POTATO LEAF DISEASE DETECTION

A Course Project report submitted

in partial fulfillment of requirement for the award of degree

BACHELOR OF TECHNOLOGY

in

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

by

A.SHRESHTA	**(2103A52043)**
K.KOUSHIK	**(2103A52054)**
G.RAHUL	**(2103A52049)**

Under the guidance of

Mr. Ramesh Dadi

Assistant Professor, Department of CSE.

SR UNIVERSITY

Department of Computer Science and Artificial Intelligence

SR UNIVERSITY

Department of Computer Science and Artificial Intelligence

CERTIFICATE

This is to certify that project entitled **“POTATO LEAF DISEASE DETECTION”** is the bonafied work carried out by **SHRESHTA , KOUSHIK , RAHUL** as a Course Project for the partial fulfillment to award the degree **BACHELOR OF TECHNOLOGY** in **ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING** during the academic year 2023-2024 under our guidance and Supervision.

Mr. D.Ramesh
Asst. Professor,
(CSE), S R University,
Ananthasagar, Warangal.

Dr. M.Sheshikala
Assoc. Prof. & HOD
S R University,
Ananthasagar, Warangal.

ACKNOWLEDGEMENT

We express our thanks to Course co-coordinator **Mr. D.Ramesh, Asst. Prof.** for guiding us from the beginning through the end of the Course Project. We express our gratitude to Head of the department CS&AI, **Dr. M.Sheshikala, Associate Professor** for encouragement, support and insightful suggestions. We truly value their consistent feedback on our progress, which was always constructive and encouraging and ultimately drove us to the right direction.

We wish to take this opportunity to express our sincere gratitude and deep sense of respect to our beloved Dean, School of Computer Science and Artificial Intelligence, **Dr C. V. Guru Rao**, for his continuous support and guidance to complete this project in the institute.

Finally, we express our thanks to all the teaching and non-teaching staff of the department for their suggestions and timely support.

ABSTRACT

With the enhancement in agricultural technology and the use of artificial intelligence in diagnosing plant diseases, it becomes important to make pertinent research to sustainable agricultural development. Various diseases like early blight and late blight immensely influence the quality and quantity of the potatoes and manual interpretation of these leaf diseases is quite time-taking and cumbersome. As it requires tremendously a good level of expertise, efficient and automated detection of these diseases in the budding phase can assist in ameliorating the potato crop production.

Previously, various models have been proposed to detect several plant diseases. In this paper, a model is presented that uses pre-trained models like for fine-tuning(transfer learning) to extract the relevant features from the dataset. Then, with the help of multiple classifiers results were perceived among which logistic regression outperformed others by a substantial margin of classification accuracy obtaining 97.8% over the test dataset.

The survey also examines how different machine learning technologies like k-nearest neighbor (KNN), Support vector machine(SVM),Random forest and Decision Tree can be used in potato leaf disease detection.

In this article, we have performed a comprehensive review of the literature to machine learning models from their datasets so that the ML algorithms can detect disease more efficiently and correctly. We have proposed the future work of using transfer learning combined with federated knowledge that could help the medical institutions and hospitals form a combined approach of performing medical image detection using real-time datasets. We have also explored the scope, future work and limitations of the proposed solution.

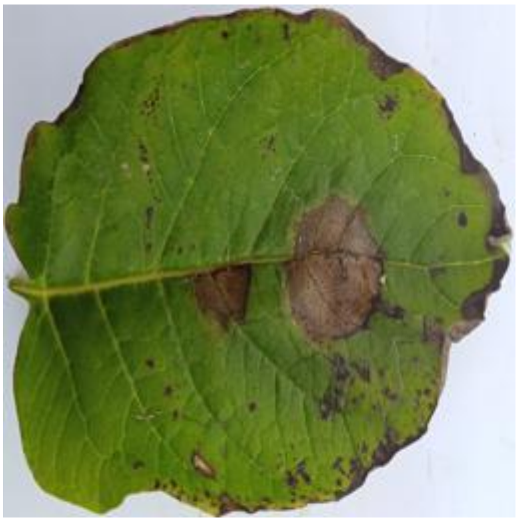
TABLE OF CONTENTS

INTRODUCTION

1.1 OVERVIEW

We will develop an end to end project which is based on pure machine learning. The reason behind building this project is to detect or identify potato leaf diseases, having a variety of illness. With our naked eyes we can’t classify them, but K means can do easily. You can’t believe when I tell you the error of some pre-trained Neural Network Architectures because it is approximately 3%, which is even less than the top 5% error of human vision. On the large-scale images, the human top-5 error has been reported to be as 5.1%, which is higher than pre-trained networks.

