Journal of Oil Palm Research
DOI: https://doi.org/10.21894/jopr.2022.0046

CROSS-VALIDATION AND RECEIVER OPERATING CHARACTERISTIC ANALYSES FOR OIL PALM LEAF METABOLOME DATASET

NUR AIN ISHAK[1]; NOOR IDAYU TAHIR[1]; NURUL LIYANA ROZALI[1]; ZAIN NURAZAH[1]; NUR RAIHAN ABD RAHIM[2]; ABRIZAH OTHMAN[1] and UMI SALAMAH RAMLI[1]*

ABSTRACT

The advancement of systems biology research has emphasized the capabilities of statistical analysis tools in distinguishing many factors associated with oil palm including genetic vs. *environment (GxE) components from omics data. The availability of an efficient and robust ecometabolomics workflow has a high potential in augmenting oil palm precision agriculture. In this study, we employed cross-validation (CV) and receiver operating characteristic (ROC) methodologies to evaluate the performance of an oil palm metabolome dataset linked to GxE factors for its predictive ability and integrity. The specificity and sensitivity of identified metabolite candidates contributing to the demarcation of the two oil palm groups in the dataset were found to be distinctive and were of discrimination quality. The dataset showed no overfitting and exhibited excellent predictive power. This work provides fundamental information and a guideline for universal metabolome data exploration toward oil palm phenotyping and precision agriculture.*

Keywords: CV, metabolome, oil palm, phenotyping, ROC.

Received: 3 February 2022; **Accepted:** 17 June 2022; **Published online:** 8 August 2022.

INTRODUCTION

The high throughput and big data domain of systems biology provide attractive omics data collection for data mining and further utilisation in precision agriculture (Li and Yan, 2020). Defined as a holistic, eco-efficient and innovative tool to assist crop management sustainably, precision agriculture is viewed as a solution for the increasing human population and managing the implications of climate change (Lee *et al.*, 2021). One of the key areas in the omics studies is metabolomics; the in-depth analysis of a set of metabolites within an organism, cell or tissue under a given set of conditions at a specific time (Goodacre *et al.*, 2004; Reinke and Broadhurst, 2012). It is regarded as an important crop phenotyping tool that revolutionises the traditional assessment of phenotype by observing and measuring an organism's physical characteristics (Razzaq *et al.*, 2019). Crop phenotyping facilitates the selection efficiency of breeding programs, expedites genetic gains and helps automate or mechanise the monitoring of crop vigour status (Chawade *et al.*, 2019).

In plants, especially oil palm, the use of metabolomics is aimed at shedding light on several significant biological traits and responses to biotic and abiotic stimuli in numerous environmental conditions affecting plants' phenotypic performance, *e.g.*, yield, oil quality and disease resistance. Various metabolomics investigations involving oil palm leaf (Rozali *et al.*, 2021; Tahir *et al.*, 2016; Vargas *et al.*, 2016), root (Muhammad *et al.*, 2021; Nurazah *et al.*, 2021), seedling (Dzulkafli *et al.*, 2019) and

[1] Malaysian Palm Oil Board,
6 Persiaran Institusi, Bandar Baru Bangi,
43000 Kajang, Selangor, Malaysia.

[2] Faculty of Arts and Science,
International University of Malaya-Wales (IUMW),
City Campus, Ground Floor,
Block A, Administration Wing, Jalan Tun Ismail,
50480 Kuala Lumpur, Malaysia.

