National Conference on Engineering Innovations in Emerging Technologies
In association with International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET219416

A Machine Learning Approach to Prediction of Soybean Disease

Dr. Nanda Ashwin[1], Uday Kumar Adusumilli[2], Prof. Kemparaju N[3], Lakshmi Kurra[4]

[1]Professor, Department of Information Science and Engineering, East Point College of Engineering and Technology, Bangalore, Karnataka, India

[2]Product Support Analyst, Associate, Infor, Bangalore, Karnataka, India

[3]Head, Department of Information Science and Engineering, East Point College of Engineering and Technology, Bangalore, Karnataka, India

[4]Student, Dept. of Information Science and Engineering, East Point College of Engineering and Technology, Bangalore, Karnataka, India

ABSTRACT

In this paper, an analysis of several machine learning and prediction techniques was conducted on 2,000 infected and healthy soybean plants to assess how these techniques can predict charcoal rot diseases. It is critical for agriculture to prevent the spread of disease by predicting pathogen infestations in advance. There are several causes of charcoal rot and among them are Macrophomina phaseolina (Tassi) Goid significantly lowers the productivity of the plants. Soybeans are at risk of a serious disease called charcoal rot. Traditional methods of disease prediction in soybeans are very time-consuming and not practical. There has been substantial interest in Machine Learning (ML) techniques across a variety of domains in recent years. Plant diseases can be detected by ML methods, even before symptoms appear fully. Inputs to ML models are a set of morphological and physiological features. Almost all of the machine learning models that have been developed achieved an accuracy of more than 90%. Among the methods used in this study, Graded Tree Boosting (GBT) achieved the best performance regarding sensitivity and specificity, with scores of 96.25 and 97.33% respectively. Based on our results, we were able to demonstrate that ML, specifically GBT, can be a successful tool for predicting the incidence of charcoal rot in real world situations. We also illustrated how it is crucial to incorporate physiological features into the learning process in order to ensure a successful outcome.

Keywords : Macrophomina phaseolina, Goid significantly lowers, Graded Tree Boosting

I. INTRODUCTION

Growing crops globally must double by 2050 to satisfy the growing population. Agricultural industry losses around the world are largely caused by plant diseases. Plant diseases, weeds, insects, and a decline in yields around the world have all been responsible for a 14% decline in worldwide crop yield. It is therefore of paramount importance that diseases are detected early in order to prevent their spread and reduce crop losses. Approximately 700 plant species are affected by Macrophomina phaseolina (Tassi) Goid. Several

crops, including cotton, grains, oilseeds, legumes, jute, and fruits and vegetables are damaged by this soil-borne fungus. It has a broad range of morphological, physiological, and pathogenic features. In a variety of climatic conditions, phaseolina adapts. Further, the fungus has the ability to survive in the soil for a long period of time as a result of its sclerotia and chlamydospore structures. It was described in Gaige et al. (2010) that the disease is dispersed by residuals, which were left in the soil, wind, and plants after the damage was completed. As a result of infection with the M Phaseolina pathogen, symptoms can appear at any growth stage. However, symptoms typically appear after midseason or at maturity, which is growth stage R7, when yellowing can be observed on the leaves and in the pods. Besides weaker plants with reduced productivity, one of the other symptoms of blacklegs is weak plants. There are various factors involved in the eventual death of infected plants, including vascular obstructions that prevent nutrient transport or exposure to phytotoxic compounds.

Agriculture has for decades used cultural practices to protect the crops from pests and diseases, mostly in conjunction with solarization and rotation of crops to provide some protection from harmful pests, as well as the development of pesticide-resistant cultivars, in conjunction with the use of biological agents. It is most commonly controlled with fumigants and fungicides, which are two of the most common control methods. Inefficient, ineffective and inefficient agriculture can be caused by various environmental factors that result in plagues of phaseolina infections. A higher fungicide dose was suggested to increase yield. Since pesticides contaminate the water sources of the environment and may have long-lasting adverse impacts on human health and the environment, it is possible that agricultural run-off may contain pesticides that might pollute the water resources. As these toxic compounds are being ingested by both bird populations and other mammals, they are accumulating in the food chain and accumulating.

A healthy agroecosystem depends on the effective detection of diseases. With the development of molecular biology and biotechnology, plant diseases can now be detected more effectively. As a matter of fact, some of the most common methods available today are RT-PCR, enzyme-linked immunosorbent assay (ELISA) and reverse transcription-polymerase chain reaction (RT-PCR). Despite their diagnostic efficiency, these procedures do not guarantee a predictability of the fungal disease. In addition, cost-effectiveness, efficiency, and accuracy can all be considered to be limitations of RT-PCR, ELISA, and Western blot.

So, preventing and controlling soybean diseases requires an automated diagnostic system. This is because it will minimize yield losses and economic losses caused by pesticide residues in the land and by improving the quality of the crops. To predict soybean diseases at an early stage, it is necessary to classify the diseases effectively.

A number of learning algorithms have also been applied to predict pest attacks and disease infestations in crops.

