ILKOM Jurnal Ilmiah Vol. 18, No. 1, April 2026, pp. 165-179
Accredited 2nd by KEMDIKTISAINTEK No. 10/C/C3/DT.05.00/2025; E-ISSN 2548-7779 | P-ISSN 2087-1716

A Hybrid Deep Learning–Machine Learning Approach for the Identification of Active Compounds in Blumea balsamifera (Sembung Leaves)

Kusnaeni [a,1,*]; Prihatin [a,2]; Rahmatullah [b,3]; Mega Sartika Hafid [a,4]; Muhammad Rifki Nisardi [a,5]; Nurmalasari [a,6]; Afif Budi Andy B [c,7]

[a] *Institut Teknologi Bacharuddin Jusuf Habibie, Jln. Balaikota No.1, Bumi Harapan, Parepare, 91122, Indonesia*
[b] *Pemerintah Kabupaten Mamuju, Jl. Pemuda No.2, Binanga, Mamuju, 91511, Indonesia*
[c] *Universitas Sulawesi Barat, Jalan Prof. Dr. Baharuddin Lopa, S.H, Majene, 91214, Indonesia*
[1] *kusnaeni25@ith.ac.id;* [2] *prihatinsugeng@ith.ac.id;* [3] *rahmatullah43@gmail.com;* [4] *megasartikahafid@gmail.com;* [5] *muhammadrifkinisardi@ith.ac.id;* [6] *nurmalasari@ith.ac.id;* [7] *afifbudiandy.b@unsulbar.ac.id*
** Corresponding author*

Article history: Received December 06, 2025; Revised February 10, 2026; Accepted April 01, 2026; Available online April 20, 2026

Abstract

Blumea balsamifera (sembung) is a medicinal plant with well-documented antibacterial, anti-inflammatory, and analgesic properties. However, the systematic identification of its bioactive compounds remains a significant challenge due to the complexity and high dimensionality of LC–MS (Liquid Chromatography–Mass Spectrometry) data. This study aims to develop a robust computational framework for automated compound identification using a hybrid modeling approach.A hybrid model integrating Long Short-Term Memory (LSTM) and Extreme Gradient Boosting (XGBoost) was employed to enhance feature extraction and classification performance. The LSTM component was utilized to capture sequential dependencies in spectral data, while XGBoost performed optimized classification through gradient boosting. This integration enables efficient handling of complex spectral patterns and improves predictive accuracy.The proposed model achieved an accuracy of 91%, demonstrating strong performance in classifying and identifying bioactive compounds. Feature importance analysis identified several key compounds contributing to the model predictions, including Luteolin-7-methyl-ether, Umbelliferone, Blumeatin, Dihydroquercetin-7,4′-dimethylether, Chrysosplenol C, Blumealactone B, and Blumeaene E. These compounds are associated with known pharmacological activities, supporting the therapeutic relevance of B. balsamifera.The proposed hybrid LSTM–XGBoost framework provides an effective and scalable approach for LC–MS-based compound identification. This method reduces analytical complexity, enhances classification reliability, and offers a data-driven strategy for accelerating phytochemical research and bioactive compound validation.

Keywords: Hybrid Method; LSTM; Sembung Leaves; XGBoost

Introduction

Blumea balsamifera, commonly known as sembung, is an indigenous medicinal plant from Indonesia that grows optimally in tropical climates [1]. The part of the plant most commonly used for medicinal purposes is its leaves [2]. This plant is known for its various therapeutic properties, such as relieving rheumatism, flu, and bronchitis [3]. Despite the existence of tens of thousands of medicinal plant species worldwide, only a small proportion has been extensively utilized or scientifically investigated as ingredients in traditional medicine [4], [5]. The high costs and lengthy duration of clinical trials remain major challenges in the development of herbal medicines [6]. Liquid Chromatography–Mass Spectrometry (LC–MS) is widely employed to analyze the chemical constituents of medicinal plants [7]. However, it presents significant challenges due to high data dimensionality, complexity, and extended analysis time [8]. Therefore, the integration of artificial intelligence is essential to automate the identification process of active compounds with improved efficiency and accuracy. A Hybrid Deep Learning–Machine Learning method can serve as an effective solution to identify the active compounds in Blumea balsamifera leaves using LC–MS data. This method combines Deep Learning algorithms (Long Short-Term Memory, LSTM) with Machine Learning techniques (Extreme Gradient Boosting, XGBoost). Such a hybrid approach enables deeper feature extraction from LC–MS spectra, thereby enhancing the model's performance in identifying and classifying active compounds with greater precision and

http://dx.doi.org/10.33096/ilkom.v18i1.3195.165-179

efficiency [9]. Moreover, it improves identification accuracy, reduces computational time in big data LC–MS analysis, and enhances result interpretability, especially in highlighting the most influential compounds.

Previous studies on the identification of active compounds in *Blumea balsamifera* leaves have predominantly employed variable selection and spectroscopic approaches, particularly FTIR-based wavelength analysis [10]–[12]. While effective for feature reduction, these methods are inherently limited in structural elucidation because vibrational spectra do not provide direct molecular confirmation. In contrast, recent advances in LC–MS-based metabolomics integrated with machine learning have significantly enhanced compound annotation accuracy and structural discrimination. For example, Alanazi (2025) developed a machine learning-driven LC–MS database that improved phytochemical identification and isomer resolution [13], while Brittin et al. (2025) achieved high predictive performance (>93%) in LC–MS/MS-based bioactivity classification [14]. Moreover, Jin et al. (2025) highlighted the growing role of deep learning and hybrid ML models in non-targeted LC–MS analysis for high-dimensional feature extraction and spectral interpretation [15].Despite these developments, hybrid Deep Learning–Machine Learning (DL–ML) frameworks have been largely applied to general medicinal plant datasets rather than species-specific investigations [16]–[20]. To date, no study has implemented a hybrid DL–ML approach for identifying active compounds in *Blumea balsamifera* leaves using LC–MS data. Considering the superior sensitivity and structural resolution of LC–MS compared to conventional spectroscopic techniques [21], this absence represents a clear methodological gap. Accordingly, this study addresses this gap by developing a hybrid DL–ML framework tailored to LC–MS metabolomic data of *Blumea balsamifera*, aiming to enhance identification accuracy while improving model interpretability.

Method

A. Data

The data used in this study were obtained from an experiment conducted by the Center for Biopharmaca Research, Institute for Research and Community Service (LPPM), IPB University, in collaboration with the Department of Statistics, IPB University. The dataset was generated through Liquid Chromatography–Mass Spectrometry (LC–MS) analysis, aimed at identifying compounds that significantly contribute to antioxidant activity in Blumea balsamifera (sembung) extract. The dataset comprises 35 observations and 2,098 explanatory variables. In this study, the explanatory variables (X) represent mass spectrometry intensity values at specific m/z ratios, while the response variable (Y) corresponds to the antioxidant level of the extract, categorized as either "Strong" or "Weak". Additional metadata include molecular mass, chemical formula, and predicted compound identities. The structure of the dataset is presented in **Table 1**.