* *Corresponding author e-mail: umi@mpob.gov.my*

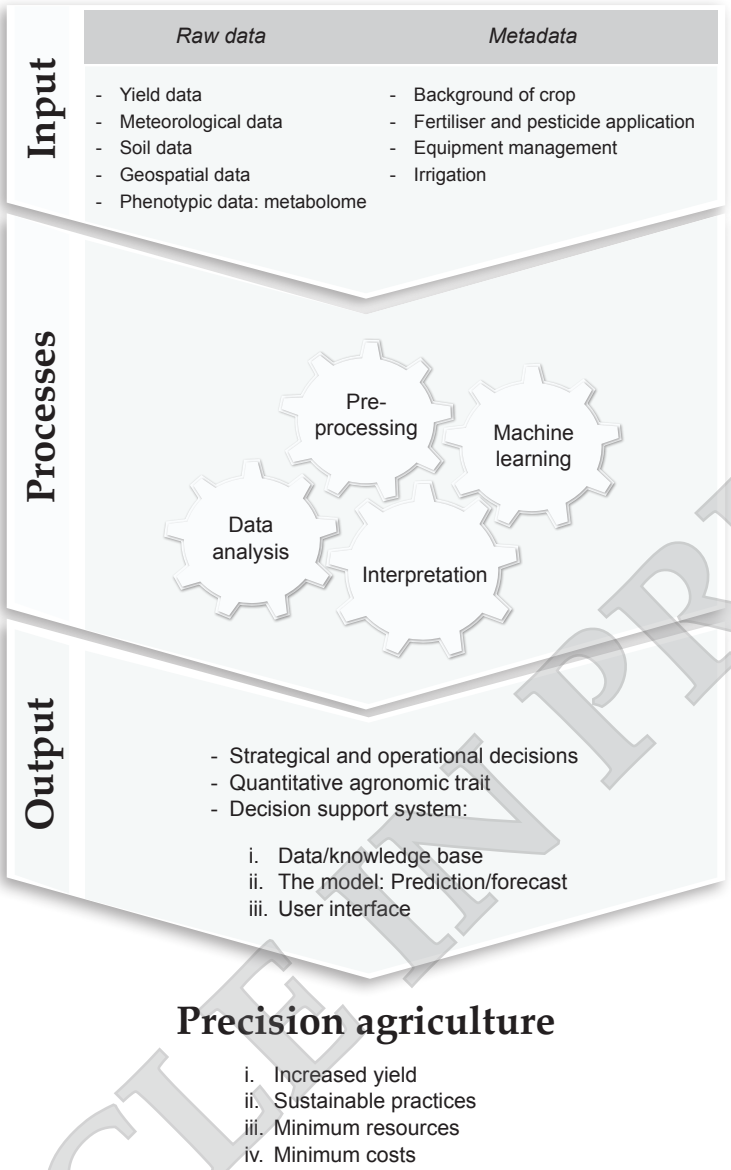


Figure 1. The general data-driven workflow for precision agriculture.

mesocarp (Teh *et al.*, 2013) were performed using various analytical approaches to elucidate the crop's biochemical characteristics. Among the metabolomics platforms, liquid chromatography paired with mass spectrometry (LC-MS) is recognised as the most frequently utilised in metabolomics (Sindelar and Patti, 2020; Xiao *et al.*, 2012) especially plant ecometabolomics strategy and can cover the separation and detection of diverse and inclusive plant metabolites (Sardans *et al.*, 2020). To systematically analyse and interpret a metabolomics dataset, chemometrics is employed (Ishak *et al.*, 2021). Chemometrics is an arm of data science for the extraction and evaluation of analytical-chemical data. Within the context of precision agriculture, the machine learning of chemometrics data collected in the field will allow automated extraction of information and further provide a model for characteristic prediction as outlined in *Figure 1.* Statistical techniques and machine learning have only lately gained popularity in oil palm research, for example, in remote sensing (Jia *et al.*, 2019), palm oil and fruit quality (Goggin *et al.*, 2021; Goh *et al.*, 2021). Here, the application of cross-validation (CV) and receiver operating characteristic (ROC) approaches to an oil palm metabolome dataset were evaluated and verified for the first time as figures of merit in the endeavour to develop a screening and predictive workflow for metabolomics-assisted crop phenotyping.

MATERIALS AND METHODS

The raw data from LC-MS of the oil palm spear leaf metabolome from an ecometabolomics study

is similar to the settings demonstrated by Tahir *et al*. (2016) and guided by mass-spectrometry-based metabolomics recommendations (Alseekh *et al*., 2021) was organised into a tabular dataset format with rows of retention times against columns of peak intensities from 1.0-59.5 min analysis time with 'binned' components containing peak intensities from mass-to-charge ratio (*m*/*z*) of 50-1000. A signal-to-noise (S/N) threshold of 5.0 was applied in which a signal must surpass the set value to be used in the peak detection. The metabolome was extracted from oil palm leaves consisting of top, middle and basal leaflets on a selected frond '0' (spear leaf) from one specific clonal oil palm line, sampled from two different planting sites of mineral and peat soil types.

Oil Palm Metabolomics Workflow

The general plant metabolomics workflow from oil palm tissue sampling for chemometrics is described in *Figure 2*, involving sample collection, storage and extraction before the analysis using analytical platforms to detect and measure the small molecules. Consideration for photo-, heat-, and temporal-sensitive specimens must be taken into account to avoid deterioration and to minimise metabolome variations. The inspection of the generated data for inconsistencies and missing values is crucial in data interpretation as it will affect subsequent statistical analysis, and poor data processing may result in or further aggravate unwanted variance (Engel *et al*., 2013). Uni- and multivariate analysis reduce the dimensions of the data and provide visualisation for interpretation. For definitive performance assessment of the statistical classification, the metabolome dataset is further validated for machine learning using CV and ROC approaches.

