



Research article

Exploring frogeye leaf spot disease severity in soybean based on hyperspectral data analysis and machine learning with Orange data mining

Yuhao Ang^a, Helmi Zulhaidi Mohd Shafri^{b,*}, Mohammed Mustafa Al-Habshi^b

^a Faculty of Sustainable Agriculture, Universiti Malaysia Sabah Sandakan Campus, Locked Bag No. 3, 90509 Sandakan, Sabah, Malaysia

^b Department of Civil Engineering and Geospatial Information Science Research Centre (GISRC), Faculty of Engineering, Universiti Putra Malaysia (UPM), 43400 Serdang, Selangor, Malaysia

Article Info

Article history:

Received 14 May 2024

Revised 19 August 2024

Accepted 24 December 2024

Available online 28 April 2025

Keywords:

Feature selection,

Hyperspectral remote sensing,

Machine learning,

Orange data mining software

Abstract

Importance of the work: The advancement of hyperspectral remote sensing technology has facilitated the examination of its potential for categorizing frogeye leaf spot (FLS) severity in soybean. No study has yet investigated the Orange mining tool as a visual programming approach to analyze hyperspectral reflectance data, especially in crop disease detection.

Objectives: To classify the severity level of FLS disease in soybean using hyperspectral reflectance data and machine learning algorithms.

Materials and Methods: Hyperspectral reflectance data were used from healthy and FLS-affected soybeans. Initially, the data were smoothed by applying the Savitzky-Golay filtering technique to remove spectrum noise. The ReliefF feature selection technique was used to determine the most influential wavelengths for the classification of FLS disease severity in soybean. Next, machine learning methods (decision tree, gradient boosting, random forest, stacking and neural network) were used to classify FLS severity in soybean. The performance was evaluated using overall accuracy, F1, precision and the curve metric receiver-operating characteristic. All these steps were conducted using the Orange data mining software.

Results: Neural network scored the highest overall accuracy (98.6%) after conducting the filtering technique. Furthermore, the ReliefF-gradient boosting and the random forest algorithms achieved promising overall levels of accuracy (97.4% and 96.9%, respectively) after implementing the filtering and feature selection techniques.

Main finding: The integration of the workflow and the specially designed spectroscopic widget in the Orange data mining software made it possible to process the hyperspectral reflectance data and to determine the severity level of the disease on the affected crop samples.

* Corresponding author.

E-mail address: helmi@upm.edu.my/hzms2312@gmail.com (H.Z.M. Shafri)

online 2452-316X print 2468-1458/Copyright © 2025. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), production and hosting by Kasetsart University Research and Development Institute on behalf of Kasetsart University.

<https://doi.org/10.34044/j.anres.2025.59.2.01>

Introduction

Soybean is the second largest crop producing primarily edible oils and a major source of proteins (Dreoni et al., 2022). Globally, the majority of soybeans are fed to pigs (Parrini, 2023). The global total soybean production in 2021 was 367.76 million tons with the forecast that soybean production would decline by 4.63% or 17.04 million tons in the near future (Tetrault, 2023). However, the seed and the quality of soybean can be damaged by pathogenic microorganisms. Frogeye leaf spot (FLS) is a soybean foliar disease caused by the fungus *Cercospora sojina* Hara (CSH), which can lead to reduced photosynthetic leaf area, premature defoliation and reduced seed weight, resulting in yield losses of 31–60% (Phillips et al., 2021; Barro et al., 2023). FLS lesions can affect the leaves, pods and stems at any stage of plant development (Kim et al., 2013). The lesions start as small, dark spots that turn tan-to-brown in color, with a narrow, purple-brown border margin (Borah & Deb, 2022). Thus, it is critical to identify the spread of FLS disease and to effectively combat it with proper strategies to achieve more sustainable production.

Currently, the identification of the disease in its early stages of infestation is based on visual assessment to prompt a response action, since the disease can present in any stage. This involves assessing aspects such as the size of the lesion area, the color patterns, the distribution and shape of the leaves, the number of stems and branches and the density of the plants (Vishnoi et al., 2021). Performing uniform planting inspections can be challenging when cultivating large areas, as it requires more human resources. However, this technique is prone to bias and can be manipulated by the observations of experts due to randomness (Xie et al., 2015). Therefore, an accurate, effective and non-destructive technology is required to assess the severity of FLS disease on soybean.

With the advancement of remote sensing, hyperspectral remote sensing has potential for use in classifying the severity of soybean disease. Hyperspectral data encompass many channels or bands with narrow bandwidths that have the ability to identify subtle abnormalities of crops (Lu et al., 2020). Frequently, hyperspectral techniques are utilized to identify the biophysical properties of crops, such as nutrient deficiency, moisture content, chlorophyll level and cell structure (Berger et al., 2020; Bruning et al., 2020). Most studies have proved the effectiveness and the efficiency of hyperspectral techniques in classifying crop disease severity, such as late blight with tomato (Zhang et al., 2003), *Ganoderma* with oil palm

(Lelong et al., 2010) and powdery mildew disease with wheat (Khan et al., 2021).

Rapid development of an automation system for classifying a diseased crop is a progressing area in precision agriculture. Machine learning (ML) methods have been used to build an accurate classification model using hyperspectral reflectance data for disease detection such as support vector machine, random forest (RF) and artificial neural network. In fact, hyperspectral data contain redundant, extra and highly correlated wavelengths that may increase the burden of computation. Therefore, data dimensionality reduction and feature selection are required, which transform and reduce the wavelengths and eventually optimize the detection accuracy. Various methods have been used effectively to reduce the dimensionality of hyperspectral data, including principal component analysis (PCA; Liu et al., 2010), recursive feature elimination (Wei et al., 2021) and the successive projection algorithm (SPA; Al-Saddik et al., 2019). Wei et al. (2021) tested several feature selections methods and ML classifiers to identify diseased peanut infected with *Athelia rolfsii*. Based on their results, they reported that recursive feature elimination with random forest and support vector machine outperformed the chi-square and select from model approach, with both implemented using random forest and support vector machine procedures. Navrozidis et al. (2023) conducted a study on the use of the feature selection technique and ML algorithms for the detection of disease in olive trees using hyperspectral data. Based on their results, recursive feature elimination and mutual information were effective in optimizing the classification accuracy. The RF and XGBoost algorithms achieved optimal performance, with a reduced number of hyperspectral features, resulting in a relative operating characteristic-area under the curve score of 1.00 in both cases.