Table 1. Data Structure

Observation	Predicted compound	Formula	RT [Min]	IC_{50} Value (µg/mL)	Antioxidant content level
Water replicates 1 ⋮ Water replicate 7	*Lutelion* ⋮ *Blumeaene E*	C15 H10 O6 ⋮ C20 H30 O6	11.175 ⋮ 15.009	372.34 ⋮ 344.1	Weak ⋮ Weak
30% ethanol replicate 1 ⋮ 30% ethanol replicate 5	*Aromadendrene oxide* ⋮ 5,7-*Dihydroxychromone*	C15 H24 O ⋮ C9 H6 O4	14.589 ⋮ 8.979	143.97 ⋮ 146.37	Strong ⋮ Strong
⋮	⋮	⋮	⋮	⋮	⋮
70% ethanol replicate 1	3,5-*Dihydroxy*-3',4',7-*trimethoxyflavone*	C18 H16 O7	15.351	52.96	Strong

B. LSTM

The Long Short-Term Memory (LSTM) network was first introduced by Sepp Hochreiter and Jürgen Schmidhuber in 1997 as a solution to the limitations faced by conventional Recurrent Neural Networks (RNNs) [22], particularly the vanishing and exploding gradient problems that occur during backpropagation through time (BPTT). The LSTM architecture was specifically designed to store and utilize information over long time intervals, making it highly suitable for time series forecasting and dynamic system modeling [23]. These issues have hindered standard RNNs in learning

and retaining long-term dependencies, which are crucial in sequential data processing. The key innovation of LSTM lies in the presence of memory cells [24] that are capable of maintaining their states over time, along with gate mechanisms that regulate the flow of information. This design enables LSTM networks to overcome gradient-related challenges by ensuring more stable and efficient information propagation across long sequences [25].

The Long Short-Term Memory (LSTM) network was first introduced by Sepp Hochreiter and Jürgen Schmidhuber in 1997 as a solution to the limitations faced by conventional Recurrent Neural Networks (RNNs) [22], particularly the vanishing and exploding gradient problems that occur during backpropagation through time (BPTT). The LSTM architecture was specifically designed to store and utilize information over long time intervals, making it highly suitable for time series forecasting and dynamic system modeling [23]. These issues have hindered standard RNNs in learning and retaining long-term dependencies, which are crucial in sequential data processing. The key innovation of LSTM lies in the presence of memory cells [24] that are capable of maintaining their states over time, along with gate mechanisms that regulate the flow of information. This design enables LSTM networks to overcome gradient-related challenges by ensuring more stable and efficient information propagation across long sequences [25].

Mathematically, consider a traditional RNN, where the hidden state $\boldsymbol{h_t}$ at time step $\boldsymbol{t}$ is given by Equation (1):

$$h_t = \tanh(W_h h_{t-1} + W_x x_t + b) \tag{1}$$

During the backpropagation process, the gradient $\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{h_t}}$, where $\boldsymbol{L}$ denotes the loss function, is computed to update the model parameters [26]. However, in long data sequences, repeated multiplication of gradients can lead to two major issues: vanishing gradients (where the values approach zero) or exploding gradients (where the values grow uncontrollably) [27]. Both conditions result in unstable and inefficient model training. LSTM addresses these challenges by introducing the cell state $\boldsymbol{C_t}$, which functions like a conveyor belt, allowing gradients to flow with minimal modification [28]. The update of the cell state is mathematically expressed in Equation (2):

$$C_t = f_t . C_{t-1} + i_t . \widetilde{C}_t \tag{2}$$

In this context, $\boldsymbol{f_t}$, $\boldsymbol{i_t}$, and $\widetilde{\boldsymbol{C_t}}$ refer to the forget gate, input gate, and candidate cell state, respectively, as previously described. The forget gate f_tdetermines the extent to which information from the previous cell state $\boldsymbol{C_{t-1}}$ should be retained, while the input gatei_tregulates the proportion of the candidate cell state $\widetilde{\boldsymbol{C_t}}$ to be added to memory. This selective update mechanism allows the LSTM to preserve relevant information over long sequences while gradually discarding less important details [29].

C. XGBoost

XGBoost (Extreme Gradient Boosting) is a scalable tree boosting algorithm that has been proven effective and widely adopted for various predictive data tasks. This algorithm utilizes decision trees as base learners and applies boosting techniques to enhance prediction accuracy [30]. XGBoost is equipped with features such as regularization, parallel processing, and the ability to handle missing values. Additionally, the system is optimized through cache-aware access patterns, data compression, and data sharding, forming an efficient and scalable boosting framework [31]. Due to the combination of these techniques, XGBoost is capable of handling large-scale data with more efficient resource utilization compared to previous methods [32].

Traditional tree boosting models generally rely only on the first-order derivative information of the loss function [33]. During the training process of the nth tree, this approach complicates the implementation of distributed training because the residuals from the previous tree $k(n-1)$ must be used. XGBoost addresses this limitation by applying a second-order Taylor expansion to the loss function, thereby enabling the use of second-order derivative information for more precise model updates [34]. Mathematically, the XGBoost model represents the prediction as the sum of K regression trees, as shown in Equation (3), where F denotes the space of base learners:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \qquad f_k \in F \tag{3}$$

The corresponding objective function, which consists of the training loss and a regularization term, is defined in Equation (4):

$$L = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{4}$$

Here, lll represents the loss function measuring the discrepancy between predicted and actual values, while Ω denotes the regularization function used to control model complexity. The regularization term is formulated in Equation (5):

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|W\|^2 \tag{5}$$

where T is the number of leaves in the tree and W denotes the leaf weights.

After applying the second-order Taylor expansion to the objective function, the information gain obtained from splitting a node during tree construction is calculated as shown in Equation (6):

$$gain = \frac{1}{2}\left[\frac{\left(\sum_{i=I_L} g_i\right)^2}{\sum_{i=I_L} h_i + \lambda} + \frac{\left(\sum_{i=I_R} g_i\right)^2}{\sum_{i=I_R} h_i + \lambda} + \frac{(\sum_{i=I} g_i)^2}{\sum_{i=I} h_i + \lambda}\right] - \gamma \tag{6}$$

The information gain in Equation (6) quantifies the improvement obtained from a candidate split. A node is split only when the gain exceeds the threshold γ, ensuring controlled tree growth and preventing overfitting.

$gain$: The information gain calculated after each split during the tree construction process. Information gain indicates how much information is obtained from a given split.

$\sum_{i=I_L} g_i$: Total gradien dari data yang masuk ke node kiri setelah pemisahan.

$\sum_{i=I_L} h_i$: The total hessian of the data that flows into the left node after the split. The hessian is used to adjust the gradient and assign greater weight to more significant observations.

$\sum_{i=I_R} g_i$: The total gradient of the data that flows into the right node after the split.

$\sum_{i=I_R} h_i$: The total hessian of the data that flows into the right node after the split.

$\sum_{i=I} g_i$: To The total gradient of the data at the node before the split.

$\sum_{i=I} h_i$: The total hessian of the data at the node before the split.

λ : A regularization parameter used to control model complexity and prevent overfitting.

γ A threshold that controls tree growth. A node will only be split if the information gain is greater than the threshold value γ .