LC-MS Dataset Validation

CV of the dataset (n=42) was performed according to leave-p-out analyses parameters of leave-one-out, leave-5%-out and leave-10%-out CV. ROC and the confidence intervals for metabolite classifiers were calculated in MetaboAnalyst 5.0 (Pang *et al*., 2021) according to group assigned parameter of soil where the x-axis was referred to as the '1-specificity' in terms of the recorded false positives and the y-axis was referred to the sensitivity in terms of the recorded true positives. Both axes were given values between 0 and 1. A test assumption was made that the first attribute is the mineral soil and the second attribute is the peat soil. Analyses and parameters were set up in ProfileAnalysis 2.1 (Bruker Daltonics, Bremen, Germany). Supervised statistical analysis of an orthogonal partial least square-discriminant analysis (OPLS-DA) predictive model with Pareto scaling and ROC model classification accuracy was performed using SIMCA-P+ 14.1 software (Sartorius Stedim Data Analytics AB, Umeå/Malmö, Sweden).

RESULTS AND DISCUSSION

Cross-validation (CV) of Oil Palm Leaf Metabolome Dataset

The CV method keeps a portion of analyses out of the model calculation and calculates several parallel models from the reduced data before it forecasts the removed data by the different models and compares the predicted values with the actual ones (Schmidt *et al*., 2019). A model is constructed and optimised using the training data while the test set is applied to see how well the model works. The procedure will be repeated in such a way that each sample appears once and only once in the test set, and the prediction error is representative of new samples. A completely independent test set must not be pretreated, preprocessed, or scaled.

The most common types of CV are leave-one-out and k-fold (Xi *et al*., 2014). Considering a case where the sample size is n, in leave-one-out CV, n-1 samples are used as a training set for fitting a classification model, and the remaining sample is used for testing. This process is repeated n times, and each sample

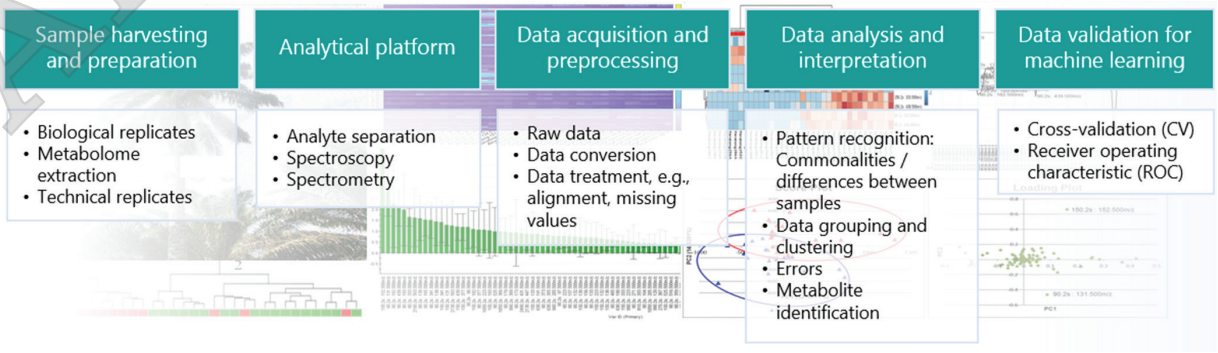


Figure 2. Oil palm metabolomics workflow.