Orange is an open-source tool for the analysis of data along using non-linear relationships, which can be implemented using ML and DM mining technique with a visual programming approach (Demšar et al., 2013). The advantage of using Orange is that it can perform ML without requiring any coding effort in the programming. Currently, no published studies have utilized this tool for the analysis of hyperspectral reflectance data, especially in crop disease detection. Therefore, the current study utilized the Orange software to classify the severity of FLS disease. This study was the first to use such a platform to perform hyperspectral analysis involving pre-processing, feature selection and the classification approach in the Orange toolbox. Therefore, the main objectives of this study were: (1) to classify the severity level of FLS disease, based on

severity classes in soybean using hyperspectral reflectance and machine learning algorithms; (2) to apply a feature selection technique to estimate the severity level of FLS disease; and (3) to evaluate the performance of the ML model in the FLS disease severity classification.

Materials and Methods

Data acquisition

The leaf hyperspectral reflectance data were collected by a group of researchers from the University of China and their work has been appropriately cited (Liu et al., 2021). The leaf hyperspectral reflectance data were acquired using a FieldSpec® HandHeld 2 spectrometer (Analytical Spectral Devices, Inc.; CO, USA). Having 512 hyperspectral bands, the hyperspectral region covered the wavelength range 325–1075 nm, with a resolution of 3 nm. There were 440 samples collected (340 diseased leaves and 100 healthy leaves). In the current study, the disease severity assessment was categorized into four distinct classes based on the percentage of leaf area affected by FLS symptoms, in accordance with the technical specifications for evaluating soybean frogeye leaf spot. Class 1 represented early, mild symptoms (0–1% of the leaf area affected), while Class 2 indicated a moderate level of disease progression (1–3% of the leaf area affected). Class 3 signified a substantial spread of the disease (3–6% of the leaf area affected). Class 4 comprised a higher percentage of lesions (6–20% of the leaf area affected). Additionally, leaves with no visible symptoms were classified as Class 0, representing healthy, unaffected plants.

Pre-processing

The wavelengths with noise were removed using ‘cut (remove)’ in a spectral pre-processing widget (Fig. 1). As a result, the wavelengths were reduced from 981 nm to 1001 nm,

with the remaining wavelengths in the range 450–980 nm, with a total of 531 hyperspectral bands. The Savitzky-Golay filter was used as a pre-processing technique before wavelength selection to smooth the original data and eliminate any parts of the spectrum containing noise (Sun et al., 2021). The configuration settings, such as length of the window and the order of polynomial interpolation, were set based on the formation of first and second derivatives. The selection of the length of the window and the order of polynomial interpolation were determined based on an experimental trial to optimize the noise reduction while minimizing the distortion of the spectral characteristics. For the first derivative, the window length and order of polynomial interpolation were 31 and 2, respectively, while in the second derivative, the window length and order of polynomial interpolation were 51 and 3, respectively.

Processing

Implementing feature selection technique

In this study, the ReliefF algorithm was used as a feature selection technique to select the important wavelengths and eliminate the wavelengths that were not relevant. The ReliefF algorithm applies a filter approach that calculates a proxy statistic for each feature, which can be used to estimate feature weights or feature ‘relevance’ to determining the endpoint value (Urbanowicz et al., 2018). ReliefF depends on a ‘number of neighbors’ user parameter k that specifies the use of the K nearest hits and the k nearest misses in the scoring updates for each target instance.

ReliefF determines the k nearest misses from each ‘other’ class, and averages the weight update based on the prior probability of each class in every iteration. As a result, the weight estimation accuracy is improved, especially when dealing with noisy data.

Model training

Before starting the training process, it was necessary to conduct the ReliefF technique to select the important wavelengths. ML models were trained in this study using cross-validation, which was used as a resampling strategy. The original data were first split into K folds, referring to the number of groups into which a particular data sample was to be divided. Then, the portion of $k-1$ (k minus 1) folds was included into the model for training, while the remaining K th fold was assigned automatically for validation. The k -fold value was 10 in this study. The trial-and-error approach was used to optimize the parameters of the ML algorithms during the training process, which contributed to achieving the highest level of accuracy.

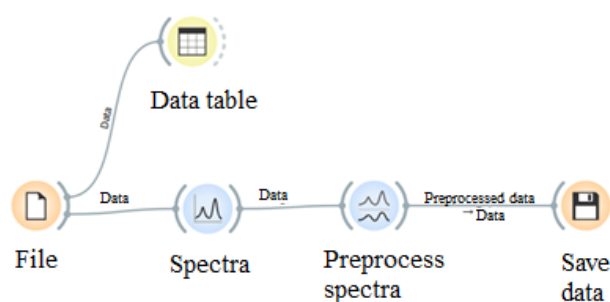


Fig.1 Workflow of pre-processing used

Modeling methods

Several ML algorithms were applied to classify the severity of FLS disease in soybean. Decision tree, random forest, stacking, gradient boosting and neural network algorithms were chosen, since they have been used widely in hyperspectral

reflectance analysis for crop disease detection (Lowe et al., 2017; Huang et al., 2022; Mustafa et al., 2022). Each of the ML algorithms was run using the spectroscopy widget in the Orange software. The workflows of the ML algorithms used for this study are presented in Figs. 2 and 3.

