Sains Malaysiana 54(8)(2025): 1913-1925
http://doi.org/10.17576/jsm-2025-5408-04

Non-Halal Gelatin Prediction: A Comparative Machine Learning Analysis between OPLS–DA and ANN Models

(Ramalan Gelatin Tidak Halal: Perbandingan Analisis Pembelajaran Mesin antara Model OPLS-DA dan ANN)

Mohd Hafis Yuswan[1,*], Norazlina Ali[2], Syaiful Izwan Ismail[2], Basyirah Muda[2], Mohamad Habeeb Helmy Idris[2], Mazidah Md Nor[2], Nur Suhadah Nawi[2], Muhamad Shirwan Abdullah Sani[3] & Lai Kok Song[4]

[1]*Halal Products Research Institute, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia*
[2]*Malaysia Halal Analysis Centre (MyHAC), Department of Islamic Development Malaysia, No. 1 Persiaran Teknologi 1, Lebuh Enstek, 71760 Bandar Baru Enstek, Negeri Sembilan, Malaysia*
[3]*International Institute for Halal Research and Training, International Islamic University Malaysia, Jalan Gombak, 53100 Kuala Lumpur, Malaysia*
[4]*Health Sciences Division, Abu Dhabi Women's College, Higher Colleges of Technology, 41012 Abu Dhabi, United Arab Emirates*

Received: 17 February 2025/Acccepted: 23 June 2025

ABSTRACT

Gelatin is derived from animal collagen, sourced primarily from bovine or porcine, and finds widespread application within the food industry. These issues raise concern over its halal status, particularly among Muslims and Jews, as they adhere to dietary laws prohibiting the consumption of pork and its derivatives. Conventional methods like quantitative Polymerase Chain Reaction (qPCR) and liquid chromatography–mass spectrometry (LC–MS) have limitations due to the deoxyribonucleic acid (DNA)'s reliability and the gelatin's complex composition, respectively. Therefore, this study aimed to explore the application of artificial intelligence (AI)–based machine learning, focusing on amino acid composition for non-halal gelatin prediction. A set of 3,780 data points enabled the analysis of the chromatographic peak areas of 18 amino acids in 210 gelatin samples. Orthogonal partial least squares discriminant analysis (OPLS–DA) and artificial neural network (ANN) compared their performance in machine learning models. The ANN employed resilient backpropagation algorithms that demonstrated high accuracy (98.5%) and regression (R^2) of 0.913, with a slightly higher Root Mean Square Error (RMSE) of 0.244. However, OPLSDA demonstrated the best accuracy (100%), R^2 of 0.997, and lower RMSE (0.130) compared to the ANN model. The ANN's robustness against outliers and direct output results provided practical advantages, while OPLS–DA offered comprehensive insights and robust discrimination. This study demonstrates the potential of AI-based machine learning in non-halal gelatin prediction, with both models showing the same capability. These findings can be integrated with existing analytical methods to complement the halal analysis, thus ensuring product integrity and upholding halal sanctity.

Keywords: Artificial neural network; gelatin; halal; machine learning; OPLS–DA

ABSTRAK

Gelatin diperoleh daripada kolagen haiwan dan biasanya diperoleh daripada lembu atau khinzir. Gelatin ini digunakan secara meluas dalam industri makanan. Hal ini menimbulkan kebimbangan mengenai status halal, terutamanya dalam kalangan umat Islam dan Yahudi, kerana mereka terikat kepada undang-undang pemakanan yang melarang pengambilan daging babi dan sumbernya. Kaedah analisis seperti tindak balas rantaian polimerase kuantitatif (qPCR) dan kromatografi cecair–spektrometri jisim (LC–MS) mempunyai had kerana kebolehpercayaan asid deoksiribonukleik (DNA) dan komposisi gelatin yang kompleks. Oleh itu, kajian ini bertujuan untuk meneroka penggunaan pembelajaran mesin berasaskan kecerdasan buatan (AI), dengan memberi tumpuan kepada komposisi asid amino untuk ramalan gelatin tidak halal. Set data yang terdiri daripada 3,780 data membolehkan analisis kawasan kromatografi bagi 18 asid amino dalam 210 sampel gelatin. Analisis diskriminan–kuasa dua separa ortogonal (OPLS–DA) dan rangkaian saraf tiruan (ANN) membandingkan prestasi masing-masing dalam model pembelajaran mesin. ANN menggunakan algoritma perambatanbalik yang menunjukkan ketepatan tinggi (98.5%) dan regresi (R^2) 0.913 dengan Ralat Purata Punca Kuasa Dua (RMSE) yang sedikit lebih tinggi iaitu 0.244. Walau bagaimanapun, OPLS–DA menunjukkan ketepatan terbaik (100%), R^2 (0.997) dan RMSE yang lebih rendah (0.130) berbanding model ANN. Ketahanan ANN terhadap pencilan dan hasil langsung memberikan kelebihan praktikal, manakala OPLS–DA memberikan pandangan yang komprehensif dan diskriminasi yang kukuh. Kajian ini menunjukkan potensi

pembelajaran mesin berasaskan AI dalam ramalan gelatin tidak halal dengan kedua-dua model menunjukkan keupayaan yang sama. Penemuan ini boleh digabungkan dengan kaedah analisis sedia ada untuk melengkapkan analisis halal, justeru memastikan integriti produk dan memelihara kesucian halal.

Kata kunci: Gelatin; halal; pembelajaran mesin; pengesahan daging; rantaian saraf tiruan; OPLS–DA

INTRODUCTION

Gelatin is a protein-rich material derived from a collagen protein found in animals' bones, skins, and connective tissues (Ahmad et al. 2024; Yuswan et al. 2021). It is widely used in the food industry, encompassing muscle foods, dairy products, confectionary and desserts, beverages, and bakery products (Ahmad et al. 2024). In 2020, the global market value for collagen and gelatin exceeded approximately USD4.7 billion, and it is anticipated to reach over USD7 billion by 2027 (Ahmad et al. 2024).

However, commercial gelatin is primarily sourced from pigs, constituting 41% of global production (Ali et al. 2018). While bovine hide, bovine bone, and fish contribute 28.5%, 29.5%, and 1%, respectively (Milovanovic & Hayes 2018). This contribution raises concerns among certain groups of people, notably Muslims and Jews, who are prohibited from consuming pig and its by-products (Uddin et al. 2021). Moreover, Hindu dietary law also forbids the consumption of bovine derivatives (Uddin et al. 2021).

Commonly, detecting the origin of gelatin involves quantitative polymerase chain reaction (qPCR) and liquid chromatography–mass spectrometry (LC–MS) methods (Jannat et al. 2018), yet both methods have limitations. The qPCR method lacks reliability because gelatin is a polypeptide by-product of collagen during partial hydrolysis in harsh industrial processing. Therefore, the possibility of obtaining so-called 'deoxyribonucleic acid (DNA)'s gelatin' depends entirely on the industrial purification process and could also originate from cross-contamination. Meanwhile, LC-MS is time-consuming and prone to false positives due to gelatin's complex molecular composition, characterized by repetitive GXY motifs, where G represents glycine and X and Y represent proline and hydroxyproline, respectively (Kleinnijenhuis, van Holthoon & Herregods 2018). Therefore, an alternative method is required to enhance the accuracy of non-halal gelatin detection.