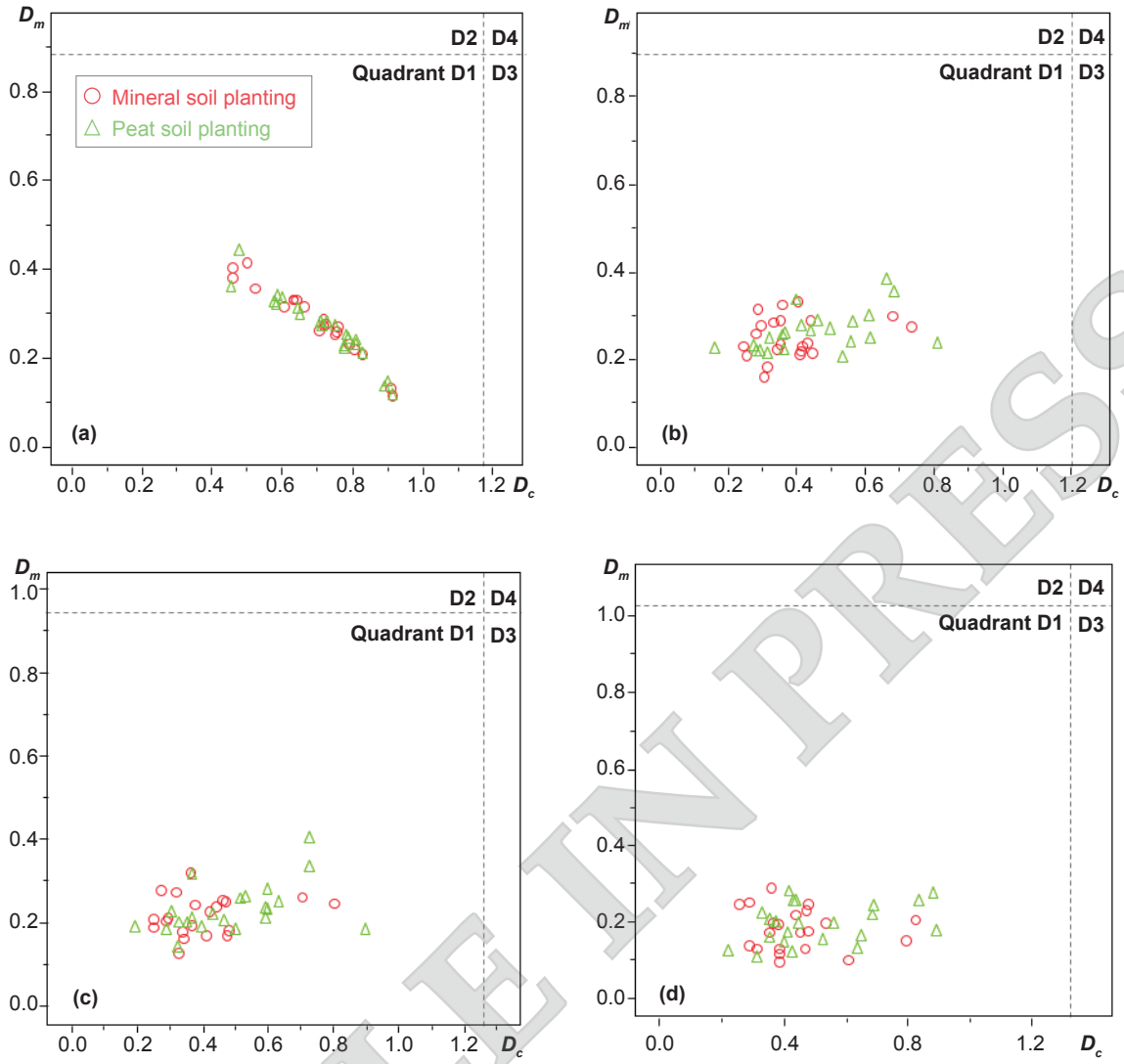
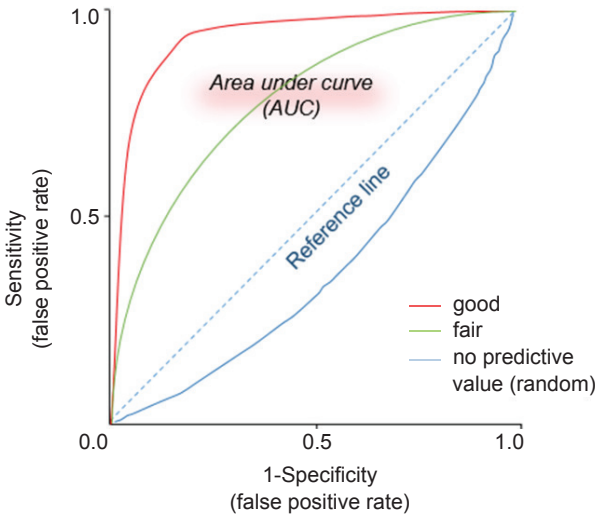


Figure 3. Influence plots of oil palm leaf metabolome dataset; (a) before cross-validation, (b) leave-one-out cross-validation, (c) leave-5%-out cross-validation and (d) leave-10%-out cross-validation (D_m - 'Distance to model' and D_c - 'Distance to centre').

is implemented just once as test data. A model constructed from *n-1* samples is approximately as accurate as one constructed from all *n* samples. The proportion of misclassified test data points is used to calculate the classification error rate. In *k*-fold CV, the entire dataset is partitioned into *k* subsets of similar size (*e.g.*, *k*=5 or *k*=10). *k-1* subsets are combined and used as a training set for each iteration, while the remaining subset is used as a test set. Every sample serves as a test data point only once. *Figure 3a* shows the influence plot of the oil palm leaf metabolome dataset before the application of any validation method while *Figures 3b, 3c* and *3d* are the influence plots after the leave-one-out, leave-5%-out and leave-10%-out CV respectively. All data points were located in the distance 1 (D1) quadrant corresponding to analyses inside the model space after the CV, indicating a good proportion of the model.



Source: Ferraris (2019).

Figure 4. Types of ROC curves for classifier accuracy.

Receiver Operating Characteristic (ROC) Curve for Discrimination Threshold

The ROC curve is an evaluation of the distinctive quality of a classifier (*i.e.*, metabolites or soil type) where the best possible prediction feature yields a curve in the direction of the upper left corner of the plot towards 100% sensitivity and 100% specificity. *Figure 4* summarises the main verdicts that can be drawn from a ROC curve.