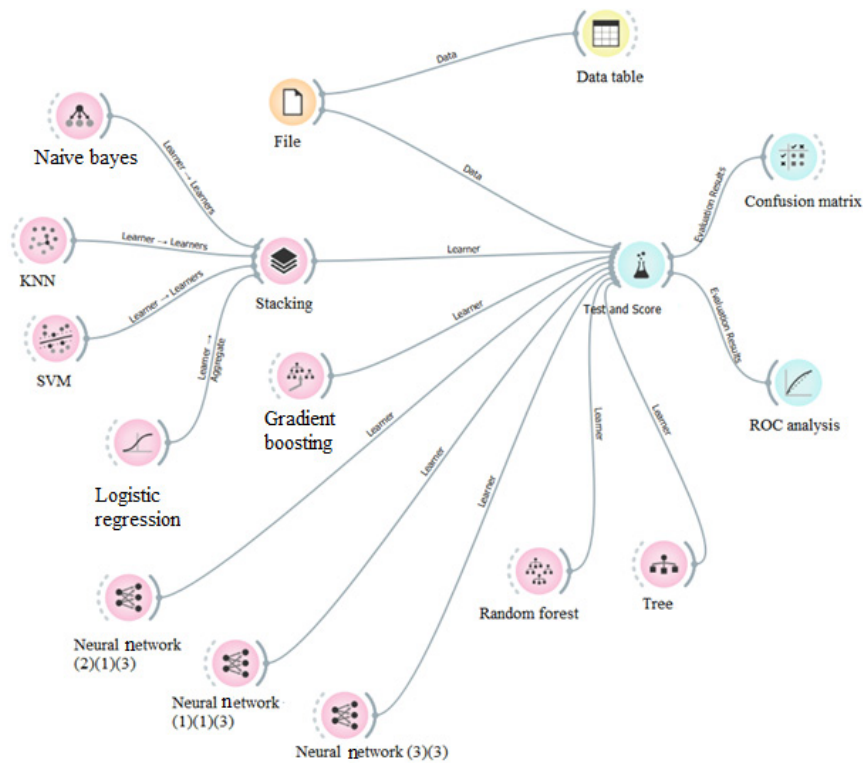


Fig.2 Workflow without feature selection, where ROC = relative operating characteristic

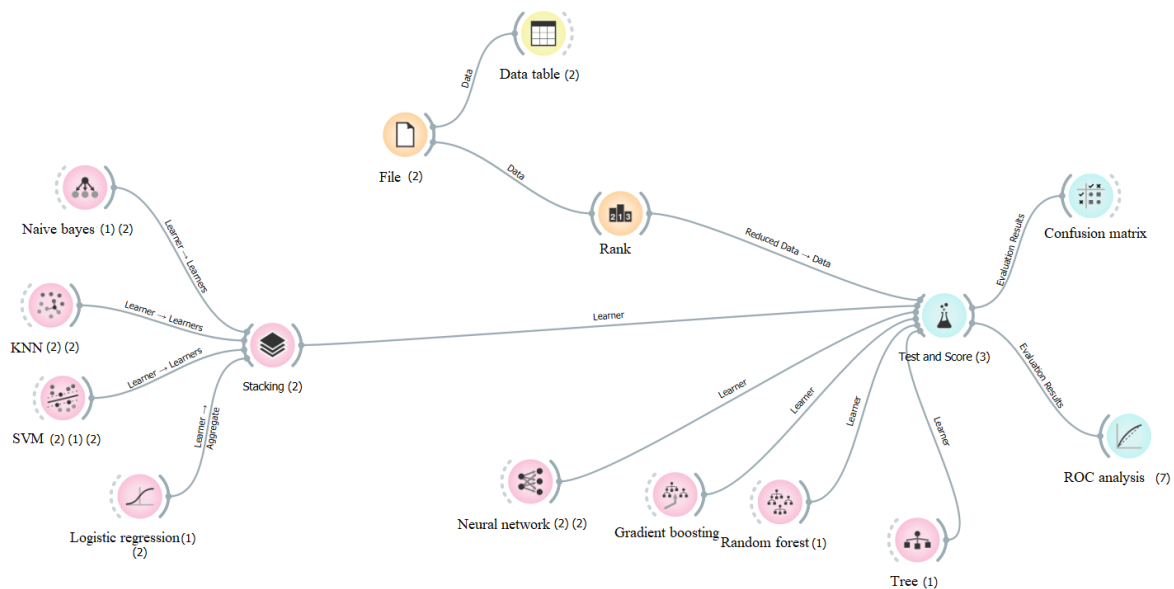


Fig.3 ML workflow using feature selection technique, where ROC = relative operating characteristic

Decision tree

A decision tree is a tree-like classifier where internal nodes represent dataset features, branches represent decision rules and leaf nodes represent outcomes (Myles et al., 2004). The tree-based model generates all possible solutions based on conditional criteria. The algorithm starts at the root node, compares the record's attribute values to the root node and then follows the corresponding branch to the next node. This process repeats, comparing attribute values at each node, until a leaf node is reached. Based on a trial-and-error approach, the parameters for the minimum sample leaf, the minimum number of samples required in a subset and the maximum tree depth were set to 2, 5 and 100, respectively. The parameter values of the algorithm used in this study are shown in Table 1.

Table 1 Parameter values of decision tree algorithm used in this study

Parameter	Value
Minimum sample leaf	1, 2, 5, 10
Minimum number of samples required in a subset	2, 5, 10, 20
Maximum tree depth	3, 5, 10, 50, 100
Stop when majority reaches [%]	95

Stacking

A stacking technique involves training heterogeneous base learners in parallel, using different ML models. By training a meta-learner based on the predictions of varied base learners, heterogeneous base-learners are aggregated into a prediction by combining the output from each individual base learner. Prediction accuracy is improved by combining the input predictions made by the base-learners with the output generated by the training dataset (Zenko et al., 2001). In this study, Naïve Bayes, k-nearest neighbor and support vector machine were used as base learners, whereas logistic regression was regarded as a meta-learner.

Random forest

The random forest approach contains multiple trees that are combined and decision trees that are assembled through a process called 'bagging' in order to generate a final model (Belgiu and Drăgu, 2016). Bootstrapping is used in this method to select 'k' samples by chance from the original dataset. These samples are used to construct decision trees and generate an output. The final output is generated by averaging the predictions of each decision tree. The parameters of the model were adjusted accordingly based on a trial-and-error approach. Parameters, such as number of trees, number of attributes considered at each split and the minimum number of samples required in a subset were set to 3, 5 and 5, respectively.

The parameter values of this algorithm set in this study are indicated in Table 2.

Table 2 Parameter values of random forest used in this study

Parameter	Value
Number of trees	3, 5, 20, 50, 100
Number of attributes considered at each split	5, 10, 15, 20
Minimum number of samples required in a subset	2, 5, 7, 10

Gradient boosting

Gradient boosting is a popular ensemble algorithm that fits boosted decision trees by minimizing the gross prediction error (Natekin and Knoll, 2013). A bunch of decision models was first added and constructed and fitted to correct the prediction errors made by prior models through a series of weight updates. In this algorithm, an additive model is built in a forward stage-wise fashion to optimize any differentiable loss function. Parameters such as the number of trees, maximum tree depth, minimum number of samples required in a subset and the fraction of training instances, were set to 100, 3, 2 and 1, respectively, after conducting a trial-and-error process. The learning rate was set to 0.1. The parameter values of this algorithm used for this study are shown in Table 3.

