BAREKENG: Journal of Mathematics and Its Applications
March 2025 Volume 19 Issue 1 Page 0675–0686
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol19iss1pp0675-0686

IMPLEMENTATION OF FEATURE IMPORTANCE XGBOOST ALGORITHM TO DETERMINE THE ACTIVE COMPOUNDS OF SEMBUNG LEAVES (BLUMEA BALSAMIFERA)

Kusnaeni [1*], Nurul Fuady Adhalia H[2], Abdul Khaliq Zulfattah[3]

[1,2]*Department of Mathematics, Institut Teknologi Bacharuddin Jusuf Habibie*
[3]*Department of Information System, Institut Teknologi Bacharuddin Jusuf Habibie*
Jln. Balaikota No.1, Bumi Harapan, Parepare, 91122, Indonesia

*Corresponding author's e-mail: * **kusnaeni25@ith.ac.id***

Article History:

Received: 29th August 2024
Revised: 16th November 2024
Accepted: 16th November 2024
Published: 13th January 2025

Keywords:

Sembung Leaves;
Feature Importance;
XGBoost;
Active Compound.

ABSTRACT

Sembung is a medicinal plant native to Indonesia that grows optimally in tropical climates. The secondary metabolite compounds found in the leaves of sembung are biopharmaceutical active ingredients. Fourier Transform Infrared (FTIR) spectroscopy can identify the functional compounds in sembung leaves by analyzing unique peaks in the spectrum, which correspond to specific functional groups of the compounds. In this research, 35 observations were made with 1,866 explanatory variables (wavelengths). Data in which the number of explanatory variables surpasses the number of observations is known as high-dimensional data. One method that can handle high-dimensional problems is to select important variables that affect the objective variable. The XGBoost algorithm can calculate the feature importance score that affects the goal variable so that it does not have to include all variables in the modeling, this can overcome problems in high-dimensional data. The results of the calculation of feature importance found Lignin Skeletal Band, CH, and CH2 aliphatic Stretching Group, C=C, C=N, C–H in ring structure, DNA and RNA backbones, NH2 Aminoacidic Group, C=O Ester Fatty Acid that the active compounds contained in the leaves of sembung.

How to cite this article:

Kusnaeni, N. F. Adhalia and A. K. Zulfattah., "IMPLEMENTATION OF FEATURE IMPORTANCE XGBOOST ALGORITHM TO DETERMINE THE ACTIVE COMPOUNDS OF SEMBUNG LEAVES (BLUMEA BALSAMIFERA)," *BAREKENG: J. Math. & App.*, vol. 19, iss. 1, pp. 0675-0686, March, 2025.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

A common problem in the big data era is the large-scale data problem with more variables than observations [1]. In cases where multicollinearity is present within the data, there is a notable increase in the variability of the estimates for regression coefficients. This phenomenon occurs because multicollinearity introduces high correlations between predictor variables, which can lead to instability and imprecision in the estimation process. As a result, the regression coefficients may exhibit considerable fluctuations, making it challenging to accurately determine their true values and assess their contributions to the model [2]. One commonly employed method to address the challenges posed by high-dimensional problems involves the selection of variables. This process entails identifying and selecting those variables that are statistically significant and have a notable impact on the objective variable. In other words, it involves searching for and isolating the key variables that exert a meaningful influence on the outcome of interest. By focusing on these important variables, the complexity of the model can be reduced, thereby enhancing the efficiency and effectiveness of the analysis while improving the interpretability of the results.

Sembung is a medicinal plant native to Indonesia, which flourishes in tropical climates. This plant is highly regarded in traditional medicine, where its various parts are utilized for their therapeutic properties. The leaves, stems, and roots of the sembung plant are particularly esteemed, being frequently used in the preparation of remedies aimed at addressing a variety of health concerns. The sembung plant is renowned for its numerous medicinal benefits, which extend to the treatment and relief of a wide range of health conditions. Among its many uses, the plant has been traditionally employed to alleviate symptoms associated with rheumatism, the common flu, mouth ulcers, and fevers [3]. The secondary metabolite compounds found in the leaves of sembung are biopharmaceutical active ingredients. Secondary metabolites consist of flavonoids, alkaloids, tannins, terpenoids, and saponins. Flavonoids have been shown to inhibit the growth of cancer cells in humans [4]. The FTIR spectroscopy was used to detect the active compounds present in sembung leaf metabolites. This tool is employed to identify the functional groups of secondary metabolites by examining the peaks and the spectrum's shape. The identification process involves observing the peaks that signify the functional group type in a compound [5]. In this research, the data from the FTIR spectrometer results in 35 observations and 1866 variables or high dimensional data.