An algorithm has been developed to compare aerial parts of healthy and diseased plants using spectral imaging data. In addition to monitoring morphological traits, there is also a substantial amount of evidence that ML methods are successful as well. Nevertheless, the varying nature of plant changes may lead to inaccurate predictions due to changes in symptomatology. Thus, appearance-based disease identification cannot be relied on to identify diseases reliably, especially at the early stages of growth. To identify the causative agent of charcoal rot, it is important to have an appropriate detection method since the symptoms do not appear until midseason. Our method for predicting the occurrence of charcoal rot disease also includes morphological characteristics (including characteristics that are related to growth and yield) as well as physiological features. In order to train and assess machine learning algorithms, a hybrid set of features derived from

healthy and diseased soybean plants is used. Experimental setups, as well as field conditions, are presented in the available dataset. It uses machine learning to distinguish healthy plants from unhealthy ones.

We do not currently have a public dataset for the classification of soybean charcoal rot. It has further been shown that supervised machine learning algorithms can be used to model disease in other diseases, based on mainly image datasets, and such algorithms can be used for disease prediction. It is the objective of this study to propose a set of features for enhancing the prediction of charcoal rot diseases, as well as to compare several machine learning techniques to enhance their accuracy.

II. RESEARCH METHODS AND MATERIALS

A. Data Collection and Analysis

Across Mazandaran province, Iran's dominant soybean producing region, soybean plants have been gathered from 10 different spots in the province. We selected healthy soybean plants using leaves, stems, and roots that were symptomless when ripe. Plants with bright gray stems and roots, sclerotia on their roots, or suspected disease were identified as infection-related varieties such as R7. During the past few months, we have collected samples from various farms throughout Mazandaran (Iran) and have stored them at 4°C while they are taken for analysis. At random, we chose 2,500 infected and healthy plants.

B. Symptoms of Infected Soybean Samples

There are various symptoms that can be produced by infected soybean plants, including symptoms in the branches, leaves, pods, petioles, roots, stems, and seeds, but the most common symptoms occur after flowering, or after the R7 stage, particularly during high temperature conditions. Chlorosis, premature defoliation, and sluggish growth are some of the symptoms of infection, as are low productivity, sterility of pods, and the production of crinkled and small seeds. A brown discoloration of the root's vascular tissues that extends into the stem is seen in plants that are infected. Infected plants display powdery black sclerotia beneath the epidermis and at the root during seed germination. Plant symptoms associated with this disease can be misconstrued as the effects of other abiotic stresses on plants, including drought and cyst nematodes, so it is difficult to detect the disease from aerial plant parts based on morphology.

C. Dry and Fresh Weights Determination

In order to calculate the fresh weights of soybean seedling stems and roots, a digital scale was used to calculate the fresh weights (FW) using an electronic top pan balance. In order to calculate dry weights (DW), the samples were dehydrated in an oven at 70 degrees Celsius for 72 hours. The dosage is expressed in grams per plant as DW and FW.

D. Feature Selection

Model performance and generalization were enhanced by designing and optimizing features. Analyzing variance and performing the F-test were used to select relevant features. Using a p-value framework, we were able to perform these analyses by excluding irrelevant attributes from the data set and focusing on the features that mattered. We calculated the F-test and the p-value for a difference in means in order to arrive at a meaningful 5% level of significance. In the end, we chose 21 features for analysis with machine learning. Using the F-test, the overall most important parameters for predicting charcoal rot in soybean were morphological and physiological characters (Table 1).

Features		
Morphological		
Growth attributes	**Yield- related**	**Physiological**
Stem length	Germination percentage	Seed oil content
Root length	Seedling vigor index	Amount of chlorophyll
Thickness of seed	Thousand seed weight	Amount of carotenoid
Stem bark thickness	Number of pods per plant	Protein content
Root bark thickness	Number of seeds per plant	
Stem fresh weight	Empty pods per plant	
Stem dry weight		
Root fresh weight		
Root dry weight		
Seedling fresh weight		
Seedling dry weight		

Table 1. There are a number of factors that predict the occurrence of charcoal rot in soybeans.

E. Computational Methods for Predicting Infected Soybeans

Four steps are involved in predicting the success of a prediction algorithm: collecting data, extracting features, training, and assessing performance of the infected soybean: (1) gathering data; (2) analyzing features; (3) training the predictors; and (4) evaluating performance. As illustrated in Figure 1, these steps are outlined and schematically depicted. Predicting healthy soybeans starts with gathering data (Figure 1A). The initial datasets were created and the incomplete instances were removed. We selected a randomly selected subset of samples from the negative dataset as well as from the positive dataset, in order to ensure that both datasets contained a similar proportion of positive samples. Optimal and nonoptimal samples (soybeans) are going to be incorporated into numerical feature vectors - and these numeric feature vectors will be used in building the classification models that will be used to analyze the data.