Table 3 Parameter values of gradient boosting used for this study

Parameter	Value
Number of trees	3, 5, 20, 50, 100
Learning rate	0.01, 0.1, 0.3
Maximum tree depth	3, 5, 20, 50, 100
Minimum number of samples required in a subset	2, 5, 10, 20
Fraction of training instances	0.2, 0.4, 0.6, 1.0

Neural network

A neural network is composed of an input layer, hidden layers and output layers. The input layer receives data from the user and applies transformations to the input data in the hidden layers using activation functions and weights. Lastly, the output layer produces the final classification prediction for the given inputs. The input data are transferred from the input layer to the hidden layer to train the network. Hidden layers are assigned to enable the network to apply transformations to the data and process intermediate representations. A loss function is minimized through backpropagation, which iteratively adjusts and updates the weights and biases during the training process (Szandala, 2021). The neural network can make predictions by passing the data to the output layer through the network, leveraging the learned neuron connections and weights in the hidden layers. The activation function and solver used in

this study were Relu and Adam, respectively. The Relu function introduces non-linearity into the network by preserving positive input values and setting negative values to 0, which facilitates efficient learning of complex patterns and prevents the vanishing gradient problem (Dubey et al., 2022). After a process of trial and error, the neural network was configured with three hidden layers, containing 100, 50 and 30 neurons in each respective layer. The maximum number of iterations was set to 200. The parameter values of the neural network used in this study are indicated in Table 4.

Table 4 Parameter values of neural network used in this study

Parameter	Value
Hidden layer	3
Neuron in hidden layers	100,50,30
Maximal number of iterations	50, 100, 200, 300
Activation	Relu
Solver	Adam

Model evaluation

The model performance was assessed and evaluated for each of the classifiers used in the prediction set based the testing data. The overall accuracy was calculated from the confusion matrix by adding the number of correctly classified classes and dividing it by the total number of the severity of soybean FLS. Other evaluation metrics, such as F1 and precision, were also used in the model evaluation. The relative operating characteristic (ROC) measure was used to quantify the prediction accuracy of the predictive model, being based on the trade-off between the true positive rate and the false positive rate when a probability threshold is used.

Results and Discussion

Spectral reflectance for raw and derivative transformation of frogeye leaf spot disease on soybean

The spectral reflectance profile of FLS severity (classes 0, 1, 2, 3 and 4) is indicated in Fig. 4. Generally, the spectral profile had lower reflectance in the visible region with a small peak in the green region, followed by a sudden increase starting at 690 nm and reaching a peak in the near infrared (NIR) region. A comparison of profiles indicated there were larger differences in reflectance between each severity FLS class at some wavelengths compared to others.

Despite the fact that the reflectance of different severity classes of soybeans was very similar in the visible spectrum (607–690 nm), healthy soybeans had the lowest reflectance in the red region (620–700 nm), due to energy absorption by chlorophyll pigments from active photosynthetic activity. The increase in reflectance observed in the red region for diseased leaves was caused by the breakdown of chlorophyll pigments as a result of the changes in the concentration of anthocyanins, carotenoids, anthocyanins and the photoprotective role of xanthophyll pigments (Devadas et al., 2009). The spectra of FLS with different severity classes varied in the infrared region (750–998 nm). In the 750–984 nm range, healthy soybean (class 0) had the highest reflectance compared to the others (Fig. 4), followed by classes 1, 2 and 3 of soybeans FLS severity. Class 4 of FLS severity, which was the worst, had the lowest reflectance.

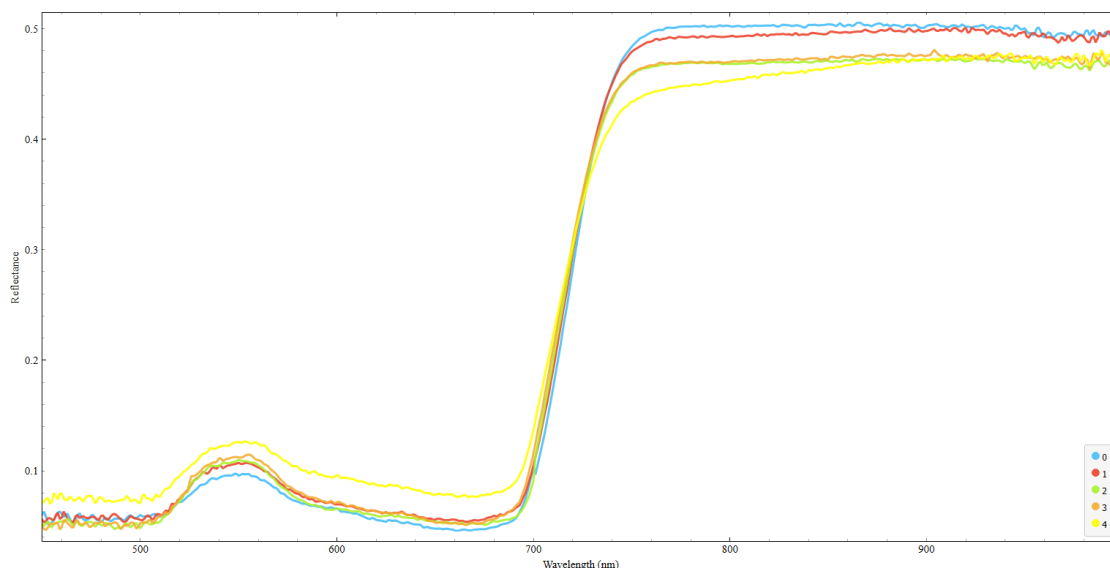


Fig.4 Original spectral reflectance of soybean frogeye leaf spot severity, where legend indicates severity class

The first derivative equation is essential for plant disease detection with a fixed position of the lesion density (Gregory, 1968). As the number of disease symptoms rises, the spectrum become more intricate and complicated. Variations in the collected spectrum are attributable to various stages of disease severity when plants in a particular location are infected. Therefore, the first derivative can eliminate both additive and multiplicative spectrum effects (Tsai and Philpot, 1998). Generally, the spectral reflectance had a modest reflectance peak at 515 nm and a large first derivative reflectance peak in the red-edge region (679–714 nm). Healthy soybean (Fig. 5) with class 0 had the highest peak (717.87 nm) in the

red-edge region. A red-edge region in crop spectra (in the range 650–800 nm) indicates the structure of the plant's cell, with conformational changes affecting the shape of the red-edge region. Furthermore, chlorophyll emits fluorescence within the range of the red-edge, indicating the chemical structure and chlorophyll content of a crop as an indicator of its biophysical properties (Zarco-Tejada et al., 2000). The spectral reflectance of the second derivative (Fig. 6) of FLS disease severity contained two distinct “windows” at 685 nm and 732 nm. Due to their insensitivity to the soil background, the first and second derivatives are generally useful indicators in estimating crop disease (Sankaran et al., 2011).