Normal



Diseased

1.2 PROBLEM STATEMENT

Farmers who grow potatoes suffer from serious financial standpoint losses each year which cause several diseases that affect potato plants. The diseases Early Blight and Late Blight are the most frequent. Early blight is caused by fungus and late blight is caused by specific micro-organisms and if farmers detect this disease early and apply appropriate treatment then it can save a lot of waste and prevent economical loss. The treatments for early blight and late blight are a little different so it’s important that you accurately identify what kind of disease is there in that potato plant. Behind the scene, we are going to use K means.

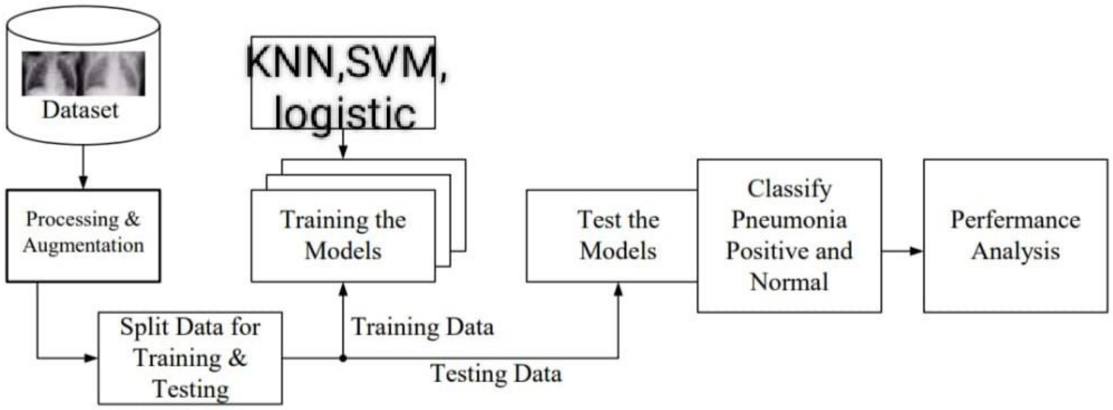
EXISTING SYSTEM

According to the search results, there are different types of existing systems for leaf disease detection based on machine learning. Some of them are:

- **Faster R-CNN**: A two-stage object detector that uses a region proposal network (RPN) to generate candidate regions and then classifies them using a K means.
- **ResNet**: A CNN architecture that uses residual connections to overcome the problem of vanishing gradients and improve accuracy. It can have different depths such as 50 or 101 layers
- **CheXNet**: A CNN model that is trained on a large dataset of leaf diseased plants.
- **DECNET**: A CNN model that uses a deformable convolution layer to adapt to the shape and size of the lung regions and improve the detection of leaf disease.

1.3 PROPOSED SYSTEM

We have anaylsed the various work done on medical image detection in the previous section. The experiments were performed based on available datasets. It has been observed that the machine learning models effectively detects medical images when the model is fed with a larger quantity of data. The use of ML algorithms has been proven effective in detection while compared to the traditional procedures mentioned in the literature review. ML models need a higher volume of data for effective training capable of achieving higher accuracy in detection.