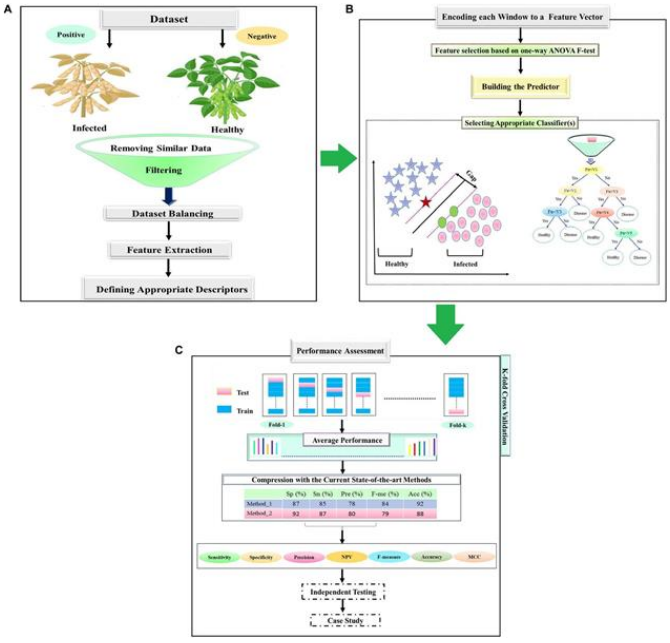

Figure 1. This flow chart outlines how statistical and ML methods are used to detect charcoal rot disease. (A) Creation of a data set. (B) Choosing and designing features. (C) ML models are trained and tested.

III. MACHINE LEARNING METHODS

We developed and applied six machine learning techniques to the training set after collecting the data and extracting its features (LR-L1, LR-L2, MLP, RF, GBT, and SVM). Based on the training data, we set the threshold based on cross-validation using 10-fold cross-validation. Scikit-learn (version 0.20.1) was used as the open-source machine learning toolkit in Python 3.6.7 for all the ML analysis. For optimising the model parameters (i.e. the number of ensembles for RF or GBT), we used internal cross validation. In order to generate an area under the receiver-operator characteristic (ROC) curve (AUC), a grid search was performed across a range of values and then the parameters that produced the best area were chosen under the ROC curve. The performance of the model reported in the paper was the best among the models reported.

A. Logistic regression with regularized coefficients (LR-L1 and LR-L2).

In LRs, the explanation variables (features) are combined with a set of explanatory variables to yield binary outcome interpretation. It is based on the LIBLINEAR model, which contains regularizations of L1 or L2 (LR-L1 and LR-L2). In many cases, regularization on the L1 or L2 axis can be used as a way of reducing model over-fitting.

B. Multilayer Perceptron (MLP)

A latent representation is developed by mapping out the input data using MLP. MLP consists of several layers of nodes that are connected directly to each other, with non-linear activation functions except at the input layer. Researchers tend to prefer MLP over similar networks because it uses back-propagation and has demonstrated wide application. A hidden layer of size 10 was considered in this study along with an Adam optimization.

C. Random Forest (RF)

As part of an ensemble algorithm, RF uses multiple decision trees to generate a non-linear outcome. A better prediction is made with the improved generalization ability. Based on the results of the analysis of each tree, a final prediction can be made. For RF to be generalized effectively, each tree's accuracy must be correlated with the others. RF is not typically sensitive to parameter selection.

D. Gradient Tree Boosting (GBT)

In addition to GBT, decision trees can be employed for their ensemble algorithm. A weighting strategy, instead of RF, is used to build decision trees sequentially by emphasizing the hardest samples. In order to make the final prediction, a weighted majority of votes are used.

E. Support Vector Machines (SVM)

By minimizing structural risk in kernel space, SVM seeks to minimize the risk of kernel crashes. In terms of kernel functions, those which are most commonly used are horizontal basis functions, linear functions, and polynomials. A kernel coefficient *, as well as a penalty parameter C, are two important hyperparameters used by SVM. As a result, we would like to discover which model is simple enough to separate the data for improved generalizability. In this study, linear kernels were taken into account.

IV. MODEL EVALUATION CRITERIA

Several parameters are measured in order to evaluate the master-learning models, including true positive (TP), true negative (TN), false positive (FP), and false negative (FN), which are essentially measures of how many healthy plants are mistakenly classified as infected ones. There are many measurements of classification performance including precision, specificity, sensitivity, Negative Predictive Value (NPV), F1 score, and Matthews Correlation Coefficient (MCC). But one of the most frequently used measures is the AUC. I believe that this value is especially useful when attempting to measure the learning curve. A threshold was established on the train data in order to provide an evaluation of the quality of the classification. As a result of this work, all the performance criteria are explained as follows:

A. Model Verification and Evaluation

We used a 10-fold cross-validation to evaluate independently and compare the strengths and reliability of each model, as a model developed by one random scale tends to overfit or only occasionally make sense. The results of six ML models on the test sets can be seen in Table 2 and Figure 2. As regards all of the evaluation criteria, MLP performed the worst in terms of accuracy (94.88%), sensitivity (94.83%), specificity (94.92%), precision (94.72%), NPV (95.06%), F1 score (94.77%), and MCC (89.76%). It was determined that GBT's classifier was ranked first by the metrics of classification accuracy (96.79%), specificity (97.33%), precision (97.16%), NPV

(96.49%), F1 score (96.68%), and MCC (93.62%). There are indeed several top classifications with more than 95% accuracy in TP (1 150) and specificity (95.83%), precision (95.61%), NPV (96.29%), F1 score (96.03%), and MCC (92.09%). As described previously, SVM and GBT had almost similar sensitivity, placing them at the top. With sensitivity levels exceeding 95 percent, they performed not only well but were also only marginally better than GBT and SVM. According to our observation, the GBT and SVM models appeared to outperform the other models during our study period. Accuracy of data (Acc) refers to the percentage of samples correctly classified out of the total samples.