In *Figure 5*, the accuracy of dopamine ($C_8H_{11}NO_2$, *m/z* 152.0719 $[M\text{-}H]^-$) and asparagine ($C_4H_8N_2O_3$, *m/z* 131.0434 $[M\text{-}H]^-$) in discriminating the two planting locations were demonstrated by the ROC curves which were plotted closer to the left and the top border of the ROC space, consistent with previous findings (Tahir *et al.*, 2016). The jagged shape of the ROC curves could be improved by adding more measurements to get smoother arcs (Xia *et al.*, 2013). According to Bünger and Mallet (2016) and Ferraris (2019), any classifier designated as a test criterion displaying a high area under the ROC curve (AUC) value of close to 1.0 indicates high diagnostic capability in discriminating two populations. As dopamine and asparagine showed high AUC values of 0.852 (0.725-0.955 confidence interval) and 0.781 (0.634-0.911 confidence interval) respectively, these metabolites are excellent classifiers for the different groups of oil palm specimens.

Assessment of Predictive Model from Oil Palm Leaf Metabolome Dataset

An orthogonal partial least square-discriminant analysis (OPLS-DA) model was constructed using the dataset based on its suitability in separating predictive from non-predictive (orthogonal) variation (Bylesjö *et al.*, 2006). Ranging from 0-1, a low value of 0.1822 root-mean-square CV error (RMSECV) was obtained for the model indicating a reliable and predictive ability (Lee *et al.*, 2018; Liu *et al.*, 2020). The statistical model from the training set was found to be significantly robust with a total accuracy of 100% correct classification and a low Fisher value (p<0.01) as tabulated in *Table 1* (Tarapoulouzi *et al.*, 2020). An external validation using three unknown samples from different planting trials as a test set (n=3) which were subjected to similar metabolome extraction, data acquisition and the preprocessing protocol was able to prove that the model can successfully predict the planting soil type of the unknown oil palm samples from three different mineral soil trials. The misclassification table (*Table 1*) of the constructed model indicated that the blind samples were grouped with the mineral group samples.

TABLE 1. OPLS-DA MODEL MISCLASSIFICATION TABLE OF OIL PALM METABOLOME SPECIMENS FROM DIFFERENT PLANTING TRIALS.

	No. of specimen	Correct classification (%)	Mineral	Peat
Mineral	21	100%	21	0
Peat	21	100%	0	21
Unknown specimen	3		3	0
Total specimen	45	100%	24	21
Fisher's probability	1.9 x10^{-12}			

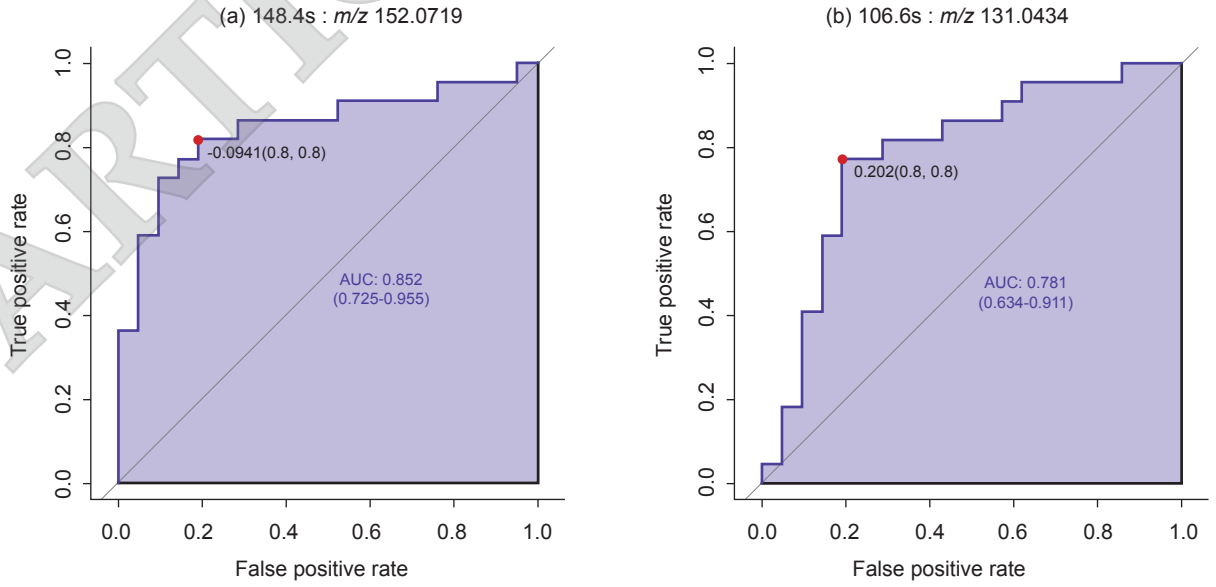


Figure 5. ROC curves with AUC values for (a) dopamine with an AUC value of 0.852 and (b) asparagine with an AUC value of 0.781.

Figure 6 presents the predicted score scatter plot of the test set merged with the training set samples from the mineral and peat soils. From the test set, two unknown samples (SampleX1 and SampleX2) were plotted on the left-hand side of the plot together with the mineral soil cluster while one unknown sample (SampleX3) was found in the outer region of the 95% mineral soil confidence ellipse. The three unknown samples are classified as specimens of mineral soil type, with SampleX3 exhibiting slight variation from the others.