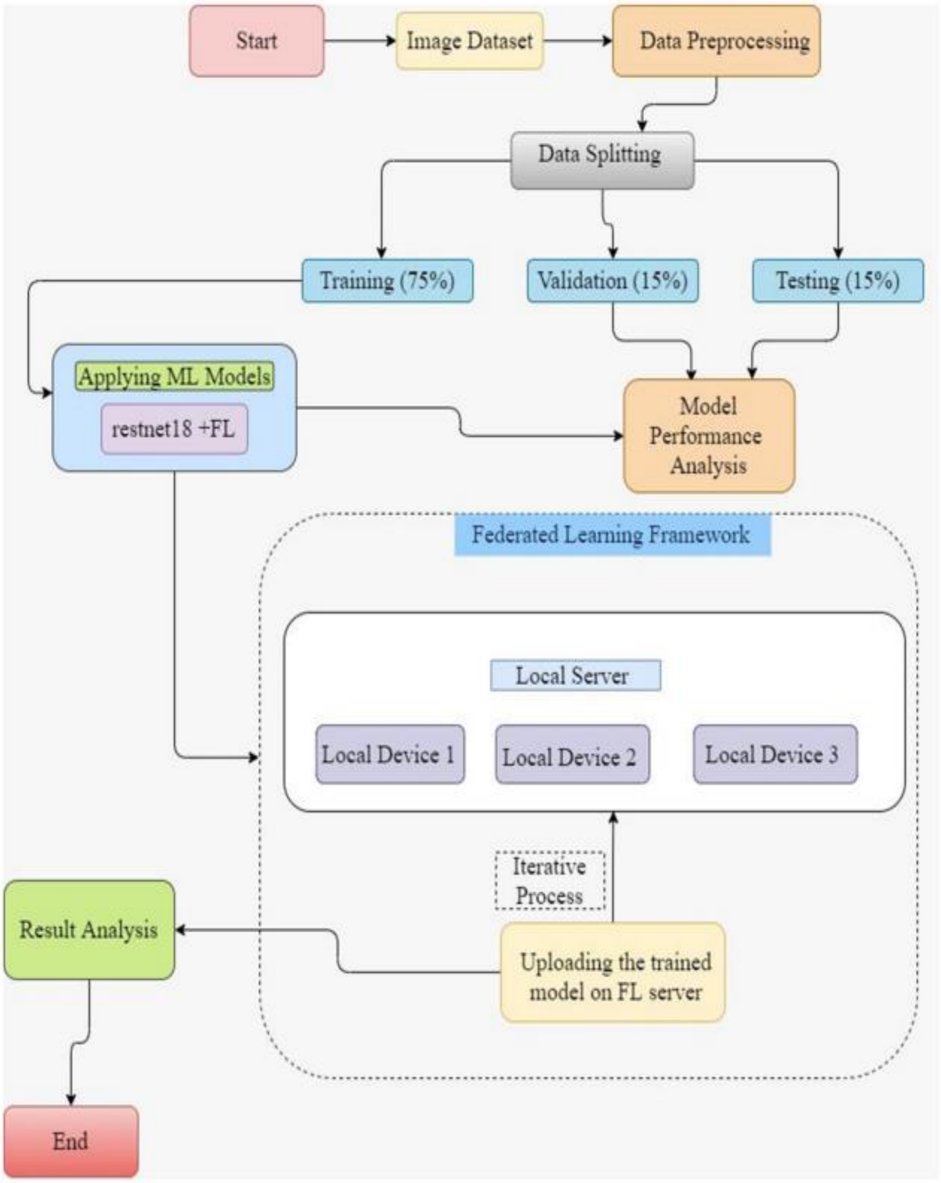
Proposed system

1.5 OBJECTIVES

The main objective of this research is to develop a potato leaf disease detection. The objective of this project is to build a model that can accurately detect the healthy ,bright ,light . This study focuses on building a machine learning model based that could detect whether the potato leaf is suffering from disease or not. It helps to avoid human interaction and also reduce the cause of disease and gives accurate result whether there is leaf has disease or not.

A feature can appear anywhere in a digital image, pixel values are stored in a two-dimensional (2D) grid, i.e., an array of numbers, and a small grid of parameters called kernel, an optimizable feature extractor, is applied at each image position, K means are highly efficient for image processing. Extracted features can evolve hierarchically and progressively more complicated as one layer feeds its output into the next layer. Training is the process of adjusting parameters like kernels to reduce the discrepancy between outputs and ground truth labels using optimization algorithms like backpropagation and gradient descent, among others.