D. Data Processing Procedure

To efficiently identify active compounds in Blumea balsamifera (sembung), this study adopts a hybrid approach by integrating deep learning and machine learning models. The data processing steps are outlined as follows:

1. Data Acquisition and Preparation

   LC–MS analysis produced retention time–intensity spectra for each sample, stored as numerical arrays of size $N \times T$, where N is the number of samples and T represents spectral time points. Samples were labeled as Strong or Weak antioxidant activity. Data were split using an 80:20 stratified train–test ratio with a fixed random seed to ensure reproducibility.

2. Data Preprocessing

   Several preprocessing steps were performed to ensure the data were in optimal condition for modeling:

   - Exploratory Data Analysis (EDA): Inspection of distribution and class imbalance.

- Data standardization: Z-score normalization based on training data statistics.
- Data balancing: SMOTE (k = 5) applied only to the training set to prevent data leakage.

3. Pattern Extraction Using LSTM

   The preprocessed sequential data were fed into an LSTM architecture:

   - The input layer received spectral data as time-series sequences.
   - One or more LSTM layers were used to capture temporal patterns and correlations among spectral points.
   - The output from the LSTM layers was processed through dense layers and flattened into representative feature vectors for each sample.

   The training configuration of the LSTM model is summarized in **Table 2**.

Table 2. Hyperparameter of LSTM

Parameter	Value
Optimizer	Adam
Learning rate	0.001
Activation	tanh
Batch size	32
Epochs	100
Early stopping	Patience = 10

4. Classification Using XGBoost

   Feature vectors extracted by the LSTM model were then used as inputs to the XGBoost algorithm:

   - The XGBoost model was trained using labeled training data (Strong/Weak antioxidant activity).
   - Model hyperparameters such as learning_rate, max_depth, and n_estimators were optimized using grid search or random search techniques.
   - The training process continued until the model achieved optimal performance on the validation data.

   The training configuration of the XGBoost model is summarized in **Table 3**.

Table 3. Hyperparameter of XGBoost

Parameter	Value
n_estimators	200
max_depth	6
learning_rate	0.1
Subsample	0.8
colsample_bytree	0.8
λ	1

5. Model Performance Evaluation

   The trained model was tested on an independent test set to assess prediction accuracy:

   - Evaluation metrics included accuracy, precision, recall, and F1-score.
   - Class-wise performance (Strong vs. Weak) was analyzed.
   - A confusion matrix was constructed to provide detailed insights into correct and incorrect classifications.

Despite the promising performance of the proposed hybrid DL–ML framework, several limitations should be acknowledged. First, the model was trained and evaluated using data derived solely from *Blumea balsamifera*, which may limit generalizability to other plant species. Second, while LSTM improves feature representation, its internal feature learning process remains partially opaque, potentially limiting interpretability. Third, model performance may

be influenced by dataset size and preprocessing quality. Future work should validate the framework on external datasets and explore explainable AI techniques to enhance interpretability and robustness.

Results and Discussion

A. Exploration Data Analysis

Preliminary data exploration in this study aimed to understand the characteristics and distribution patterns of the data before modeling was conducted. This stage involved visualizing several relevant variables to obtain a general overview and to detect potential issues such as data imbalance, outliers, or inter-feature correlations. Various visualization techniques were employed, including distribution plots of individual variables, boxplots to identify extreme values and data dispersion, heatmaps to illustrate correlations among variables, and scatterplots to visually evaluate relationships between pairs of variables. The results of this exploratory analysis served as a critical foundation for the pre-modeling process and guided the selection of the most relevant features to be used during model training.

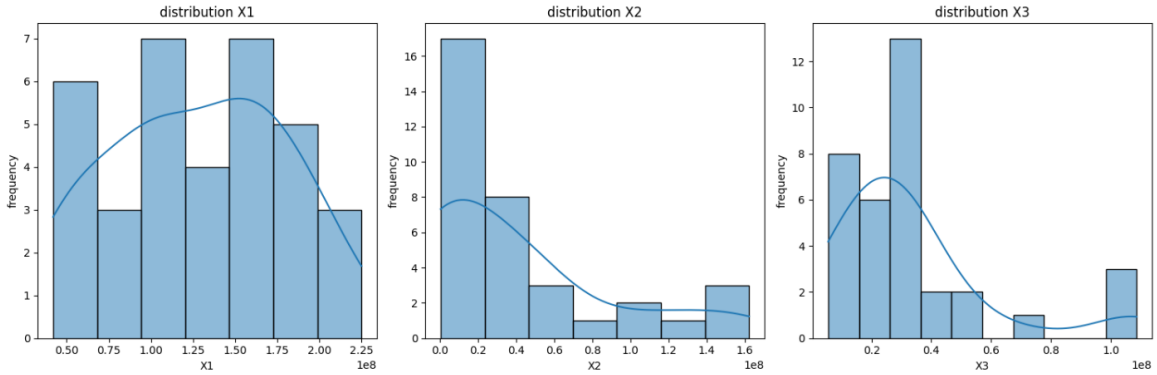


Figure 1. Visualization of the Distribution of Variables X1, X2, and X3

Figure 1 illustrates the distribution of three representative features (X1, X2, and X3). Feature X1 shows a relatively uniform distribution, suggesting broad variability that may support discriminative learning. In contrast, X2 and X3 exhibit right-skewed patterns dominated by low-intensity values with limited high-value peaks. Such skewness is typical in spectral datasets, where a few high-intensity signals may correspond to biologically relevant compounds. While skewed distributions can affect model stability, especially in gradient-based learning, these tail values may contain critical information and therefore warrant preservation through appropriate normalization rather than removal.

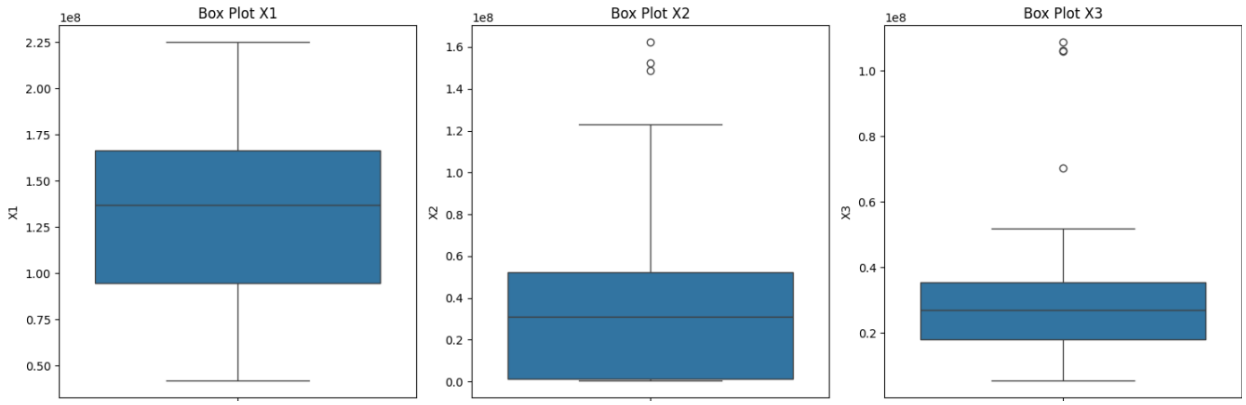


Figure 2. Boxplot of the Distribution of Variables X1, X2, and X3

Figure 2 shows the boxplot distributions of X1, X2, and X3. X1 appears relatively symmetric with stable variability, whereas X2 and X3 exhibit positive skewness and several upper outliers. Such patterns are typical in spectral data, where a limited number of high-intensity peaks may represent biologically relevant compounds. Although these extreme values can influence model behavior, especially in tree-based classifiers, their preservation is important to retain meaningful chemical information, justifying the use of normalization rather than removal.