Numerous studies have been undertaken to identify the compounds present in sembung leaves. Notably, in 2022, Kusnaeni conducted research employing both Sparse Group Lasso and Overlapping Group Lasso techniques to identify the active compounds within these leaves. This study revealed that several significant compounds, including SiO2, polyphenols, the CN Amide II band, and C=C unsaturated compounds, are present and active in sembung leaves [6]. In 2022, Cahya conducted a detailed study that identified several active compounds present in sembung plants through the application of advanced analytical techniques, specifically LAD-LASSO and WLAD-LASSO. The research findings revealed that the compounds Umbelliferone, 12-Hydroxyjasmonic Acid I, C22H14N8O2, and Acetyleugeno are notably present in sembung plants. These compounds were identified as active constituents, contributing to the plant's medicinal properties [7]. In 2023, Rochyati used Robust Lasso to find significant Importance variables for the antioxidant content of sembung leaves, namely Umberlliferone (7-hydroxycoumarin)/C9H6O3, Quercetin/C15H10O7, Austroinulin/C20H34O3, Gurjunene/C15H24, C5H11NO2, C7H7NO2, C8H16N2O9, C12H14O [8]. Previous research used the Selection Variable method in determining the Importance Variable, but no one has used the classification method in Machine Learning to find the Importance Variable for the identification of the active compounds of Sembung leaves. Many studies have found that the Machine Learning classification method is more accurate in determining Variable Importance than the Selection Variable method, as shown by Hassan in 2023, who found that the Machine Learning classification method has better performance than the Selection Variable method in terms of accuracy in the case of selecting the Importance Variable in breast cancer diagnosis [9]. In 2020, Embark showed that the Machine Learning classification method performs better than the Selection Variable method in terms of accuracy and AUC in the case of selecting significant variables on credit scoring [10]. In 2019, Zhu also showed that the Machine Learning classification method performs better than the Selection Variable method in terms of accuracy and recall in the case of predicting customer churn [11]. In addition, there have been many studies using Machine Learning classification methods in drug development, and health such as Mirka's study on comparing methods for determining which features are most important for explaining classification models in 2021 [12], Jiaju's research on developing a model for predicting and screening products using a combination of fusion regression and XGBoost classification in 2022 [13], Trevor's study comparing methodologies for selecting features and learning algorithms to develop an estimator for telomere length

based on DNA methylation in 2023 [14], Davide's research on practical recommendations for applying gradient boosting techniques to predict molecular properties in 2023 [15], and Rooshree's study on using machine learning methods to map soil suitability for medicinal plants in 2024 [16].

The XGBoost algorithm is capable of processing large-scale data, which is commonly encountered in the screening of active compounds derived from medicinal plants [17]. Moreover, the XGBoost algorithm can effectively manage a wide variety of feature types, including those that are complex and diverse. This Machine Learning algorithm demonstrates exceptional performance in both classification and regression tasks, offering a high level of accuracy and strong generalization capabilities [18]. The XGBoost algorithm, with its ability to handle high-dimensional and complex data, is a very suitable tool for analyzing FTIR spectra of sembung leaves. Through the feature importance analysis provided by the XGBoost algorithm, this study aims to identify key bioactive compounds that contribute to the medicinal properties of sembung. By revealing the molecular mechanisms behind the therapeutic effects of sembung, this research is expected to contribute to the development of new drugs significantly. In addition, the findings of this study can also pave the way for the more optimal utilization of sembung as a potential source of bioactive compounds in the pharmaceutical field.

2. RESEARCH METHODS

This research aims to identify functional groups that significantly contribute to antioxidant activity using variable importance from the XGBoost classification machine learning algorithm. The procedures to be implemented include: (1) standardizing the data, (2) balancing the data, (3) determining variable importance using the XGBoost classification algorithm, (4) identifying functional groups significant to antioxidant activity based on the variable importance values generated by XGBoost algorithm, and (5) evaluating the effectiveness of the XGBoost variable importance algorithm by calculating the Mean Absolute Error (MAE). The choice of MAE as the evaluation metric was made because it provides an intuitive, straightforward measure of average prediction error in the same units as the target variable, which enhances interpretability for antioxidant activity prediction. Unlike Mean Squared Error (MSE), which magnifies the impact of larger errors due to squaring, MAE treats all errors equally, reducing sensitivity to outliers in the data. This is particularly beneficial for our analysis, where outlier influence on antioxidant activity can vary due to the diverse chemical composition of functional groups. MAE allows a direct, linear assessment of prediction accuracy, aligning with our focus on minimizing average deviation in variable importance evaluation.

2.1 Data

The data used in this study is data from the results of bioactivity tests and spectrometry and chromatography experiments conducted at the Central Laboratory of Biopharmaceutical Studies, IPB, from May to June 2021. The antioxidant content of sembung leaves was examined using data obtained from FTIR spectroscopy. The dataset comprised 35 observations involving extracts of sembung leaves with five different solvents: water, 30% ethanol, 50% ethanol, 70% ethanol, and pure ethanol. Each extraction type was repeated seven times. The explanatory variable (X) in this study is the absorbance at a wavelength of 1866 for each observation. The response variable (Y) indicates the antioxidant content of the sembung leaves for each observation. All observation, including wavelength, absorbance, and Antioxidant content, can be seen in Table 1.