1.6 ARCHITECTURE



2.LITERATURE SURVEY

S.NO	AUTHOR	YEAR	APPROACH	RESULT
1.	Sakshi sharma, Vatsala Anand	2021	Support vector machine	92% accuracy
2.	Monzurul Islam , Anh Dinh	2017	Support vector machine	95% accuacy
3.	Priyadarshini Patil, Nagarthana Yalgir	2017	Random forest	79% accuracy
4.	Md ashiqur Rahaman Nishad , Meherabin Akter Mitu	2022	K means	97% accuracy
5.	Jaskaran Singh , Harpreet Kaur	2019	Knn	97% accuracy
6.	Donna Henderson, Christopher J Williams , Jeffrey S Miller	2007	Logistic regression	80.8% accuracy

3. DATA PRE-PROCESSING

3.1 DATASET DESCRIPTION

About the dataset:

- The dataset is organized into 3 folders (train, test, val) and contains subfolders for each image category. There are 4025 X-Ray images (JPEG) and 3 categories (late bright/healthy bright/early bright).
- For the analysis of leaf detection images, all images are radiographs, were initially screened for quality control by removing all low quality or unreadable scans.
- There are 3254 training images and 408 testing images.
- There are 419 validation images.

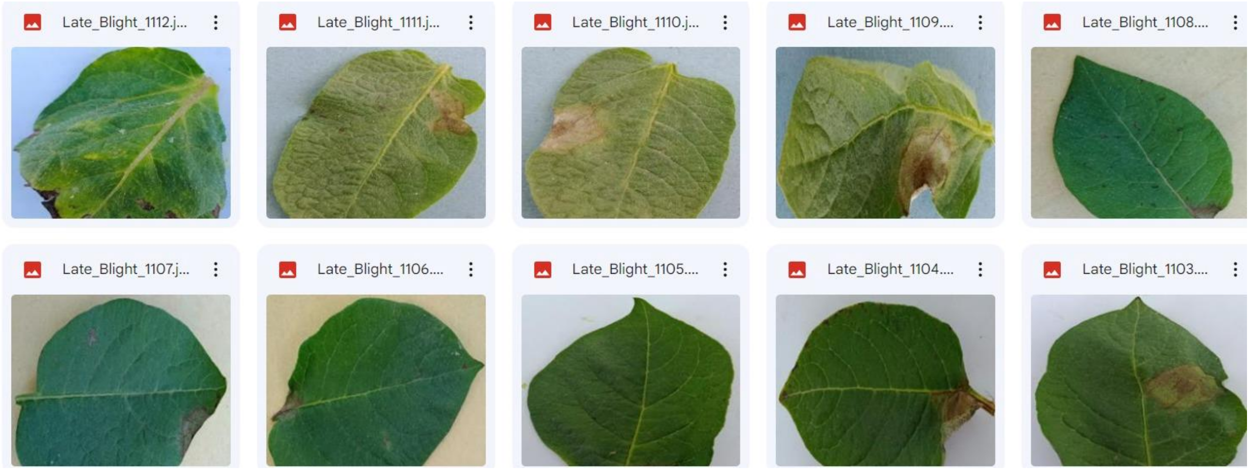
Target:

1.Target Variable: Late_Blight,Healthy,Early_Blight

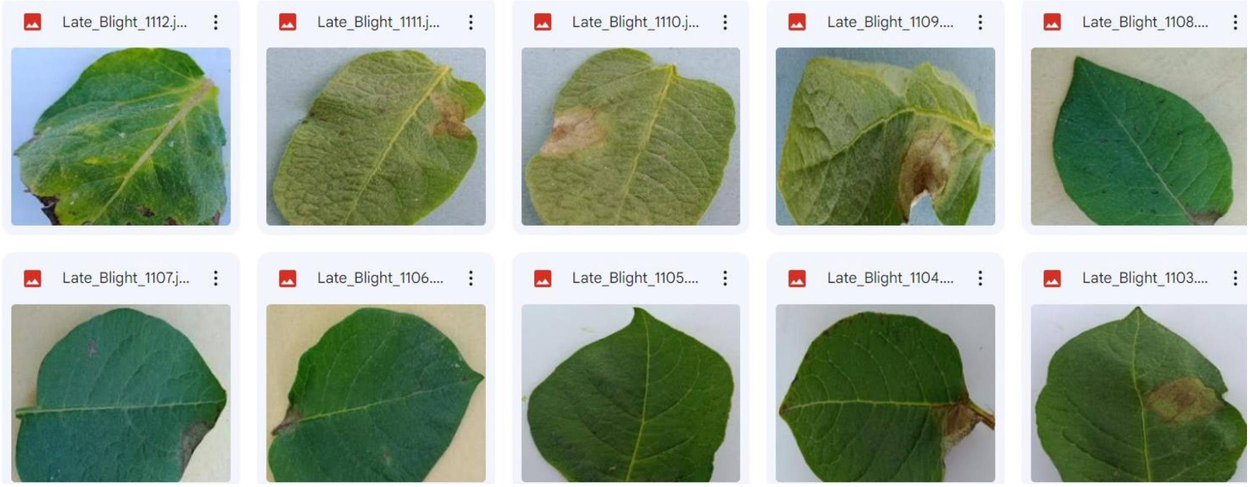
2. Late_Blight,Healthy,Early_Blight : 0 for Early_Blight , 1 for Healthy,2 for Late_Blight .

