



Classification of soybean diseases by logistic regression with LASSO using phenotypic and environmental features

Zhang H¹, Zhang H.H²

Received: May 20, 2023, Revised: version 1, July 15, 2023, version 2, July 31, 2023

Accepted: July 31, 2023

Abstract

The soybean plant is an important food and oil cash crop – soybeans and their byproducts are the most-traded agricultural commodities. However, crop losses sustained by the onset of disease can lessen yield and product quality. In the United States alone – the world’s primary soybean producer and second-largest exporter – economic losses due to soybean disease amounted to more than \$95 billion. However, few efforts have been conducted using statistical analysis and machine learning methods on the extraction of phenotypic parameters and analysis of growth patterns to diagnose and predict soybean diseases. The most common method is still the standard manual measurement of phenotypic features, combined with molecular methods, which requires excess time and effort. The goal of this study is to develop a statistical method *via* machine learning that accurately differentiates between five soybean diseases, i.e. brown stem rot, brown spot, anthracnose, alternaria leaf spot, and frogeye leaf spot. The Least Absolute Shrinkage and Selection Operator (LASSO) regression was used to classify 160 observations of diseased soybean plants. An overall accuracy of 87.5% was obtained with cross validation in detecting soybean diseases. In addition, we assessed which phenotypic, pathological, and environmental factors most influence the onset of each disease.

Keywords

LASSO, Soybean, Soybean disease, Machine learning, Variable selection, Logistic regression, Brown stem rot, Anthracnose, Alternaria leaf spot, Frogeye leaf spot

¹Hannah Zhang, Basis Oro Valley, 11155 N Oracle Rd, Oro Valley, AZ 85737, USA.
hannah.zhang05@gmail.com.

²Hao Helen Zhang, Department of Mathematics, The University of Arizona, Tucson, AZ 85721, USA. hzhang@math.arizona.edu.

Introduction

The soybean plant is a major economic crop for food and oil. Soybean products are the most-traded agricultural commodities, contributing more than 10% to global agricultural trade (1). However, losses sustained by the onset of disease can lessen yield and grain quality, diminishing high-quality production of soybean-related food, feed, and oil (2). Multiple factors influence disease development, such as environmental conditions, production practices, field and disease history, and cultivar selection (3).

Accurate classification of fungal disease is essential for effective disease control, surveillance efforts, and resistance breeding, ensuring the long-term health and productivity of soybean crops. Because fungal pathogens have different susceptibilities to varying fungicides, the identification of disease-causing pathogens enables the selection of appropriate fungicides for effective and streamlined treatment. While broad-spectrum fungicides can be effective, their use should be balanced with sustainable disease management practices composed of integrated strategies, rather than the singular practice of universal fungicide application. Understanding the fungal pathogen at hand allows for the implementation of this comprehensive management approach that includes resistant cultivars, crop rotation, and other non-chemical methods – strategies that can be tailored based on the pathogen's attributes and life cycle (4, 5). This approach can reduce the overuse of broad-spectrum fungicides, which often lead to fungicide resistance and environmental damage (6). By tracking the prevalence and distribution of

identified fungal pathogens, researchers are able to carry out effective surveillance and monitoring efforts, providing prompt information to farmers about possible outbreaks and disease management strategies (7). This identification is also crucial for breeding programs that aim to develop soybean cultivars resistant to certain diseases. By understanding the pathogens and their mechanisms of infection, researchers can focus on breeding for resistance traits that target those specific pathogens, enhancing the long-term sustainability of disease management (8, 9).

With the development of molecular biology and biotechnology, plant diseases can be detected more effectively, e.g. reverse transcription-polymerase chain reaction, enzyme-linked immunosorbent assay (10). However, while these procedures do procure high detection accuracies, they do not guarantee a predictability of soybean diseases, additionally possessing limitations of cost-effectiveness and time efficiency (11). Hence, preventing soybean diseases requires an automated diagnostic system with readily available data such as phenotypic features and environmental conditions. Such diagnosis has the potential to increase yield, minimize economic losses and pesticide residues, and enhance product quality while being an accessible, efficient tool (12). Machine learning (ML) techniques have already found application in several areas of agricultural research, such as crop management, yield prediction, disease detection, and weed detection crop quality (13, 14, 15). Researchers have also applied learning algorithms to predict

pest attacks and disease infestation in crops (16). These ML algorithms learn through observations *via* training data to generalize and predict the unseen data, or test data (17). In addition, statistical techniques have been largely incorporated in a variety of ML domains in recent years.

The goal of this study was to develop a statistical regression model with machine learning to analyze phenotypic features and environmental conditions to differentiate between five common soybean crop diseases, i.e. brown stem rot, brown spot, anthracnose, alternaria leaf spot, and frogeye leaf spot. There are seed varieties that are resistant to most of these fungal diseases; however, fungal pathogens have the ability to evolve and overcome resistance traits in plant varieties (18). Continuous monitoring and identification of fungal infections can help identify shifts in pathogen populations or the emergence of new pathotypes that can overcome resistance traits in certain soybean varieties. By staying informed of the specific fungal pathogens, farmers and breeders can rapidly respond to changes and select new resistant varieties as needed. Additionally, soybean seed varieties may perform differently in specific regions or under certain environmental conditions (19). By considering the unique fungal pathogens

and the resistance profiles of each seed variety, farmers can select locally adapted varieties that not only are resistant but also exhibit good agronomic traits, yielding potential in their specific growing conditions. Hence, while resistant seed varieties are a valuable tool, classifying fungal diseases helps monitor changes in pathogen populations and tailor the choice of resistant varieties to the specific fungal pathogens present.