B. Precision

The positive predictive value (PPV) of a test measures the probability of correctly diagnosing samples in relation to the number of positives predicted.

C. Area Under the ROC Curve (AUC)

It has been over a decade since Fawcett (2006) evaluated the ratio of true positives to false positives in machine learning utilizing ROC charts. ROC curve quality can be characterized as a function of an area under the curve (AUC), which is a metric that measures both the accuracy of the classification as well as the effectiveness of the ROC model. The AUC is our primary evaluation metric as it does not depend upon the decision to select an arbitrary threshold for classification.

V. RESULTS

A. Model Verification and Evaluation

In order to avoid over-fitting or sporadic outcomes when developing a model by random scale, we used a tenfold factorisation cross-validation to assess and relate the strength and trustworthiness of all models. Figure 2 and Table 2 show the results of the six ML models that the test sets were fitted with. As regards all of the evaluation criteria, we concluded that MLP had the worst quality scores for accuracy (94.88%), sensitivity (94.83%), specificity (94.92%), precision (94.72%), NPV (95.06%), F1 score (94.77%), and MCC (89.76%). It was determined that GBT's classifier was ranked first by the metrics of classification accuracy (96.79%), specificity (97.33%), precision (97.16%), NPV (96.49%), F1 score (96.68%), and MCC (93.62%). There are indeed several top classifications with more than 95% accuracy in TP (1 150) and specificity (95.83%), precision (95.61%), NPV (96.29%), F1 score (96.03%), and MCC (92.09%). SVM and GBT were almost identical in terms of their sensitivity and the two came in second place. Despite only slightly lower sensitivity than GBT and SVM, they performed well and reached a sensitivity of over 95%. According to our observation, the GBT and SVM models appeared to outperform the other models during our study period.

Method	TP	FP	TN	FN	Accuracy	Sensitivity	Specificity	Precision	NPV	F1 score	MCC	AUC
LR-L1	1153	47	1149	51	95.92 ± 8.32	95.75 ± 8.74	96.08 ± 7.93	96.01 ± 8.05	95.84 ± 8.60	95.88 ± 8.39	91.84 ± 16.63	97.37 ± 5.82
LR-L2	1143	57	1151	49	95.58 ± 9.00	95.92 ± 8.90	95.25 ± 9.16	95.31 ± 9.11	95.87 ± 8.90	95.61 ± 8.99	91.18 ± 17.99	97.05 ± 6.58
MLP	1139	61	1138	62	94.88 ± 9.58	94.83 ± 10.66	94.92 ± 8.51	94.72 ± 8.97	95.06 ± 10.16	94.77 ± 9.83	89.76 ± 19.14	96.69 ± 7.39
RF	1143	57	1147	53	95.42 ± 9.64	95.58 ± 9.27	95.25 ± 10.01	95.34 ± 9.80	95.50 ± 9.47	95.46 ± 9.54	90.83 ± 19.28	97.20 ± 6.35
GBT	1168	32	1155	45	96.79 ± 6.49	96.25 ± 7.88	97.33 ± 5.16	97.16 ± 5.55	96.49 ± 7.29	96.68 ± 6.75	93.62 ± 12.90	98.42 ± 3.42
SVM	1150	50	1155	45	96.04 ± 7.55	96.25 ± 7.88	95.83 ± 7.26	95.81 ± 7.36	96.29 ± 7.76	96.03 ± 7.61	92.09 ± 15.10	97.86 ± 4.65

Table 2. Analysis of the performance of several machine learning algorithms on a dataset of full features as a method for predicting soybean charcoal rot.

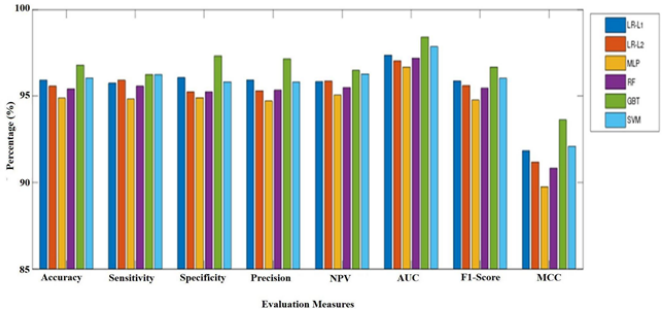

Figure 2. Using different machine-learning algorithms, we compare different evaluation criteria for infected and healthy soybean plants with charcoal rot.

