulations of many image formats, and it is easy to implement using python 3. Image data takes time to process, and it needs some powerful machines if there are more images. Another problem with image processing is time, image conversion takes more processing time, and it will slow down the overall process time. To counter these problems, we can save the vector formats and other information in the local system. Once the information is saved in the local system, we can use it for many purposes and in the future with reduced process time.

In this study, the data is saved in different sizes in separate folders. The data is converted to different vector forms and then data is saved in the local system using two formats for each vector form. The pixel data of different channels are saved in a separate NumPy array and the data about different features are stored in the CSV file. We can save the image vectors in a CSV file if there are images with fewer pixels and fewer dimensions. We are dealing with more than five thousand images in all stages and there are images of 227\*227 size. If this data is saved in the CSV format, It will take more memory space and take a long time to process. Due to these problems, the CSV format is not a good pick for saving the image vector data in this case. The simple and efficient method of saving the vector data is by using the NumPy array file npy. This will save the memory and we can read the data as an array easily. The main advantage of saving the data to vector form is memory saving and ease of use. We need a CSV file for saving the disease data to implement the classification of images and perform the machine learning algorithm on the data. The steps involved in the data pre-processing are explained in Figure 2.

Images placed with correct folder names of their diseases

Resizing Images to different sizes and save to different Folders

To check the count bias: Images with a maximum count of 600 is taken for all the diseases

Converting images to different vector formats and saving it in a NumPy array

Saving the features and disease in a CSV file

Figure 3. Data Pre-processing

The images are already separated using folders based on the disease type by the contributor. The name of the folder contains the disease name and some numbers. So, the first step in pre-processing is to save the folders with the correct name. Later the images are resized to 64\*64 and 32\*32 sizes and saved in a separate folder. There are images with uneven counts, so images with a maximum count of 600 are taken from all the subfolders to reduce the bias in the count of the images.

Text

Description automatically generated

Figure 4. Data stored in different folders based on the type

The images are converted to vectors and saved into a NumPy file, and the diseases are saved into a CSV file. By using the pillow library functions, the images are converted into Grayscale, RGB format, red channel, green channel, and blue channel. The converted vector data from the images are in the array form and the data is flattened to the one-dimensional array before saving them. We can even use this data and reconstruct the image of each vector.

1. The below figures show how the images are converted to different vectors (Machine Learning - Going Furthur with CNN Part 2 - DEV Community n.d.).

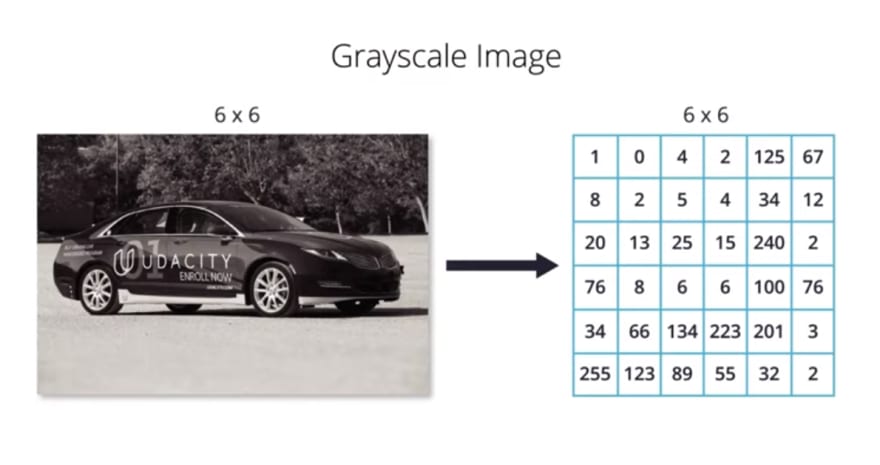
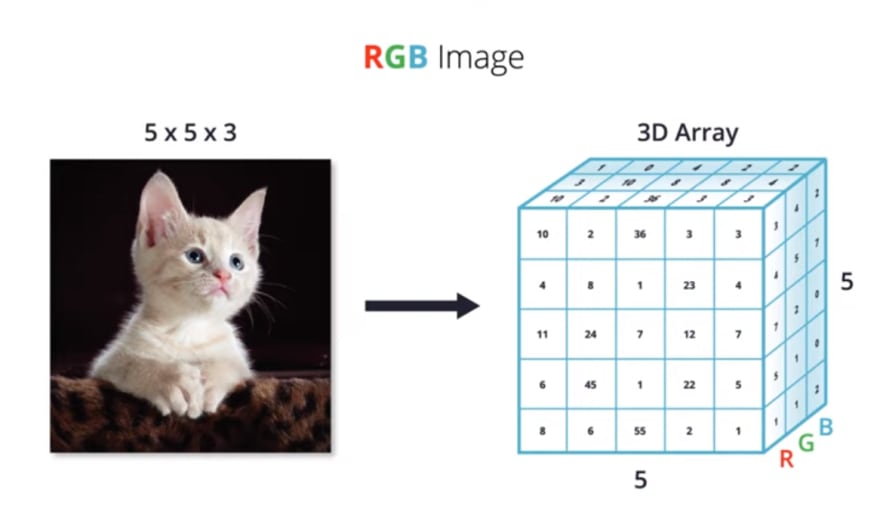
 