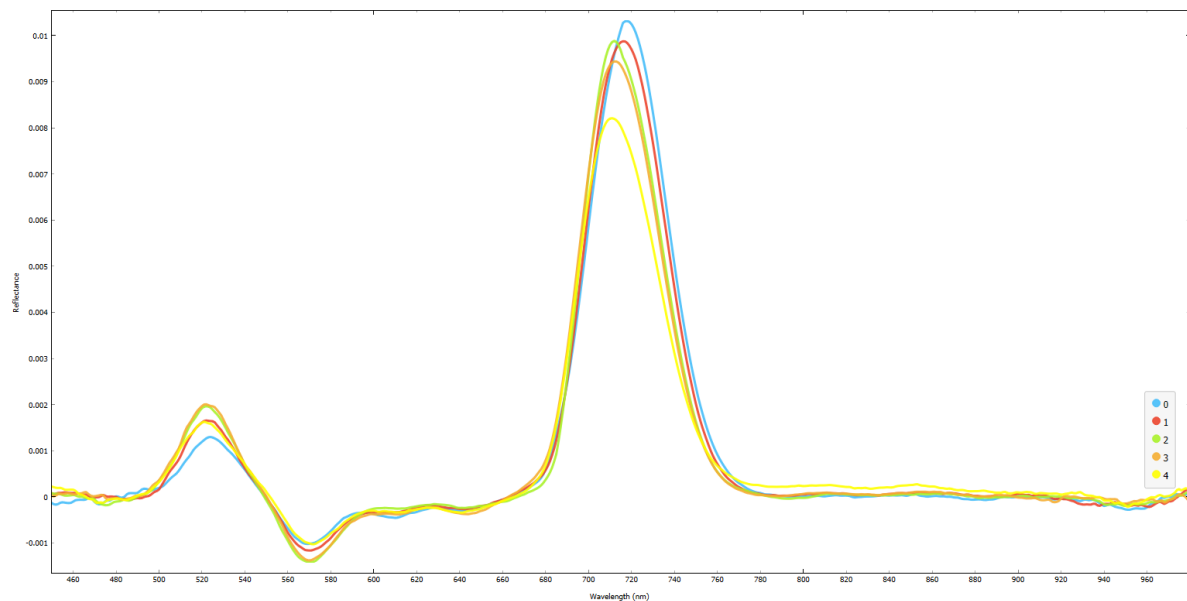


Fig.5 First derivatives of soybean frog eye leaf spot severity, where legend indicates severity class

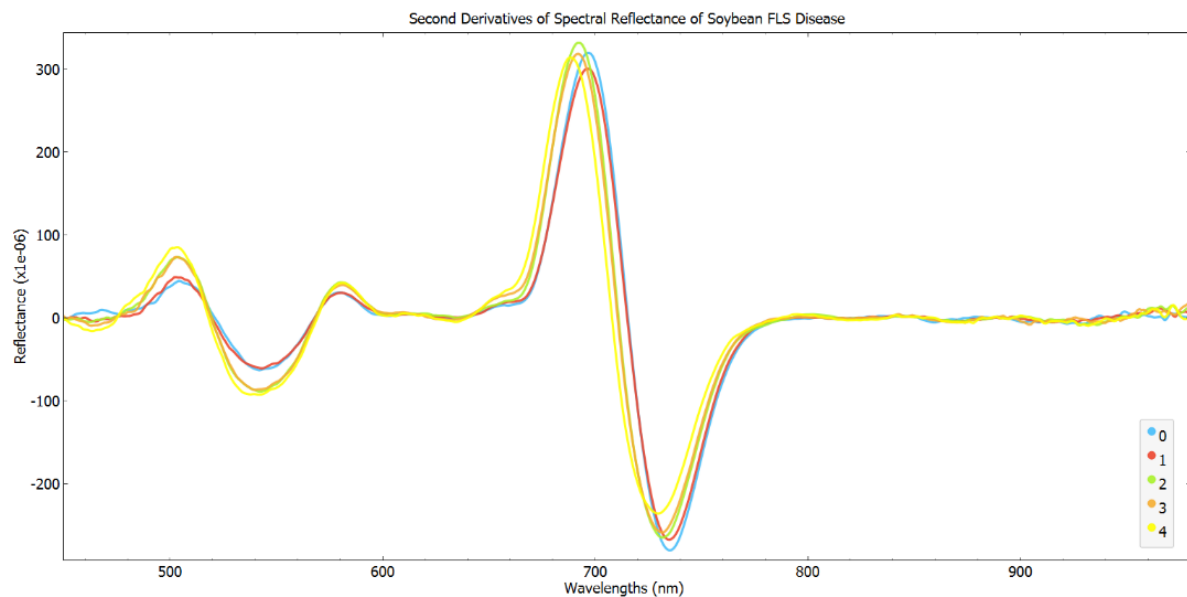


Fig.6 Second derivatives of soybean frog eye leaf spot severity, where legend indicates severity class

Classification of soybean frogeye leaf spot disease severity using mathematical learning model

The accuracy of the ML algorithms utilized in the classification of FLS disease severity is shown in Table 5. The neural network had the highest accuracy (98.6%), F1 (98.6%) and precision (98.5%) using the full first derivative data. This was followed by gradient boosting that had an overall accuracy of 97.4% and precision of 97.3% using the first derivative data with feature selection. Random forest achieved 96.9% and 91.9% accuracy, respectively, for first and second derivatives with feature selection of spectral reflectance, increasing from 88.9% and 85.2% for the full first and second derivatives of spectral reflectance, respectively. For selected first and second and second derivatives of spectral reflectance, the stacking algorithm improved to 94.9% and 88.7%, respectively, from the full first and second derivatives of spectral reflectance of 91.1% and 88.6%, respectively. Despite being subjected to feature selection techniques, the decision tree had the lowest accuracy levels of 77.6% and 73.5% for the selected first and second derivatives of spectral reflectance, respectively.