This statistical regression model can be a successful tool for diagnosing the incidence of soybean diseases in real world situations, simultaneously incorporating physiological, pathological, growth, and environmental features into the classification process to produce an effective outcome.

Materials and Methods

Dataset

The Soybean (Large) Dataset was extracted from the UC Irvine Machine Learning Repository (20). There are 19 disease classes and 35 categorical attributes – the values for attributes are numerically encoded, with the first value as "0," the second as "1," etc. An unknown value is encoded as "?" and the value "dna" equates to "does not apply." Table 1 shows the dataset used for this study.

Table 1: The Soybean (Large) Dataset with all 35 categorical attributes and their corresponding values

Variable #	Attribute	Value
1	date	april,may,june,july,august,september,october,?
2	plant-stand	normal,lt-normal,?
3	precip	lt-norm,norm,gt-norm,?
4	temp	lt-norm,norm,gt-norm,?
5	hail	yes,no,?
6	crop-hist	diff-1st-year,same-1st-yr,same-1st-two-yrs, same-1st-sev-yrs,?
7	area-damaged	scattered,low-areas,upper-areas,whole-field,?
8	severity	minor,pot-severe,severe,?
9	seed-tmt	none,fungicide,other,?
10	germination	90-100%,80-89%,lt-80%,?
11	plant-growth	norm,abnorm,?
12	leaves	norm,abnorm
13	leafspots-halo	absent,yellow-halos,no-yellow-halos,?
14	leafspots-marg	w-s-marg,no-w-s-marg,dna,?
15	leafspot-size	lt-1/8,gt-1/8,dna,?
16	leaf-shread	absent,present,?
17	leaf-malf	absent,present,?
18	leaf-mild	absent,upper-surf,lower-surf,?
19	stem	norm,abnorm,?
20	lodging	yes,no,?
21	stem-cankers	absent,below-soil,above-soil,above-sec-nde,?
22	canker-lesion	dna,brown,dk-brown-blk,tan,?
23	fruiting-bodies	absent,present,?
24	external decay	absent,firm-and-dry,watery,?
25	mycelium	absent,present,?
26	int-discolor	none,brown,black,?
27	sclerotia	absent,present,?
28	fruit-pods	norm,diseased,few-present,dna,?
29	fruit spots	absent,colored,brown-w/blk-specks, distort,dna,?
30	seed	norm,abnorm,?
31	mold-growth	absent,present,?
32	seed-discolor	absent,present,?
33	seed-size	norm,lt-norm,?
34	shriveling	absent,present,?
35	roots	norm,rotted,galls-cysts,?

Statistical model

Regularization is one of the methods widely used to increase generalization of a model for a more accurate prediction. The Least Absolute Shrinkage and Selection Operator (LASSO) procedure leads to simpler models, or models with fewer parameters. LASSO regularization adds a penalty, i.e. lambda, that shrinks the coefficient value (which indicates the relevance of the corresponding variable to the model) of each data point towards zero. This regularization type results in sparse models with less coefficients, as the less contributing coefficients become zero and are eliminated from the model. Larger penalties result in coefficient values closer to zero, which are optimal for producing simpler models. This method is well-suited for models that show a high degree of multicollinearity. Additionally, this method is effective when certain parts of model selection need to be automated, such as variable selection or parameter elimination (21).

R program

We used R language and environment version 4.2.2 (2022-10-31) to conduct this study. All code was written in R Studio. R is a statistical computing language and environment. It is a free software in source code and is highly extendable, providing numerous statistical and graphical techniques, such as linear and nonlinear modeling, classical statistical tests, etc (22).

Data Pre-processing

To begin, we installed and loaded the glmnet library into R studio. Glmnet is a package that fits generalized linear and logistic regression models with the penalized maximum likelihood. The regularization path is determined for the lasso penalty at a grid of values for the parameter lambda (23). We loaded in our unedited soybean data with 307 observations across 19 classes. After analyzing the data and determining the missing value rate for each variable per class, we decided to remove the classes that were missing more than 50% of values across all variables. For example, our class herbicide injury had 80 missing values out of 140 total values with a missing value rate of 57.14%. We omitted those incomplete classes and continued with classes where the model could analyze all or most variables for a given class. Additionally, we removed classes that had less than 20 observations. Our edited dataset contained 160 total observations across 5 diseases, which we separated into 5 classes. There were 20 instances of brown stem rot (class 1) and anthracnose (class 3) and 40 instances of brown spot (class 2), alternaria leaf spot (class 4), and frogeye leaf spot (class 5).

Part 1

To conduct Part 1 of the study – using a single fixed lambda value of 0.001 (without tuning) to perform a five-class classification problem – we turned our edited data into a matrix of variable “x” and created training labels for each class by generating the class number repeatedly for each observation. We generated the number 1 twenty times for the 20 observations of class

1, continuing for the next four classes and saving those labels in the variable “y.” We arbitrarily allocated 80% of observations for each class into the training set and the remaining 20% of observations into the testing set. We created our *xtrain*, *ytrain*, *xtest*, and *ytest* variables and stored the data accordingly, with the x variables representing the observations and the y variables representing the labels.