Hence, the objective of this study was to explore the utilization of artificial intelligence (AI)–based machine learning in forecasting non-halal gelatin by analyzing its biomolecular structure, focusing on amino acid composition. The chromatographic peak areas of amino acids served as input data. AI is a branch of computer science that simulates human cognitive functions such as reasoning, learning, and knowledge acquisition. It has been widely adopted across various sectors, including gaming, weather forecasting, food processing, the medical industry, data mining, and stem cell research (Mavani et al. 2022). The performance of AI is enhanced through machine learning which involves specific algorithms for learning and improving from experience without being explicitly programmed. Machine learning techniques are generally categorized into three main types: supervised learning, unsupervised learning, and reinforcement learning (Mavani et al. 2022). In the food industry, numerous machine learning algorithms have been applied, such as regression analysis, classification analysis, cluster analysis, dimensionality reduction, association rule learning, reinforcement learning, and artificial neural networks as well as deep learning (Sarker 2021).

In this study, Principal Component Analysis (PCA), a dimensionality reduction technique, was employed to eliminate outliers from the dataset before implementing the Orthogonal Partial Least Squares–Discriminant Analysis (OPLS–DA), a classification method, and Artificial Neural Network (ANN) to evaluate the potential of AI-based machine learning approaches in predicting the presence of non-halal gelatin. The performance of OPLS–DA and ANN was assessed for their suitability in this application. Previously, OPLS–DA an advancement of PLS–DA, enhances discrimination between multiple groups (Boccard & Rutledge 2013), while ANN has shown promising results in predicting, particularly for protein structure and function (Tsuchiya & Tomii 2020). This innovative AI–driven methodology offers a complementary tool to existing analytical techniques, thus ensuring product integrity and upholding halal sanctity.

MATERIALS AND METHODS

MATERIALS

Two batches of porcine (G1890 and G2625), bovine (G9382 and G6650), and fish (G7041 and G7765) gelatin were purchased from Sigma-Aldrich. To ensure a diverse sample set, an additional batch of gelatin was obtained from Millipore; however, only porcine gelatin (G48722) was available from this supplier. The gelatin samples consisted of 90 type A gelatin samples from porcine skin (30 samples from each of the batches G1890 Sigma Aldrich, G2625 Sigma Aldrich, and G48722 Millipore), 60 type B gelatin samples from bovine skin (30 samples from each of the batches G9382 Sigma Aldrich and G6650 Sigma Aldrich), and 60 gelatin samples from cold water fish skin (30 samples from each of the batches G7041 Sigma Aldrich and G7765 Sigma Aldrich). All gelatin samples were subjected to acid hydrolysis for chromatographic amino

acid separation. The datasets were constructed based on the peak area of 18 amino acids, including aspartic acid, serine, glutamic acid, glycine, hydroxyproline, histidine, arginine, threonine, alanine, proline, cystine, tyrosine, valine, methionine, lysine, isoleucine, leucine, and phenylalanine across 210 gelatin samples.

AMINO ACID HYDROLYSIS

Each gelatin sample was subjected to amino acid hydrolysis and derivatization, as described previously (Yuswan et al. 2021). Each sample was weighed approximately 0.2 g and incubated with 5 mL of 6 N HCl at 110 °C for 24 h in an oven for hydrolysis. Then, the sample was transferred into a 100-mL volumetric flask, and 4 mL of 2.5 mM L-2-aminobutyric acid (AABA) was added. The 100-mL volumetric flask volume was made up of ultrapure water. Next, 2 mL of the diluted sample was filtered into an Eppendorf tube using a 0.45-µm MS PTFE syringe filter. For derivatization, 70 µL of borate buffer was added to a new clean Eppendorf tube, followed by 10 µL of filtered sample. The mixture was vortexed immediately for 5 s. Then, 20 µL of AccQ-Fluor Reagent was added and vortexed for 5 s. The sample was transferred into an HPLC insert vial before chromatographic separation.

CHROMATOGRAPHIC SEPARATION

Chromatographic separation was performed as described previously (Yuswan et al. 2021). The HPLC system consisted of a Waters e2695 separation module, a Waters column compartment, and a Waters 2475 multi λ fluorescence detector (Maple Street Milford, MA, USA). The sample was injected into a Waters AccQ Tag reversed-phase column (3.9 × 150 mm, 4 µm, 60 Å) at a flow rate of 1 mL/min at 36 °C for 50 min. Mobile phases A, B, and C were AccQ Tag HPLC Eluent A (1:10), 100% acetonitrile, and 100% ultrapure water, respectively. The column was equilibrated for 10 column volumes before sample injection at a volume of 10 µL. The chromatographic gradient was initially set to 98% A: 0.8% B: 1.2% C for 0.5 min, then 92% A: 3.2% B: 4.8% C from 0.5 to 15 min, then 85% A: 6% B: 9% C from 15 to 19 min, then 65% A: 14% B: 21% C from 19 to 32 min, maintained for 32 to 33 min, then 40% B: 60% C from 33 to 35 min, maintained from 35 to 38 min, then back to the initial conditions from 38 to 39 min before equilibration for the next injection from 39 to 50 min. The excitation and emission wavelengths of fluorescence detection were set at 250 and 395 nm, respectively.

STATISTICAL ANALYSIS AND MACHINE LEARNING MODELLING

All data are presented as a means with standard deviations. The suitability of the dataset was assessed through Kaiser–Meyer–Olkin (KMO) analysis, while the intercorrelation among 18 amino acids within the dataset was evaluated using Bartlett's test of sphericity. The relationship among the 18 amino acids was also examined through Pearson correlations. Principal Component Analysis (PCA) was employed to identify the most significant factors contributing to non-halal gelatin prediction and to remove potential outliers. Subsequently, the *sample()* function in RStudio Team was used to randomly select sample replicates for balanced sample analysis. Subsequently, the factors were utilized in modelling using orthogonal partial least squares discriminant analysis (OPLS–DA) (Jadhav et al. 2021; Liu et al. 2023) technique with some modifications. The OPLS–DA model was assessed based on four parameters: variance explained by the X matrix (R^2X), variance explained by the Y matrix (R^2Y), goodness of predictive power (Q^2), and the root mean squared error (RMSE). The random cross–validation for groups was set to 7. Meanwhile, the artificial neural network (ANN) model was evaluated for ideal network architecture based on resilient backpropagation algorithms, activation functions (hyperbolic tangent vs logistic), hidden layer as well as its neuron numbers, determination coefficient (R^2), accuracy, and the RMSE. To develop the OPLSA–DA and ANN models, the dataset was randomly split into two subsets at a 7:3 ratio: 70% for training (model development) and 30% for testing (model validation) using the *sample()* function in RStudio Team. The PCA and OPLS–DA models were constructed using MKS Umetrics AB SIMCA software, version 14.1.0.2047 (Umea, Sweden), while the data subset randomization (7:3 ratio), KMO analysis, Bartlett's test of sphericity, Pearson correlation, and ANN model was developed using RStudio Team (Version 1.4.1717): Integrated Development Environment for R, 2015. A significance level of $\alpha = 0.05$ was used.

