Data-driven modelling of *Aquilaria* essential oils via dual GC profiling and multicollinearity diagnostics

Nur Athirah Syafiqah Noramli, Noor Aida Syakira Ahmad Sabri, Muhammad Ikhsan Roslan, Nurlaila Ismail, Zakiah Mohd Yusoff, Mohd Nasir Taib
Advanced Signal Processing Research Interest Group, Faculty of Electrical Engineering, Universiti Teknologi MARA, Shah Alam, Malaysia

Article Info

Article history:

Received Jun 16, 2025
Revised Oct 21, 2025
Accepted Dec 6, 2025

Keywords:

Aquilaria essential oils
Chemical composition
Gas chromatography
Multicollinearity
Regression analysis
Species differentiation
Variation inflation factor

ABSTRACT

Aquilaria-derived essential oils are chemically diverse and hold significant value in pharmaceuticals, fragrances, and traditional medicine. However, the complexity of their chemical composition presents challenges in statistical modelling, particularly due to multicollinearity among biosynthetically related compounds. This study investigates the extent of multicollinearity in *Aquilaria* essential oil data using multiple linear regression (MLR) and variance inflation factor (VIF) analysis. A regression model was constructed using three compounds, δ-guaiene, 10-epi-γ-eudesmol, and γ-eudesmol, across 360 samples, with VIF and collinearity diagnostics applied to assess model validity. The model explained 93% of the variance in species classification, which is substantially higher than values typically reported in earlier chemometric studies of *Aquilaria* oils. This demonstrates that even a limited number of carefully selected compounds, when supported by diagnostic safeguards, can achieve strong classification accuracy. These findings emphasize the importance of applying multicollinearity diagnostics to improve the interpretability and reliability of chemometric analyses. The study contributes a robust analytical framework for future research and practical applications in species authentication, essential oil quality control, and conservation of *Aquilaria* resources.

Corresponding Author:

Nurlaila Ismail
Advanced Signal Processing Research Interest Group, Faculty of Electrical Engineering
Universiti Teknologi MARA
Shah Alam, Selangor 40450, Malaysia
Email: nurlaila0583@uitm.edu.my

1. INTRODUCTION

Essential oils are complex mixtures of volatile organic compounds that play important roles in pharmaceuticals, cosmetics, perfumery, and aromatherapy [1]–[3]. Among essential oil–producing plants, *Aquilaria* is particularly significant because it yields agarwood, a rare and highly valuable resinous wood [4]. The essential oils derived from *Aquilaria* species contain diverse terpenoid and sesquiterpene compounds whose concentrations vary according to species, geographical origin, and extraction method [3], [5]. This chemical diversity contributes to their high economic and medicinal value but also presents analytical challenges for classification and quality control.

Previous studies have investigated the bioactive properties and chemical compositions of *Aquilaria* oils, often identifying characteristic compounds such as γ-eudesmol and δ-guaiene as significant indicators of species differentiation [6]. In parallel, metabolomics and chemometric approaches have been applied to essential oil data, demonstrating the potential of multivariate statistical models for capturing complex chemical variation [7]. These studies established important foundations for linking chemical compounds to biological or taxonomic outcomes.

Despite these advances, a major limitation remains: most existing analysis treat chemical compounds as independent variables, without accounting for multicollinearity. Multicollinearity, which arises when predictors are strongly correlated due to shared biosynthetic pathways, can inflate standard errors, distort regression coefficients, and reduce model interpretability [8]–[10]. As a result, models built without addressing collinearity may yield misleading results or overlook meaningful interactions among compounds. This limitation constrains the reliability of chemometric methods for *Aquilaria* species authentication and essential oil quality assessment [11], [12].

To address this problem, the present study integrates multicollinearity diagnostics into the statistical analysis of *Aquilaria* essential oils. Using 360 samples from four species, a multiple linear regression (MLR) model was constructed based on three consistently reported compounds: δ-guaiene, 10-epi-γ-eudesmol, and γ-eudesmol. Variance inflation factor (VIF), condition indices, and eigenvalue decomposition were employed to evaluate the extent of multicollinearity and its effect on model stability. This approach provides a means of testing whether regression models remain robust despite intercorrelated predictors.

The contributions of this research are threefold. First, the study demonstrates that a model based on these three compounds explains more than 90% of the variation in species classification. Second, the findings confirm that moderate multicollinearity exists among biosynthetically related compounds, highlighting the necessity of applying diagnostics in chemometric research. Third, the study establishes a methodological framework that can enhance future applications in species authentication, essential oil quality control, and conservation monitoring. The remainder of this article is structured as follows. Section 2 describes the experimental design and statistical methods. Section 3 presents the regression results and multicollinearity diagnostics together with their analysis. Section 4 provides discussion of the findings and concludes the study by summarizing the main contributions and implications.

2. METHOD

This section outlines the procedures used to assess multicollinearity among chemical constituents in essential oils from *Aquilaria* species. Data collection and compound quantification were performed using chromatographic techniques, followed by MLR with three predictor compounds for species classification. Multicollinearity diagnostics, including VIF and condition index analysis, were applied to ensure interpretability and robustness of the results.