Figure 3 shows the correlation between features (X1, X2, X3) and the target variable (Y). X1 has the strongest positive correlation with Y (r = 0.77), followed by X3 (r = 0.52), while X2 shows weak association (r = 0.25), indicating

that X1 and X3 are more informative for classification. A strong inter-feature correlation between X1 and X3 (r = 0.81) suggests shared information and potential multicollinearity. Although this may affect linear models, tree-based methods such as XGBoost are relatively robust to correlated features. Thus, the correlation analysis supports the relevance of X1 and X3 within the proposed hybrid framework while highlighting the limited contribution of X2.

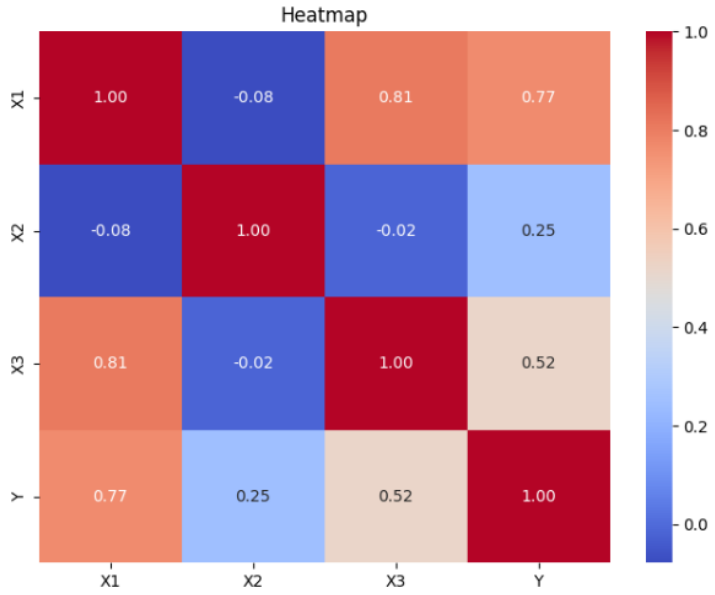
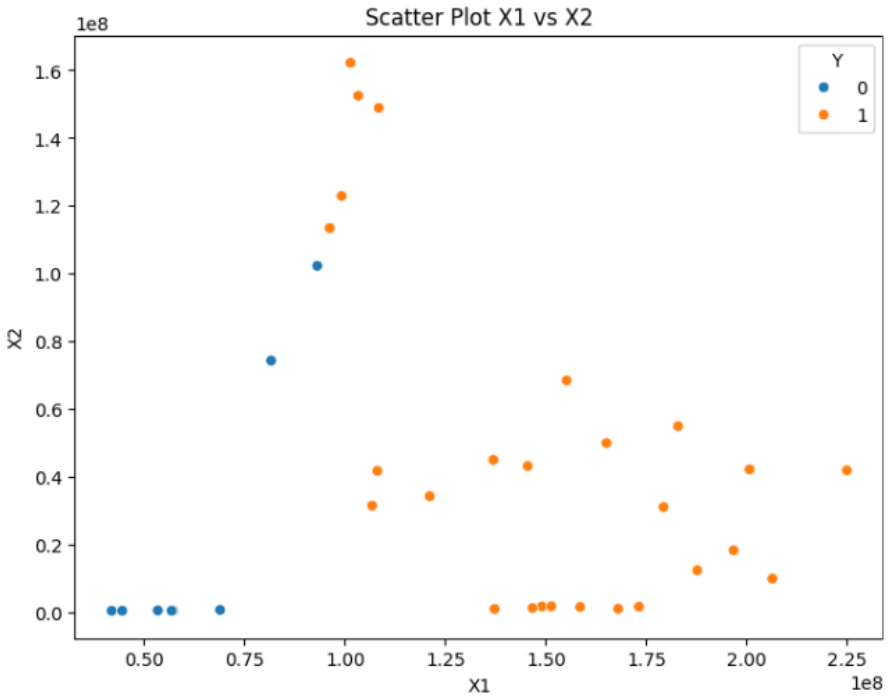


Figure 3. Heatmap of the correlation matrix among three features (X1, X2, X3) and the target variable (Y).

Figure 4 presents a scatter plot of X1 versus X2 shows partial class separation, with class 1 distributed across a broader range of X1 values, while class 0 is concentrated at lower X1 levels. This supports the stronger discriminative role of X1 identified in the correlation analysis. However, noticeable overlap between classes indicates that the relationship is not linearly separable. This justifies the use of nonlinear models such as LSTM and XGBoost, which can capture complex feature interactions. Although X2 shows limited independent separation, it may still contribute through interaction effects within the hybrid framework.



.**Figure 4.** Scatter plot illustrating the relationship between features X1 and X2 and the target variable Y.

B. Balancing Data

Figure 5 illustrates the class distribution of the target variable prior to balancing. The dataset exhibits class imbalance, with 26 samples in class 1 (strong activity) and only 9 samples in class 0 (weak activity). Such imbalance can bias classification models toward the majority class, potentially inflating overall accuracy while reducing minority-class recall. This issue is well-documented in imbalanced learning theory, where classifiers may prioritize dominant patterns at the expense of minority representation. To mitigate this risk, SMOTE was applied to the training data to improve class representation. However, while oversampling can enhance minority detection, it may also introduce synthetic noise, necessitating careful validation through cross-validation and confusion matrix analysis to ensure robust generalization.

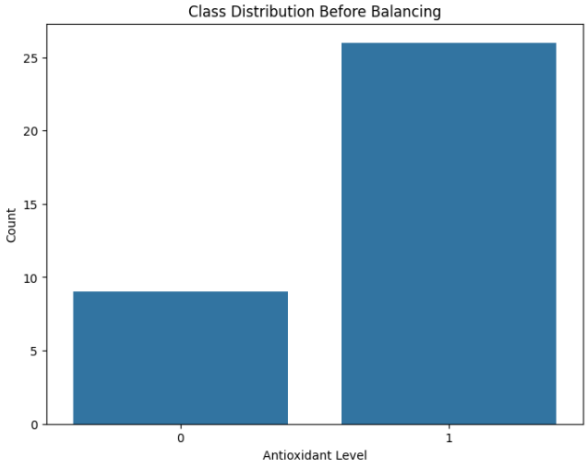


Figure 5. Class Distribution Before Balancing

Figure 6 illustrates the class distribution before and after applying SMOTE. Initially, the dataset was imbalanced, with the strong antioxidant class dominating the weak class, potentially biasing the model toward majority-class predictions. Such imbalance is known to reduce minority-class sensitivity and distort performance metrics, particularly accuracy. After applying SMOTE to the training data, the class distribution became balanced, allowing the model to learn more representative decision boundaries. While oversampling improves minority-class representation, it may also introduce synthetic variability; therefore, model robustness was further assessed using cross-validation and confusion matrix analysis. This balanced setup supports more reliable evaluation of the hybrid model's classification performance.