Each variable value in the generated matrix “x” was type “character.” We transformed the “character” type into “numeric” to perform the regression function. We fit the model using *glmnet* function with family type “multinomial” to correctly represent the multi-classification problem of sorting the data into five classes. We used the *predict()* function to test the “unseen” test data using families “response” for the prediction probability and “coefficient” for the variables’ coefficients per class. We utilized the *apply()* function with our variable “*prob*” and “*which.max*” to select the greatest prediction value per test observation for each class out of total value 1, which we saved into the variable “yhat.” The class with the greatest probability for a given test observation out of 1 would be the class the model selected to sort that observation into. We took the mean of where yhat and ytest were not equal to find the prediction error of the untuned model for the testing set.

Part 2

Part 2 of the study utilized cross validation to tune the model. A regularization regression procedure like LASSO involves a tuning parameter called lambda, which balances the

trade-off between the data fit and the final model complexity. If the value of lambda is too small (e.g., close to zero), the procedure might overfit the data and result in a complex model; if the value is too large, the procedure might underfit the data and return an overly simple model. Therefore, the choice of lambda is critical to ensure a good balance. The process of choosing a proper lambda is called tuning. An untuned model uses a default value provided in the package, which is not tailored for the dataset to give the best performance. A tuned model employs a data-adaptive procedure such as cross validation to select the value of lambda optimal for the data set. In this study, we used cross validation (CV) for tuning in two ways. In the first method, we selected the lambda value that minimized the cross validation error, the resulting value was hence called lambda.min. In the second method, we chose the largest lambda value whose CV error fell within one-standard error of the CV error of lambda.min, and the resulting value was designated lambda 1se (24). In practice, lambda.1se tended to produce a simpler model than lambda.min and achieved higher sparsity. We used the same training and testing sets as in Part 1. To compute the cross validation error, we set the seed to 100 and used *cv.glmnet* to perform cross-validation for *glmnet* and return a value for lambda.min of 0.062 and lambda.1se of 0.082. We used both lambda.min and lambda.1se values to predict test data and achieved the same accuracy of 87.5%. We obtained the accuracy values for each disease class by calculating the percentage of samples in the testing set that the algorithm correctly classified out of the total testing set. For overall accuracy values, we combined the testing sets

from all disease classes and performed the same calculation.

Results

The cross-validated LASSO regression model classified brown stem rot with 100% accuracy with a $\lambda_{\min}$ value of 0.062 and λ_{1se} value of 0.082. After performing variable selection with LASSO regression for the untuned model, we discovered that attribute variable 26 (internal discoloration) had the greatest non-zero coefficient of 15.48. A positive non-zero coefficient symbolizes a direct correlation between the increase of an attribute value's encoded number and the probability of disease. In other words, as the attribute value's number increases, the potential for disease also increases. For negative non-zero coefficients, the opposite can be said – there is an inverse correlation between the attribute value's encoded number and the probability of disease. For example, variable 26 (internal discoloration) has values of “none, brown, and black” encoded as “1, 2, 3” as seen in Table 2. Variable 26 was given a positive non-zero coefficient, so as the

variable's values correspond with higher numbers (brown- 2 and black- 3), they are increasingly positively correlated to the onset of brown stem rot. For the tuned model utilizing $\lambda_{\min}$ value, variable 26 was also the most important feature for class 1 with a value of 5.48. For the tuned model utilizing λ_{1se} , the most prominent feature was variable 26 with a value of 4.76. All other variables displayed a coefficient of 0. Table 2 shows the feature coefficients for brown stem rot disease with the untuned and tuned models. Brown stem rot is caused by the soil-borne fungus *Cadophora gregata*. Foliar symptoms manifest when pods begin to fill, but the pathogen infects the roots early in the season (25). Depending on environment and pathogen genotype, soybean leaf necrosis may occur alongside vascular browning. Additionally, the pith of longitudinally split stems is discolored as a dark, chocolate-brown (26). Figure 1 depicts brown stem rot. Our findings suggest that paying attention to internal/vascular discoloration, e.g. brown/black/no discolor may help with earlier identification of the disease.



Figure 1: Photo of brown stem rot

Table 2: Feature coefficients for brown stem rot disease

Variable	Coefficient (untuned)	Coefficient (lambda.min)	Coefficient (lambda.1se)
26	15.48	5.48	4.76

The model classified brown spot with 87.5% accuracy with a lambda.min value of 0.062 and lambda.1se value of 0.082. We discovered that attribute variable 23 (fruiting bodies) had the greatest positive non-zero coefficient of 8.66. Variable 1 (date planted) had the greatest negative non-zero coefficient, which meant that the later the date planted (August-October), the less probable the onset of brown spot became. Brown spot is favored by warm and humid conditions, with temperatures ranging from 77-86°F (25-30°C) being optimal for disease development, and it is more commonly observed in areas with warmer climates and months with higher temperatures (April-July) (27). For the tuned model utilizing lambda.min

value, variable 16 (leaf shred) was the most important feature for class 2. For the tuned model utilizing lambda.1se, the most prominent feature was also variable 16. Table 3 shows the feature coefficients for brown spot disease with the untuned and tuned models. Soybean brown spot is caused by the fungal pathogen *Septoria glycines* and manifests as irregular brown patches of dead tissue on leaves, causing defoliation in severe cases (28). The fruiting bodies of fungus develop in lesions on infected cotyledons and unifoliate leaves, providing spores for later infections of trifoliate leaves, stems, pods, and other neighboring plants (29). Figure 2 depicts brown spot.