A neural network has the ability to interpret the relationships between wavelengths without the implementation of feature selection because of some configurations in the layers (Ahmad et al., 2019). Because of the network architecture and specific components within it, the network learns to extract and understand the features that represent the relationships and dependencies between wavelengths, which are crucial for accurate classification. By removing certain useful wavelengths through feature selection, the interdependencies between the remaining features may be disrupted, causing the model's performance to deteriorate. Based on the current findings, using the ReliefF algorithm as a feature selection increased the overall accuracy level for the majority of the ML algorithms, because the ReliefF algorithm could effectively select the major wavelengths for soybean FLS disease severity classification, reducing the number of wavelengths in the models and generating higher classification accuracy. The current findings were consistent with Meng et al. (2020), who demonstrated that ReliefF algorithms, such as feature selection, could identify wavelengths with high discriminative strength in the green, red-edge, NIR and shortwave infrared regions when assessing Southern Corn Rust-infected leaves of varying severity levels. In addition, the results showed that the ML algorithm using feature selection for the second derivative did not improve accuracy, because the second derivative contained noise (Antonov and Stoyanov, 1996) and the threshold for the ReliefF technique resulted in a loss of information. Therefore, the threshold for ReliefF should be further adjusted to determine the accuracy.

Table 5 Accuracy of mathematical learning algorithms in soybean frogeye leaf spot disease severity classification

Classifier	Raw data (%)			First derivative (%) (Savitzky-Golay) (full)			Second derivative (%) (Savitzky-Golay) (full)			First derivative (%) (Savitzky-Golay) + Feature selection (ReliefF algorithm)			Second derivative (%) (Savitzky-Golay) + Feature selection (ReliefF algorithm)		
	Accuracy	F1	Precision	Accuracy	F1	precision	Accuracy	F1	precision	Accuracy	F1	precision	Accuracy	F1	precision
Decision tree	71.3	70.3	71.2	75.2	75.4	75.1	74.3	74.2	74.0	77.6	77.4	77.5	73.5	73.2	73.3
Random forest	80.6	80.5	80.6	88.9	88.5	88.1	85.2	85.1	85.2	96.9	96.4	96.9	91.9	91.5	91.7
Neural network	85.0	84.9	85.1	98.6	98.6	98.5	93.6	93.6	93.6	86.7	86.7	86.7	86.0	86.0	86.0
Stacking (KNN + Naïve + svm as base learner; logistic as meta-)	79.5	79.2	79.1	91.1	91.2	91.0	88.6	88.6	88.4	94.9	94.7	94.8	88.7	88.6	88.6
Gradient boosting	82.0	81.9	82.1	91.1	91.0	91.1	89.8	89.8	89.7	97.4	97.2	97.3	95.5	95.5	95.4

Fig. 7 shows the ROC curve for the neural network model that had the highest accuracy in classifying the severity of FLS disease. This analysis has provided insight into the accuracy of model classification of the various severity classifications

(0, 1, 2, 3 and 4) of FLS. Overall, the metric area under the curve had an acceptable range in the ROC analysis, indicating that the model was effective at distinguishing the FLS severity classes on soybean.

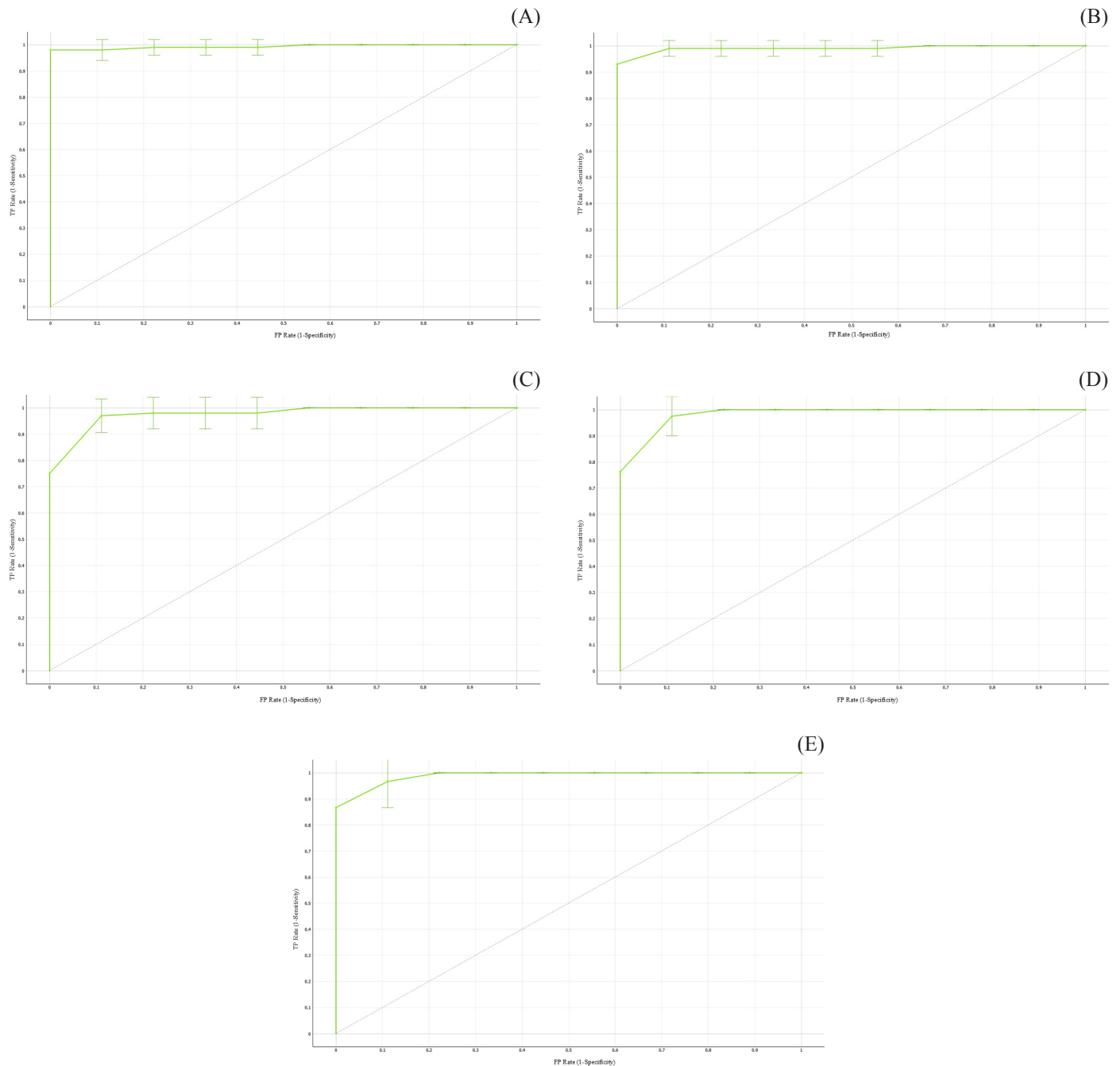


Fig.7 Best relative operating characteristic from neural network model for soybean frog eye leaf spot disease severity classification: (A) class 0; (B) class 1; (C) class 2; (D) class 3; (E) class 4