B. Effectiveness Analysis of Feature Selection

A total of 12 morphological features were taken and analyzed using algorithms (LR-L1, LR-L2, MLP, RF,

GBT, and SVM) as part of our ML project in an attempt to further evaluate the effectiveness of a more comprehensive feature set on ML performance for prediction of charcoal rot. On the basis of the 12 features, the prediction results are evaluated using cross-validation 10-fold. Among the three models, GBT performed best with an accuracy level of 96.13 percent, followed by SVM and LR-L1 which both performed only slightly lower than GBT with an accuracy level of 95.58%. According to the results of the MLP algorithm, our classification accuracy was 94.50%. On Table 3, we have displayed the averaged prediction performance of the full feature set and compared it to that of the averaged prediction performance. In Table 2, it is possible to see that the accuracy, sensitivity, specificity, precision, NPV, F1 score, MCC, AUC, and NPV of the full features are slightly higher than those of the morphological features. Based on the results shown in Tables 2 and 3, we can conclude that the GBT algorithm has a higher performance when the hybrid feature set is considered as opposed to morphological features (96.79 vs. 96.13).

Method	Accuracy	Sensitivity	Specificity	Precision	NPV	F1 score	MCC	AUC
LR-L1	95.58 ± 8.51	95.25 ± 10.27	95.92 ± 6.97	95.66 ± 7.59	95.61 ± 9.28	95.41 ± 8.97	91.21 ± 16.92	97.24 ± 5.84
LR-L2	94.96 ± 9.63	94.50 ± 12.51	95.42 ± 7.32	94.99 ± 8.32	95.15 ± 10.55	94.64 ± 10.54	90.03 ± 18.99	96.96 ± 6.40
MLP	94.50 ± 8.64	94.17 ± 12.04	94.83 ± 5.88	94.46 ± 6.57	94.90 ± 10.32	94.17 ± 9.41	89.18 ± 16.98	97.29 ± 5.45
RF	95.46 ± 9.33	95.08 ± 10.33	95.83 ± 8.41	95.64 ± 8.88	95.31 ± 9.73	95.35 ± 9.62	90.93 ± 18.63	97.12 ± 6.35
GBT	96.13 ± 7.64	95.92 ± 8.34	96.33 ± 6.97	96.22 ± 7.22	96.05 ± 8.04	96.06 ± 7.78	92.26 ± 15.27	98.00 ± 4.28
SVM	95.58 ± 7.73	95.67 ± 9.05	95.50 ± 6.53	95.34 ± 6.92	95.88 ± 8.49	95.48 ± 7.99	91.20 ± 15.41	97.46 ± 5.51

Table 3. Performance comparison of various ML techniques based on 12 morphological features for prediction of soybean charcoal rot disease.

C. Feature Ranking

This table illustrates how features play an important role in terms of the ML models when considering the features. The features of the classification system were ranked according to how important they were to each type. A classification system's performance is enhanced by the presence of incrementally useful features, which in turn have a relevance from the perspective of feature ranking. In Table 4, the top 10 features in this work as ranked by each ML algorithm are displayed by different colors that represent their ranking. The SHAP analysis (Shapley Additions and Explanations) was done on the GBT model in order to gain a better understanding of the effect of individual features on the model predictions, and the results are shown in Figure 4. We can interpret the SHAR values by comparing them with the value of a given feature and its effect on the model prediction in terms of comparing it to the values in a given baseline, based on the differences in their values.

Number	Features	LR-L1	LR-L2	MLP	RF	GBT	SVM
1	Stem length	13	13	21	2	2	13
2	Root length	3	3	9	14	3	3
3	Thousand seed weight	18	18	19	16	5	19
4	Stem bark thickness	19	19	18	8	13	18
5	Root bark thickness	6	6	16	5	14	6
6	Thickness of seed	14	9	14	13	19	14
7	Stem fresh weight	17	1	6	18	21	1
8	Stem dry weight	1	14	12	11	7	16
9	Root fresh weight	16	17	13	17	17	4
10	Root dry weight	4	16	7	21	16	17
11	Seedling fresh weight	15	15	2	19	11	15
12	Seedling dry weight	21	4	3	10	10	21
13	Protein content	9	20	20	20	4	9
14	Seed oil content	12	21	17	4	1	11
15	Germination percentage	20	7	11	7	11	20
16	Amount of chlorophyll	5	12	15	1	15	7
17	Amount of carotenoid	7	11	5	3	9	5
18	Empty pods per plant	11	5	8	9	8	12
19	Number of seeds per plant	10	10	10	15	6	10
20	Seedling vigor index	8	8	1	6	20	8
21	Number of pods per plant	2	2	4	12	18	2

Colors indicate the top 10 features ranked from high (blue) to low (red).

Table 4. Feature ranking results for various ML techniques.