RESULTS AND DISCUSSION

ASSESSMENT OF THE GELATIN BIG DATA

The dataset employed in this study consisted of 18 rows representing the chromatographic peak area of amino acids and 210 columns of gelatin samples, resulting in 3,780 data points. The Kaiser-Meyer-Olkin (KMO) analysis assesses the adequacy of the dataset, showing an overall measure of sampling adequacy (MSA) of 0.86. The KMO index ranges from 0 to 1, where an index exceeding 0.5 indicates the appropriateness of samples (Jameel & Al-Salami 2023). Before conducting the Principal Component Analysis (PCA), Bartlett's test ascertains the validity of PCA (Jameel & Al-Salami 2023) by verifying the significant interconnections among 18 amino acids across 210 gelatin samples. Bartlett's test yielded a p–value < 0.001, denoting significant intercorrelations. Subsequently, the relationships among the 18 amino acids were examined through Pearson correlation, as depicted in Table 1. The Pearson correlation analysis showed that, out of 153 correlations,

only eight were not statistically significant. Among these, six involved hydroxyproline's relationships with arginine, cystine, tyrosine, valine, isoleucine, and leucine. This high prevalence of significant correlations can be attributed to the use of a single parameter (chromatographic peak area), in contrast to another study that implements multiple parameters such as color, smell, taste, hardness, viscosity index, adhesiveness, and acidity (Stangierski, Weiss & Kaczmarek 2019). Additionally, aspartic acid and glutamic acid exhibited the strongest positive correlation ($r = 0.986$, p–value < 0.001), suggesting that both amino acids are derived from the deamination of asparagine and glutamine, respectively (Ahmad et al. 2024). Conversely, proline and hydroxyproline showed the weakest positive correlation ($r = 0.146$, $p < 0.05$), which may reflect the conversion of proline into hydroxyproline through hydroxylation (Amira Aqilah et al. 2019).

To reduce the dimensionality of the dataset while retaining important information, a PCA model was developed, and Figure 1 illustrates the PCA score plot. Consequently, 25, 22, and 14 outliers corresponded to porcine, bovine, and fish gelatin, respectively (Figure 1(a)). Outlier selection was based on the sample location, wherein any sample located outside the tolerance ellipse (Hotelling's T^2 statistic at $p = 0.05$) and clustering with a different group, was considered an outlier. Furthermore, there was unclear discrimination among the porcine and bovine gelatin samples. This is due to the highly similar amino acid compositions of porcine and bovine gelatins, both of which contain high concentrations of hydroxyproline, resulting in minimal differences that make them indistinguishable in the PCA score plot (Yuswan et al. 2021). A similar observation has been reported in a previous study on authentication and quantification of porcine adulterant in gelatin and marshmallow (Muhamad Shirwan et al. 2021). After outliers' exclusion (Figure 1(b)), the PCA model underwent refinement, showing the first two principal components (PCs). Notably, PC1 and PC2 contributed 0.843 and 0.108, respectively. Both PC1 and PC2 accounted for a total variance (R^2X) of 0.95, elucidating the dataset with a predictive ability (Q^2) of 0.933. A previous study on gelatin and collagen as halal-critical food ingredients reported low R^2X (0.643) and Q^2 (0.547) (Yuswan et al. 2021), which might be due to the presence of diverse samples such as carrageenan and various test samples. In contrast, this study exclusively focused on gelatin samples for non-halal gelatin prediction.

ORTHOGONAL PARTIAL LEAST SQUARES DISCRIMINANT ANALYSIS (OPLS-DA) MODEL

An OPLS-DA model predicts non-halal gelatin. Initially, the clean dataset, consisting of 149 gelatin samples (after outliers' exclusion), was partitioned into training and testing datasets based on subset randomization at a 7:3 ratio. The model was constructed using the training dataset (100 gelatin samples), and subsequently, the testing dataset (49 gelatin samples) assessed the model's recognition accuracy and predictive ability (Oliveri et al. 2021). Figure 2 illustrates the score and loading plot of the training and testing OPLS-DA model. As anticipated, the training OPLS-DA score plot exhibited distinct discrimination (Figure 2(a)), effectively segregating all gelatin samples into three distinct groups. When incorporating the testing dataset into the established training OPLS-DA model, a 100% number of correct classifications was achieved for the same gelatin sample groups on the predicted OPLS-DA testing score plot (Figure 2(b)). However, large intra-variance for porcine gelatin samples was observed as compared to others gelatin samples. Several factors may contribute to the observed intra-variance among porcine gelatin samples, including source heterogeneity (arising from differences in animal age, geographic origin, and supplier-related factors), processing differences (such as variations in extraction and purification methods), and analytical variability (such as instrument sensitivity, operator technique, and sample preparation inconsistencies) (Liu & Locasale 2017). This intra-variance does not significantly affect the OPLS-DA model, as the R^2X values for both the training and testing models are nearly identical. Conversely, the loading plot of the OPLS-DA model delineated the distribution of 18 amino acids with their respective contributions to predicted gelatin samples (Figure 2(c) and 2(d)). Notably, only hydroxyproline demonstrated a negative correlation with others among these amino acids. A previous study reported that hydroxyproline was a signature amino acid in gelatin and collagen for the initial detection of halal-critical food ingredients (Yuswan et al. 2021).

Table 2 presents a summary of discrimination results from the OPLS-DA model of the source of gelatin samples. For the training of the OPLS-DA model, the score plot effectively segregated gelatin samples with 99.7% explanatory power for a variation on X ($R^2X = 0.997$), thereby indicating that species constitutes the primary factor influencing the distinct sources of gelatin. Furthermore, the OPLS-DA model training demonstrated 99.2% goodness of fit ($R^2Y = 0.992$) with a predictive accuracy of 99.1% ($Q^2 = 0.991$). To ensure that the OPLS-DA model training avoided overfitting, random cross-validation was conducted during its establishment using the training dataset (Liu et al. 2023). As inferred from the random cross-validation, the Root Mean Square Error (RMSE) for porcine, bovine, and fish gelatin were estimated at 0.059, 0.045, and 0.026, respectively. These values indicate the accuracy of the OPLS-DA model training in predicting the non-halal gelatin samples. The RMSE can be defined as a dimensionless statistic and serves as a valuable metric for model evaluation (Hang et al. 2022).