Table 1. Data Structure

Observation	Wavelength	Absorbance	Antioxidant content level
1st Water Measurement	399.2374	0.209643	
1st Water Measurement	401.1661	0.2243725	372.34
⋮	⋮	⋮	
1st Water Measurement	3996.231	0.06065752	
⋮	⋮	⋮	⋮
7st Ethanol 70% Measurement	399.2374	0.209643	
7st Ethanol 70% Measurement	401.1661	0.2243725	68.40
⋮	⋮	⋮	

Observation	Wavelength	Absorbance	Antioxidant content level
7st Ethanol 70% Measurement	3996.231	0.08768222	
1st Pure Ethanol Measurement	399.2374	0.0727334	54.61
1st Pure Ethanol Measurement	401.1661	0.09190677	
⋮	⋮	⋮	
1st Pure Ethanol Measurement	3996.231	0.01976719	

***Data source**: The Central Laboratory of Biopharmaceutical Studies, IPB University.*

2.2 Balancing Data

Numerous real-world fields, such as financial management, medical diagnosis, and telecommunications, have encountered challenges related to imbalanced data [19]. Class imbalance occurs when the majority class significantly outnumbers the minority class. For example, this could be illustrated by a data ratio of 1:100, where 1 represents the minority class and 100 represents the majority class [20]. Imbalanced datasets can have a detrimental effect on the performance of machine learning algorithms. Different resampling methods have been created recently to tackle class imbalance. When datasets have a large number of features, they can often be improved by reducing dimensionality, which in turn decreases the number of features [21]. Imbalanced datasets can adversely affect the performance of machine learning algorithms. To tackle class imbalance, a variety of resampling methods have been introduced recently. Additionally, datasets with a large number of features often benefit from dimensionality reduction, which helps to decrease the number of features [22].

2.3 XGBoost Algorithm

The XGBoost algorithm, introduced in 2015, rapidly became one of the most popular algorithms in machine learning. Its primary functions are for regression (predicting continuous values) and classification/segmentation tasks [23]. The XGBoost algorithm is made up of several base learners, which are regression trees. When training, it employs a forward stepwise algorithm to gradually enhance the performance of each base learner results.

The objective function in the XGBoost algorithm is crucial for optimizing the model's performance. It is expressed mathematically as:

$$Obj^{(t)} = \sum_{i=1}^{N} L\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{j=1}^{t} \Omega(f_j) \quad (1)$$

Information:

$Obj^{(t)}$: This objective function in the XGBoost algorithm

$\hat{y}_i^{(t)}$: This represents the predicted value of the i^{th} sample at iteration t. It indicates the model's prediction after adding t trees.

f_j : This denotes the j^{th} regression tree added to the model. Each tree contributes to refining the model's predictions.

L : This is the loss function that measures the difference between the actual value y_i and the predicted value $\hat{y}_i^{(t)}$ for the i^{th} sample. The choice of loss function can vary based on the specific task (e.g., regression or classification).

Ω : This term represents the regularization function for the j^{th} tree, which penalizes the complexity of the model to prevent overfitting. It typically includes parameters that control the size and structure of the trees.

The boosting technique involves sequentially adding new models to fix the errors made by previous models, and this process continues until no additional improvement is possible. This process continues until no further improvement can be achieved. The classification and regression trees (CART) that make up the tree ensemble model are used in the ensemble technique. Since a single CART typically lacks significant predictive power, the ensemble approach is employed. XGBoost uses a tree ensemble model, such as a CART set, to make predictions by considering multiple trees. In XGBoost, the new model forecasts the residuals from the previous model and then combines these predictions to produce the final outcome. The XGBoost

algorithm starts with an initial leaf that contains the probability values of the predicted attributes. The steps to work on the gradient Boosting Regression Algorithm:

Gradient Boosting Regression Algorithm

a. Input:

Data $\{(x_i - y_i)\}_{i=1}^{n}$ and a differentiable loss function $L\big(y_i, F(x)\big)$

Data: A dataset consisting of n observations, where each observation is represented as a pair (x_i, y_i) for $i = 1,2,3, \dots, n$

x_i : Features or input variables.

y_i : Actual target values or responses.

Loss Function: A differentiable loss function $L\big(y_i, F(x)\big)$ that measures the difference between the actual target values y_i and the predictions made by the model $F(x)$.

b. First Step

Initialize model with a constant value is:

$$F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma^2) \tag{2}$$

Information:

F_0: This represents the initial prediction function. It is a constant value at the start of the algorithm.