The results indicate that the features that were most prominent were mainly physiological ones, which showed more important biological features can help predict the onset of charcoal rot disease in soybeans when compared with morphological features. Based on the observation that seed oil, protein content, and chlorophyll content are all in the top 10 features of the list, these three factors must be predictive for all of the ML methods. The amount of carotenoid and empty pods per plant is listed within the top ten of the results of all of the methods listed above, with the exception of MLP and GBT. The seed weight, seed thickness, and the number of seeds in the plant are selected using four different machine-learning algorithms based on the thousand seed weight, seed thickness, and seed number per plant. In general, the least informative features are those focused on the

length of the roots, the thickness of the bark on the stem, the thickness of the bark on the roots, and the seedling vigor index. Meanwhile, features that are not listed or only selected based on one method would be deemed less informative because they do not fit this category. It is important to note that this information can evaluate the seeds by picking out their main characteristics.

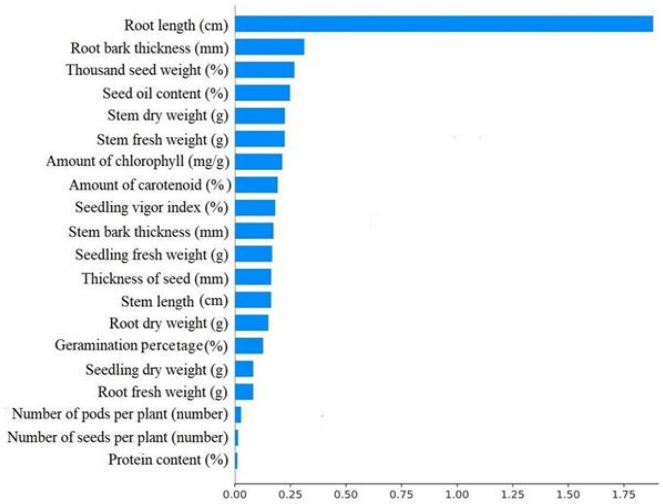

Figure 4. GBT model's Mean SHAP values.

VI. DISCUSSION

Through either direct or indirect methods, it is possible to predict fungi diseases. Among the direct procedures are in situ hybridization, polymerase chain reaction, fluorescence in situ hybridization, enzyme-linked immunosorbent assay, flow cytometry, liquid chromatography mass spectrometry, as well as western blotting. The high throughput systems that are used can be used for analyzing large numbers of samples simultaneously in order to get precise information by doing so. The indirect method for estimating the disease of a plant can be calculated by observing the changes in the shape and morphology of the plant or by measuring the compounds released by the infected plant as a defense mechanism. It has become increasingly popular to use mathematical logic techniques to detect plant diseases indirectly, which has led to a wide range of techniques that have been developed.

In conclusion, there are a variety of ways in which ML methods are used in agriculture, however the most important use is to identify, detect, as well as predict crop diseases and the effects of stresses on plants. As part of the plant protection management process, it is critical to be able to predict plant diseases efficiently and accurately. As a matter of fact, in detecting diseases early, humans are less likely to interfere, a method that has been successfully implemented recently. The diagnosis and classification of charcoal rot disease are of fundamental importance for future agriculture because of its effects, but more importantly is the prediction and quantification of the disease. A project such as this could provide an early preventative measure against disease outbreaks and reduce the use of pesticides for pest control.

By analyzing a series of symptoms exhibited by different parts of soybean plants, and analyzing the ML models fabricated, we developed special models to identify the disease of charcoal rot. We have, in order to compensate for the lack of datasets available on this condition, created our own. The dataset and all details are explained in the dataset section. Among the main advantages of the method we propose, is that it can detect soybean charcoal rot disease at an early stage, prior to the onset of symptoms.

For this study, we established a database of 2,000 soybean plants in natural field conditions. A supervised neural network classifier that classifies plants based on their L1L2 MLP data was trained to distinguish healthy plants from those infected. With the analysis of the suggested feature set, GBT classification achieved a success rate of 95.99% with a success rate of 96.79%.

As we all know, charcoal rot is a regular occurrence in soybean production, and as soybeans grow, we are more likely to encounter a variety of kinds of the disease. Therefore, different treatment techniques can be developed based on the developed methodologies in this study in order to identify and treat the different symptoms of the charcoal rot disease.

Moreover, it is possible to create an automated method of predicting charcoal rot by combining an identification model with the information about soya disease. Image processing methods and computer vision algorithms have been reported in the past to be helpful in the identification of plant diseases. A 90.5% recognition rate can be achieved by combining the accuracy of classification and the pre-processing of images. The onset of the charcoal rot disease development has yet to be predicted using any scientific studies that have been conducted so far. Most of the time, an algorithm used for identifying diseased charcoal rot would be an algorithm for classification and segmentation of images. This algorithm is used in a soybean plant to distinguish between healthy and unhealthy leaves and stems. Based on the results of the SVM approach, the highest level of accuracy was 95.76 percent, and the F-score for identifying the presence of the charcoal rot disease in soybeans was 87%.

The effect of illumination conditions on the performance of image processing and AI algorithms has been well documented in the early detection of disease. In order to tackle this challenge, physiological evaluations can be of great help. As a result of the presented results, it has been demonstrated that the physiological features could be used to predict charcoal rot in soybeans. This was determined by comparing the result of the hybrid features to that of the morphological features, as shown in Table 3. As shown in Table 3, the hybrid features gave slightly better results. Moreover, Table 4 also shows that, based on the ranking of features based on different ML models, physiological features play a very important role in the prediction of diseases.