Figure 2: Photo of brown spot

Table 3: Feature coefficients for brown spot disease

Variable	Coefficient (untuned)	Coefficient (lambda.min)	Coefficient (lambda.1se)
1	-2.80	-0.72	-0.64
5	0.07	0	0
6	0.22	0	0
7	0.78	0.02	0
16	0.20	0.94	0.83
23	8.66	0.50	0.10
28	-2.14	0	0
30	-1.29	0	0
31	-0.02	0	0

Anthracnose was classified with 100% untuned and tuned models. Anthracnose is accuracy utilizing both a lambda.min value of 0.062 and lambda.1se value of 0.082. After performing variable selection, we discovered that attribute variable 31 (leaf spots margin) had the greatest non-zero coefficient of 2.56. For the tuned model utilizing lambda.min value, variable 15 (leaf spot size) was the most important feature for class 3. For the tuned model utilizing lambda.1se, the most prominent feature was also variable 15. Table 4 shows the feature coefficients for anthracnose with the

most commonly caused by the fungus *Colletotrichum truncatum*, which produces crowded, black acervuli on the infected tissues. When infected seeds are planted, sunken, dark brown lesions develop on the cotyledons of seedlings, which may then expand to the stem and kill the plant. Plants can become infected at any stage of developments, and the most common symptoms include irregular-shaped brown spots on the stem, pods, and petioles (30). Figure 3 depicts anthracnose.

Table 4: Feature coefficients for anthracnose disease

Variable	Coefficient (untuned)	Coefficient (lambda.min)	Coefficient (lambda.1se)
3	5.18×10^{-1}	0	0
13	-3.22×10^{-1}	0	0
14	9.34×10^{-17}	0	0
15	1.04×10^1	3.62	3.06
21	6.93×10^{-1}	0	0
30	5.98×10^{-1}	0.29	0.14
31	2.56	0.54	0.53
33	8.25×10^{-1}	0	0



Figure 3: Photo of Anthracnose

The model classified the alternaria leaf spot with 100% accuracy with a lambda.min value of 0.062 and lambda.1se value of 0.082. Variable 9 (seed tmt) had the greatest negative non-zero coefficient of -1.70. This meant that seeds that were treated with fungicide had a lesser chance of developing alternaria leaf spot, as the factor of seed treatment was inversely correlated with the onset of disease. Variable 1 (date planted) had the greatest positive non-zero coefficient, indicating that the probability of developing alternaria leaf spot became greater in the later months. This supports research showing that alternaria leaf spot typically thrives in climates with cool to moderately warm temperatures and high humidity, mainly being observed on mature plants in late summer or early fall when temperatures begin to drop (31). For the tuned model utilizing lambda.min value, variable 19 (stem) was the most important feature for class 4. For the tuned model utilizing lambda.1se, the most prominent feature was variable 19. Table 5 shows the feature coefficients for alternaria leaf spot with the untuned and tuned models. Alternaria leaf spot is caused by the fungus *Alternaria brassicicola*. Physical symptoms include large brown spots on leaves, around 0.5 to 0.75 inches in diameter. These spots turn black as the fungus produces spores (32). Alternaria leaf spot is one of the most destructive foliar diseases in soybean growing areas around the world. Their pathogenicity has also increased due to the present climate-change scenario (33). Figure 4 depicts alternaria leaf spot.

Table 5: Feature coefficients for alternaria leaf spot disease

Variable	Coefficient (untuned)	Coefficient (lambda.min)	Coefficient (lambda.1se)
1	1.49	0.16	0.06
4	1.09	0	0
6	-0.37	0	0
7	-0.18	0	0
9	-1.70	0	0
10	0.17	0	0
19	-1.06	-1.87	-1.57
21	-1.60	0	0
32	1.35	0	0

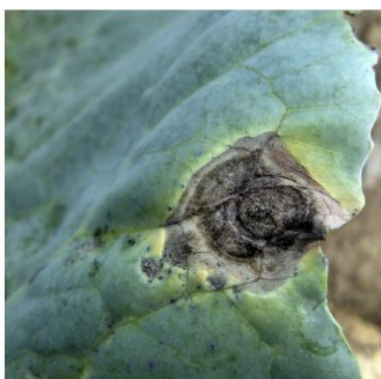


Figure 4: Photo of alternaria leaf spot

Frogeye leaf spot was classified with 62.5% accuracy utilizing a lambda.min value of 0.062 and lambda.1se value of 0.082. Variable 16 (leaf shred) had the greatest negative non-zero coefficient of -4.83. This meant that leaf shred – the manual shredding of leaves to prevent them from packing down when wet and smothering the plants they are meant to protect – and disease probability were inversely correlated, so the greater the leaf shred, the lesser the probability that the soybean plant would develop frogeye leaf spot. Variable 28 (fruit pods) had the greatest positive non-zero coefficient, with diseased fruit pods – which exhibited unusual features such as deformities in shape/size and discoloration – and few-present fruit pods having the greatest effect on the probability of frogeye leaf spot presence. For the tuned model utilizing lambda.min value, variable 28 was the most important feature for class 5. For the tuned model utilizing lambda.1se, the most prominent feature was variable 28. Table 6 shows the feature coefficients for frogeye leaf spot with

the untuned and tuned models. Frogeye leaf spot is caused by the fungus *Cercospora sojina* and is a common foliar disease of soybeans. Infection typically occurs during reproductive growth stages but can occur during any stage of soybean development. Symptoms begin as dark water-soaked lesions on younger leaves with centers that become ash-gray to light-brown. The lesions later become circular to angular with a purple to dark brown margin around the center. On leaf undersides, the lesion center may have a dark area where spores are being produced (24). Figure 5 depicts frogeye leaf spot.