$\sum_{i=1}^{n} L(y_i, \gamma^2)$: This is the total loss computed over all n data points. Here, L is the loss function that measures how well the constant prediction γ matches the actual values y_i.

c. Second Step: Iterative Boosting

For each iteration $m = 1$ to M, where M is the total number of boosting iterations:

i. Compute Pseudo-Residuals:

$$r_{im} = -\left[\frac{\partial L\big(y_i, F(x_i)\big)}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \tag{3}$$

for $i = 1 \dots n$

Here, $r_{i,m}$ represents the direction and magnitude of correction needed for the current predictions.

ii. Fit a regression tree to the computed pseudo-residuals r_{im}. This tree will identify terminal regions R_{jm} for $j = 1 \dots . Jm$, where Jm is the number of terminal nodes in the tree.

iii. Compute Optimal Leaf Values:

For each terminal regionR_{jm}, compute the optimal valueγ_{jm} that minimizes the loss:

$$\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{ij}}^{n} L(y_i, F_{m-1}(x_i) + \gamma) \tag{4}$$

This step determines how much to adjust the predictions for observations in each terminal region.

iv. Update the Model:

Update the predictions by adding the contributions from the new regression tree:

$$F_m(x) = F_{m-1}(x) + \gamma \sum_{j=1}^{Jm} \gamma_{jm} I\left(x \in R_{jm}\right) \quad (5)$$

Here, $I\left(x \in R_{jm}\right)$ is an indicator function that equals 1 if x belongs to the terminal region R_{jm} and 0 otherwise.

d. Output:

The final model after M iterations is given by: $F_m(x)$.

This represents the combined predictions from all the weak learners added during the boosting process.

2.4 Mean Absolute Error (MAE)

One measure of calculating popular model error is MAE [24]. Mean Absolute Error (MAE) is commonly employed to assess the accuracy of recommender systems and is given by:

$$MAE = \frac{\sum_{n=1}^{N} |\widehat{s_n} - s_n|}{N} \quad (6)$$

Where $\widehat{s_n}$ represents the predicted rating, s_n denotes the actual rating in the test dataset, and N is the total number of rating prediction pairs between the test data and the predicted results [25].

3. RESULTS AND DISCUSSION

3.1 Data Standardization

Based on the FTIR spectroscopy data consisting of 1866 predictor variables and 35 response variables, Figure 1 displays the absorbance spectra for a number of samples from both the "water" and "ethanol" groups, measured across the wavenumber range of 500 to 4000 cm^{-1}. The visualization results indicate significant variations in absorbance intensity at certain wavenumber ranges, particularly between 1000-2000 cm^{-1} and 2800-3600 cm^{-1}.

The "water" group, represented by the darker lines, tends to exhibit higher and more varied absorbance patterns compared to the "ethanol" group, which shows a more uniform absorbance pattern throughout the spectrum. These differences suggest variations in chemical composition or concentration of substances within each sample. This data provides an initial overview of the heterogeneity in the absorbance spectra of each group before the data standardization process is applied.

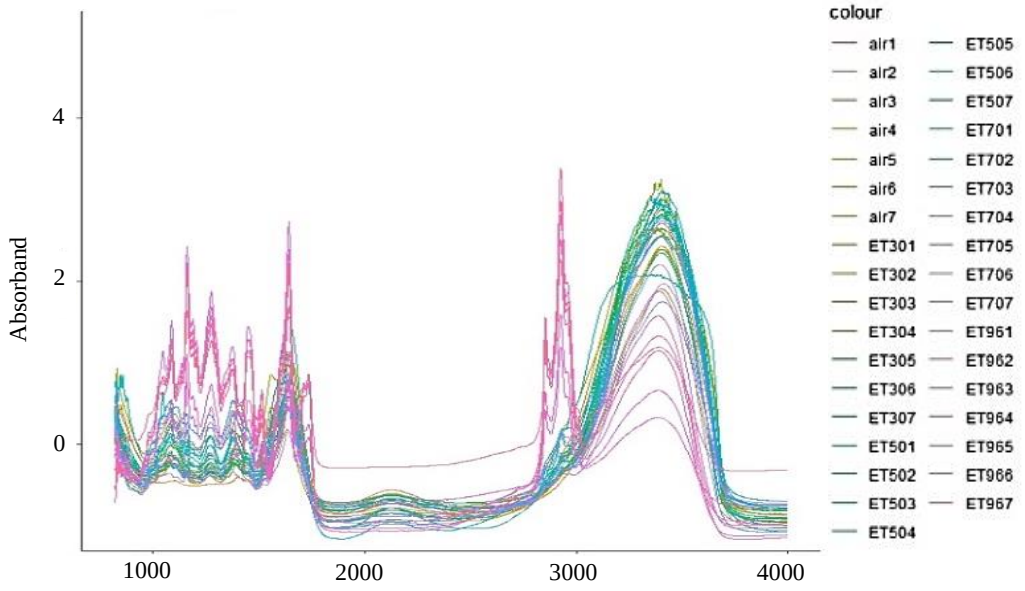

Figure 1. FTIR Data Plot Before Standardization Process

Figure 2 shows the absorbance spectra after data standardization for samples from the "water" and "ethanol" groups within the wavenumber range of 500 to 4000 cm^{-1}. Following the standardization process, it is evident that variability between the spectra has been significantly reduced, particularly in the wavenumber ranges of 1000-2000 cm^{-1} and 2800-3600 cm^{-1}.