Nevertheless, the OPLS-DA model testing decreased, although 100% of predictions were achieved for the testing dataset (Table 2). The accuracy was computed by class prediction rates (Supplementary 1). Both explanatory

TABLE 1. Pearson correlation between 18 amino acids based on chromatographic peak area

	Aspartic. Acid	*Serine*	*Glutamic. Acid*	*Glycine*	*Hydroxyproline*	*Histidine*	*Arginine*	*Threonine*	*Alanine*	*Proline*	*Cystine*	*Tyrosine*	*Valine*	*Methionine*	*Lysine*	*Isoleucine*	*leucine*	*Phenylalanine*
Aspartic. Acid																		
Serine	0.721***																	
Glutamic. Acid	0.986***	0.653***																
Glycine	0.798***	0.747***	0.812***															
Hydroxyproline	-0.425***	-0.519***	-0.289***	-0.084														
Histidine	0.541***	0.415***	0.513***	0.409***	-0.271***													
Arginine	0.747***	0.735***	0.775***	0.923***	0.055	0.483***												
Threonine	0.727***	0.946***	0.670***	0.819***	-0.411***	0.484***	0.842***											
Alanine	0.972***	0.710***	0.989***	0.867***	-0.244***	0.498***	0.846***	0.739***										
Proline	0.815***	0.386***	0.893***	0.762***	0.146*	0.457***	0.782***	0.449***	0.894***									
Cystine	0.474***	0.133	0.514***	0.340***	-0.004	0.246***	0.335***	0.141*	0.476***	0.543***								
Tyrosine	0.628***	0.420***	0.621***	0.606***	-0.046	0.759***	0.679***	0.540***	0.611***	0.646***	0.356***							
Valine	0.857***	0.455***	0.925***	0.787***	0.066	0.491***	0.807***	0.513***	0.926***	0.992***	0.562***	0.661***						
Methionine	0.762***	0.954***	0.679***	0.744***	-0.608***	0.497***	0.726***	0.971***	0.725***	0.381***	0.143*	0.497***	0.457***					
Lysine	0.783***	0.509***	0.790***	0.607***	-0.271***	0.382***	0.570***	0.510***	0.777***	0.689***	0.353***	0.433***	0.711***	0.535***				
Isoleucine	0.812***	0.601***	0.877***	0.846***	0.060	0.340***	0.875***	0.653***	0.921***	0.902***	0.406***	0.458***	0.912***	0.568***	0.683***			
leucine	0.829***	0.430***	0.906***	0.781***	0.120	0.459***	0.806***	0.490***	0.912***	0.995***	0.542***	0.631***	0.997***	0.423***	0.699***	0.926***		
Phenylalanine	0.364***	0.317***	0.432***	0.667***	0.503***	0.417***	0.819***	0.484***	0.497***	0.650***	0.246***	0.728***	0.629***	0.298***	0.279***	0.610***	0.645***	

Statistically significant correlations are denoted by an asterisk. The symbols * and *** represent p-values < 0.05 and 0.001, respectively

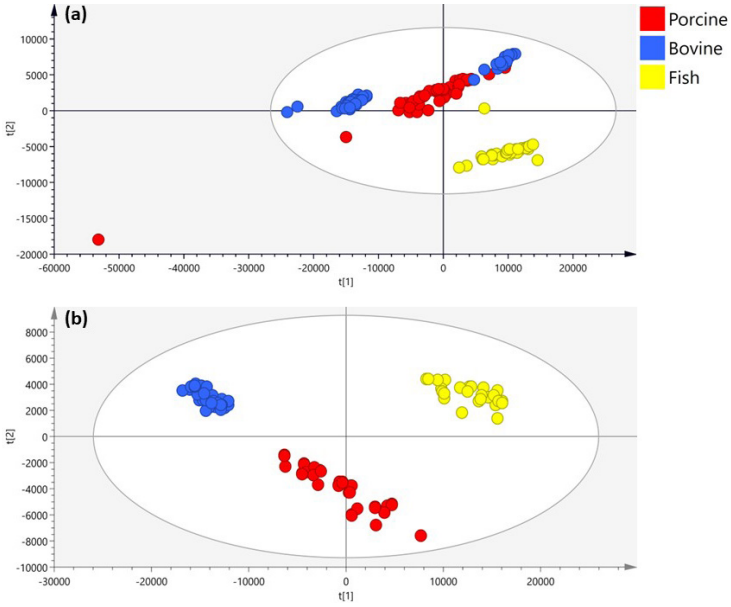

FIGURE 1. Score plot for (a) the Principal Component Analysis (PCA) model including outliers and (b) the PCA model excluding outliers

power for a variation on X and goodness of fit were dropped to 88.4% (R^2X = 0.884) and 38.3% (R^2Y = 38.3%), respectively. The predictive accuracy also dropped to 37.9% (Q^2 = 0.379), indicating that the OPLS-DA model performed poorly in predicting non-halal gelatin. R^2Y represents the proportion of variance in the response variable (Y) explained by the model, whereas Q^2 reflects the model's predictive ability based on cross-validation. A large discrepancy between R^2Y and Q^2 typically indicates overfitting. In this study, although both R^2Y and Q^2 values dropped, the difference between them remained small ($R^2Y - Q^2 = 0.004$). This suggests that the decrease in these values is more likely due to limited model generalizability caused by class imbalance after outlier removal, rather than overfitting. This interpretation is supported by a previous metabolomics-based comparative analysis, in which one of the OPLS-DA models reported R^2X = 0.227, R^2Y = 0.477, and Q^2 = 0.129 (Zhang et al. 2022).

ARTIFICIAL NEURAL NETWORK (ANN) MODEL

This study employed an ANN model to predict non-halal gelatin based on the chromatographic peak area of 18 amino acids in 210 gelatin samples utilized as input data. The ANN was trained using an independent training dataset comprising 143 gelatin samples and subsequently validated using another independent testing dataset consisting of 67 gelatin samples. Table 3 shows the parameters utilized for developing the ANN model. There are 18 combinations of parameters in constructing and training the ANN model.

The ANN model was a multilayer input feed-forward architecture, incorporating a single hidden layer and output. The number of multilayer inputs was the same as that of 18 amino acids. A previous study indicated that a single layer is sufficient for the ANN model to approximate any complex nonlinear functions (Oliveri et al. 2021). Meanwhile, determining the number of neurons within the hidden layer significantly influenced the estimated parameter (NP) count, encompassing weights and biases. In this study, the range of neurons tested within the hidden layer extended from 1 to 9, ensuring that the total count of NP remained below the original dataset size (210) for optimum performance (Gonçalves Neto et al. 2021). As the ANN model was constructed for cluster prediction, the output was a single layer with possible outputs being porcine, bovine, fish, or unknown. Figure 3 illustrates the schematic model of the ANN model.

The resilient backpropagation (rprop) was the algorithm to determine the optimal performance of the ANN model. Generally, the rprop algorithm with the hyperbolic tangent (tanh) activation function yielded the best predictions for the non-halal gelatin, with an R^2 value of 0.913 and an accuracy of 0.985. An R^2 value exceeding 0.98 denotes a strong alignment between observed and predicted data (Stangierski, Weiss & Kaczmarek 2019). This algorithm and activation function also achieved the lowest RMSE of 0.244 among other combinations. This RMSE signifies the deviation for the ANN model, where a low value indicates model robustness (Stangierski, Weiss & Kaczmarek 2019). In this case, the ANN architecture of 1, 2 and 4 are the best combination parameters (Table 3). These ANN models accurately predict gelatin samples except for one sample of fish gelatin, resulting in an RMSE value of 0.471 for fish gelatin.