Table 6: Feature coefficients for frogeye leaf spot disease

Variable	Coefficient (untuned)	Coefficient (lambda.min)	Coefficient (lambda.1se)
2	-0.11	0	0
4	-0.96	0	0
5	-2.28	0	0
8	-1.13	0	0
10	-0.40	0	0
12	0.51	0	0
16	-4.83	0	0
23	-0.57	-0.27	-0.07
24	0	0.20	0.23
28	3.71	0.99	0.73
32	-1.11	0	0



Figure 5: Photo of frogeye leaf spot

Discussion

In this study, we performed experiments based on a novel approach of classifying common soybean diseases. Our machine learning model provided a quick, accurate diagnosis of soybean diseases based on a combination of plant-specific and surrounding features, supplementing traditional methods of utilizing hand tools and the naked eye to detect soybean diseases. These traditional methods lead to inaccuracies given a general limited scope of disease knowledge. Our machine learning model was capable of replacing biological methods – which generate high costs due to additional capital and sample preparation. Our tests, implemented with a publicly available and widely reputable dataset, utilized a combination of phenotypic, pathological, and environmental features to yield a comprehensive and well-informed diagnosis based on a multitude of diverse factors affecting the soybean crop. Our approach was able to identify patterns based on the interconnection between different phenotypic and environmental variables, and, with the implementation of machine learning, could successfully draw conclusions based on those relationships. However, it is worth noting that this also became a limitation of our study. For fungal diseases where the model must solely rely on plant phenotype factors, our model would not be able to predict disease, since it utilized a combination of environmental, growth, and phenotypic factors.

A key component of our study included feature selection, an important step of the machine learning pipeline that allowed us to identify the variables that influenced the model's decision

making processes. LASSO regression allowed us to accomplish this feat by giving us an interpretation of feature importance. Statistical models carry more interpretability from the human perspective when compared to machine learning algorithms such as neural networks, which do not offer any tangible insight into the reasoning behind the model's decisions. This is a phenomenon known as the “black box,” which represents the opaqueness behind a model's results, hence adding a layer of doubt to the model's credibility. LASSO regression fixes this issue by selecting the features useful to the model's assembly of results and discarding the features lacking influence, giving the investigators insight into why and how the model makes the decisions it does. The idea of using LASSO regression for variable selection is also relatively straightforward – we fitted the regression method onto our scaled dataset and only considered the features that had a non-zero coefficient as being useful. Feature selection through LASSO regression allowed us to understand the important relationships between the variables and the target function so as to choose the best model.

In this study, we used LASSO regression to classify five types of common soybean diseases based on independent variables from a public database. We chose to use a multi-target logistic regression model aimed at concurrently predicting output variables based on a diverse set of input variables. Logistic regression is the relation of qualitative variables to other variables through a logistic functional form (34). It works by measuring, i.e. estimating the probabilities, of the relationship between the dependent variable and independent variables.

For brown stem rot disease, the model accurately diagnosed 100% of cases with the influence of a single variable – internal discoloration. Hence, internal discoloration is the most important feature in influencing the onset of brown stem rot. For brown spot disease, this model accurately classified 87.5% of cases with a combination of environmental factors, such as date, and phenotypic factors, such as area damaged, leaf shred, and fruiting bodies. To diagnose anthracnose disease, the model relied on phenotypic features, such as leaf spot size, seed, and mold growth. The diagnosis accuracy for anthracnose was 100% accurate. For alternaria leaf spot disease, the

model relied on date and stem attributes. The diagnosis accuracy was 100%. For frogeye leaf spot disease, the model relied on several attributes including fruit pods, fruiting bodies, and external decay– all together the diagnosis accuracy was 62.5%. The model's predictions for frogeye leaf spot disease were the lowest of all disease categories. In the future, we plan to focus on increasing detection accuracy for this disease. However, with our high detection accuracies for brown stem rot, anthracnose, and alternaria leaf spot, we can conclude that our model is suitable for detecting these diseases. Figure 6 depicts a comparison chart of the five soybean disease prediction accuracies.