The spectra of each group now appear more uniform, indicating that differences caused by external factors or undesired variations have been minimized. This suggests that the standardization successfully normalized the data, allowing for more accurate comparisons between the spectra of different samples. This standardization is crucial to ensure that subsequent analyses focus on differences genuinely caused by variations in the sample composition rather than technical or instrumental discrepancies. These results provide a solid foundation for more in-depth analysis of the normalized spectral data.

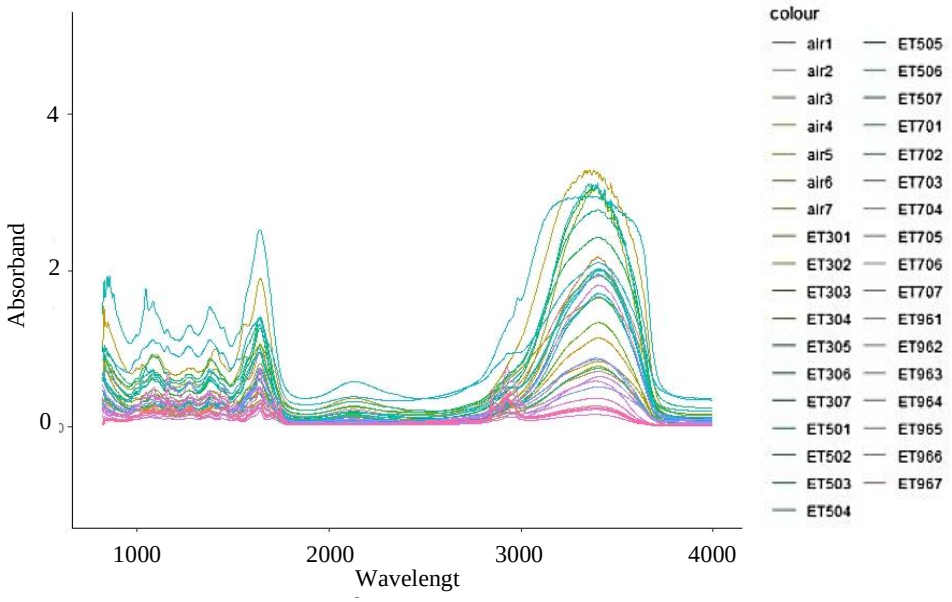

Figure 2. FTIR Data Plot After Standardization Process

3.2 Data Balancing

A preliminary examination of the class distribution highlighted a substantial imbalance. **Figure 3** demonstrates that the majority class (class 1) comprised 70% of the dataset, whereas the minority class (class 0) constituted only 30%. Such a class imbalance poses a risk of bias in the generated model, as it may be inclined to favor the majority class. To mitigate this issue, random oversampling was selected as the data balancing strategy. By randomly replicating samples from the minority class, the technique aims to achieve a balanced class distribution, thereby enhancing the model's ability to accurately classify both classes.

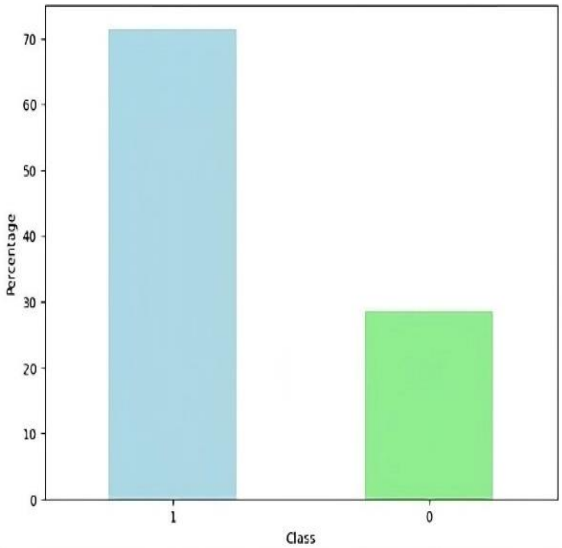

Figure 3. Percentage of Number Classes Before Balancing

Figure 4 depicts the class distribution following data balancing using the random oversampling technique. It can be observed that the percentage of class 0 and class 1 is now balanced at 50% each. Prior to balancing, there was a significant imbalance with the majority class dominating. By randomly duplicating samples from the minority class, the random oversampling technique achieved parity between the two classes. This indicates that the class imbalance issue has been successfully addressed, leading to the expectation of improved and unbiased classification model performance.