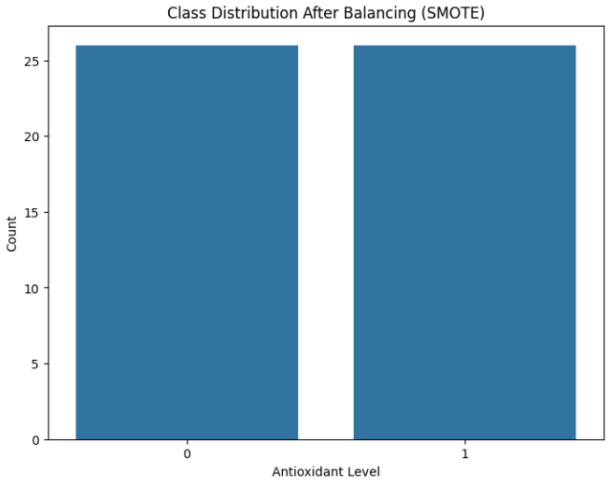


Figure 6. Class Distribution After Balancing (SMOTE)

C. Feature Extraction Using the LSTM Algorithm

Figure 7 illustrates the LSTM-extracted features (1–5) exhibit predominantly right-skewed and sparse distributions, with most activations concentrated near zero and limited high-value responses. This pattern aligns with representation learning theory, where deep models generate sparse latent embeddings that highlight informative spectral patterns while

suppressing redundancy. Features 1, 2, and 4 show strong sparsity, suggesting selective activation by specific biochemical signals, whereas Feature 5 displays broader variability, indicating potentially greater discriminative capacity. Feature 3 appears more centered but with limited dispersion, implying lower separability. Although the LSTM successfully transformed raw spectra into structured latent representations, the dominance of near-zero activations suggests that only certain dimensions meaningfully contribute to prediction, justifying further evaluation through feature importance analysis.

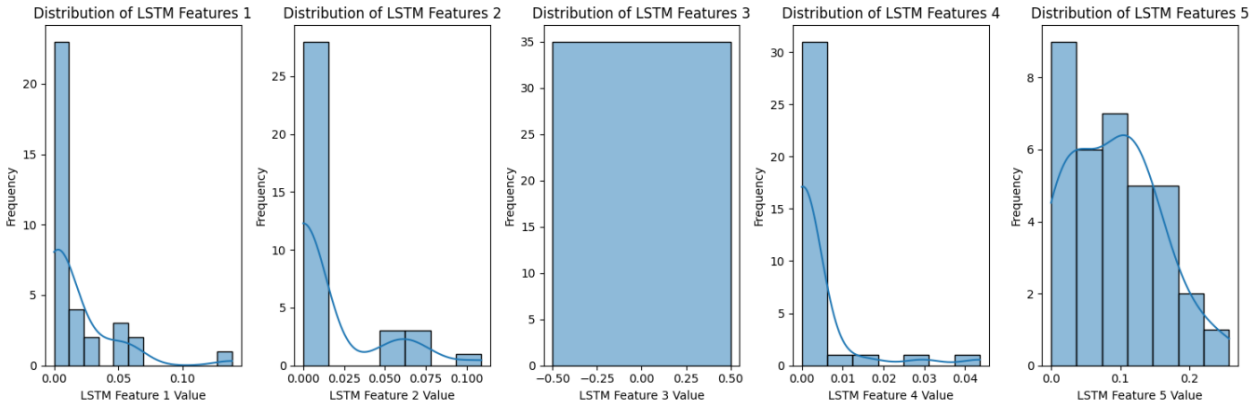


Figure 7. Distribution of five representative features extracted by the Long Short-Term Memory (LSTM) model.

Figure 8 presents a visualization of the PCA projection of the LSTM-extracted features reveals the emergence of distinct clustering patterns in the reduced two-dimensional space. Several samples form relatively compact groups, while others appear more dispersed, indicating that the latent representations capture structured variability within the data.

The observable separation suggests that the LSTM successfully encoded discriminative information from the original spectral sequences, supporting the effectiveness of deep sequential feature extraction. However, the clusters are not perfectly separated, implying that class boundaries are not strictly linear. This justifies the subsequent use of a nonlinear classifier such as XGBoost to model complex decision boundaries.

From a dimensionality reduction perspective, PCA highlights that meaningful variance is concentrated along the first two principal components, consistent with representation learning theory where deep models compress high-dimensional inputs into informative latent spaces. Nevertheless, partial overlap between groups indicates that additional nonlinear structure may exist beyond linear projections, reinforcing the need for ensemble-based classification.

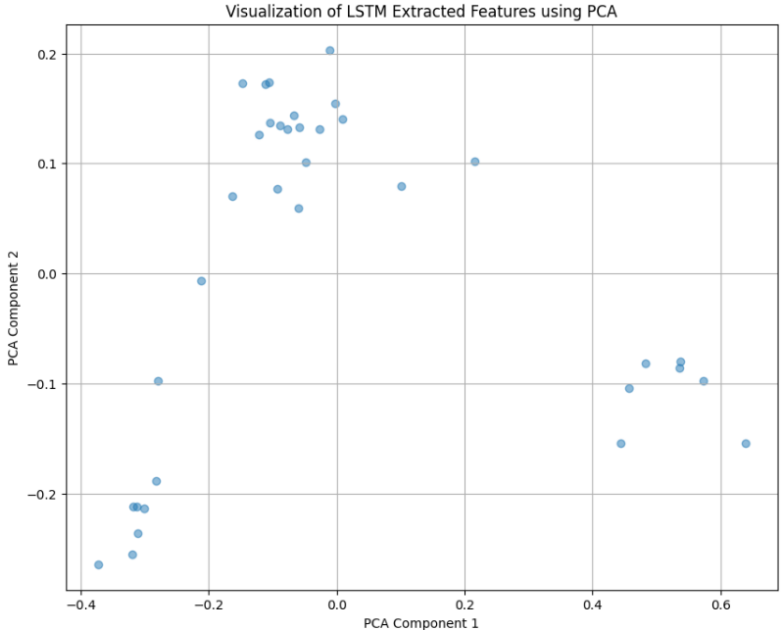


Figure 8. Long Short-Term Memory (LSTM) Extracted using PCA.