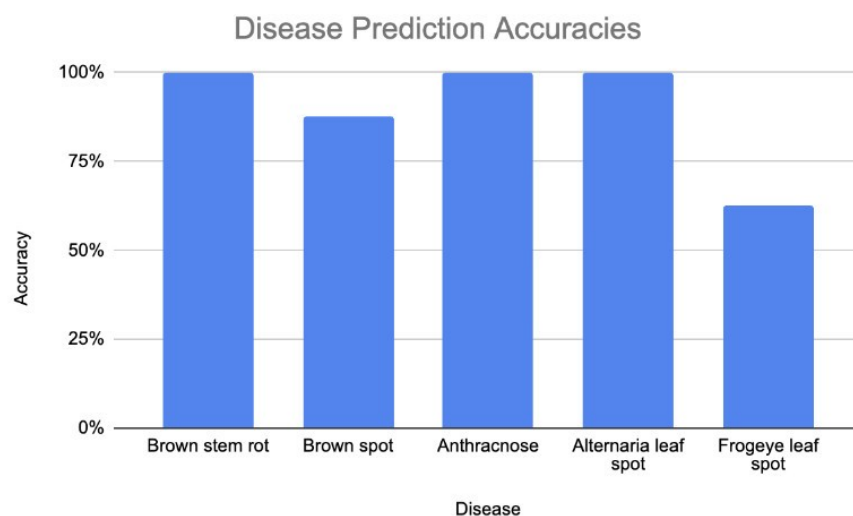


Figure 6: Prediction accuracies for 5 soybean disease classes

Figure 7 shows a column graph of each significant factor's influence across the five diseases using coefficients from λ_{1se} , which gives the most regularized model. From analyzing the column graph, we can see that internal discoloration only factors as an important attribute when identifying brown

stem rot, carrying the largest positive coefficient present on the graph. Brown stem rot infects the vascular tissues of soybean plants and induces a browning or discoloration of the internal tissues – a key diagnostic feature of the disease. Additionally, the factor “fruiting bodies” appeared as a significant factor for

both brown spot and frogeye leaf spot, which are diseases often caused by fungal pathogens that produce visible spore-bearing structures called fruiting bodies, frequently bearing unique characteristics, such as shape, size, and

color that are specific to the pathogen causing disease (35). Fruiting bodies do not affect the identification of brown stem rot, alternaria leaf spot, and anthracnose according to our model.

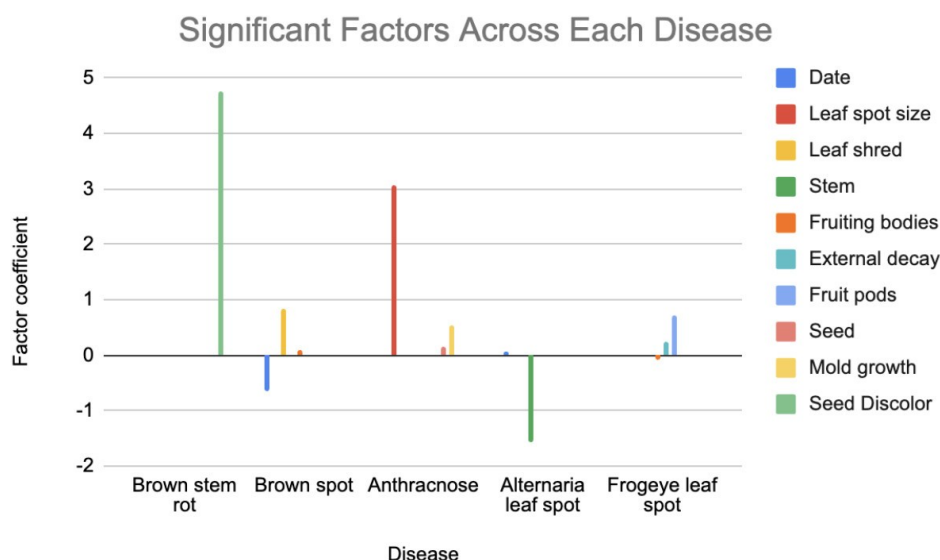


Figure 7: Coefficient comparison for significant factors across the five diseases using $\lambda_{\text{bda.1se}}$

There are some features the model did not deem important to the model's classification process that we found surprising. One particular feature was crop rotation – the practice of planting different crops sequentially on the same plot of land to improve soil health – which is widely recognized as an effective strategy for managing plant diseases, including those affecting leaves, stems, and roots (36). However, research shows that the effect of crop rotation on stem and root diseases becomes apparent after 6 years of rotation and the disease index is lowest in the most diverse crop rotation (37). In our dataset, we only tested

against soybean plants that had a rotation history of 1 year, or no rotation history at all.

Our dataset was limited and did not contain soybean plants that had rotation histories over a 6 year or greater span. In future studies, we hope to obtain soybean data samples with a longer rotation history and perform these tests again to gain a more accurate sense of the role crop rotation plays in preventing common soybean diseases. Soybean seed fungicide – which can help reduce stem, root, and leaf diseases in soybean plants – also did not manifest as an important factor for the algorithm. Fungicide seed treatments are commonly used to protect seeds and seedlings

from soil-borne pathogens and early-season diseases by providing a protective barrier around the seed and seedling, helping to prevent or minimize infection by fungal pathogens present in the soil (38). This protection is crucial during the vulnerable early stages of plant development when diseases can have a significant impact. However, fungicide seed treatments may not provide complete control of all fungal diseases throughout the entire growing season. Factors such as environmental conditions, pathogen variability, and the duration of protection provided by the fungicide can influence their efficacy (39). Local regulations, specific product recommendations, and pathogen selective pressure in different regions may also influence the effectiveness and use of fungicide seed treatments.

Our next step is to add more observations for each of the five classes, which will allow the model to generalize based on more diverse data. Currently, the maximum number of observations our study utilizes for any given disease category is 40 – we plan to increase this number by at least another order of magnitude.