2.1. Experimental design and data acquisition

A chemometric approach was applied to investigate statistical correlations in the chemical profiles of essential oils obtained from multiple *Aquilaria* species. The dataset was generated by the Bio Aromatic Research Centre of Excellence (BARCE) at Universiti Malaysia Pahang Al-Sultan Abdullah (UMPSA) using advanced chromatographic methods [13]–[15]. The overall workflow for sample collection, hydro-distillation, compound selection, chemical analysis, and chemometric modelling is summarized in Figure 1.

A total of 360 essential oil samples were collected from four species: *A. beccariana* (AB), *A. malaccensis* (AM), *A. crassna* (AC), and *A. subintegra* (AS). Three compounds were selected for detailed analysis: δ-guaiene (Compound C), 10-epi-γ-eudesmol (Compound D), and γ-eudesmol (Compound E). The selection was based on three criteria:

- Consistency: these compounds were detected reliably across all four sampled species, ensuring comparability [6].
- Chemical and practical importance: they are widely recognized contributors to agarwood aroma, oil quality, and pharmacological activity [6].
- Chemotaxonomic relevance: prior studies identified these sesquiterpenes as diagnostic markers for differentiating *Aquilaria* species [6], [13]–[15].

By focusing on compounds with both biochemical significance and established presence in the literature, the analysis ensured that the predictors carried practical utility for authentication while also offering a robust test case for examining multicollinearity [9], [10], [12]. Their relative concentrations, expressed as peak area percentages, were used as predictor variables in the regression model, as shown in Table 1. This compound selection process ensured alignment with the study objective of linking statistical analysis to chemotaxonomic interpretation.

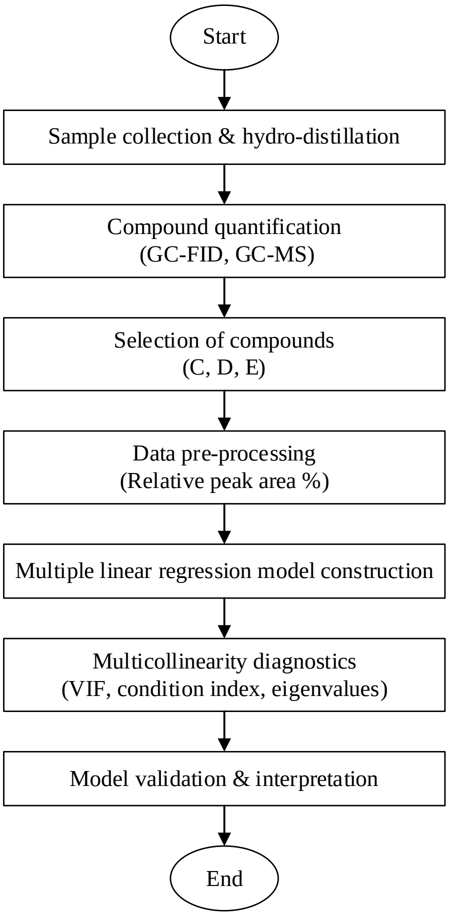
Start
Sample collection & hydro-distillation
Compound quantification
(GC-FID, GC-MS)
Selection of compounds
(C, D, E)
Data pre-processing
(Relative peak area %)
Multiple linear regression model construction
Multicollinearity diagnostics
(VIF, condition index, eigenvalues)
Model validation & interpretation
End

Figure 1. Workflow for *Aquilaria* essential oil analysis and chemometric modelling

Table 1. Chemical compound profiles and corresponding peak area percentages across different *Aquilaria* species

Code	Compounds	Indent. mode	Peak area (%)			
			AB	AM	AC	AS
C	δ-guaiene	MS and FID	0.74	2.02	0.21	0.35
D	10-epi-γ-eudesmol	MS and FID	0.34	6.73	2.54	2.16
E	γ-eudesmol	MS and FID	0.26	2.17	0.95	1.85

Essential oils were extracted by hydro-distillation, following standard practice for volatile oil recovery [6], [13]. Agarwood chips were pre-soaked in water for 2–3 days to soften resin glands and improve extraction efficiency, and distillation was then carried out for 3–5 days. Hydro-distillation was selected because it is both traditional and reproducible [6], while ensuring recovery of the full range of volatile constituents. The extracted oils were diluted with analytical-grade dichloromethane (DCM) prior to analysis to maintain consistency in injection and detection.

Chemical profiling employed gas chromatography–flame ionization detection (GC-FID) for quantification and gas chromatography–mass spectrometry (GC-MS) for compound identification [14]. The combined use of GC-FID and GC-MS was chosen because it ensures reproducibility while providing complementary strengths, with FID enabling precise quantification and MS allowing unambiguous structural confirmation. This dual approach is widely used because FID provides accurate quantitation of volatile compounds [15], while MS enables structural identification based on library matching, as depicted in Figure 2. Identification was confirmed by comparison with the National Institute of Standards and Technology (NIST) mass spectral library, using a similarity threshold of ≥80%. These procedures align with previously established protocols for essential oil characterization [6], [13]–[15] and are summarized in Table 1.