Dataset:

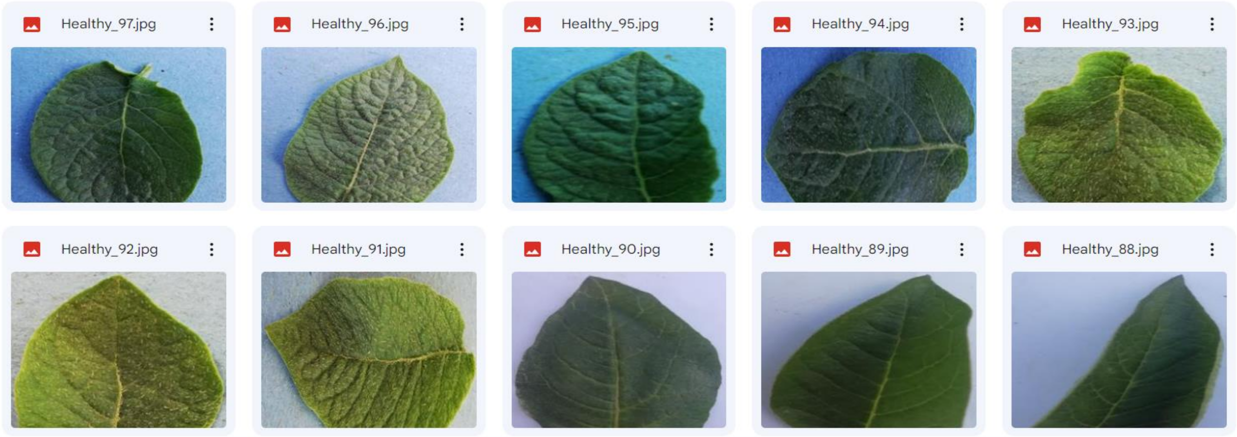
Training dataset



Testing dataset:



Validation dataset:



Features:

X:

[255, 217, 219], [255, 216, 218],[255, 216, 218],...,[206, 166, 168],[205, 165, 167], [208, 168, 170]],[[255, 216, 218],[255, 216, 218],[254, 215, 217],...,[207, 167, 169],[206, 166, 168], [209, 169, 171]],[[254, 215, 217],[254, 215, 217],[253, 214, 216],...,[208, 168, 170],[208, 168, 170],[210, 170, 172] ...,...,

Y:

[0, 0, 0, 0, 0 , 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 11, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,]…….

3.2 DATA CLEANING

Data quality has become an important issue. This issue becomes more and more important, where the need for effective decision making is high. In this context, the need for data cleaning to improve data quality is becoming crucial. Duplicate records elimination is a challenging data cleansing task. Here, we present a duplicate records elimination approach to improve the quality of data. Also, we propose an algorithm for duplicated records correction. Then, we apply the proposed duplicate records elimination approach to analyse the effect of data cleaning on the quality of decisions.

- In sum: There are 4,025 X-Ray images (JPEG) and 3 categories (Late_Blight,Healthy,Early_Blight).
- There are 3254training images for leaf detection
- There are 408 testing images .
- There are 419 validation images**.**

3.3 DATA AUGMENTATION

We are going to use only 4000 images to train our model and 300 images for validation. As we all know, training a deep learning model requires a lot of data. To overcome this problem we will use one of the simple and effective methods

3.4 DATA VISUALISATION

There are 4,025 X-Ray images (JPEG) and 3 categories (Late_Blight,Healthy,Early_Blight).