Figure 4. grayscale to vector. Figure 5. RGB image to vector format

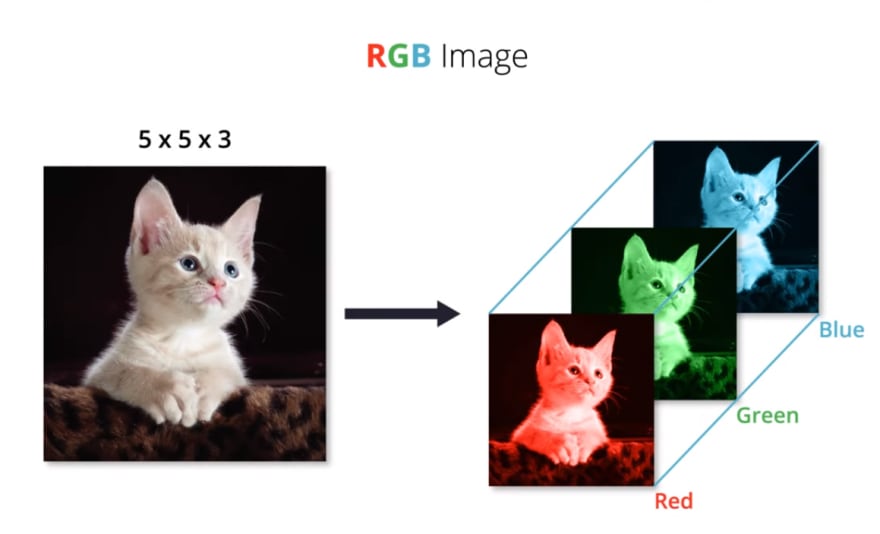
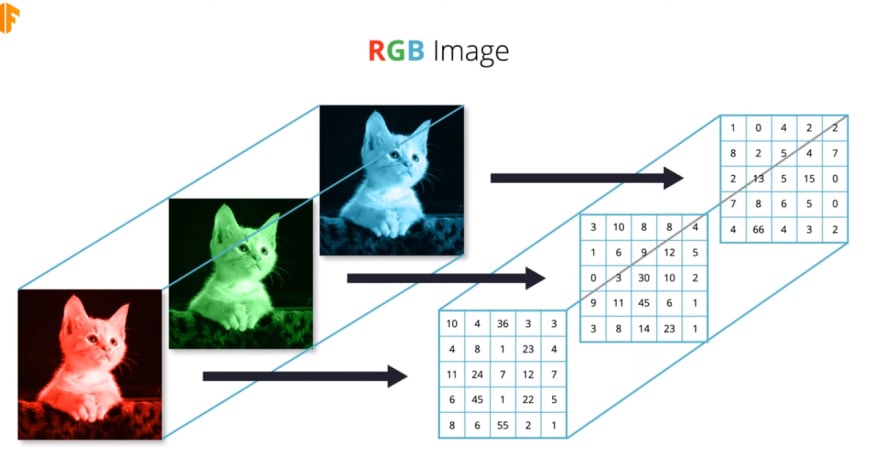
 

Figure 5. Image to RGB vector format. Figure 6. RGB to individual vectors.