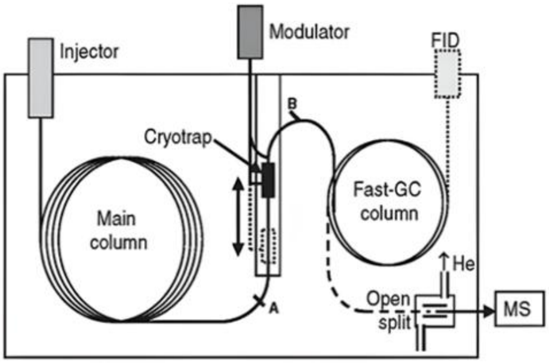

Figure 2. GC×GC system with dual detection: MS and FID [15]

2.2. Statistical analysis and multicollinearity diagnostics

The statistical framework was designed to directly address the problem of multicollinearity outlined in the introduction. MLR was used to model species classification as a function of compounds C, D, and E. The predictors were entered simultaneously using the "Enter" method, enabling evaluation of each variable's contribution while controlling for the others [16].

Linear regression was selected over other modelling methods such as principal component analysis (PCA), partial least squares (PLS), or artificial neural networks (ANN) because the objective was to quantify the direct contribution of individual compounds while testing for multicollinearity. PCA and PLS are effective for dimensionality reduction but obscure variable-level effects, while ANN models, though flexible, function largely as black-box predictors with limited interpretability [17]. In contrast, linear regression provides transparent coefficients, direct interpretability, and compatibility with multicollinearity diagnostics, making it the most appropriate choice for the aims of this study [16].

Prior to modelling, compound concentrations were standardized as relative peak area percentages to ensure comparability across samples. Homoscedasticity and normality of residuals were assessed through residual histograms and normal probability (P–P) plots. These checks confirmed the suitability of the dataset for parametric modelling, allowing coefficients to be interpreted without transformation.

Multicollinearity was assessed primarily through the VIF, calculated as shown in (1). VIF values close to 1 indicate low collinearity, while values greater than 5 or 10 are generally considered indicative of moderate or severe multicollinearity [8]–[12], [18]–[20]. The formula for the VIF is expressed as (1):

$$VIF(x_i) = \frac{1}{1-R_i^2} \qquad (1)$$

where R_i^2 is the coefficient of determination obtained by regressing x_i on the remaining predictors. Because VIF alone may not fully capture collinearity structures, additional diagnostics were applied, including condition indices and eigenvalue decomposition [18], [19]. A condition index above 10, particularly when combined with high variance proportions for multiple variables, was considered evidence of problematic multicollinearity [19]. This dual approach provided both global and variable-level insights into the correlation structure.

All analyses were performed in IBM SPSS Statistics version 26. SPSS was selected for its established reliability in regression modelling and collinearity diagnostics, as well as its ability to produce reproducible outputs. This combination of standard statistical software, validated diagnostic tools, and transparent procedures ensures that the analysis can be replicated by other researchers.

3. RESULTS AND DISCUSSION

This section presents the outcomes of the regression analysis and multicollinearity diagnostics, followed by a critical interpretation of their significance. The model's performance metrics, including R^2 values and analysis of variance (ANOVA) results, are reported to assess the explanatory power of the selected predictor variables. Individual regression coefficients are examined to determine the relative contribution of each compound to species classification. In addition, multicollinearity is assessed using VIF values and collinearity diagnostics to evaluate potential interdependencies among predictors. The findings are then contextualised within existing literature to highlight their methodological and practical implications for essential oil analysis and species authentication.

3.1. Model fit and summary statistics

The MLR model demonstrated a strong overall fit, as summarized in Table 2. The coefficient of determination (R²) was 0.930, indicating that approximately 93% of the variability in the dependent variable (species classification) could be accounted for by the three predictor compounds: C, D, and E. The adjusted R² value of 0.928 further validated the model's robustness by compensating for the number of predictors included. Additionally, the high correlation coefficient (R=0.964) underscored the strength of the linear relationship between the independent variables and the dependent outcome.

Table 2. Regression model summary

Model summary									
Model	R	R^2	Adjusted R^2	Std. error of the estimate	Change statistics				
					R^2 change	F change	df1	df2	Sig. F change
1	0.964[a]	0.930	0.928	0.300	0.930	515.546	3	116	0.000

a: Predictors: (constant), Compound_E, Compound_C, Compound_D

The standard error of the estimate was recorded at 0.300, reflecting low dispersion between predicted and observed values. This indicates that the model performed with a high degree of accuracy in capturing the variation in species classification. Together, these metrics support the model's adequacy and the relevance of its predictors. The overall statistical significance of the regression model was confirmed by the ANOVA results presented in Table 3. The F-statistic was 515.546 with a p-value <0.001, indicating that the model predicts the outcome variable significantly. The large F-value suggests that the variation explained by the model is substantially greater than the unexplained variance. No violations of underlying regression assumptions were observed at this stage of analysis. The combination of high R², low standard error, and significant F-value supported the reliability and precision of the model, justifying further examination of individual regression coefficients and multicollinearity diagnostics.

Table 3. ANOVA for model significance

ANOVA[a]						
	Model	Sum of squares	df	Mean square	F	Sig.
1	Regression	139.535	3	46.512	515.546	0.000[b]
	Residual	10.465	116	0.090		
	Total	150.000	119			

a: Dependent variable: species

b: Predictors: (constant), Compound_E, Compound_C, Compound_D

3.2. Regression coefficients

The individual contributions of each compound to the model are detailed in Table 4. The intercept (constant) was -0.199, with a standard error of 0.090, resulting in a t-value of -2.213 and a p-value of 0.029, indicating statistical significance. This establishes that the model has a meaningful baseline even in the absence of predictor variables. Compound_C had a positive unstandardized coefficient of 1.080 and a standardized beta of 0.646. The t-value of 21.209 and p-value <0.001 confirmed the significance of this variable in explaining the dependent variable. Its contribution, while strong, was moderate in comparison to the other predictors.