TABLE 2. Summary of discrimination results from the OPLS-DA model of the source of gelatin samples

Model OPLS-DA	Number of samples			Σ	Number of correct classifications			Σ	Root mean square error (RMSE)			R^2X	R^2Y	Q^2
	Porcine	Bovine	Fish		Porcine	Bovine	Fish		Porcine	Bovine	Fish			
Training	43	26	31	100	43	26	31	100	0.059	0.045	0.026	0.997	0.992	0.991
Testing	22	12	15	49	22	12	15	49	0.318	0.241	0.260	0.884	0.383	0.379
				149				149						

The training and testing dataset is a subset of the 7:3 ratio from each source of gelatin samples. Σ = total. R^2X = variance explained by the X matrix, R^2Y = variance explained by the Y matrix, Q^2 = goodness of predictive power, RMSE = measure of the difference between a model's predicted and actual values as estimated from a cross-validation

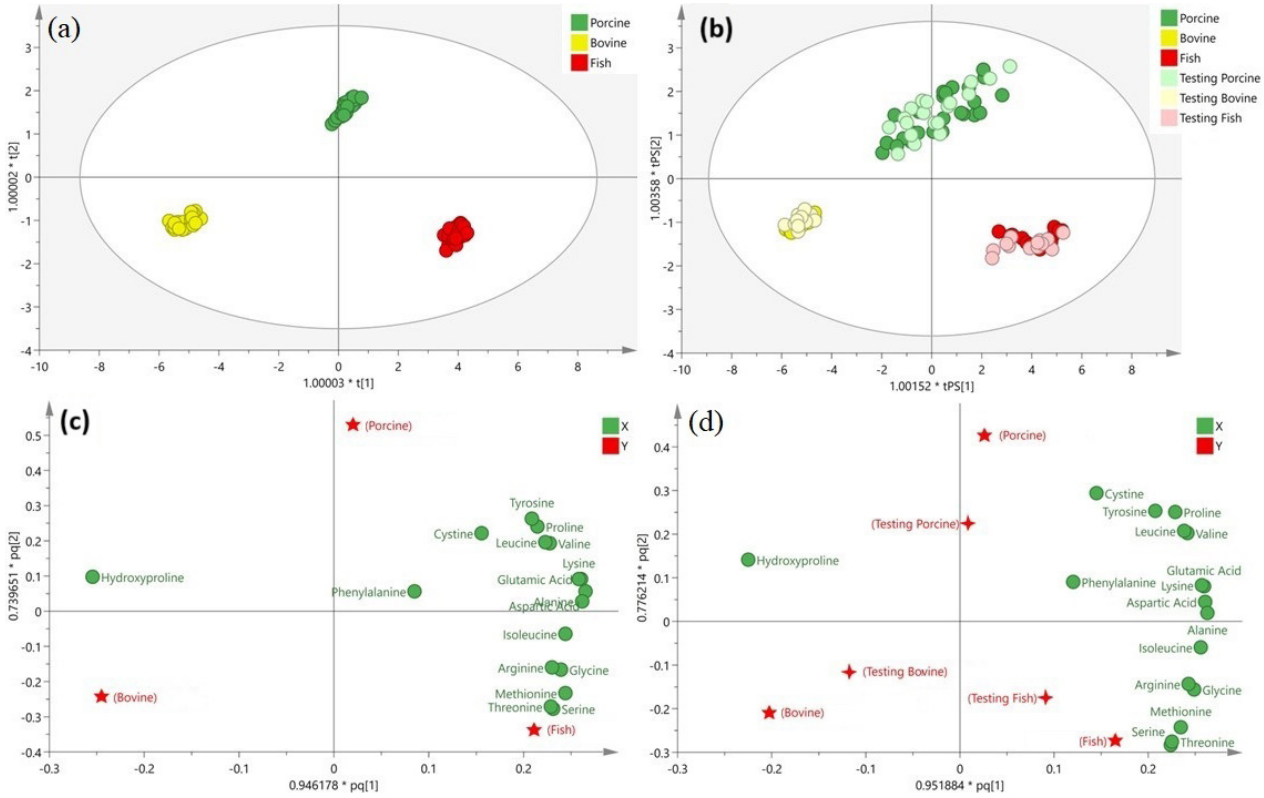

FIGURE 2. OPLS-DA consists of (a) training score plot shows a clear discrimination of training dataset, (b) testing score plot shows a clear prediction of testing dataset, (c) training loading plot shows a prediction of training (5-points star) dataset based on species among the 18 amino acids, and (d) loading plot shows a prediction of training (5-points star) and testing (4-points star) dataset based on species among the 18 amino acids

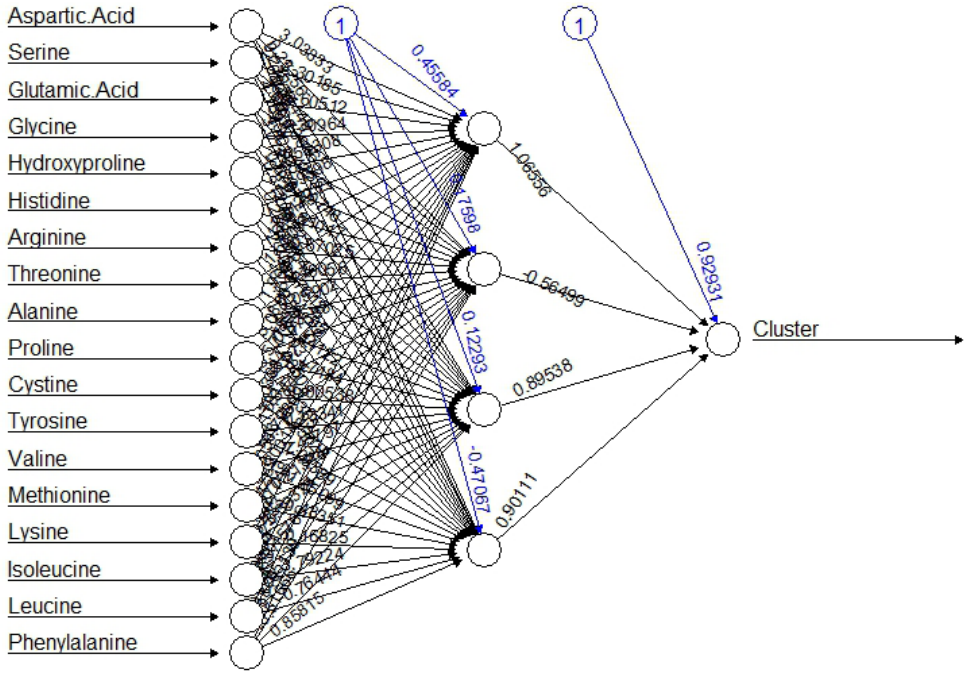

FIGURE 3. The schematic model of an artificial neural network (ANN). Circles symbolize neurons, with 18 multilayer inputs (amino acids), a single hidden layer consisting of 4 neurons in the middle, and a single output neuron representing cluster for porcine gelatin, bovine gelatin, fish gelatin, or unknown. Neurons labelled with number 1 represent biases, while arrows annotated with number represent weights. The biases serve as additional input to neurons (excluding input neurons) and represent threshold of the neuron's activation function to minimize the error of the ANN. The weights reflect the strength of connection between two neurons and are adjusted iteratively to minimize the error between the predicted and actual output. These numerical parameters are optimized for the ANN model after 15029 iterations through the training dataset