Overall, all methods performed well as far as classification performance is concerned. As a classifier, GBT was the best performer as it tries to sequentially improve the performance and in the learning process, it also includes a feature interaction. There was a low performance of MLP among the other players. As neural networks usually require larger quantities of data in order to operate well, this could be due to the small size of the data in our experiments. We would like to point out that this study is limited by the small size of the dataset and the fact that it allocates the same importance to all features.

VII.CONCLUSION

Using morphological and physiological features for detecting and classifying soybean charcoal rot disease, this paper studied different machine learning algorithms. We present an evaluation and comparative study of six machine learning methods for predicting charcoal rot disease as part of this research effort. In terms of performance, the results indicated that the performances of various ML approaches, when evaluating them via various evaluation metrics, were slightly different. The results of the quantitative analysis showed that GBT and SVM performed almost the same as LR-L1, LR-L2 , MLP, and RF, and demonstrated a better performance as compared with these approaches. In addition, the ranking of features has demonstrated to us that it's important to incorporate a variety of features into the educational process. As part of the research, we will investigate how other feature types, for instance, molecular structure and chemical composition, can be incorporated into the learning process.

VIII. REFERENCES

[1]. Abbas, H. K., Bellaloui, N., Accinelli, C., Smith, J. R., and Shier, W. T. (2019). Toxin production in soybean (Glycine max L.) plants with charcoal rot disease and by Macrophomina phaseolina, the fungus that causes the disease. Toxins 11:645. doi: 10.3390/toxins11110645

[2]. Ambrosio, M. M., Dantas, A. C., Martinez-Perez, E., Medeiros, A. C., Nunes, G. H., and Pico, M. B. (2015). Screening a variable germplasm collection of Cucumis melo L. for seedling

resistance to Macrophomina phaseolina. Euphytica 206, 287–300. doi: 10.1007/s10681-015-1452-x

[3]. Ashfaq, R. A. R., Wang, X. Z., Huang, J. Z., Abbas, H., and He, Y. L. (2017). Fuzziness based semi-supervised learning approach for intrusion detection system. Inform. Sci. 378, 484–497. doi: 10.1016/j.ins.2016.04.019

[4]. Azlah, M. A. F., Chua, L. S., Rahmad, F. R., Abdullah, F. I., and Wan Alwi, S. R. (2019). Review on techniques for plant leaf classification and recognition. Computers 8, 4–77. doi: 10.3390/computers8040077

[5]. Barbedo, J. G. (2018). Factors influencing the use of deep learning for plant disease recognition. Biosyst. Eng. 172, 84–91. doi: 10.1016/j.biosystemseng.2018.05.013

[6]. Basu, S., Kumbier, K., Brown, J. B., and Yu, B. (2018). Iterative random forests to discover predictive and stable high-order interactions. Proc. Natl. Acad. Sci. U. S. A. 115, 1943–1948. doi: 10.1073/pnas.1711236115

[7]. Boughorbel, S., Jarray, F., and El-Anbari, M. (2017). Optimal classifier for imbalanced data using matthews correlation coefficient metric. PloS One 12:e0177678. doi: 10.1371/journal.pone.0177678

[8]. Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern. Recognit. 30, 1145–1159. doi: 10.1016/S0031-3203(96)00142-2

[9]. Breiman, L. (2001). Random forests. Mach. Learn. 45, 5–32.

[10]. Brevik, E. C., Slaughter, L., Singh, B. R., Steffan, J. J., Collier, D., Barnhart, P., et al. (2020). Soil and human health: current status and future needs. Air. Soil. Water. Res. 13, 3–21. doi: 10.1177/1178622120934441

[11]. Brzezinski, C. R., Henning, A. A., Abati, J., Henning, F. A., França-Neto, J. D. B., Krzyzanowski, F. C., et al. (2015). Seeds treatment times in the establishment and yield performance of soybean crops. J. Seed. Sci. 37, 147–153. doi: 10.1590/2317-1545v37n2148363

[12]. Chamorro, M., Miranda, L., Domínguez, P., Medina, J. J., Soria, C., Romero, F., et al. (2015). Evaluation of biosolarization for the control of charcoal rot disease (Macrophomina phaseolina) in strawberry. Crop. Prot. 67, 279–286. doi: 10.1016/j.cropro.2014.10.021

[13]. Cheng, J., Li, G., and Chen, X. (2018). Research on travel time prediction model of freeway based on gradient boosting decision tree. IEEE Access 7, 7466–7480. doi: 10.1109/ACCESS.2018.2886549

[14]. Chlingaryan, A., Sukkarieh, S., and Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: a review. Comput. Electron. Agric. 151, 61–69. doi: 10.1016/j.compag.2018.05.012

[15]. Couronne, R., Probst, P., and Boulesteix, A. L. (2018). Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics 19, 270–284. doi: 10.1186/s12859-018-2264-5

[16]. Czarnecki, W. M., and Tabor, J. (2015). Multithreshold entropy linear classifier: theory and applications. Expert Syst. Appl. 42, 5591–5606. doi: 10.1016/j.eswa.2015.03.007