An additional ROC analysis performed to assess the predictive OPLS-DA model's ability to appropriately categorise the oil palm specimens resulted in an AUC value of 1.0 (*Figure 7*), indicating its excellent classification power in separating the oil palm leaf metabolome samples to their corresponding planting soils (Ruisánchez *et al.*, 2021).

CONCLUSION

This work demonstrated the validation and evaluation methods of CV and ROC on an oil palm metabolome dataset in an attempt to obtain a screening and predictive workflow for metabolomics-assisted crop phenotyping. The demarcation of oil palm leaf samples from the peat and mineral soil planting sites by dopamine and asparagine respectively was verified as statistically significant for discriminating and predicting the two classes/groups of samples. The predictive statistical model constructed from the chemometric analysis also demonstrated excellent predictive ability. These findings established the provision of metabolome data exploration for oil palm phenotyping and its potential utilisation for machine learning which is highly prospective for oil palm integrated precision agriculture.

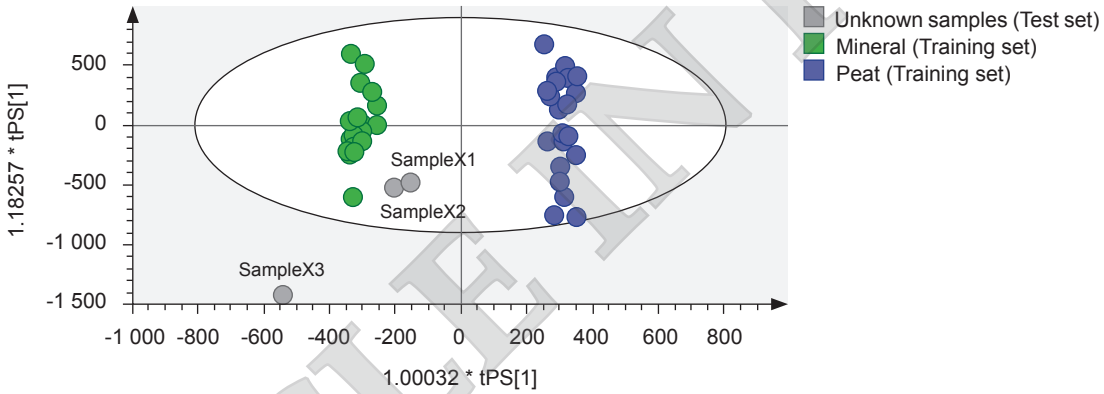


Figure 6. Predicted score plot with training set merged with the test set for oil palm spear leaf samples from different planting sites.

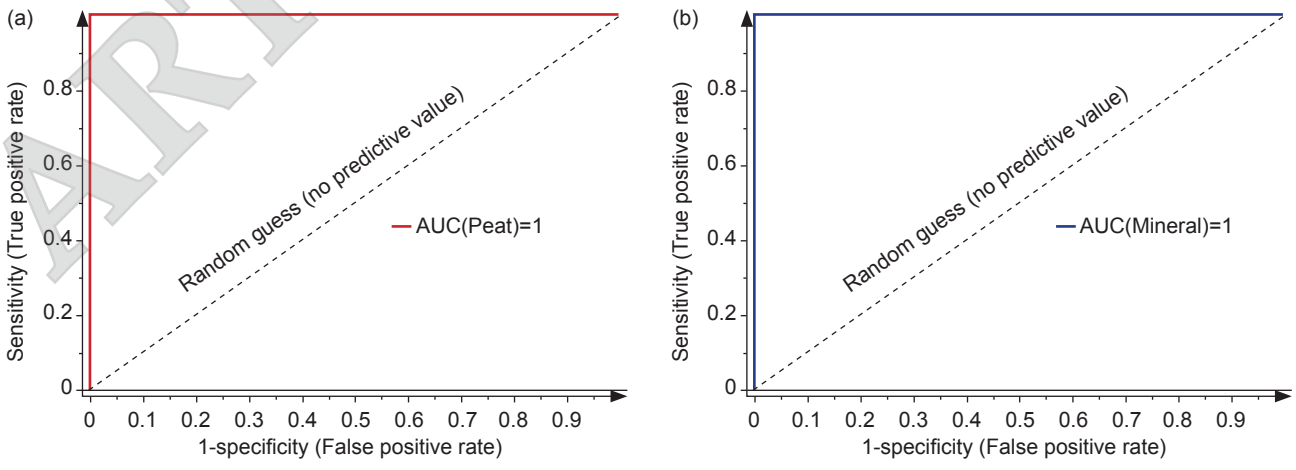


Figure 7. ROC curves for the classification accuracy of the OPLS-DA model with the area under curve (AUC) values of 1.0 for (a) peat soil and (b) mineral soil planting trials.