For each category, different vector types of information are saved to different files individually. This information can be used for further analysis and to implement the machine learning algorithms in the data. The below figure shows how the data is saved in the local system after taking vectors in the npy file and features in the CSV files.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 7. data of different vector forms

The below table shows how the data is saved in all the folders after Pre-processing:

|  |  |
| --- | --- |
| Folder | Sub-Folders and data available |
| 32 | data(images with 32\*32 size), vector(.npy and CSV files of images in data), no\_bias(images with maximum count 600 of each disease), no\_bias\_vector(.npy and CSV files of no\_bias folder images). |
| 64 | data(images with 64\*64 size), vector(.npy and CSV files of images in data), no\_bias(images with maximum count 600 of each disease), no\_bias\_vector(.npy and CSV files of no\_bias folder images). |
| no\_bias | data(contains images with a maximum count of 600 for each disease with 227\*227 size). |
| Preprocessed\_data | data(images from contributor), vector(.npy and CSV files of images in data folder). |

Table 2. Data stored in the system after pre-processing

**(V) Ethical Considerations and SWOT Analysis:**

Ethical considerations are crucial and mandatory for any research. Following the data scientist’s code of ethics will give more value to the research and it will help to reduce the problems of the data. In some cases, the project intends to rectify a problem, but it will fail due to the lack of ethical knowledge and improper ethics. In practice, negligence of ethics will create a life and death situation. It is the responsibility of the researcher to evaluate the ethical considerations of a project before taking the data and following the ethics throughout the process. In this study, images of tomato leaves are used for disease prediction and image data also have many benefits and harms. It is easy to understand from an image for humans than the text or tables. The images used in this project are taken from a public website and except for credits, there are not many permissions needed for the usage of these images. The images are contributed by Mei-Ling Huang and Ya-Han Chang. The information is available on the Mendeley data website(Huang and Chang 2020). It is free to use, and modifications of the data are also allowed.

So, there is no ethical issue with the usage of the data but there are some ethically significant benefits and harms related to this data.

The major ethical benefits of using this image data are Human understanding, Timesaving, predictive accuracy, and Economic efficiency. Any person can see the images used in this research and get a basic idea about the diseases. Different diseases affect the plant leaves differently and we can see the effects of each disease using these images. By using this model in real-time, we can reduce the time needed to wait for the agriculture professional and the farmer can start the implementation of the precautions in the early stage. Since this is a classification problem, the accuracy of the model is one of the major metrics and the percentage of accuracy is easily understandable to most people. If a farmer wants to re-examine the crops after some time, then he must again pay for the testing. This model can be built once, and we can use it as many times as required without spending more money. By implementing this model, we can reduce the cost of testing the crops.

The major harms with data are harm to privacy, harm to security harms to fairness and justice, and harms to transparency and autonomy. Images data will be predicted wrongly if there are colour spots or light fields on the images(Dufaux 2021). This will cause the model to predict the data with less accuracy and this is harmful to the fairness and justice rules. The farmer needs to use pesticides based on the disease and the wrong prediction will destroy the crops. Overfit model may produce wrong outputs and using precautions without thinking may harm the crops. The researcher has to take care of these issues while building the model for detecting the tomato disease on the leaves. These are the major challenges with this data. There is no human-related information or unethical images are used in this analysis. The images of tomato leaves are used for the analysis and the credits are given to the contributor. So, there is no need to consider the privacy and security harms of the data.

**SWOT Analysis:**

SWOT is a strategy that organizes project-related ideas in a list. SWOT means Strengths, Weakness, Opportunities, and Threats. This helps in analyzing big problems systematically and more easily(SWOT Analysis: What Is It and How to Do It for Your Business n.d.). It is an initial step for properly organizing all the ideas and examine what are the internal and external points. Strengths and weaknesses come under internal origin and we have control over them. Weakness and opportunities are external origins, and we don’t have control over them(What Is a SWOT Analysis and How to Do it Right in 2021 (With Examples) n.d.). By applying this strategy, there is a clear idea about the effectiveness and impacts of this project.

|  |  |
| --- | --- |
| **Strengths**  1. Cost-effective: this can be implemented with less amount.  2. Accuracy: Normally, image processing models give good accuracy.  3. Time reduction: This will reduce the time taken for examining the disease.  4. Flexible vectors: The vector information is stored in the form of lists. So, this data less time to process the information. | **Weakness**  1. Time-consuming: Takes more time for converting the data to vector form.  2. Noise Effect: If there is noise in the images then it is not capable to predict correctly.  3. Reach financials: This model is built intended to help farmers. So, this needs financials to educate farmers about this model. |
| **Opportunities**  1. Extendable: This model can be used for other crops also.  2. Geographical: More than 150 countries farm tomatoes.  3. Partnership: Fertilizer companies invest in these types of technologies.  4. Trending technology: Image processing is a reliable and attractive technology. | **Threats**  1. New Technologies: Deep Learning provides more accurate predictions if more data is provided.  2. Services: Requires lots of services according to the crop. |

Table 3. SWOT Analysis of this project