Figure 9 illustrates the t-SNE visualization of LSTM-extracted features in a two-dimensional space. The projection reveals clearer nonlinear grouping patterns compared to PCA, where several samples form relatively compact local clusters while others remain moderately dispersed, indicating preserved neighborhood structures in the latent space. Unlike PCA, which captures linear variance, t-SNE emphasizes nonlinear relationships. The improved cluster separation suggests that the LSTM features encode nonlinear discriminative information not fully observable under linear projection. However, partial overlap and scattered points indicate remaining inter-class similarity or intra-class variability. Therefore, the use of a nonlinear classifier such as XGBoost is theoretically justified to model complex feature interactions beyond linear decision boundaries

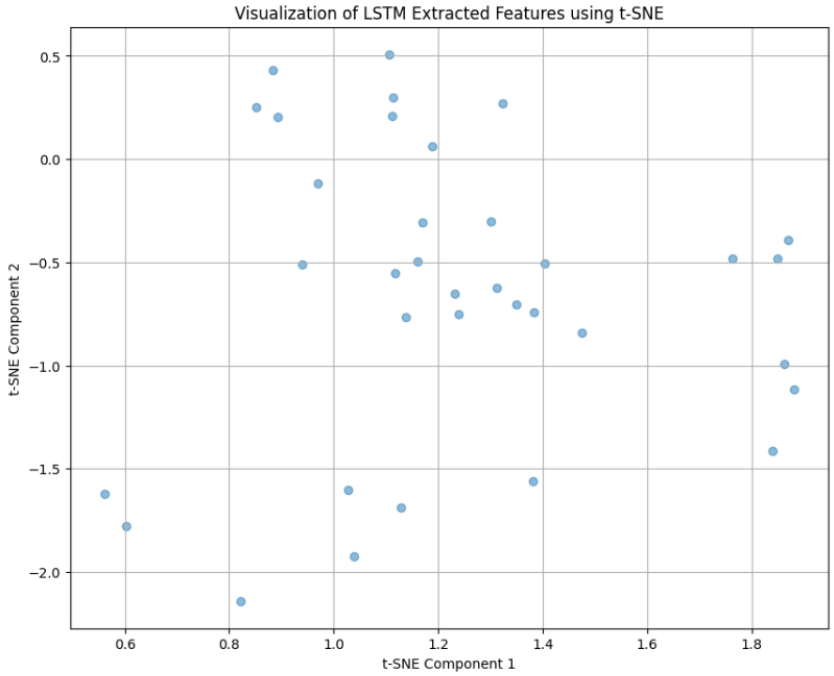


Figure 9. LSTM Extracted Features using t-SNE

D. Feature Importance in the XGBoost Algorithm

Figure 10 presents the feature importance visualization of the XGBoost model applied to the representation features previously extracted using the Long Short-Term Memory (LSTM) algorithm. This process constitutes a downstream phase of the hybrid pipeline designed to identify active compounds in Blumea balsamifera leaves that contribute to antioxidant activity levels. The visualization reveals that LSTM_Feature_13 holds the highest predictive contribution to the model, followed by LSTM_Feature_4 and LSTM_Feature_38. These importance scores reflect the extent to which each feature influences the model's classification decisions regarding antioxidant activity levels, categorized in this study as either "strong" or "weak." The findings indicate that certain LSTM-derived features are capable of capturing relevant patterns from spectral data that likely correspond to the presence of bioactive compounds with significant antioxidant potential. Consequently, these results support the effectiveness of the hybrid deep learning approach in facilitating a more targeted and data-driven identification of functional compounds based on natural product spectral data.

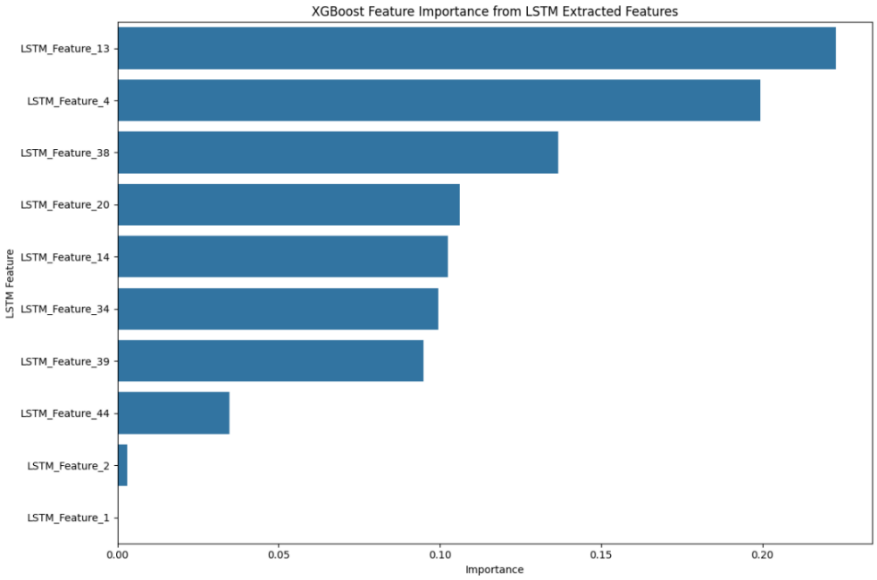


Figure 10. XGBoost Feature Importance from LSTM Extracted Features

Table 4 presents the selected variables along with the names and chemical formulas of the active compounds in *Blumea balsamifera* leaves that are correlated with antioxidant activity

Table 4. Data Structure

No	Selected Explanatory Variable	Name of Compound/ Formula	Accuracy
1	X13	*Luteolin-7-methyl-ether*/ C16 H12 O6	0.91
2	X4	*Umberlliferone* (*7-hydroxycoumarin*)/ C9 H6 O3	
3	X38	*Blumeatin* (*5,3',5'-trihydroxy-methoxydihydro-flavone*)/ C16 H14 O6	
4	X20	*Dihydroquercetin-7,4'-dimethylether*/ C17 H16 O7	
5	X14	*Luteolin-7-methyl-ether*/ C16 H12 O6	
6	X34	*Chrysosplenol C*/ C18 H16 O8	
7	X39	*Blumealactone B*/ C20 H28 O6	
8	X44	*Blumeaene E*/ C20 H30 O6	
9	X2	*Umberlliferone (7-hydroxycoumarin)*/ C9 H6 O3	

E. Accuracy Model

Figure 11 presents the evaluation results of the XGBoost model using five-fold cross-validation. Each bar in the diagram represents the accuracy achieved by the model on a particular fold, with values ranging approximately from 0.72 to 1.00. The overall average accuracy obtained across all five folds is 0.94, as indicated by the horizontal red line in the plot. These results suggest that the model demonstrates consistently high classification performance in predicting antioxidant activity levels based on the extracted features. Although Fold 1 exhibits slightly lower accuracy compared to the others, the overall variance among the folds remains relatively small. This indicates that the model does not suffer from significant overfitting or underfitting during the training process. Therefore, the hybrid LSTM–XGBoost approach employed in this study has proven effective in providing good generalization performance when applied to previously unseen data.

Table 5 presents the evaluation metrics of the classification model. The model demonstrates strong performance in detecting class 1, achieving a precision of 0.88 and a perfect recall of 1.00. This indicates that all actual instances of class 1 were correctly identified, with no false negatives, and only a small number of false positives. The F1-score for class 1 reaches 0.93, reflecting a well-balanced trade-off between precision and recall. In contrast, for class 0, the model achieves perfect precision (1.00), indicating no false positives, but only manages to recall 75% of actual class 0 instances, resulting in an F1-score of 0.86. This suggests that although the model is highly precise when predicting class

0, some instances of this class are misclassified as class 1. Overall, the model exhibits satisfactory performance, particularly in scenarios where the accurate detection of class 1 (high antioxidant activity) is prioritized.