Table 4. Coefficients and VIF values

Coefficients[a]								
Model		Unstandardized coefficients		Standardized coefficients	t	Sig.	Collinearity statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-0.199	0.090		-2.213	0.029		
	Compound_C	1.080	0.051	0.646	21.209	0.000	0.648	1.543
	Compound_D	-0.782	0.024	-1.636	-32.851	0.000	0.242	4.125
	Compound_E	3.049	0.078	2.070	39.168	0.000	0.215	4.645

a: Dependent variable: species

Compound_D had a negative coefficient of -0.782 and a standardised beta of -1.636, with a t-value of -32.851 and a p-value <0.001. These results indicate a strong inverse association with the dependent variable, making it a key driver in the regression model. The magnitude of this coefficient suggested a substantial impact. Compound_E yielded the highest positive unstandardized coefficient of 3.049 and a standardised beta of 2.070. With a t-value of 39.168 and a p-value <0.001, it was identified as the most influential predictor in the model. This suggests that Compound_E plays a dominant role in determining species classification.

The strong effect of γ-eudesmol (Compound_E) is consistent with its role as a dominant oxygenated sesquiterpene in *Aquilaria* oils. This compound is known to contribute to the resin's characteristic fragrance and bioactivity, and it is frequently reported as abundant and chemically stable across multiple species [5], [6]. Its biochemical prominence likely explains its large predictive weight in the model. By contrast, the negative association of 10-epi-γ-eudesmol (Compound_D) may indicate species-specific differences in sesquiterpene biosynthesis, where shifts in enzymatic pathways influence compound ratios more than absolute concentrations.

3.3. Variance inflation factor and collinearity diagnostics

The collinearity diagnostics presented in Table 5 provide important insights into the interrelationships among the predictor variables within the regression model [18], [19]. Multicollinearity arises when independent variables exhibit high correlations with one another, potentially resulting in biased regression coefficients, increased standard errors, and unstable parameter estimates [8], [10]. To assess the presence and severity of multicollinearity, VIF, and tolerance values were utilized [19]. Evaluating these metrics is essential for determining the extent of multicollinearity, thereby supporting the reliability and robustness of the model's parameter estimations [8], [12], [16], [19]. Multicollinearity was assessed using the VIF values presented in Table 6.

Table 5. Collinearity statistics description

Collinearity statistics	Range	Description
VIF	1	Variables are not correlated
	1-5	Variables are moderately correlated
	>5	Variables are highly correlated
Tolerance	<0.1	There is significant multicollinearity
	>0.1	There is no significant multicollinearity

Table 6. Assessment of collinearity diagnostics

Collinearity diagnostics[a]							
Model	Dimension	Eigenvalue	Condition index	Variance proportions			
				(Constant)	Compound_C	Compound_D	Compound_E
1	1	3.323	1.000	0.01	0.01	0.01	0.00
	2	0.491	2.601	0.01	0.27	0.03	0.02
	3	0.158	4.584	0.28	0.19	0.20	0.01
	4	0.028	10.908	0.70	0.53	0.76	0.96

a: Dependent variable: species

Compound_C exhibited a VIF of 1.543, suggesting a low degree of multicollinearity. This indicates that Compound_C is relatively independent from the other predictors and unlikely to introduce instability in the model. Compound_D recorded a VIF value of 4.125, which, while below the critical threshold of 5, indicates moderate multicollinearity. Its shared variance with the other variables may be due to biochemical similarity or co-expression in the essential oil profile. Although not a critical concern, this finding suggests some overlap that could affect coefficient interpretation.

Compound_E had the highest VIF value of 4.645, which is close to the threshold typically used to flag multicollinearity issues. This suggests a higher level of correlation with one or more predictors, warranting caution in interpretation. However, the value remains within an acceptable range for practical applications. To further assess multicollinearity, a collinearity diagnostics table was generated. This included eigenvalues and condition indices, with one dimension showing a condition index of 10.908 and high variance proportions across multiple variables. These patterns confirmed moderate multicollinearity, reinforcing the VIF findings and indicating that model refinement may be warranted in future analyses with expanded variable sets.

The observed collinearity between γ-eudesmol and 10-epi-γ-eudesmol reflects their shared derivation from the farnesyl diphosphate pathway, a common precursor in sesquiterpene biosynthesis [6], [21]. Enzymatic branching within this pathway likely results in correlated production of these compounds, which manifests statistically as moderate multicollinearity. This reinforces the idea that chemometric correlations can mirror underlying metabolic linkages [22]. The normal P-P plot of the regression standardized residuals depicted in Figure 3 provides evidence supporting the appropriateness of the regression models employed. The observed linearity within the plot indicates that the residuals follow a normal distribution, thereby satisfying a key assumption underlying regression analysis.