ACKNOWLEDGEMENT

The authors thank the Director-General of MPOB for permission to publish this article and Dr. Zainab Idris for her technical comments. The authors appreciate Prof. Jianping Yang (Zhejiang Sci-Tech University, China) and Mr. Kathiresan Gopal (Universiti Putra Malaysia) for technical contributions. We are grateful to Dr. Mohamad Arif Abd Manaf for his consistent support and the Proteomics and Metabolomics (PROMET) research team for their invaluable assistance.

REFERENCES

Alseekh, S; Aharoni, A; Brotman, Y; Contrepois, K; D'Auria, J; Ewald J; Ewald, J C; Fraser, P D; Giavalisco, P; Hall, R D; Heinemann, M; Link, H; Luo, J; Neumann, S; Nielsen, J; Perez de Souza, L; Saito, K; Sauer, U; Schroeder, F C; Schuster, S; Siuzdak, G; Skirycz, A; Sumner, L W; Snyder, M P; Tang, H; Tohge, T; Wang, Y; Wen, W; Wu, S; Xu, G; Zamboni, N and Fernie, A R (2021). Mass spectrometry-based metabolomics: A guide for annotation, quantification and best reporting practices. *Nat. Methods, 18(7)*: 747-756.

Bünger, R and Mallet, R T (2016). Metabolomics and receiver operating characteristic analysis: A promising approach for sepsis diagnosis. *Crit. Care Med., 44(9)*: 1784-1785.

Bylesjö, M; Rantalainen, M; Cloarec, O; Nicholson, J K; Holmes, E and Trygg, J (2006). OPLS discriminant analysis: Combining the strengths of PLS-DA and SIMCA classification. *J. Chemom., 20(8-10)*: 341-351.

Chawade, A; van Ham, J; Blomquist, H; Bagge, O; Alexandersson, E and Ortiz, R (2019). High-throughput field-phenotyping tools for plant breeding and precision agriculture. *Agronomy, 9(5)*: 258.

Dzulkafli, S B; Abrizah, O; Syahanim, S; Nurazah, Z; Manaf, M A A; Idris, A S; Amiruddin, M D; Tahir, N I and Ramli, U S (2019). Identification of chelidonic acid and asparagine in *Ganoderma boninense*-inoculated oil palm seedlings. *J. Oil Palm Res., 31*: 53-66.

Engel, J; Gerretzen, J; Szymańska, E; Jansen, J J; Downey, G; Blanchet, L and Buydens, L M C (2013). Breaking with trends in pre-processing? *Trends Analyt. Chem., 50*: 96-106.

Ferraris, V A (2019). Commentary: Should we rely on receiver operating characteristic curves? From submarines to medical tests, the answer is a definite maybe! *J. Thorac. Cardiovasc. Surg., 157(6)*: 2354-2355.

Goggin, K A; Brodrick, E; Wicaksono, A; Covington, J A; Davies, A N and Murphy, D J (2021). A proof-of-concept study: Determining the geographical origin of crude palm oil with the combined use of GC-IMS fingerprinting and chemometrics. *J. Oil Palm Res., 33(2)*: 227-234.

Goh, J Q; Mohamed Shariff, A R and Mat Nawi, N (2021). Application of optical spectrometer to determine maturity level of oil palm fresh fruit bunches based on analysis of the front equatorial, front basil, back equatorial, back basil and apical parts of the oil palm bunches. *Agriculture, 11(12)*: 1179.

Goodacre, R; Vaidyanathan, S; Dunn, W B; Harrigan, G G and Kell, D B (2004). Metabolomics by numbers: Acquiring and understanding global metabolite data. *Trends Biotechnol., 22(5)*: 245-252.

Ishak, N A; Tahir N I; Mohd Sa'id, S N; Kathiresan, G; Abrizah, O and Ramli, U S (2021). Comparative analysis of statistical tools for oil palm phytochemical research. *Heliyon, 7(2)*: e06048.

Jia, X; Khandelwal, A; Carlson, K; Gerber, J S; West, P C and Kumar, V (2019). Plantation mapping in Southeast Asia. *Front. Big Data, 2*: 46.

Lee, B J; Zhou, Y; Lee, J S; Shin, B K; Seo, J A; Lee, D; Kim, Y S and Choi, H K (2018). Discrimination and prediction of the origin of chinese and korean soybeans using fourier transform infrared spectrometry (FT-IR) with multivariate statistical analysis. *PLoS ONE, 13*: e0196315.

Lee, C L; Strong, R and Dooley, K E (2021). Analyzing precision agriculture adoption across the globe: A systematic review of scholarship from 1999-2020. *Sustainability, 13(18)*: 10295.

Li, Q and Yan, J (2020). Sustainable agriculture in the era of omics: Knowledge-driven crop breeding. *Genome Biol., 21*: 154.