Conclusion

The neural network was the optimal model, achieving 98.6% accuracy in the classification of FLS disease severity on soybean, utilizing the full first derivative data. In addition, the ReliefF-gradient boosting and random forest algorithms had promising levels of accuracy (97.4% and 96.9%, respectively) in classifying the severity of soybean FLS disease. Future enhancements will explore various feature selection strategies to assess hyperspectral reflectance for crop disease classification. In addition, different pre-processing techniques will be evaluated to ensure that spectrum noise is effectively removed. Furthermore, more advanced ML algorithms may be designed and implemented for hyperspectral reflectance analysis. The Orange software, with its spectroscopic widget, provides extra capabilities such as extensive hyperspectral pre-processing (Savitzky-Golay) and the implementation of advanced ML algorithms. Therefore, there needs to be further study of the capability of this software as a visual programming platform for assessing hyperspectral reflectance, to ensure its practicability and use. The findings of the current study should facilitate the rapid deployment of processes, including model development, for assessing crop disease severity levels using hyperspectral reflectance data.

Conflict of Interest

The authors declare that there are no conflicts of interest.

Acknowledgements

The Malaysian Ministry of Higher Education (MOHE) provided financial support through the Long-Term Research Grant Scheme (LRGS) of the Malaysian Research University Network (MRUN), with funding under Grant No. 203.PKOMP.6770007 with UPM Vote No. 6300268-10801. Liu et al. (2021) gave permission for the use of their data.

References

Ahmad, J., Farman, H., Jan, Z. 2019. Deep learning methods and applications. In: Deep Learning: Convergence to Big Data Analytics. SpringerBriefs in Computer Science. Springer, Singapore. pp. 31–42.

- Al-Saddik, H., Simon, J.C., Cointault, F. 2019. Assessment of the optimal spectral bands for designing a sensor for vineyard disease detection: the case of 'Flavescence dorée'. *Precis. Agric.* 20: 398–422. doi.org/10.1007/s11119-018-9594-1
- Antonov, L., Stoyanov, S. 1996. Noise reduction in second derivative UV-VIS spectroscopy. *Spectrosc. Lett.* 29: 231–239. doi.org/10.1080/00387019608001599
- Barro, J.P., Neves, D.L., Del Ponte, E.M., Bradley, C.A. 2023. Frogeye leaf spot caused by *Cercospora sojina*: A review. *Trop. Plant Pathol.* 48: 363–374. doi.org/10.1007/s40858-023-00583-8
- Belgiu, M., Drăgu, L. 2016. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* 114: 24–31. doi.org/10.1016/j.isprsjprs.2016.01.011
- Berger, K., Verrelst, J., Féret, J.B., Wang, Z., Woche, M., Strathmann, M., et al. 2020. Crop nitrogen monitoring: Recent progress and principal developments in the context of imaging spectroscopy missions. *Remote Sens. Environ.* 242: 1–18. doi.org/10.1016/j.rse.2020.111758
- Borah, M., Deb, B. 2022. Fungi causing leaf spot diseases of soybean: their epidemiology and integrated management strategies. *Int. J. Econ. Plants* 9: 88–94. doi.org/10.23910/2/2022.0437a
- Bruning, B., Berger, B., Lewis, M., Liu, H., Garnett, T. 2020. Approaches, applications, and future directions for hyperspectral vegetation studies: An emphasis on yield-limiting factors in wheat. *Plant Phenome J.* 3: 1–22. doi.org/10.1002/ppj2.20007
- Dems̄ar, J., Curk, T., Erjavec, A., et al. 2013. Orange: Data mining toolbox in Python. *J. Mach. Learn. Res.* 14: 2349–2353.
- Devadas, R., Lamb, D.W., Simpfendorfer, S., Backhouse, D. 2009. Evaluating ten spectral vegetation indices for identifying rust infection in individual wheat leaves. *Precis Agric.* 10: 459–70. doi.org/10.1007/s11119-008-9100-2
- Dreoni, I., Matthews, Z., Schaafsma, M. 2022. The impacts of soy production on multi-dimensional well-being and ecosystem services: A systematic review. *J. Clean. Prod.* 335: 130182. https://doi.org/10.1016/j.jclepro.2021.130182
- Dubey, S.R., Singh, S.K., Chaudhuri, B.B. 2022. Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing*. 503: 92–108. doi.org/10.1016/j.neucom.2022.06.111
- Gregory, P.H. 1968. Interpreting plant disease dispersal gradients. *Annu. Rev. Phytopathol.* 6: 189–212. doi.org/10.1146/annurev.py.06.090168.001201
- Huang, L., Liu, Y., Huang, W., Dong, Y., Ma, H., Wu, K., Guo, A. 2022. Combining random forest and XGBoost methods in detecting early and mid-term winter wheat stripe rust using canopy level hyperspectral measurements. *Agriculture* 12: 1–16. doi.org/10.3390/agriculture12010074
- Khan, I.H., Liu, H., Li, W., et al. . 2021. Early detection of powdery mildew disease and accurate quantification of its severity using hyperspectral images in wheat. *Remote Sens.* 13: 1–18. doi.org/10.3390/rs13183612
- Kim, H., Newell, A.D., Cota-Sieckmeyer, R.G., Rupe, J.C., Fakhoury, A.M., Bluhm, B.H. 2013. Mating-type distribution and genetic diversity of *Cercospora sojina* populations on soybean from Arkansas: evidence for potential sexual reproduction. *Phytopathol.* 103: 1045–1051. doi.org/10.1094/PHYTO-09-12-0229-R