This finding enhances the credibility and applicability of the models. Overall, the models demonstrate strong potential by effectively integrating the complementary influences of Compounds C, D, and E. The consistently high R² values, together with the confirmation of normally distributed residuals, underscore the robustness and validity of the regression outcomes. For future research, emphasis should be placed on selecting compounds that exhibit minimal multicollinearity while contributing significantly to predictive accuracy, in order to maximize model efficacy [12], [18], [19].

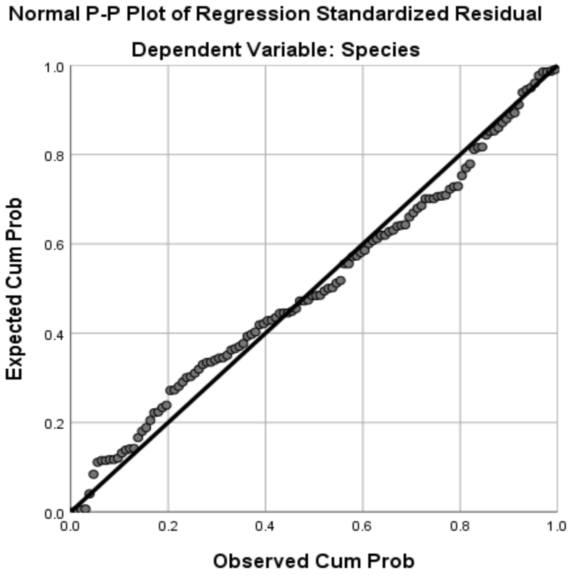

Figure 3. Normal P-P plot illustrating standardized residuals in the regression model for species

3.4. Discussion

Previous studies on *Aquilaria* essential oils have highlighted sesquiterpenes such as γ-eudesmol and δ-guaiene as chemotaxonomic markers [6]. However, most prior work focused on compound identification and classification models without explicitly addressing multicollinearity among biosynthetically related constituents [7], [17], [23]–[26]. Since multicollinearity can distort regression coefficients and weaken interpretability [8], this omission represents a critical gap. The present study addressed it by integrating VIF and condition index diagnostics into regression modelling, providing a safeguard often absent from fingerprinting research.

A regression model based on δ-guaiene, 10-epi-γ-eudesmol, and γ-eudesmol explained 93% of species variation. γ-eudesmol was the strongest positive predictor, δ-guaiene contributed moderately, and 10-epi-γ-eudesmol showed a strong negative effect. Diagnostics revealed moderate interdependence between γ-eudesmol and 10-epi-γ-eudesmol, consistent with their biosynthetic proximity, but within acceptable thresholds, confirming model validity and interpretability.

This work advances fingerprinting research by showing that i) a small, well-chosen set of compounds can match the predictive power of larger untargeted datasets and ii) safeguards such as VIF and condition indices enhance model reliability. Earlier studies often overlooked predictor correlations, risking inflated coefficients and reduced reproducibility [17], [27], [28]. By directly addressing multicollinearity, this study strengthens methodological rigor and interpretive reliability in chemometric applications.

Beyond regression, multicollinearity also influences discriminant analysis (PLS-DA), support vector machines (SVM), and ANN, commonly used in metabolomic and essential oil classification [17], [28], [29]. While effective for large, correlated datasets, these methods may reduce transparency by obscuring predictor importance. In PLS-DA, collinearity inflates latent variable weights [29]; in SVM and ANN, redundancy can increase training complexity and reduce generalizability [17], [28]. By contrast, simple regression models with diagnostics can reveal biosynthetic linkages (e.g., between γ-eudesmol and 10-epi-γ-eudesmol) while preserving interpretability. Table 7 situates this regression–diagnostics framework alongside other modelling approaches.

Table 7. Comparison of modelling approaches for chemical fingerprinting of essential oils

Approach	Strengths	Limitations	Relevance to multicollinearity
MLR with diagnostics (this study)	Transparent coefficients; interpretable; and collinearity checked	Limited for nonlinear relationships; fewer predictors	Directly detects and quantifies collinearity (VIF and condition indices)
PCA [25], [28]	Reduces dimensionality; shows variance patterns	Principal components hard to interpret biologically	Collinearity absorbed in components, obscuring individual contributions
PLS-DA [23], [29]	Handles many predictors; accurate classification	Prone to overfitting; loadings hard to interpret	Collinearity inflates weights, complicates interpretation
SVM [25], [28]	High accuracy; robust to nonlinearities	Black-box; limited interpretability	Correlated predictors reduce transparency, add redundancy
ANN [17]	Captures complex nonlinear interactions	Requires large datasets; low interpretability	Collinearity adds redundancy; needs feature selection/regularization

Importantly, multicollinearity is not merely statistical noise but reflects biosynthetic relationships with ecological meaning. Bridging diagnostics with chemical ecology, this study promotes fingerprinting methods that are both accurate and interpretable. Future work should extend the framework to larger compound sets, broader sampling, and advanced machine learning, while keeping multicollinearity checks central to maintain interpretability.

In conclusion, δ-guaiene, 10-epi-γ-eudesmol, and γ-eudesmol are robust predictors of *Aquilaria* species. Moderate multicollinearity, though present, was effectively managed, illustrating how statistical safeguards can link biochemical insight with methodological rigor. This framework thus supports species authentication, conservation, and quality control.