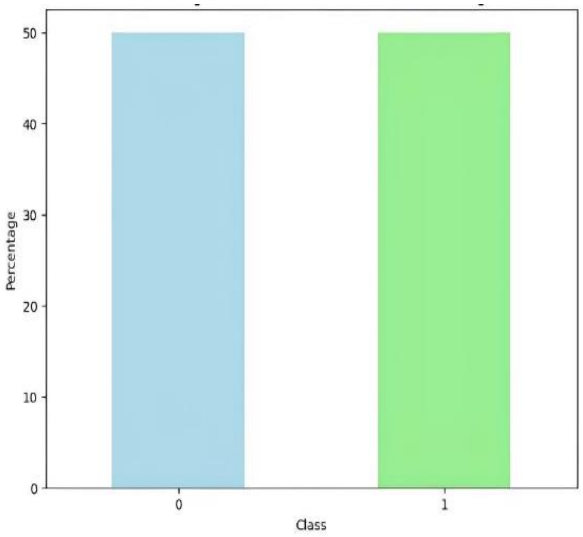

Figure 4. Percentage of Number Classes After Balancing

3.3 Result of Feature Importance

The boosted trees' construction in gradient boosting offers the advantage of easily obtaining importance scores for each feature. Generally, these scores show the usefulness or importance of each feature in constructing the boosted decision trees within the model [26]. Features that are used more frequently to make decisions in the trees are assigned higher importance scores. These scores are calculated for each feature in the dataset and can be compared to evaluate their relative significance. The importance of a single decision tree is determined by how much each attribute improves the performance measure at each partition index and is based on the number of observations present at each node. A measure of performance is the white (Gini index) or other specific error function used to select the segmentation point. The importance of the feature is then produced in all decision trees of the model. The results of feature importance calculation using the XGBoost machine learning algorithm can be seen in **Figure 5**.

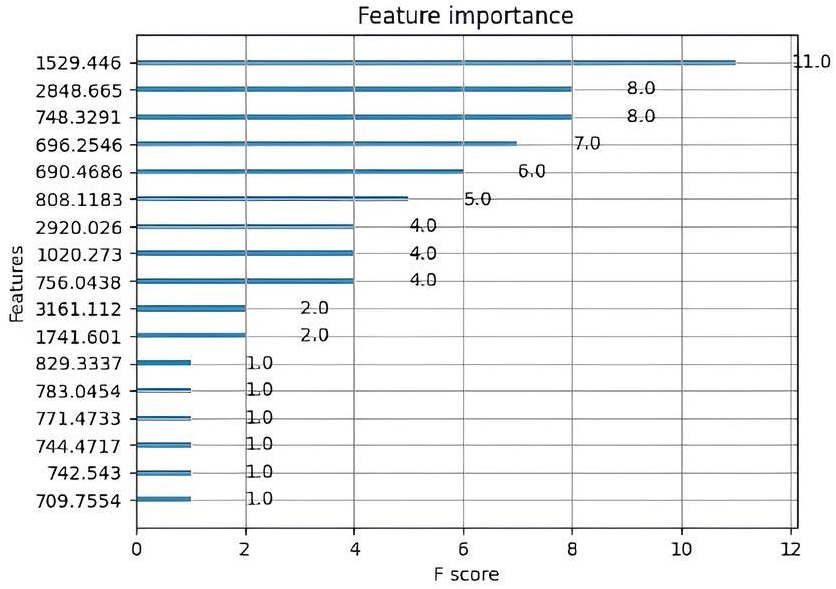

Figure 5. Wavelength Result of Feature Importance

The following are the results of the feature selection using the XGBoost algorithm Importance score feature. Some wavelengths persist after the calculation of the feature importance score: 1529.446, 2848.665, 748.3291, 696.2546, 690.4686, 808.1183, 2920.026, 1020.273, 756.0438, 3161.112, 1741.601, 829.3337, 783.0454, 771.4733, 744.4717, 742.543, 709.7554. The wavelength is significant to the objective variable, in this case, the antioxidant level of constipated leaves. based on [27] The wavelength of the feature important score selection results is defined in the functional groups as listed in **Table 2**.

Table 2. Variable Selection Results of Feature Importance

Wavenumber	Functional Group
1529.446	Lignin skeletal band
2848.665	CH and CH2 aliphatic Stretching Group
808.1183	C=C, C=N, C–H in ring structure, DNA and RNA backbones
2920.026	CH and CH2 aliphatic Stretching Group
3161.112	NH2 Aminoacidic Group
1741.601	C=O Ester Fatty acid Group
829.3337	C=C, C=N, C–H in ring structure, DNA and RNA backbones

The MAE value for the XGBoost algorithm in estimating the antioxidant levels of sembung leaves, based on various significant functional groups such as Lignin Skeletal Band, CH and CH2 aliphatic stretching groups, C=C, C=N, C–H in ring structures, DNA and RNA backbones, NH2 Aminoacidic Group, and C=O Ester Fatty Acid Group, was 0.22. The Skeletal Band Lignin has potential antioxidant properties. C=C, C=N, C–H in ring structures, and DNA and RNA backbones are functional groups relevant to antioxidants, while NH2 Aminoacidic Group is known for its antioxidant properties. Additionally, the C=O Ester Fatty Acid Group is also recognized as significant for antioxidants.