Table 5. Model Evaluation Results

	Precision	Recall	F1-Score
0	1.00	0.75	0.86
1	0.88	1.00	0.93

Figure 12 presents the confusion matrix used to evaluate the performance of the binary classification model implemented in this study. Based on the visualization, the model successfully classified 3 actual negative instances as negative (True Negatives) and 7 actual positive instances as positive (True Positives). There was one misclassification, where one actual negative instance was incorrectly predicted as positive (False Positive), and no positive instances were misclassified as negative (no False Negatives). These results demonstrate that the model exhibits strong classification performance, particularly in accurately identifying the positive class. Overall, the confusion matrix indicates that the model achieves a low error rate and high reliability, making it suitable for similar classification tasks. This study further shows that the application of a hybrid deep learning–machine learning method using LC–MS data from Blumea balsamifera extract can improve accuracy in identifying active compounds in comparison to previous studies.

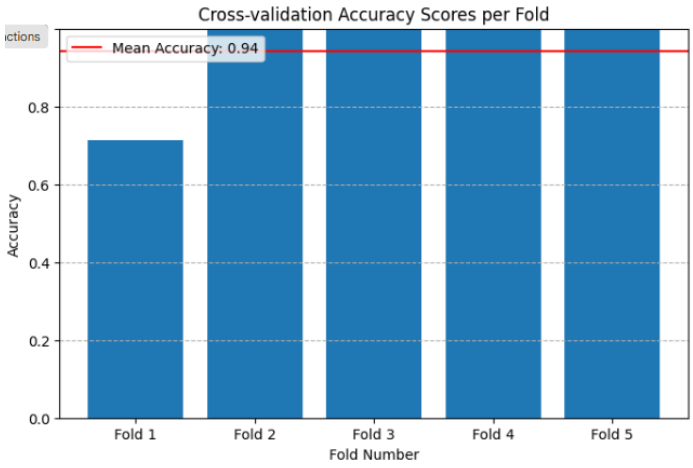


Figure 11. Cross-validation Acuracy Scores per Fold

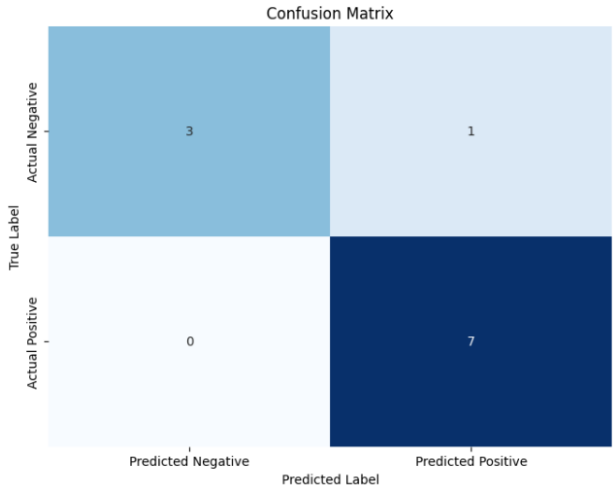


Figure 12. Confusion matrix evaluation model

Conclusion

This study developed a hybrid LSTM–XGBoost framework to identify antioxidant-related active compounds in *Blumea balsamifera* using LC–MS spectral data. The model achieved an accuracy of 0.91, demonstrating strong capability in classifying bioactive compounds. The integration of sequential feature extraction through LSTM with

ensemble-based classification proved effective for handling high-dimensional spectral data. From a computational standpoint, this work highlights the value of combining deep representation learning with gradient boosting to enhance compound identification tasks. Several compounds, including Luteolin-7-methyl-ether, Umbelliferone (7-hydroxycoumarin), Blumeatin (5,3',5'-trihydroxy-methoxydihydro-flavone), Dihydroquercetin-7,4'-dimethylether, Chrysosplenol C, Blumealactone B, and Blumeaene E, were strongly associated with antioxidant activity, supporting the biological relevance of the model predictions. Although the framework shows promising performance, validation on larger and more diverse datasets is required to ensure generalizability. Future research should incorporate explainable AI techniques and experimental vazlidation to strengthen interpretability and practical applicability. Overall, the proposed approach demonstrates the potential of hybrid deep learning models in advancing AI-driven natural product discovery.

Acknowledgement

The author expresses sincere gratitude to the Ministry of Higher Education, Science and Technology (Kemendiktisaintek) for the financial support provided through Penelitian Dosen Pemula (PDP) 2025. This support has been essential in facilitating the smooth progress and success of this research. The author also thanks all parties who have contributed support and assistance throughout the research process.

References

[1] I. G. Widhiantara and I. M. Jawi, "Phytochemical composition and health properties of Sembung plant (Blumea balsamifera): A review," *Vet. World*, vol. 14, no. 5, pp. 1185–1196, 2021, doi: www.doi.org/10.14202/vetworld.2021.1185-1196.

[2] A. Ruhardi and M. Handoyo Sahumena, "Identifikasi Senyawa Flavanoid Daun Sembung (Blumea balsamifera L.)," *J. Syifa Sci. Clin. Res.*, vol. 3, no. 1, pp. 29–36, 2021, doi: www.doi.org/10.37311/jsscr.v3i1.9925.

[3] X. Huang, D. Wang, Y. Liu, and Y. Cheng, "Diterpenoids from Blumea balsamifera and Their Anti-Inflammatory Activities," *Molecules*, vol. 27, p. 2890, 2022, doi: https://doi.org/10.3390/molecules27092890.

[4] Y. D. Derso *et al.*, "Composition , medicinal values , and threats of plants used in indigenous medicine in Jawi District , Ethiopia : implications for conservation and sustainable use," *Sci. Rep.*, vol. 14, no. 23638, pp. 1–18, 2024, doi: https://doi.org/10.1038/s41598-024-71411-5.

[5] S. L. Chen, H. Yu, H. M. Luo, Q. Wu, C. F. Li, and A. Steinmetz, "Conservation and sustainable use of medicinal plants : problems , progress , and prospects," *Chin. Med.*, vol. 11, no. 37, pp. 1–10, 2016, doi: https://doi.org/10.1186/s13020-016-0108-7.

[6] C. C. Davis and P. Choisy, "Medicinal plants meet modern biodiversity science," *Curr. Biol.*, vol. 34, no. 4, pp. R158–R173, 2024, doi: www.doi.org/10.1016/j.cub.2023.12.038.

[7] X. Zhang, X. Yu, X. Sun, X. Meng, J. Fan, and F. Zhang, "Comparative study on chemical constituents of different medicinal parts of Lonicera japonica Thunb . Based on LC-MS combined with multivariate statistical analysis," *Heliyon*, vol. 10, no. 12, p. e31722, 2024, doi: https://doi.org/10.1016/j.heliyon.2024.e31722.

[8] H. Jiang, Y. Zhang, Z. Liu, X. Wang, J. He, and H. Jin, "Advanced applications of mass spectrometry imaging technology in quality control and safety assessments of traditional Chinese medicines," *J. Ethnopharmacol.*, vol. 284, no. October 2021, p. 114760, 2022, doi: www.doi.org/10.1016/j.jep.2021.114760.