4. CONCLUSION

This study demonstrated that δ-guaiene, 10-epi-γ-eudesmol, and γ-eudesmol can reliably classify *Aquilaria* species, with regression models explaining 93% of observed variance and multicollinearity diagnostics confirming moderate but manageable interdependence. The strong predictive weight of γ-eudesmol reflects its biochemical prominence as a dominant oxygenated sesquiterpene, while the negative association of 10-epi-γ-eudesmol highlights species-specific variations in sesquiterpene biosynthesis. Beyond identifying key chemotaxonomic markers, the explicit integration of VIFs and condition indices advances chemometric methodology by addressing the often-overlooked issue of multicollinearity, thereby improving interpretability and robustness of models in natural product research. These findings carry practical significance for species authentication, essential oil quality control, and conservation monitoring, while also offering a framework that can be extended to metabolomic studies where correlated variables are the norm. Future research should expand analyses to include additional compounds, geographically diverse datasets, and advanced modelling approaches such as machine learning to enhance predictive power, reveal nonlinear interactions, and strengthen the application of chemometrics in sustainable resource management and pharmaceutical innovation.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support provided by the Ministry of Higher Education Malaysia through the Fundamental Research Grant Scheme (FRGS), under grant number FRGS/1/2023/ICT11/UITM/02/1, which was instrumental in facilitating this research. Sincere appreciation is also extended to the Advanced Signal Processing Research Interest Group (ASPRIG) for their valuable insights and continuous support throughout the study. The authors further wish to thank the BARCE at UMPSA for their technical expertise and crucial role in the data extraction process, which significantly contributed to the success of this work.

FUNDING INFORMATION

The authors gratefully acknowledge the financial support provided by the Ministry of Higher Education Malaysia through the Fundamental Research Grant Scheme (FRGS), under grant number FRGS/1/2023/ICT11/UITM/02/1, which was instrumental in facilitating this research.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Nur Athirah Syafiqah Noramli	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	
Noor Aida Syakira Ahmad Sabri	✓	✓	✓			✓		✓		✓	✓			
Muhammad Ikhsan Roslan	✓	✓	✓	✓			✓			✓				
Nurlaila Ismail	✓			✓						✓	✓	✓		✓
Zakiah Mohd Yussoff	✓			✓			✓	✓		✓				
Mohd Nasir Taib	✓	✓		✓						✓	✓	✓	✓	

C	: **C**onceptualization	I	: **I**nvestigation	Vi	: **Vi**sualization
M	: **M**ethodology	R	: **R**esources	Su	: **Su**pervision
So	: **So**ftware	D	: **D**ata Curation	P	: **P**roject administration
Va	: **Va**lidation	O	: Writing - **O**riginal Draft	Fu	: **Fu**nding acquisition
Fo	: **Fo**rmal analysis	E	: Writing - Review & **E**diting		

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] J. Sharmeen, F. Mahomoodally, G. Zengin, and F. Maggi, “Essential Oils as Natural Sources of Fragrance Compounds for Cosmetics and Cosmeceuticals,” *Molecules*, vol. 26, no. 3, p. 666, Jan. 2021, doi: 10.3390/molecules26030666.

[2] P. Shivanand, N. F. Arbie, S. Krishnamoorthy, and N. Ahmad, “Agarwood—The Fragrant Molecules of a Wounded Tree,” *Molecules*, vol. 27, no. 11, p. 3386, May 2022, doi: 10.3390/molecules27113386.

[3] I. D. Thompson, T. Lim, and M. Turjaman, *A review of the agarwood-producing genera Aquilaria and Gyrinops: CITES considerations, trade patterns, conservation, and management*. 2022. [Online]. Available https://cites.org/sites/default/files/documents/E-CoP19-Inf-12.pdf. (Accessed: Mar. 17, 2025).

[4] Y. V. Thi *et al.*, “An Updated Review of *Aquilaria* Species: Distribution, Chemical Constituents and Authentication Methods,” *Asian Journal of Plant Sciences*, vol. 23, no. 2, pp. 146–167, Mar. 2024, doi: 10.3923/ajps.2024.146.167.

[5] X. Wang, S. W. Chan, N. Singaram, M.-L. Teoh, S. H. Mah, and C. Y. Looi, “Essential oil from *Aquilaria* spp. (agarwood): a comprehensive review on the impact of extraction methods on yield, chemical composition, and biological activities,” *Journal of Essential Oil Research*, vol. 37, no. 1, pp. 110–144, Jan. 2025, doi: 10.1080/10412905.2024.2447706.

[6] N. A. S. Noramli, N. A. S. A. Sabri, M. I. Roslan, N. Ismail, Z. M. Yusoff, and M. N. Taib, “Unraveling the relationships among essential oil compounds in Aquilaria species using GC-MS and GC-FID techniques,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 39, no. 1, pp. 167-177, Jul. 2025, doi: 10.11591/ijeecs.v39.i1.pp167-177.

[7] S. M. H. M. Huzir *et al.*, “Stepwise regression of agarwood oil significant chemical compounds into four quality differentiation,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 29, no. 2, p. 735–741, Feb. 2023, doi: 10.11591/ijeecs.v29.i2.pp735-741.

[8] T. Kyriazos and M. Poga, “Dealing with Multicollinearity in Factor Analysis: The Problem, Detections, and Solutions,” *Open Journal of Statistics*, vol. 13, no. 03, pp. 404–424, 2023, doi: 10.4236/ojs.2023.133020.