We plan to use data from multiple geographic locations to mimic real world situations from multiple soybean production areas. Additionally, we hope to test more soybean diseases to enhance the model's versatility. For example, soybean rust – one of the most common soybean diseases – was not tested on in this study because of our limited dataset. Our hope is that the final product/algorithm will be deployed to soybean farmers and researchers to aid them in efficiently and effectively detecting and predicting soybean diseases.

Conclusion

In this study, we used logistic regression with LASSO to successfully differentiate five common soybean crop diseases, i.e. brown stem rot, brown spot, anthracnose, alternaria leaf spot, and frogeye leaf spot, with 87.50% accuracy. We illustrated how it is crucial to incorporate physiological and pathological features, as well as growth and environmental conditions into the classification process in order to ensure a successful outcome. This method is easily interpretable, allowing farmers to efficiently diagnose common soybean diseases in real world situations.

References

1. Lee T., Tran A., Hansen J., Ash M. Major Factors Affecting Global Soybean and Products Trade Projections. *USDA Economic Research Service*, 2 May 2016.
<https://www.ers.usda.gov/amber-waves/2016/may/major-factors-affecting-global-soybean-and-products-trade-projections/>.
2. Martinelli F., Scalenghe R., Davino S., Panno S., Scuderi G., Ruisi P., et al. (2015). Advanced methods of plant disease detection. *Rev. Agron. Sustain. Dev.* 35 1–25.
<https://doi.org/10.1007/s13593-014-0246-1>.

3. Holmes G. J., Mansouripour S. M., Hewavitharana S. S. (2020). Strawberries at the Crossroads: management of soilborne diseases in california without methyl bromide. *Phytopathology* 110 956–968. <https://doi.org/10.1094/phyto-11-19-0406-ia>.
4. Bhunjun C.S., Phillips A.J.L., Jayawardena R.S., Promptuttha I., Hyde K.D. (2021). Importance of Molecular Data to Identify Fungal Plant Pathogens and Guidelines for Pathogenicity Testing Based on Koch's Postulates. *Pathogens*. <https://doi.org/10.3390/pathogens10091096>.
5. Venbrux M., Crauwels S., and Rediers H. (2023). Current and emerging trends in techniques for plant pathogen detection. *Frontiers in Plant Science*. <https://doi.org/10.3389/fpls.2023.1120968>.
6. Baćmaga M., Wyszowska J., Kucharski J. (2016). The effect of the Falcon 460 EC fungicide on soil microbial communities, enzyme activities and plant growth. *Ecotoxicology*. <https://doi.org/10.1007/s10646-016-1713-z>.
7. Tsui C.K., Woodhall J., Chen W., Lévesque C.A., Lau A., Schoen C.D., Baschien C., Najafzadeh M.J., de Hoog G.S. (2011). Molecular techniques for pathogen identification and fungus detection in the environment. *IMA Fungus*. <https://doi.org/10.5598/imafungus.2011.02.02.09>.
8. Laugé R., Joosten M., Haanstra J., et al. (1998) Successful search for a resistance gene in tomato targeted against a virulence factor of a fungal pathogen. *Proceedings of the National Academy of Sciences*. <https://doi.org/10.1073/pnas.95.15.9014>.
9. Goates, B.J. (2012) Identification of new pathogenic races of common bunt and dwarf bunt fungi, and evaluation of known races using an expanded set of differential wheat lines. *Plant disease*. <https://doi.org/10.1094/PDIS-04-11-0339>.
10. Fang Y., Ramasamy R.P. Current and Prospective Methods for Plant Disease Detection. *Biosensors (Basel)*. 2015 Aug 6;5(3):537-61. <https://doi.org/10.3390/bios5030537>.
11. Sakudo A., Suganuma Y., Kobayashi T., Onodera T., Ikuta K. (2006). Near-infrared spectroscopy: Promising diagnostic tool for viral infections. *Biochemical and Biophysical Research Communications*, Volume 341, Issue 2, 279-284. <https://doi.org/10.1016/j.bbrc.2005.12.153>.