- There are 3254 training images for leaf detection
- There are 408 testing images .
- There are 419 validation images.

4. METHODOLOGY

4.1 LOGISTIC REGRESSION

Linear regression models are used to identify the relationship between a continuous dependent variable and one or more independent variables. When there is only one independent variable and one dependent variable, it is known as simple linear regression, but as the number of independent variables increases, it is referred to as multiple linear regression. For each type of linear regression, it seeks to plot a line of best fit through a set of data points, which is typically calculated using the least squares method.

Similar to linear regression, logistic regression is also used to estimate the relationship between a dependent variable and one or more independent variables, but it is used to make a prediction about a categorical variable versus a continuous one. The unit of measure also differs from linear regression as it produces a probability, but the logit function transforms the S-curve into straight line.

There are three types of logistic regression models, which are defined based on categorical response.

- **Binary logistic regression**: In this approach, the response or dependent variable is dichotomous in nature—i.e. it has only two possible outcomes.Within logistic regression, this is the most commonly used approach, and more generally, it is one of the most common classifiers for binary classification.

- **Multinomial logistic regression:** In this type of logistic regression model, the dependent variable has three or more possible outcomes; however, these values have no specified order. A multinomial logistic regression model can help the studio to determine the strength of influence a person's age, gender, and dating status may have on the type of film that they prefer.

- **Ordinal logistic regression:** This type of logistic regression model is leveraged when the response variable has three or more possible outcome, but in this case, these values do have a defined order.

Terminologies involved in Logistic Regression:

- **Independent variables:** The input characteristics or predictor factors applied to the dependent variable's predictions.
- **Dependent variable:** The target variable in a logistic regression model, which we are trying to predict.
- **Logistic function:** The formula used to represent how the independent and dependent variables relate to one another. The logistic function transforms the input variables into a probability value between 0,1 and 2, which represents the likelihood of the dependent variable . **[10]**

- **Coefficient:** The logistic regression model's estimated parameters, show how the independent and dependent variables relate to one another.
- **Intercept:** A constant term in the logistic regression model, which represents the log odds when all independent variables are equal to zero.
- **Maximum likelihood estimation:** The method used to estimate the coefficients of the logistic regression model, which maximizes the likelihood of observing the data given the model.

RESULTS:

Accuracy: 0.7045454545454546

4.2 K-NEAREST NEIGHBOR (KNN)

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.
- It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

The K-NN working can be explained on the basis of the below algorithm:

- **Step-1**: Select the number K of the neighbors
- **Step-2**: Calculate the Euclidean distance of K number of neighbors
- **Step-3**: Take the K nearest neighbors as per the calculated Euclidean distance.
- **Step-4**: Among these k neighbors, count the number of the data points in each category.
- **Step-5**: Assign the new data points to that category for which the number of the neighbor is maximum.
- **Step-6:** Our model is ready.

ADVANTAGES OF KNN

- It is simple to implement.
- It is robust to the noisy training data
- It can be more effective if the training data is large.

DISADVANTAGES OF KNN:

- Always needs to determine the value of K which may be complex some time.
- The computation cost is high because of calculating the distance between the data points for all the training samples.

Below are some points to remember while selecting the value of K in the K-NN algorithm:

- ✓ There is no particular way to determine the best value for “K”, so we need to try some values to find the best out of them. The most preferred value for K is 5.
- ✓ A very low value for K such as K=1 or K=2, can be noisy and lead to the effects of outliers in the model.
- ✓ Large values for K are good, but it may find some difficulties.

RESULTS:

Accuracy: 0.6038961038961039

4.3 SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

SVM can be of two types:

- **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.
- **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

Hyperplane:

There can be multiple lines/decision boundaries to segregate the classes in n-dimensional space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the hyperplane of SVM.

The dimensions of the hyperplane depend on the features present in the dataset, which means if there are 2 features (as shown in image), then hyperplane will be a straight line. And if there are 3 features, then hyperplane will be a 2-dimension plane.

We always create a hyperplane that has a maximum margin, which means the maximum distance between the data points.

Support Vectors:

The data points or vectors that are the closest to the hyperplane and which affect the position of the hyperplane are termed as Support Vector. Since these vectors support the hyperplane, hence called a Support vector.