Liu, L; Zuo, Z T; Xu, F R and Wand, Y Z (2020). Study on quality response to environmental factors and geographical traceability of wild *Gentiana rigescens* franch. *Front. Plant Sci., 11*: 1128.

Muhammad, I I; Abdullah, S N A; Saud, H M; Shaharuddin, N A and Isa, N M (2021). The dynamic responses of oil palm leaf and root metabolome to phosphorus deficiency. *Metabolites, 11(4)*: 217.

Nurazah, Z; Idris, A S; Amiruddin, M D; Manaf, M A A; Abrizah, O and Ramli, U S (2021). Metabolite fingerprinting of oil palm (*Elaeis guineensis* Jacq.) root for the identification of altered metabolic pathways associated with basal stem rot (BSR) disease. *Physiol. Mol. Plant Pathol., 115*: 101647.

Pang, Z; Chong, J; Zhou, G; de Lima Morais, D A; Chang, L; Barrette, M; Gauthier, C; Jacques, P É; Li, S and Xia, J (2021). MetaboAnalyst 5.0: Narrowing the gap between raw spectra and functional insights. *Nucleic Acids Res., 49(W1)*: W388-W396.

Razzaq, A; Sadia, B; Raza, A; Khalid Hameed, M and Saleem, F (2019). Metabolomics: A way forward for crop improvement. *Metabolites*, *9(12)*: 303.

Reinke, S N and Broadhurst, D I (2012). Moving metabolomics from a data-driven science to an integrative systems science. *Genome Med., 4(11)*: 85.

Rozali, N L; Tahir, N I; Hassan, H; Abrizah, O and Ramli, U S (2021). Identification of amines, amino and organic acids in oil palm (*Elaeis guineensis* Jacq.) spear leaf using GC- and LC/Q-TOF MS metabolomics platforms. *Biocatal. Agric. Biotechnol., 37*: 102165.

Ruisánchez, I; Jiménez-Carvelo, A M and Callao, M P (2021). ROC curves for the optimization of one-class model parameters. A case study: Authenticating extra virgin olive oil from a Catalan protected designation of origin. *Talanta, 222*: 121564.

Sardans, J; Gargallo-Garriga, A; Urban, O; Klem, K; Walker, T W N; Holub, P; Janssens, I A and Peñuelas, J (2020). Ecometabolomics for a better understanding of plant responses and acclimation to abiotic factors linked to global change. *Metabolites*, *10(6)*: 239.

Schmidt, J; Marques, M R G; Botti, S and Marques, A L (2019). Recent advances and applications of machine learning in solid-state materials science. *NPJ Comput. Mater., 5*: 83.

Sindelar, M and Patti, G J (2020). Chemical discovery in the era of metabolomics. *J. Am. Chem. Soc., 142(20)*: 9097-9105.

Tahir, N I; Shaari, K; Abas, F; Ishak, Z; Tarmizi, A; Amiruddin, M D; Parveez, G K A and Ramli, U S (2016). Metabolome analysis of oil palm clone P325 of different planting trials. *J. Oil Palm Res., 28(4)*: 431-441.

Tarapoulouzi, M; Kokkinofta, R and Theocharis, C R (2020). Chemometric analysis combined with FTIR spectroscopy of milk and Halloumi cheese samples according to species' origin. *Food Sci. Nutr., 8*: 3262-3273.

Teh, H F; Neoh, B K; Hong, M P L; Low, J Y S; Ng, T L M; Ithnin, N; Thang, Y M; Mohamed, M; Chew, F T; Yusof, H M; Kulaveerasingam, H and Appleton, D R (2013). Differential metabolite profiles during fruit development in high-yielding oil palm mesocarp. *PLoS ONE, 8(4)*: e61344.

Vargas, L H G; Neto, J C R; De Aquino Ribeiro, J A; Ricci-Silva, M E; Souza, M T; Rodrigues, C M; De Oliveira, A E and Abdelnur, P V (2016). Metabolomics analysis of oil palm (*Elaeis guineensis*) leaf: Evaluation of sample preparation steps using UHPLC–MS/MS. *Metabolomics, 12(10)*: 153.

Xi, B; Gu, H; Baniasadi, H and Raftery, D (2014). Statistical analysis and modeling of mass spectrometry-based metabolomics data. *Mass Spectrometry in Metabolomics. Methods in Molecular Biology (Methods and Protocols)* (Raftery, D ed.). Vol. 1198. Humana Press, New York, NY. p. 1198:333-353.

Xia, J; Broadhurst, D I; Wilson, M and Wishart, D S (2013). Translational biomarker discovery in clinical metabolomics: An introductory tutorial. *Metabolomics, 9(2)*: 280-299.

Xiao, J F; Zhou, B and Ressom, H W (2012). Metabolite identification and quantitation in LC-MS/MS-based metabolomics. *TrAC - Trends Anal. Chem., 32*: 1-14.