12. Nagasubramanian K., Jones S., Sarkar S., Singh A. K., Singh A., Ganapathysubramanian B. (2018). Hyperspectral band selection using genetic algorithm and support vector machines for early identification of charcoal rot disease in soybean stems. *Plant Methods* 14 86–99. <https://doi.org/10.1186/s13007-018-0349-9>.
13. Chlingaryan A., Sukkarieh S., 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. <https://doi.org/10.1016/j.compag.2018.05.012>.
14. Liakos K. G., Busato P., Moshou D., Pearson S., Bochtis D. (2018). Machine learning in agriculture: a review. *Sensors* 18:2674. <https://doi.org/10.3390/s18082674>.
15. Wang A., Zhang W., Wei X. (2019). A review on weed detection using ground-based machine vision and image processing techniques. *Comput. Electron. Agric.* 158 226–240. <https://doi.org/10.1016/j.compag.2019.02.005>.
16. Patricio D. I., Rieder R. (2018). Computer vision and artificial intelligence in precision agriculture for grain crops: a systematic review. *Comput. Electron. Agric.* 153 69–81. <https://doi.org/10.1016/j.compag.2018.08.001>.
17. Ashfaq R. A. R., Wang X. Z., Huang J. Z., Abbas H., He Y. L. (2017). Fuzziness based semi-supervised learning approach for intrusion detection system. *Inform. Sci.* 378 484–497. <https://doi.org/10.1016/j.ins.2016.04.019>.
18. Rimbaud L., et al. Assessing the durability and efficiency of landscape-based strategies to deploy plant resistance to pathogens. *PLoS computational biology* 14.4 (2018): e1006067. <https://doi.org/10.1371/journal.pcbi.1006067>.
19. Lin J., et al. Seed-germination response of *Leymus chinensis* to cold stratification in a range of temperatures, light and low water potentials under salt and drought stresses. *Crop and Pasture Science* 68.2 (2017): 188-194. <https://doi.org/10.1071/CP16402>.
20. Soybean (Large Dataset). *UCI Machine Learning Repository*. [https://archive.ics.uci.edu/ml/datasets/Soybean+\(Large\)](https://archive.ics.uci.edu/ml/datasets/Soybean+(Large)).
21. Tibshirani R, Regression Shrinkage and Selection via the Lasso, *J. R. Statist. Soc. B*, 1996 (58), No. 1, pp 267-288. <https://www.jstor.org/stable/2346178>.
22. The R Project for Statistical Computing. <https://www.r-project.org/>.

23. Hastie T., Qian J., Tay K. An Introduction to glmnet. 27 March 2023.
<https://glmnet.stanford.edu/articles/glmnet.html#:~:text=Glmnet%20is%20a%20package%20that,for%20the%20regularization%20parameter%20lambda.>
24. Klau, S. prioritylasso Vignette. 30 March 2023.
https://cran.r-project.org/web/packages/prioritylasso/vignettes/prioritylasso_vignette.html#:~:text=1se%20.,lambda.,leads%20to%20more%20sparse%20results.
25. Westphal, A., Abney, T.S., Shaner, G. Disease of Soybean. *Purdue University*.
<https://www.extension.purdue.edu/extmedia/BP/BP-41-W.pdf>.
26. Identifying Common Soybean Diseases During Vegetative Growth. 10 July 2019.
<https://www.fontanelle.com/en-us/agronomy-library/identifying-common-soybean-diseases-during-vegetative-growth.html>.
27. Brown Spots in Grass: Identification and Prevention. *This Old House*, 17 Nov. 2022.
<https://www.thisoldhouse.com/lawns/reviews/what-causes-brown-spots-in-lawns#:~:text=Brown%20patch%20disease%20is%20caused,by%20a%20smoke%20ring%20border.>
28. Ernst A., Thiessen L. Brown Spot of Soybean. *NC State Extension Publications*, 3 Sept. 2020. <https://content.ces.ncsu.edu/brown-spot-of-soybean>.
29. Dorrance, A.E., Mills D.R. Brown Spot of Soybeans. *Ohio State University Extension*, 9 Feb. 2011. <https://ohioline.osu.edu/factsheet/AC-18>.
30. Giesler L.J. Anthracnose. *UNL Cropwatch*.
<https://cropwatch.unl.edu/plantdisease/soybean/anthracnose>.
31. Adaskaveg J.E., Gubler W.D., Duncan R.A., Stapleton J.J., Holtz H.A., Trouillas F.P. Alternaria Leaf Spot. UC IPM, Aug. 2017.
<https://ipm.ucanr.edu/agriculture/almond/alternaria-leaf-spot/>.
32. Fagodiya, R.K., Trivedi, A. & Fagodia, B.L. Impact of weather parameters on Alternaria leaf spot of soybean incited by *Alternaria alternata*. *Sci Rep* 12, 6131 (2022).
<https://doi.org/10.1038/s41598-022-10108-z>.
33. Lin, B. and Kelly, H (2018). Frogeye Leaf Spot of Soybean. *The Plant Health Instructor*.
<https://www.apsnet.org/edcenter/disandpath/fungalasco/pdlessons/Pages/FrogeyeLeafSpot.aspx>.

34. S. James Press, and Sandra Wilson. Choosing Between Logistic Regression and Discriminant Analysis. *Journal of the American Statistical Association*, vol. 73, no. 364, 1978, pp. 699–705. <https://doi.org/10.2307/2286261>.
35. Pataky, N. What is a Fruiting Body? *University of Illinois Extension*, 23 June 2004. <https://hyg.ipm.illinois.edu/pastpest/200410d.html#:~:text=Fructing%20bodies%20are%20fungal%20structures,identification%20of%20the%20specific%20fungus>.
36. Crop Rotations. *Rodale Institute*. <https://rodaleinstitute.org/why-organic/organic-farming-practices/crop-rotations/>.
37. Jalli M., et al. Effects of crop rotation on spring wheat yield and pest occurrence in different tillage systems: a multi-year experiment in Finnish growing conditions. *Frontiers in Sustainable Food Systems* 5 (2021). <https://doi.org/10.3389/fsufs.2021.647335>.
38. Strunk, C. Fungicide Seed Treatment Considerations for 2019. *South Dakota State University Extension*, 9 Feb. 2022. <https://extension.sdstate.edu/fungicide-seed-treatment-considerations-2019>.
39. Kerns, Jim. "Factors affecting fungicide performance." June 2012. <https://archive.lib.msu.edu/tic/groot/article/2012may8.pdf>.