3.4 Mean Absolute Error (MAE) Calculation

The Mean Absolute Error (MAE) metric was employed to evaluate the model's performance. MAE computes the average absolute deviation between the model's predicted values and the actual values. The formula used for calculating MAE is in **Equation (6)**.

Table 3. Mean Absolute Error (MAE) Calculation

No	$\widehat{s_n}$	s_n	$\lvert\widehat{s_n} - s_n\rvert$
1	1	1	0
2	1	1	0
3	1	1	0
4	1	0	1
5	1	1	0
6	0	1	1
7	0	0	0
8	0	0	0
9	1	1	0

$$MAE = \frac{\sum_{n=1}^{N} |\widehat{s_n} - s_n|}{N}$$

$$MAE = \frac{0+0+0+1+0+1+0+0+0}{9}$$

$$MAE = \frac{2}{9}$$

$$MAE = 0.22$$

4. CONCLUSIONS

The results of variable selection using the XGBoost machine learning classification algorithm identified key active compounds significantly related to the antioxidant properties of sembung leaves, including the Lignin Skeletal Band, CH and CH2 aliphatic stretching groups, C=C, C=N, C–H in ring structures, DNA and RNA backbones, NH2 amino acid groups, and C=O ester fatty acids. These compounds contribute crucially to the antioxidant efficacy of sembung leaves. The antioxidant levels were effectively

estimated using the XGBoost algorithm, achieving a Mean Absolute Error (MAE) value of 0.22 based on the feature importance scores of these functional groups. The XGBoost algorithm proved valuable in identifying and estimating antioxidant compounds in sembung leaves, with relatively low error rates in predictions, meeting the study objectives of enhancing compound identification through feature importance in machine learning. However, this study faced limitations, particularly in potential variations in compound properties not fully captured in the FTIR spectral data, which may have affected model accuracy. For future research, expanding the dataset to include additional plant samples and environmental conditions could provide a broader understanding of the antioxidant composition. Further studies could also explore other machine learning techniques or ensemble methods to compare efficacy and improve predictive precision in phytochemical studies.

ACKNOWLEDGMENT

This research is supported by the Ministry of Education, Culture, Research, and Technology of Indonesia. The responsibility for the content of this research is entirely my own.

REFERENCES

[1] S. Suboh, I. Abdul, S. Milleana, and S. Akmar, "A Systematic Review of Anomaly Detection within High Dimensional and Multivariate Data," *JOIV Int. J. Inform. Vis.*, vol. 7, no. March, 2023.

[2] J. I. Daoud, "Multicollinearity and Regression Analysis," *J. Phys. Conf. Ser.*, vol. 949, 2017, doi: doi :10.1088/1742-6596/949/1/012009.

[3] S. Wahjuni, I. Bagus, P. Manuaba, and N. M. Puspawati, "Peningkatan Kesejahteraan Masyarakat Dimasa Pandemi Covid 19 dengan Pelatihn Pengemasan Produk Loloh Daun Sembung (Blumea Balsamifera) di Banjar Dinas Apit Yeh Kaja, Desa Manggis Kabupaten Karangasem," *J. Pengabdi. Kpd. Masy. Fak. Ekon. dan Bisnis UNMAS Denpasar*, vol. 1, no. 3, pp. 230–236, 2021.

[4] W. Wardah and E. S. Kuncari, "Kajian Etnobotani Pakundalang (Blumea balsamifera (L.) DC.) sebagai Solusi Alternatif untuk Kemandirian Kesehatan Masyarakat Banggai Kepulauan, Sulawesi Tengah," *J. Trop. Ethnobiol.*, vol. III, no. 2, 2020, doi: https://doi.org/10.46359/jte.v3i2.51.

[5] R. A. Pratiwi and A. B. D. Nandiyanto, "How to Read and Interpret UV-VIS Spectrophotometric Results in Determining the Structure of Chemical Compounds," *Indones. J. Educ. Res. Technol.*, vol. 2, no. 1, pp. 1–20, 2022.

[6] K. Kusnaeni, A. M. Soleh, F. M. Afendi, and B. Sartono, "Function Group Selection of Sembung Leaves (Blumea Balsamifera) Significant To Antioxidants Using Overlapping Group Lasso," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 16, no. 2, pp. 721–728, 2022, doi: 10.30598/barekengvol16iss2pp721-728.

[7] S. D. Cahya, B. Sartono, I. Indahwati, and E. Purnaningrum, "Performance of LAD-LASSO and WLAD-LASSO on High Dimensional Regression in Handling Data Containing Outliers," *JTAM (Jurnal Teor. dan Apl. Mat.*, vol. 6, no. 4, p. 844, 2022, doi: 10.31764/jtam.v6i4.8968.

[8] R. Rochayati, K. Sadik, B. Sartono, and E. Purnaningrum, "Study on the performance of Robust LASSO in determining important variables data with outliers," *J. Nat.*, vol. 23, no. 1, pp. 9–15, 2023, doi: 10.24815/jn.v23i1.26279.