[17]. Dimitriadis, G., Neto, J. P., and Kampff, A. R. (2018). T-SNE visualization of large-scale neural recordings. Neural. Comput. 30, 1750–1774. doi: 10.1162/neco-a-01097

[18]. Elssied, N. O. F., Ibrahim, O., and Osman, A. H. (2014). A novel feature selection based on one-way anova f-test for e-mail spam classification. J. Appl. Sci. Eng. 7, 625–638. doi: 10.19026/rjaset.7.299

[19]. Eskandari, S., and Javidi, M. M. (2016). Online streaming feature selection using rough sets. Int. J. Approx. Reason. 69, 35–57. doi: 10.1016/j.ijar.2015.11.006

[20]. Eun, A. J. C., Huang, L., Chew, F. T., Li, S. F. Y., and Wong, S. M. (2002). Detection of two orchid viruses using quartz crystal microbalance (QCM) immunosensors.). J. Virol. Methods 99, 71–79. doi: 10.1016/S0166-0934(01)00382-2

[21]. Evans, J., and Poorter, H. (2001). Photosynthetic acclimation of plants to growth irradiance: the relative importance of specific leaf area and nitrogen partitioning in maximizing carbon gain. Plant Cell Environ. 24, 755–767. doi: 10.1046/j.1365-3040.2001.00724.x

[22]. Fang, Y., and Ramasamy, R. P. (2015). Current and prospective methods for plant disease detection. Biosensors 5, 537–561. doi: 10.3390/bios50305537

[23]. Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognit. Lett. 27, 861–874. doi: 10.1016/j.patrec.2005.10.010

[24]. Fenta, B. A., Beebe, S. E., Kunert, K. J., Burridge, J. D., Barlow, K. M., Lynch, J. P., et al. (2014). Field phenotyping of soybean roots for drought stress tolerance. Agron 4, 418–435. doi: 10.3390/agronomy4030418

[25]. Friedman, J. H. (2002). Stochastic gradient boosting. Comput. Stat. Data An. 38, 367–378. doi: 10.1016/S0167-9473(01)00065-2

[26]. Gaige, A. R., Ayella, A., and Shuai, B. (2010). Methyl jasmonate and ethylene induce partial resistance in Medicago truncatula against the charcoal rot pathogen Macrophomina phaseolina. Physiol. Mol. Plants 74, 412–418. doi: 10.1016/j.pmpp.2010.07.001

[27]. Golhani, K., Balasundram, S. K., Vadamalai, G., and Pradhan, B. (2018). A review of neural networks in plant disease detection using hyperspectral data. Inf. Process. Agric. 5, 354–371. doi: 10.1016/j.inpa.2018.05.002

[28]. Grimes, M. L., Lee, W. J., Van der Maaten, L., and Shannon, P. (2013). Wrangling phosphoproteomic data to elucidate cancer signaling pathways. PloS One 3:e52884. doi: 10.1371/journal.pone.0052884

[29]. Gupta, G. K., Sharma, S. K., and Ramteke, R. (2012). Biology, epidemiology and management of the pathogenic fungus Macrophomina phaseolina (Tassi) Goid with special reference to charcoal rot of soybean (Glycine max (L.) Merrill). J. Phytopathol. 160, 167–180. doi: 10.1111/j.1439-0434.2012.01884.x

[30]. Hartman, G. L., Pawlowski, M. L., Herman, T. K., and Eastburn, D. (2016). Organically grown soybean production in the USA: constraints and management of pathogens and insect pests. Agron 6:16. doi: 10.3390/agronomy6010016

[31]. Hendry, G. A. F., and Price, A. H. (1993). “Stress indicators: chlorophylls and carotenoids,” in Methods in Comparative Plant Ecology Chapman and Hall, eds G. A. F. Hendry and J. P. Grime, (Berlin: Springer).

[32]. Holmes, G. J., Mansouripour, S. M., and Hewavitharana, S. S. (2020). Strawberries at the Crossroads: management of soilborne diseases in california without methyl bromide. Phytopathology 110, 956–968. doi: 10.1094/phyto-11-19-0406-ia

[33]. Irish, J. M. (2014). Beyond the age of cellular discovery. Nat. Immunol. 15, 1095–1097. doi: 10.1038/ni.3034

[34]. Islam, A. K. M. A., Anuar, N., and Yaakob, Z. (2009). Effect of genotypes and pre-sowing treatments on seed germination behavior of Jatropha. Asian J. Plant Sci. 8, 433–439. doi: 10.3923/ajps.2009.433.439

[35]. Jeong, J. J., Ju, H. J., and Noh, J. (2014). A review of detection methods for the plant viruses. Res. Plant Dis. 20, 173–181. doi: 10.5423/RPD.2014.20.3173

[36]. Joshi, R., Singh, J., and Vig, A. P. (2015). Vermicompost as an effective organic fertilizer and biocontrol agent: effect on growth, yield and quality of plants. Rev. Environ. Sci. Biotechnol. 14, 137–159. doi: 10.1007/s11157-014-9347-1