ADVANTAGES OF SVM:

- Effective in high-dimensional cases.
- Its memory is efficient as it uses a subset of training points in the decision function called support vectors.
- Different kernel functions can be specified for the decision functions and its possible to specify custom kernels.

RESULTS:

```
Accuracy: 0.7272727272727273
```

4.4 DECISION TREE

- Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.
- In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
- The decisions or the test are performed on the basis of features of the given dataset.
- It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.
- It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
- In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.
- A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

There are various algorithms in Machine learning, so choosing the best algorithm for the given dataset and problem is the main point to remember while creating a machine learning model. Below are the two reasons for using the Decision tree:

- Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.
- The logic behind the decision tree can be easily understood because it shows a tree-like structure.

DECISION TREE TERMINOLOGIES

- **Root Node:** Root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.
- **Leaf Node:** Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.
- **Splitting:** Splitting is the process of dividing the decision node/root node into sub-nodes according to the given conditions.
- **Branch/Sub Tree:** A tree formed by splitting the tree.
- **Pruning:** Pruning is the process of removing the unwanted branches from the tree.

- **Parent/Child node:** The root node of the tree is called the parent node, and other nodes are called the child nodes.

Steps for decision tree

- **Step-1:** Begin the tree with the root node, says S, which contains the complete dataset.
- **Step-2:** Find the best attribute in the dataset using **Attribute Selection Measure (ASM).**
- **Step-3:** Divide the S into subsets that contains possible values for the best attributes.
- **Step-4:** Generate the decision tree node, which contains the best attribute.
- **Step-5:** Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

RESULTS:

Accuracy: 0.6038961038961039

4.5 RANDOM FOREST

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier:

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations.

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram:

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

RESULTS:

Accuracy: 0.769480519480519

5.RESULTS AND DISCUSSIONS

The accuracy value for decision tree and random forest are more accurate than other machine learning models .So we can say that potato leaf disease detection values can be obtained accurately by using decision tree and random forest methods. And the others methods obtain around 99% accuracy score.

The overall accuracy values are :

LOGISTIC REGRESSION:

Total Accuracy: 0.7045454545454546

KNN:

Accuracy: 0.6038961038961039

SVM:

Accuracy: 0.7272727272727273

DECISION TREE:

Accuracy: 0.6038961038961039

RANDOM FOREST:

Accuracy: 0.7694805194805194

6.CONCLUSION AND FUTURE SCOPE

- This project is a remarkable illustration of how AI and ML can address practical issues for the given current situation.
- Several diseases in potato plants can be identified based on leaf conditions.
- By this project we can get high rate of accuracy so that it can bring many benefits to agriculture related to world food security.

7.REFERENCES

S. M. R. Kabir et al., "Potato Leaf Disease Detection Using Deep Learning Techniques," in 2021 IEEE 2nd International Conference on Advanced Computational and Communication Paradigms (ICACCP), 2021, pp. 1-6. doi: 10.1109/ICACCP51391.2021.9470866

D. K. Dash and R. C. Balabantaray, "Potato Leaf Disease Detection using Machine Learning Techniques," in 2021 International Conference on Computational Intelligence in Data Science (ICCIDS), 2021, pp. 1-5. doi: 10.1109/ICCIDS52203.2021.9486675

H. Lee et al., "Potato Late Blight Detection using Convolutional Neural Networks with Class Activation Map," in 2020 IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia), 2020, pp. 1-4. doi: 10.1109/ICCE-Asia49866.2020.9232093

S. Ashokkumar et al., "Potato leaf disease identification using improved CNNs and deep feature fusion," in Journal of Ambient Intelligence and Humanized Computing, vol. 12, no. 2, pp. 1211-1222, 2021. doi: 10.1007/s12652-019-01492-3

M. S. Faruk et al., "Detection of Potato Leaf Diseases Using Transfer Learning with Deep Convolutional Neural Networks," in 2020 IEEE International Conference on Computing, Electronics & Communications Engineering (iCCECE), 2020, pp. 1-6. doi: 10.1109/iCCECE49124.2020.9180464

https://www.ijert.org/potato-leaf-disease-detection-using-deep-learning