[9] H. O. Etaga, R. C. Ndubisi, and N. L. Oluebube, “Effect of Multicollinearity on Variable Selection in Multiple Regression,” *Science Journal of Applied Mathematics and Statistics*, vol. 9, no. 6, p. 141, 2021, doi: 10.11648/j.sjams.20210906.12.

[10] S. Gokmen, R. Dagalp, and S. Kilickaplan, “Multicollinearity in measurement error models,” *Communications in Statistics - Theory and Methods*, vol. 51, no. 2, pp. 474–485, Jan. 2022, doi: 10.1080/03610926.2020.1750654.

[11] J. Y.-L. Chan *et al.*, “Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review,” *Mathematics*, vol. 10, no. 8, p. 1283, Apr. 2022, doi: 10.3390/math10081283.

[12] S. H. Mahmood, “Estimating Models and Evaluating their Efficiency under Multicollinearity in Multiple Linear Regression: A Comparative Study,” *Zanco Journal of Humanity Sciences*, vol. 28, no. 5, pp. 264–277, Oct. 2024, doi: 10.21271/zjhs.28.5.17.

[13] Z. M. Yusoff and N. Ismail, “Datasets of chemical compounds in three different species of *Aquilaria* using GC-MS coupled with GC-FID analysis,” *Data Brief*, vol. 53, p. 110209, Apr. 2024, doi: 10.1016/j.dib.2024.110209.

[14] A. Shuttleworth and S. D. Johnson, “GC-MS/FID/EAD: A method for combining mass spectrometry with gas chromatography-electroantennographic detection,” *Frontiers in Ecology and Evolution*, vol. 10, Dec. 2022, doi: 10.3389/fevo.2022.1042732.

[15] G. Schomburg, “Two-dimensional gas chromatography: Principles, instrumentation, methods,” *Journal of Chromatography A*, vol. 703, no. 1–2, pp. 309–325, Jun. 1995, doi: 10.1016/0021-9673(95)00190-X.

[16] A. Hristozova, M. Batmazyan, K. Simitchiev, S. Tsoneva, V. Kmetov, and E. Rosenberg, “Headspace – Solid phase microextraction vs liquid injection GC-MS analysis of essential oils: Prediction of linear retention indices by multiple linear regression,” *Acta Chromatographica*, vol. 37, no. 1, pp. 76–86, Feb. 2025, doi: 10.1556/1326.2024.01207.

[17] A. A.-J. Safhadi, T. R. Noviandy, I. Irvanizam, R. Suhendra, T. Karma, and R. Idroes, “Backpropagation Neural Network-Based Prediction of Kovats Retention Index for Essential Oil Compounds,” *Infolitika Journal of Data Science*, vol. 2, no. 1, pp. 28–33, May 2024, doi: 10.60084/ijds.v2i1.197.

[18] J. Jacob and R. Varadharajan, “Robust Variance Inflation Factor: A Promising Approach for Collinearity Diagnostics in the Presence of Outliers,” *Sankhya B*, vol. 86, no. 2, pp. 845–871, Nov. 2024, doi: 10.1007/s13571-024-00342-y.

[19] A. U. Ahmad *et al*., “A Study of Multicollinearity Detection and Rectification under Missing Values,” *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 1S, pp. 399–418, 2021, doi: 10.17762/turcomat.v12i1S.1880.

[20] X. Zhang *et al.*, “Sesquiterpenoids in Agarwood: Biosynthesis, Microbial Induction, and Pharmacological Activities,” *Journal of Agricultural and Food Chemistry*, vol. 72, no. 42, pp. 23039–23052, Oct. 2024, doi: 10.1021/acs.jafc.4c06383.

[21] S. K. Niazi and Z. Mariam, “Recent Advances in Machine-Learning-Based Chemoinformatics: A Comprehensive Review,” *International Journal of Molecular Sciences*, vol. 24, no. 14, p. 11488, Jul. 2023, doi: 10.3390/ijms241411488.

[22] D. Yildirim *et al.*, “The efficacy of lavender oil on fatigue and sleep quality in patients with hematological malignancy receiving chemotherapy: a single-blind randomized controlled trial,” *Supportive Care in Cancer*, vol. 33, no. 2, p. 79, Feb. 2025, doi: 10.1007/s00520-024-09143-5.

[23] S.-Z. Qian, Y.-M. Jiang, Q.-L. Yan, D.-H. Wu, W.-X. Zhang, and J.-P. Chung, “Visualization OPLS class models of GC-MS-based metabolomics data for identifying agarwood essential oil extracted by hydro-distillation,” *Scientific Reports*, vol. 15, no. 1, p. 5421, Feb. 2025, doi: 10.1038/s41598-025-85976-2.

[24] Y. Sundaraj, A. Mediani, K. F. Rodrigues, and S. N. Baharum, “GC-MS olfactometry reveals sesquiterpenes α-humulene and δ-cadinene significantly influence the aroma of treated *Aquilaria* malaccensis essential oil,” *Australian Journal Crop Science*, vol. 17, no. 12, pp. 893–901, Dec. 2023, doi: 10.21475/ajcs.23.17.12.p3916.