COMPARISON OF ORTHOGONAL PARTIAL LEAST SQUARE DISCRIMINANT ANALYSIS (OPLS-DA) AND ARTIFICIAL NEURAL NETWORK (ANN) MODELS

OPLS-DA and ANN represent machine learning models that are widely employed in multivariate data analysis across diverse disciplines such as chemistry, biology, and engineering (Chang et al. 2022; Guiné 2019; Liu et al. 2023; Tsuchiya & Tomii 2020; Zheng et al. 2011). In this study, both OPLS-DA and ANN models demonstrate efficacy in predicting non-halal gelatin based on the chromatographic peak area of amino acids. A comparative analysis between the testing OPLS-DA and optimal ANN models (ID 1, 2, and 4 in Table 3) focuses on accuracy, coefficient determination (R^2), and variation (RMSE).

Regarding accuracy, the OPLS-DA model achieved 100% of correct classifications with no unknown prediction (Table 3). This performance is due to the exclusion of all outliers through principal component analysis before constructing the OPLS-DA model (Liu et al. 2023; Oliveri et al. 2021; Wang et al. 2023). The ANN models demonstrated robustness by maintaining high accuracy (98.5%) upon training on input datasets containing outliers. These advantages result from their nonlinear modeling capabilities, robust training algorithms that can detect and mitigate the influence of outliers, and high performance in complex data (Gbashi et al. 2023). Remarkably, the optimal ANN models (ANN ID 1, 2, and 4 in Table 3) did not generate any unknown predictions and accurately classified all gelatin samples, except for one incorrect prediction on fish gelatin samples. This observation may be attributed to the presence of outliers in the training dataset, which can induce significant oscillations in the performance of the ANN model. Therefore, the exclusion of outliers is recommended in future studies (Gbashi et al. 2023).

R^2 represents the proportion of the variance in the dependent variables that is predictable from the independent variables (Bhagya Raj & Dash 2022; Gonçalves Neto et al. 2021; Stangierski, Weiss & Kaczmarek 2019). In this case, R^2 indicates the variance in chromatographic peak area that

TABLE 3. The parameters used for the development of the artificial neural network (ANN) model. The ANN identity 1, 2, and 4 are the best combination parameter

ANN ID	Activation function	Number of neurons in single hidden layer	NP	Actual* vs prediction							RMSE				Accuracy	R^2
				Porcine (30)		Bovine (19)		Fish (18)		?	Porcine	Bovine	Fish	Σ		
				√	x	√	x	√	x							
1	tanh	1	21	30	0	19	0	17	1	0	0.000	0.000	0.471	0.244	0.985	0.913
2		2	41	30	0	19	0	17	1	0	0.000	0.000	0.471	0.244	0.985	0.913
3		3	61	30	0	19	0	17	0	1	0.000	0.000	0.707	0.367	0.985	0.804
4		4	81	30	0	19	0	17	1	0	0.000	0.000	0.471	0.244	0.985	0.913
5		5	101	29	1	19	0	17	1	0	0.183	0.000	0.471	0.273	0.970	0.891
6		6	121	29	1	19	0	17	1	0	0.183	0.000	0.471	0.273	0.970	0.891
7		7	141	29	1	19	0	17	0	1	0.183	0.000	0.707	0.386	0.970	0.782
8		8	161	29	1	19	0	16	2	0	0.183	0.000	0.527	0.299	0.955	0.869
9		9	181	29	1	19	0	16	2	0	0.183	0.000	0.333	0.212	0.955	0.935
10	logistic	1	21	1	29	19	0	0	18	0	0.983	0.000	1.000	0.838	0.299	-0.025
11		2	41	29	1	19	0	17	1	0	0.183	0.000	0.471	0.273	0.970	0.891
12		3	61	29	1	19	0	17	0	1	0.183	0.000	0.707	0.386	0.970	0.782
13		4	81	29	1	19	0	17	0	1	0.183	0.000	0.707	0.386	0.970	0.782
14		5	101	29	1	19	0	17	0	1	0.183	0.000	0.707	0.386	0.970	0.782
15		6	121	29	1	19	0	17	0	1	0.183	0.000	0.707	0.386	0.970	0.782
16		7	141	29	1	19	0	17	1	0	0.183	0.000	0.471	0.273	0.970	0.891
17		8	161	29	1	19	0	17	1	0	0.183	0.000	0.471	0.273	0.970	0.891
18		9	181	29	1	19	0	16	1	1	0.183	0.000	0.745	0.405	0.955	0.760

*Actual number represents in bracket, tanh = hyperbolic tangent, NP = total quantity of estimated parameters, √ = correct prediction, x = wrong prediction, ? = unknown prediction, Σ = total, RMSE = root mean square error, R^2 = determination coefficient

is predictable from the amino acid composition. A good model for predicting non-halal gelatin should demonstrate a high R^2 value. The OPLS-DA model has two types of R^2 values that refer to the dataset ($R^2X = 0.997$) and sample clusters ($R^2Y = 1.00$). However, the ANN models exhibited a slightly low R^2 of 0.913. This trend aligns with findings from a previous study, where the R^2 value for the partial least square regression (PLSR) model exhibits a slightly higher than the ANN model (Zheng et al. 2011).

The low RMSE value indicates the robustness of the machine learning model (Zheng et al. 2011). In this case, the OPLS-DA model demonstrated a lower RMSE value of 0.130, in contrast to the ANN models (RMSE = 0.244). A lower RMSE value typically signifies a better predictive model. However, the two models have no significant difference in RMSE values. This observation aligns with a previous study on applying ANN and PLSR to predict changes in nutritional components in red bayberry juice (Zheng et al. 2011).

CONCLUSION

This study compared the OPLS-DA and ANN models to ascertain their performance in predicting non-halal gelatin. The results showed that both models displayed competence in predicting non-halal gelatin samples. The OPLS-DA model demonstrated better accuracy, coefficient of determination (R2), and root mean square error (RMSE) compared to the ANN model. This performance was due to the inherent structure of OPLS-DA, which performed outlier removal and provided a comprehensive discrimination overview. Consequently, the OPLS-DA is practical for obtaining direct insights into the system, particularly the interaction between multiple groups of samples. However, the ANN model presents distinct advantages owing to its resilience towards outliers and capability to yield immediate output results in predicting non-halal gelatin samples. ANN offers distinct advantages in processing large, complex, and nonlinear datasets, which are characteristic of real-time industrial applications. Through robust training

algorithms, ANN can mitigate the influence of outliers during the modeling process, enabling the network to learn complex patterns even in the presence of outlier data. Further research is recommended in expanding datasets, the attribute of outliers, investigating alternative machine learning models, incorporating additional parameters, and collaborating with industry stakeholders to enhance the accuracy and reliability of non-halal gelatin prediction through artificial intelligence.