[9] M. M. Hassan *et al.*, "A comparative assessment of machine learning algorithms with the Least Absolute Shrinkage and Selection Operator for breast cancer detection and prediction," *Decis. Anal. J.*, vol. 7, p. 100245, 2023, doi: 10.1016/j.dajour.2023.100245.

[10] A. Embark, R. Y. Haggag, S. Aboul, and F. Saleh, "A Framework for Feature Selection Using XGBoost for Prediction Banking Risk," 2020.

[11] Q. Zhu, X. Yu, Y. Zhao, and D. Li, "Customer churn prediction based on LASSO and Random Forest models," in *IOP Conference Series: Materials Science and Engineering*, Nov. 2019, vol. 631, no. 5. doi: 10.1088/1757-899X/631/5/052008.

[12] M. Saarela and S. Jauhiainen, "Comparison of feature importance measures as explanations for classification models," *SN Appl. Sci.*, vol. 3, no. 2, pp. 1–12, 2021, doi: 10.1007/s42452-021-04148-9.

[13] J. Wu *et al.*, "Prediction and Screening Model for Products Based on Fusion Regression and XGBoost Classification," *Comput. Intell. Neurosci.*, pp. 1–14, 2022, doi: 10.1155/2022/4987639.

[14] T. Doherty *et al.*, "A comparison of feature selection methodologies and learning algorithms in the development of a DNA methylation-based telomere length estimator," *BMC Bioinformatics*, vol. 24, no. 178, pp. 1–30, 2023, doi: 10.1186/s12859-023-05282-4.

[15] D. Boldini, F. Grisoni, D. Kuhn, L. Friedrich, and S. A. Sieber, "Practical guidelines for the use of gradient boosting for molecular property prediction," *J. Cheminform.*, vol. 15, no. 73, pp. 1–13, 2023, doi: 10.1186/s13321-023-00743-7.

[16] S. Roopashree, J. Anitha, S. Challa, T. R. Mahesh, V. K. Venkatesan, and S. Guluwadi, "Mapping of soil suitability for medicinal plants using machine learning methods," *Sci. Rep.*, pp. 1–17, 2024, doi: 10.1038/s41598-024-54465-3.

[17] Y. Chen and J. Kirchmair, "Cheminformatics in Natural Product-based Drug Discovery," *Mol. Inform.*, vol. 39, no. 12, pp. e2000171–e2000171, Dec. 2020, doi: 10.1002/minf.202000171.

[18] X. Y. Liew, N. Hameed, and J. Clos, "An investigation of XGBoost-based algorithm for breast cancer classification," *Mach. Learn. with Appl.*, vol. 6, no. April, p. 100154, 2021, doi: 10.1016/j.mlwa.2021.100154.

[19] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, and H. Yuanyue, "Learning from class-imbalanced data : Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, 2017, doi: 10.1016/j.eswa.2016.12.035.

[20] G. Douzas, F. Bacao, and F. Last, "Improving Imbalanced Learning Through a Heuristic Oversampling Method Based on K-Means and SMOTE Georgios," *Inf. Sci. (Ny).*, 2018, doi: 10.1016/j.ins.2018.06.056.
[21] P. Mooijman, C. Catal, B. Tekinerdogan, and A. Lommen, "The effects of data balancing approaches : A case study," *Appl. Soft Comput.*, vol. 132, 2023, doi: https://doi.org/10.1016/j.asoc.2022.109853.
[22] D. Yu, J. Hu, Z. Tang, H. Shen, J. Yang, and J. Yang, "Improving protein-ATP binding residues prediction by boosting SVMs with random under-sampling," *Neurocomputing*, vol. 104, pp. 180–190, 2013, doi: 10.1016/j.neucom.2012.10.012.
[23] S. Anne and A. Gueye, "CNN and XGBoost for Automatic Segmentation of Stroke Lesions International Conference on Industry Sciences and Computer Science Innovation using CT Data," *Procedia*, vol. 237, pp. 72–79, 2024, doi: 10.1016/j.procs.2024.05.081.
[24] D. Koutsandreas, E. Spiliotis, and F. Petropoulos, "On the selection of forecasting accuracy measures," *J. Oper. Res. Soc.*, vol. 0, no. 0, pp. 1–18, 2021, doi: 10.1080/01605682.2021.1892464.
[25] W. Wang and Y. Lu, "Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 324, 2018, doi: 10.1088/1757-899X/324/1/012049.
[26] A. Alsahaf, N. Petkov, V. Shenoy, and G. Azzopardi, "A framework for feature selection through boosting," *Expert Syst. Appl.*, vol. 187, no. February 2021, p. 115895, 2022, doi: 10.1016/j.eswa.2021.115895.
[27] M. Mecozzi and E. Sturchio, "Computer Assisted Examination of Infrared and Near Infrared Spectra to Assess Structural and Molecular Changes in Biological Samples Exposed to Pollutants: A Case of Study," *J. Imaging*, vol. 3, no. 1, Mar. 2017.