- Lelong, C.C., Roger, J.-M., Brégand, S., Dubertret, F., Lanore, M., Sitorus, N., Raharjo, D., Caliman, J.-P. 2010. Evaluation of oil-palm fungal disease infestation with canopy hyperspectral reflectance data. *Sensors*. 10: 734–747. doi.org/10.3390/s100100734
- Liu, S., Yu, H., Sui, Y., Zhou, H., Zhang, J., Kong, L., Dang, J., Zhang, L. 2021. Classification of soybean frogeye leaf spot disease using leaf hyperspectral reflectance. *PLoS One*. 16: e0257008. doi.org/10.1371/journal.pone.0257008
- Liu, Z.Y., Wu, H.F., Huang, J.F. 2010. Application of neural networks to discriminate fungal infection levels in rice panicles using hyperspectral reflectance and principal components analysis. *Comput. Electron Agric.* 72: 99–106. doi.org/10.1016/j.compag.2010.03.003
- Lowe, A., Harrison, N., French, A.P. 2017. Hyperspectral image analysis techniques for the detection and classification of the early onset of plant disease and stress. *Plant Methods*. 13: 80. doi.org/10.1186/s13007-017-0233-z
- Lu, B., Dao, P.D., Liu, J., He, Y., Shang, J. 2020. Recent advances of hyperspectral imaging technology and applications in agriculture. *Remote Sens.* 12: 2659. doi.org/10.3390/rs12162659
- Meng, R., Lv, Z., Yan, J., Chen, G., Zhao, F., Zeng, L., Xu, B. 2020. Development of spectral disease indices for southern corn rust detection and severity classification. *Remote Sens.* 12: 3233. doi.org/10.3390/rs12193233
- Mustafa, G., Zheng, H., Khan, I.H., et al. 2022. Hyperspectral reflectance proxies to diagnose in-field fusarium head blight in wheat with machine learning. *Remote Sens.* 14: 2784. doi.org/10.3390/rs14122784
- Myles, A.J., Feudale, R.N., Liu, Y., Woody, N.A., Brown, S.D. 2004. An introduction to decision tree modeling. *J. Chemom.* 18: 275–285. doi.org/10.1002/cem.873
- Natekin, A., Knoll, A. 2013. Gradient boosting machines, a tutorial. *Front. Neurobot.* 7: 21. doi.org/10.3389/fnbot.2013.00021
- Navrozidis, I., Pantazi, X.E., Lagopodi, A., Bochtis, D., Alexandridis, T.K. 2023. Application of machine learning for disease detection tasks in olive trees using hyperspectral data. *Remote Sens.* 15: 5683. doi.org/10.3390/rs15245683
- Parrini, S., Aquilani, C., Pugliese, C., Bozzi, R., Sirtori, F. 2023. Soybean Replacement by Alternative Protein Sources in Pig Nutrition and Its Effect on Meat Quality. *Animals*. 13: 494. doi.org/10.3390/ani13030494
- Phillips, X.A., Kandel, Y.R., Mueller, D.S. 2021. Impact of foliar fungicides on frogeye leaf spot severity, radiation use efficiency and yield of soybean in Iowa. *Agronomy*. 11: 1785. doi.org/10.3390/agronomy11091785
- Sankaran, S., Mishra, A., Maja, J.M., Ehsani, R. 2011. Visible-near infrared spectroscopy for detection of Huanglongbing in citrus orchards. *Comput. Electron Agric.* 77: 127–134. doi.org/10.1016/j.compag.2011.03.004
- Sun, J., Yang, W., Zhang, M., Feng, M., Xiao, L., Ding, G. 2021. Estimation of water content in corn leaves using hyperspectral data based on fractional order Savitzky-Golay derivation coupled with wavelength selection. *Comput. Electron Agric.* 182: 1–8. doi.org/10.1016/j.compag.2021.105989
- Szandała, T. 2021. Review and comparison of commonly used activation functions for deep neural networks. In: Bhoi, A., Mallick, P., Liu, C.M., Balas, V. (Eds.). *Bio-inspired Neurocomputing*. Springer. Singapore, pp. 203–224.
- Tetrault, R. 2023. *World Agricultural Production*. United States Department of Agriculture. City, 2-letter state abbreviation, USA
- Tsai, F., Philpot, W. 1998. Derivative analysis of hyperspectral data. *Remote Sens. Environ.* 66: 41–51. doi.org/10.1016/S0034-4257(98)00032-7
- Urbanowicz, R.J., Meeker, M., La-Cava, W., Olson, R.S., Moore, J.H. 2018. Relief-based feature selection: Introduction and review. *J. Biomed. Inform.* 85: 189–203. doi.org/10.1016/j.jbi.2018.07.014
- Vishnoi, V.K., Kumar, K., Kumar, B. 2021. Plant disease detection using computational intelligence and image processing. *J. Plant Dis. Prot.* 128: 19–53. doi.org/10.1007/s41348-020-00368-0
- Wei, X., Johnson, M.A., Langston-Jr, D.B., Mehl, H.L., Li, S. 2021. Identifying optimal wavelengths as disease signatures using hyperspectral sensor and machine learning. *Remote Sens.* 13: 2833. doi.org/10.3390/rs13142833
- Xie, C., Shao, Y., Li, X., He, Y. 2015. Detection of early blight and late blight diseases on tomato leaves using hyperspectral imaging. *Sci. Rep.* 5: 1–11. doi.org/10.1038/srep16564
- Zarco-Tejada, P.J., Miller, J.R., Mohammed, G.H., Noland, T.L. 2000. Chlorophyll fluorescence effects on vegetation apparent reflectance: I. Leaf-level measurements and model simulation. *Remote Sens. Environ.* 74: 582–595. doi.org/10.1016/S0034-4257(00)00148-6
- Zenko, B., Todorovski, L., Dzeroski, S. 2001. A comparison of stacking with meta decision trees to bagging, boosting, and stacking with other methods. In: *Proceedings 2001 IEEE international conference on data mining*. IEEE. San Jose, CA, USA, 669 pp.
- Zhang, M., Qin, Z., Liu, X., Ustin, S.L. 2003. Detection of stress in tomatoes induced by late blight disease in California, USA, using hyperspectral remote sensing. *Int. J. Appl. Earth Obs. Geoinf.* 4: 295–310. doi.org/10.1016/S0303-2434(03)00008-4