[9] M. Jiang *et al.*, "Integration of deep neural network modeling and LC-MS-based pseudo-targeted metabolomics to discriminate easily confused ginseng species," *J. Pharm. Anal.*, vol. 15, no. 1, 2025, doi: www.doi.org/10.1016/j.jpha.2024.101116.

[10] 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, Oct. 2022, doi: www.doi.org/10.31764/jtam.v6i4.8968.

[11] 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:

www.doi.org/10.24815/jn.v23i1.26279.

[12] 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: www.doi.org/10.30598/barekengvol16iss2pp721-728.

[13] S. Alanazi, "Recent Advances in Liquid Chromatography – Mass Spectrometry (LC – MS) Applications in Biological and Applied Sciences," *Anal. Sci. Adv.*, vol. 6, no. e70024, pp. 1–12, 2025, doi: https://doi.org/10.1002/ansa.70024.

[14] N. J. Brittin, J. M. Anderson, D. R. Braun, S. R. Rajski, C. R. Currie, and T. S. Bugni, "Machine Learning-Based Bioactivity Classification of Natural Products Using LC-MS/MS Metabolomics," *J. Nat. Prod.*, vol. 88, pp. 361–372, 2025, doi: www.doi.org/10.1021/acs.jnatprod.4c01123.

[15] Z. Jin, L. Chen, Y. Wang, C. Shi, Y. Zhou, and B. Xia, "Application of machine learning in LC-MS-based non-targeted analysis," *Trends Anal. Chem.*, vol. 189, no. 118243, 2025, doi: https://doi.org/10.1016/j.trac.2025.118243.

[16] Y. Hong, S. Li, Y. Ye, and H. Tang, "FIDDLE: a deep learning method for chemical formulas prediction from tandem mass spectra," *Nat. Commun.*, vol. 16, no. 11102, pp. 1–23, Nov. 2025, doi: https://doi.org/10.1038/s41467-025-66060-9.

[17] S. Javid, A. Rahmanulla, M. G. Ahmed, R. Sultana, and B. R. Prashantha Kumar, "Machine learning & deep learning tools in pharmaceutical sciences: A comprehensive review," *Intell. Pharm.*, vol. 2949–866X, no. October 2024, pp. 1–14, 2025, doi: www.doi.org/10.1016/j.ipha.2024.11.003.

[18] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, and T. Blaschke, "The rise of deep learning in drug discovery," *Drug Discov. Today*, vol. 23, no. 6, pp. 1241–1250, 2018, doi: www.doi.org/10.1016/j.drudis.2018.01.039.

[19] S. Yoo *et al.*, "A Deep Learning-Based Approach for Identifying the Medicinal Uses of Plant-Derived Natural Compounds," *Front. Pharmacol.*, vol. 11, no. November, pp. 1–15, 2020, doi: www.doi.org/10.3389/fphar.2020.584875.

[20] F. L. Duan *et al.*, "AI-driven drug discovery from natural products," *Adv. Agrochem*, vol. 3, no. 3, pp. 185–187, 2024, doi: www.doi.org/10.1016/j.aac.2024.06.003.

[21] G. D. Urso *et al.*, "The Role of LC-MS in Profiling Bioactive Compounds from Plant Waste for Cosmetic Applications : A General Overview," *Plants*, vol. 14, no. 2284, pp. 1–26, 2025, doi: https://doi.org/10.3390/plants14152284.

[22] A. Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network," *Phys. D*, vol. 404, p. 132306, 2020, doi: www.doi.org/10.1016/j.physd.2019.132306.

[23] A. Mahmoudi, "Investigating LSTM-based time series prediction using dynamic systems measures," *Evol. Syst.*, vol. 16, pp. 1–18, 2025, doi: https://doi.org/10.1007/s12530-025-09703-y.

[24] F. Landi, L. Baraldi, M. Cornia, and R. Cucchiara, "Working Memory Connections for LSTM," *Neural Networks*, vol. 144, pp. 334–341, 2021, doi: https://doi.org/10.1016/j.neunet.2021.08.030.

[25] M. J. Allah, S. Hassouna, A. Timesli, and B. A. El Majd, "LSTM-based neural network architecture for predicting the nonlinear dynamic behavior of functional gradient viscoelastic porous plates," *Mater. Today*, vol. 42, no. 111269, 2025, doi: https://doi.org/10.1016/j.mtcomm.2024.111269.

[26] M. Candan and M. Cubukcu, "Implementation of Caputo Type Fractional Derivative Chain Rule on Back Propagation Algorithm," *Appl. Soft Comput.*, vol. 155, p. 111475, 2024, doi: https://doi.org/10.1016/j.asoc.2024.111475.

[27] M. Waqas and U. W. Humphries, "A critical review of RNN and LSTM variants in hydrological time series predictions," *MethodsX*, vol. 13, no. 102946, 2024, doi: https://doi.org/10.1016/j.mex.2024.102946.

[28] G. Liao, M. Nashrul, M. Zubir, H. J. Yap, A. H. Ali, and M. Alkhedher, "Enhanced Convolutional Neural Network (CNN) -Long Short-Term Memory (LSTM) attention model with adaptive loss for lithium-ion battery state of health estimation," *PeerJ Comput. Sci.*, vol. 11, no. e3006, 2025, doi: www.doi.org/10.7717/peerj-cs.3006.

[29] J. Duan, P. F. Zhang, R. Qiu, and Z. Huang, "Long short-term enhanced memory for sequential recommendation," *World Wide Web*, vol. 26, no. 2, pp. 561–583, 2023, doi: www.doi.org/10.1007/s11280-022-01056-9.

[30] S. Hakkal and A. A. Lahcen, "XGBoost To Enhance Learner Performance Prediction," *Comput. Educ. Artif. Intell.*, vol. 7, no. 100254, 2024, doi: www.doi.org/10.1016/j.caeai.2024.100254.

[31] N. Basha, K. Jongkittinarukorn, and K. Bingi, "XGBoost based enhanced predictive model for handling missing input parameters : A case study on gas turbine," *Case Stud. Chem. Environ. Eng.*, vol. 10, no. 100775, 2024, doi: www.doi.org/10.1016/j.cscee.2024.100775.

[32] H. Chen, "Enterprise marketing strategy using big data mining technology combined with XGBoost model in the new economic era," *PLoS One*, vol. 18, no. 6, pp. 1–22, 2023, doi: www.doi.org/10.1371/journal.pone.0285506.

[33] C. Banyong, N. Hantanong, and P. Wisutwattanasak, "A machine learning comparison of transportation mode changes from high-speed railway promotion in Thailand," *Results Eng.*, vol. 24, no. 103110, 2024, doi: www.doi.org/10.1016/j.rineng.2024.103110.

[34] Z. Zhang, Y. Zhang, Y. Wen, and Y. Ren, "Data-driven XGBoost model for maximum stress prediction of additive manufactured lattice structures," *Complex Intell. Syst.*, vol. 9, no. 5, pp. 5881–5892, 2023, doi: https://doi.org/10.1007/s40747-023-01061-z.