[25] M. Rasekh, H. Karami, A. D. Wilson, and M. Gancarz, “Classification and Identification of Essential Oils from Herbs and Fruits Based on a MOS Electronic-Nose Technology,” *Chemosensors*, vol. 9, no. 6, p. 142, Jun. 2021, doi: 10.3390/chemosensors9060142.

[26] M. A. Farag, E. M. Kabbash, A. Mediani, S. Döll, T. Esatbeyoglu, and S. M. Afifi, “Comparative Metabolite Fingerprinting of Four Different Cinnamon Species Analyzed via UPLC–MS and GC–MS and Chemometric Tools,” *Molecules*, vol. 27, no. 9, p. 2935, May 2022, doi: 10.3390/molecules27092935.

[27] N. Feizi, F. S. Hashemi-Nasab, F. Golpelichi, N. Saburouh, and H. Parastar, “Recent trends in application of chemometric methods for GC-MS and GC×GC-MS-based metabolomic studies,” *TrAC Trends in Analytical Chemistry*, vol. 138, p. 116239, May 2021, doi: 10.1016/j.trac.2021.116239.

[28] R. Hayati, A. A. Munawar, E. Lukitaningsih, N. Earlia, T. Karma, and R. Idroes, “Combination of PCA with LDA and SVM classifiers: A model for determining the geographical origin of coconut in the coastal plantation, Aceh Province, Indonesia,” *Case Studies in Chemical and Environmental Engineering*, vol. 9, Jun. 2024, doi: 10.1016/j.cscee.2023.100552.

[29] M. A. Farag, S. M. Ezzat, M. M. Salama, and M. G. Tadros, “Anti-acetylcholinesterase potential and metabolome classification of 4 Ocimum species as determined via UPLC/qTOF/MS and chemometric tools,” *Journal of Pharmaceutical and Biomedical Analysis*, vol. 125, pp. 292–302, Jun. 2016, doi: 10.1016/j.jpba.2016.03.037.

BIOGRAPHIES OF AUTHORS

Nur Athirah Syafiqah Noramli received her B.Sc. (Hons) in computer science from Universiti Teknologi MARA (UiTM) Cawangan Melaka Kampus Jasin. She is currently pursuing her studies as a postgraduate student at the Faculty of Electrical Engineering, at Universiti Teknologi MARA (UiTM) Shah Alam, Selangor, Malaysia. Her research interests include advanced signal processing, machine learning, and deep learning. She can be contacted at email: athirah.noramli1@gmail.com.

Noor Aida Syakira Ahmad Sabri received her Bachelor of Engineering (Hons) in electronic engineering from Universiti Teknologi MARA (UiTM), Shah Alam, Malaysia, in 2022. Currently, she is pursuing postgraduate studies at the Faculty of Electrical Engineering, Universiti Teknologi MARA (UiTM), Shah Alam, Malaysia. Her research interests focus on advanced signal processing and machine learning. She can be contacted at email: aidasyakiraaa01@gmail.com.

Muhammad Ikhsan Roslan earned his Master of Science in Electronic Systems Design Engineering from Universiti Sains Malaysia (USM), Penang, Malaysia, in 2022 with first-class honors. He is currently a server validation engineer specializing in IP-level validation at UST (M) Sdn. Bhd, while also pursuing full-time postgraduate studies at the Faculty of Electrical Engineering, Universiti Teknologi MARA (UiTM), Shah Alam, Malaysia. With a strong passion for research in engineering, particularly in artificial intelligence, he combines academic excellence with practical experience, showcasing a dedicated commitment to advancing the field. He can be contacted at email: muhammadikhsanroslan@gmail.com.

Assoc. Prof. Ir. Ts. Dr. Nurlaila Ismail received her Ph.D. in electrical engineering from Universiti Teknologi MARA, Malaysia. She is currently a senior lecturer at Faculty of Electrical Engineering, Universiti Teknologi MARA Shah Alam, Malaysia. Her research interests include advanced signal processing and artificial intelligence. She can be contacted at email: nurlaila0583@uitm.edu.my.

Assoc. Prof. Ts. Dr. Zakiah Mohd Yusoff received her Bachelor's Degree in electrical engineering and Ph.D. in electrical engineering from Universiti Teknologi MARA Shah Alam, in 2009 and 2014, respectively. She is a senior lecturer who is currently working at Faculty of Electrical Engineering, Universiti Teknologi MARA (UiTM) Shah Alam, Malaysia. In Mei 2014, she joined Universiti Teknologi MARA as a teaching staff. Her major interests include process control, system identification, and essential oil extraction systems. She can be contacted at email: zakiah9018@uitm.edu.my.

Prof. Ir. Ts. Dr. Haji Mohd Nasir Taib received the degree in electrical engineering from the University of Tasmania, Hobart, Australia, the M.Sc. degree in control engineering from Sheffield University, UK, and the Ph.D. degree in instrumentation from the University of Manchester Institute of Science and Technology, UK. He is currently an Honorary Professor at Universiti Teknologi MARA (UiTM), Malaysia. He heads the Advanced Signal Processing Research Interest Group, Faculty of Electrical Engineering, UiTM. He has been a very active researcher and over the years had author and/or co-author many papers published in refereed journals and conferences. He can be contacted at email: dr.nasir@uitm.edu.my.