ACKNOWLEDGEMENTS

The study was supported by the Islamic Economics Research and Innovation Fund (IERIF) 2024 (I-RISE/IERIF/AWARD/2025/BATCH 2/71) with account number 6300564–10205.

REFERENCES

Abdullah Sani Ahmad, M.I., Li, Y., Pan, J., Liu, F., Dai, H., Fu, Y., Huang, T., Farooq, S. & Zhang, H. 2024. Collagen and gelatin: structure, properties, and applications in food industry. *International Journal of Biological Macromolecules* 254(3): 128037. https://doi.org/10.1016/j.ijbiomac.2023.128037

Ali, E., Sultana, S., Abd Hamid, S.B., Hossain, M., Yehya, W.A., Kader, A. & Bhargava, S.K. 2018. Gelatin controversies in food, pharmaceuticals, and personal care products: Authentication methods, current status, and future challenges. *Critical Reviews in Food Science and Nutrition* 58(9): 1495-1511.

Amira Aqilah, S., Nur Azira, T., Haizatul Hadirah, G., Nurul Azarima, M.A. & Siti Nur Syahirah, Z. 2019. Analytical methods for gelatin differentiation. *Journal of Halal Industry & Services* 2(1): a0000048. http://www.journals.hh-publisher.com/index.php/JHIS/article/view/57

Bhagya Raj, G.V.S. & Dash, K.K. 2022. Comprehensive study on applications of artificial neural network in food process modeling. *Critical Reviews in Food Science and Nutrition* 62(10): 2756-2783. https://doi.org/10.1080/10408398.2020.1858398

Boccard, J. & Rutledge, D.N. 2013. A consensus orthogonal partial least squares discriminant analysis (OPLS-DA) strategy for multiblock omics data fusion. *Analytica Chimica Acta* 769: 30-39. https://doi.org/10.1016/j.aca.2013.01.022

Chang, L., Mu, G., Wang, M., Zhao, T., Tuo, Y., Zhu, X. & Qian, F. 2022. Microbial diversity and quality-related physicochemical properties of spicy cabbage in Northeastern China and their correlation analysis. *Foods* 11(10): 1511. https://doi.org/10.3390/foods11101511

Gbashi, S., Maselesele, T.L., Njobeh, P.B., Molelekoa, T.B.J., Oyeyinka, S.A., Makhuvele, R. & Adebo, O.A. 2023. Application of a generative adversarial network for multi-featured fermentation data synthesis and artificial neural network (ANN) modeling of bitter gourd–grape beverage production. *Scientific Reports* 13(1): 11755. https://doi.org/10.1038/s41598-023-38322-3

Gonçalves Neto, J., Ozorio, L.V., Campos de Abreu, T.V., dos Santos, B.F. & Pradelle, F. 2021. Modeling of biogas production from food, fruits and vegetables wastes using artificial neural network (ANN). *Fuel* 285: 119081. https://doi.org/10.1016/j.fuel.2020.119081

Guiné, R.P.F. 2019. The use of artificial neural networks (ANN) in food process engineering. *International Journal of Food Engineering* 5(1): 15-21. https://doi.org/10.18178/ijfe.5.1.15-21

Hang, J., Shi, D., Neufeld, J., Bett, K.E. & House, J.D. 2022. Prediction of protein and amino acid contents in whole and ground lentils using near-infrared reflectance spectroscopy. *LWT* 165: 113669. https://doi.org/10.1016/j.lwt.2022.113669

Jadhav, S.R., Shah, R.M., Karpe, A.V., Barlow, R.S., McMillan, K.E., Colgrave, M.L. & Beale, D.J. 2021. Utilizing the food-pathogen metabolome to putatively identify biomarkers for the detection of shiga toxin-producing *E. coli* (stec) from spinach. *Metabolites* 11(2): 67. https://doi.org/10.3390/metabo11020067

Jameel, A.M. & Al-Salami, Q.H. 2023. Principal component analysis technique for finding the best applicant for a job. *Cihan University-Erbil Journal of Humanities and Social Sciences* 7(1): 121-125. https://doi.org/10.24086/cuejhss.v7n1y2023.pp121-125

Jannat, B., Ghorbani, K., Shafieyan, H., Kouchaki, S., Behfar, A., Sadeghi, N., Beyramysoltan, S., Rabbani, F., Dashtifard, S. & Sadeghi, M. 2018. Gelatin speciation using real-time PCR and analysis of mass spectrometry-based proteomics datasets. *Food Control* 87: 79-87. https://doi.org/10.1016/j.foodcont.2017.12.006

Kleinnijenhuis, A.J., van Holthoon, F.L. & Herregods, G. 2018. Validation and theoretical justification of an LC-MS method for the animal species specific detection of gelatin. *Food Chemistry* 243: 461-467. https://doi.org/10.1016/j.foodchem.2017.09.104

Liu, K., Jin, Y., Gu, L., Li, M., Wang, P., Yin, G., Wang, S., Wang, T., Wang, L. & Wang, B. 2023. Classification and authentication of Lonicerae japonicae flos and Lonicerae flos by using 1H-NMR spectroscopy and chemical pattern recognition analysis. *Molecules* 28(19): 6860. https://doi.org/10.3390/molecules28196860

Liu, X. & Locasale, J.W. 2017. Metabolomics reveals intratumor heterogeneity - Implications for precision medicine. *EBioMedicine* 19: 4-5. https://doi.org/10.1016/j.ebiom.2017.04.030

Mavani Nidhi Rajesh, Jarinah Mohd Ali, Suhaili Othman, M.A. Hussain, Haslaniza Hashim & Norliza Abd Rahman. 2022. Application of artificial intelligence in food industry - A guideline. *Food Engineering Reviews* 14: 134-175. https://doi.org/10.1007/s12393-021-09290-z

Milovanovic, I. & Hayes, M. 2018. Marine gelatine from rest raw materials. *Applied Sciences* 8(12): 2407.

Muhamad Shirwan Abdullah Sani, Azilawati Mohd Ismail, Azman Azid & Mohd Saiful Samsudin. 2021. Establishing forensic food models for authentication and quantification of porcine adulterant in gelatine and marshmallow. *Food Control* 130: 108350. https://doi.org/10.1016/j.foodcont.2021.108350

Oliveri, P., Malegori, C., Mustorgi, E. & Casale, M. 2021. Qualitative pattern recognition in chemistry: Theoretical background and practical guidelines. *Microchemical Journal* 162: 105725. https://doi.org/10.1016/j.microc.2020.105725

Sarker, I.H. 2021. Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science* 2: 160. https://doi.org/10.1007/s42979-021-00592-x

Stangierski, J., Weiss, D. & Kaczmarek, A. 2019. Multiple regression models and artificial neural network (ANN) as prediction tools of changes in overall quality during the storage of spreadable processed gouda cheese. *European Food Research and Technology* 245(11): 2539-2547. https://doi.org/10.1007/s00217-019-03369-y

Uddin, S.M.K., Motalib Hossain, M.A., Sagadevan, S., Al Amin, Md. & Johan, M.R. 2021. Halal and kosher gelatin: Applications as well as detection approaches with challenges and prospects. *Food Bioscience* 44(PA): 101422. https://doi.org/10.1016/j.fbio.2021.101422

Wang, Y., Wang, X., Huang, Y., Yue, T. & Cao, W. 2023. Analysis of volatile markers and their biotransformation in raw chicken during *Staphylococcus aureus* early contamination. *Foods* 12(14): 2782. https://doi.org/10.3390/foods12142782

Tsuchiya, Y. & Tomii, K. 2020. Neural networks for protein structure and function prediction and dynamic analysis. *Biophysical Reviews* 12(2): 569-573. https://doi.org/10.1007/s12551-020-00685-6

Yuswan, M.H., A. Jalil, N.H., Mohamad, H., Keso, S., Mohamad, N.A., Tengku Md. Yusoff, T.S., Ismail, N.F., Abdul Manaf, Y.N., Mohd Hashim, A., Mohd Desa, M.N., Yusof, Y.A. & Mustafa, S. 2021. Hydroxyproline determination for initial detection of halal-critical food ingredients (gelatin and collagen). *Food Chemistry* 337: 127762. https://doi.org/10.1016/j.foodchem.2020.127762

Zhang, R-Z., Zhao, J.T., Wang, W.Q., Fan, R.H., Rong, R., Yu, Z.G. & Zhao, Y.L. 2022. Metabolomics-based comparative analysis of the effects of host and environment on *Viscum coloratum* metabolites and antioxidative activities. *Journal of Pharmaceutical Analysis* 12(2): 243-252. https://doi.org/10.1016/j.jpha.2021.04.003

Zheng, H., Jiang, L., Lou, H., Hu, Y., Kong, X. & Lu, H. 2011. Application of artificial neural network (ANN) and partial least-squares regression (PLSR) to predict the changes of anthocyanins, ascorbic acid, total phenols, flavonoids, and antioxidant activity during storage of red bayberry juice based on fractal analysis and red, green, and blue (RGB) intensity values. *Journal of Agricultural and Food Chemistry* 59(2): 592-600. https://doi.org/10.1021/jf1032476

*Corresponding author; email: hafisyuswan@upm.edu.my

SUPPLEMENTARY 1. The accuracy was computed by class prediction rates

Obs ID (Primary)	Obs ID ($ Class ID)	M2. YVar PS ($M2. DA (Testing Porcine))	M2. YPred PS[2] ($M2. DA(Testing Porcine))	M2. YVar PS ($M2. DA(Testing Bovine))	M2. YPred PS[2] ($M2. DA(Testing Bovine))	M2. YVar PS ($M2. DA(Testing Fish))	M2. YPred PS[2] ($M2. DA(Testing Fish))
P1	Testing Porcine	0	0.532168	0	0.0583115	0	0.0857265
P6	Testing Porcine	0	0.468441	0	0.154884	0	0.0504751
P7	Testing Porcine	0	0.413008	0	0.2233	0	0.0356546
P19	Testing Porcine	0	0.513032	0	0.147077	0	0.0143999
P20	Testing Porcine	0	0.58584	0	0.0554164	0	0.0356947
P26	Testing Porcine	0	0.585743	0	0.0463305	0	0.0450248
P35	Testing Porcine	0	0.911005	0	-0.279291	0	0.05493
P36	Testing Porcine	0	0.640257	0	0.066292	0	-0.0290676
P47	Testing Porcine	0	0.660823	0	0.092811	0	-0.0763173
P53	Testing Porcine	0	0.817858	0	-0.136144	0	0.0013819
P59	Testing Porcine	0	0.700641	0	0.0337175	0	-0.0555593
P60	Testing Porcine	0	0.599965	0	0.13411	0	-0.0582242
P61	Testing Porcine	0	0.682411	0	-0.0191571	0	0.0161714
P62	Testing Porcine	0	0.696338	0	-0.0286442	0	0.0120669
P63	Testing Porcine	0	0.753475	0	-0.103061	0	0.0313046
P66	Testing Porcine	0	0.640257	0	0.066292	0	-0.0290676
P67	Testing Porcine	0	0.580299	0	0.137041	0	-0.0417928
P71	Testing Porcine	0	0.841537	0	-0.196033	0	0.0388777
P78	Testing Porcine	0	0.553935	0	0.196357	0	-0.0760562
P83	Testing Porcine	0	0.817858	0	-0.136144	0	0.0013819
P88	Testing Porcine	0	0.702331	0	0.0468373	0	-0.0705612

continue to next page

continue from previous page

P89	Testing Porcine	0	0.700641	0	0.0337175	0	-0.0555593
B2	Testing Bovine	0	0.0589384	0	0.60026	0	0.00200574
B18	Testing Bovine	0	-0.000308424	0	0.661311	0	-0.00156473
B24	Testing Bovine	0	0.0321463	0	0.600757	0	0.0279451
B27	Testing Bovine	0	-0.0282305	0	0.669257	0	0.017918
B34	Testing Bovine	0	0.0258904	0	0.627964	0	0.0064678
B44	Testing Bovine	0	0.0593272	0	0.636199	0	-0.0349032
B45	Testing Bovine	0	-0.0019311	0	0.698813	0	-0.0380781
B46	Testing Bovine	0	0.0910209	0	0.606746	0	-0.0362509
B49	Testing Bovine	0	0.0442484	0	0.633978	0	-0.017763
B52	Testing Bovine	0	0.0107719	0	0.664249	0	-0.0154871
B55	Testing Bovine	0	0.0295166	0	0.65099	0	-0.0205126
B58	Testing Bovine	0	0.000937194	0	0.659886	0	-0.00134601
F7	Testing Fish	0	-0.0225064	0	-0.00856388	0	0.701152
F8	Testing Fish	0	-0.0201115	0	-0.0185483	0	0.708935
F12	Testing Fish	0	-0.0482744	0	0.0816723	0	0.634876
F14	Testing Fish	0	-0.0142483	0	-0.0169536	0	0.701528
F25	Testing Fish	0	-0.0545754	0	0.0250485	0	0.698643
F27	Testing Fish	0	-0.0484214	0	0.0020093	0	0.715984
F29	Testing Fish	0	-0.0551773	0	-0.0364854	0	0.761775
F32	Testing Fish	0	-0.0779026	0	0.138225	0	0.606643
F33	Testing Fish	0	-0.00355253	0	-0.0423986	0	0.716831
F34	Testing Fish	0	-0.0297909	0	-0.0189659	0	0.718913
F36	Testing Fish	0	-0.000722051	0	0.0582066	0	0.611791
F37	Testing Fish	0	-0.0225064	0	-0.00856388	0	0.701152
F50	Testing Fish	0	0.0342072	0	-0.100714	0	0.73883
F52	Testing Fish	0	-0.036506	0	0.0841553	0	0.620737
F56	Testing Fish	0	-0.116594	0	0.156334	0	0.626426