MOBILE AND FORENSICS
Vol. 8, No. 2, 2026, pp. 77-92 | DOI: 10.12928/mf.v8i2.16102
ISSN: 2714-6685 (Online)
ISSN: 2656-6257 (Print)

Machine Learning-Based Lifestyle Analysis for Health Risk Detection in Coffee Drinkers

Sri Winiarti[a,1,*], Ahmad Azhari[a,2], Nathaniela Isya Nur Rofiah[a,3]

[a] Universitas Ahmad Dahlan, Ringroad Selatan, Tamanan Bantul, Yogyakarta, and 55191, Indonesia
[1] sri.winiarti@tif.uad.ac.id*; [2] ahmad.azhari@tif.uad.ac.id; [3] 2300018309@webmail.uad.ac.id

* Corresponding Author

ARTICLE INFO

ARTICLE HISTORY

Received: Mar 30, 2026
Revised: May 10, 2026
Accepted: Jul 06, 2026

KEYWORDS

Random Forest, Gradient Boosting, Machine Learning, Lifestyle Analysis, and Health Risks.

ABSTRACT

This study aims to develop an intelligent machine learning system for detecting health risks among coffee drinkers through lifestyle analysis. The research is motivated by the increasing coffee consumption associated with modern lifestyles, which may lead to health problems when combined with unhealthy habits. The dataset was collected through surveys covering variables such as coffee consumption frequency, sugar intake, sleep duration, physical activity, body mass index (BMI), and blood pressure. Health risk levels were classified into tiga categories: Very Healthy, Healthy, and Unhealthy. The research process included data preprocessing, normalization, and classification using two ensemble machine learning algorithms: Random Forest and Gradient Boosting. Model performance was evaluated using accuracy, precision, recall, and F1-score. Experimental results showed that both models achieved strong performance, especially in the Very Healthy and Healthy categories. Random Forest outperformed Gradient Boosting in the Healthy class with 90% accuracy compared to 87%, while both achieved 95% accuracy in the Very Healthy category. The findings indicate that coffee consumption frequency, sleep duration, and sugar intake were the most influential variables in determining health risks.

1. INTRODUCTION

Coffee consumption has become an integral part of modern lifestyles and has seen a significant increase in recent years. Changes in digital work patterns, increased social activities, and the growth of the coffee industry have driven coffee consumption to become a daily habit for people of all ages. Moderate coffee consumption is known to provide benefits such as improved focus, cognitive performance, and energy metabolism. However, excessive consumption that is not balanced with a healthy lifestyle has the potential to increase health risks such as sleep disorders, hypertension, and metabolic disorders [1].

Health risks among coffee drinkers are influenced not only by the amount of caffeine consumed, but also by other lifestyle factors such as sugar intake, sleep duration, physical activity, body mass index

(BMI), and blood pressure history. The complex interactions among these variables are difficult to analyze using conventional approaches. Therefore, an artificial intelligence-based approach is needed that can more accurately model the non-linear relationships among lifestyle variables [2], [3]. However, some recent studies have also confirmed that excessive consumption, especially when combined with high sugar intake and irregular sleep patterns, can potentially increase the risk of hypertension, sleep disorders, anxiety, and metabolic syndrome [4], [5], [6]. In Indonesia, the increase in coffee consumption is most pronounced among the working-age population, particularly the urban youth. A national study reports that the majority of coffee consumers also tend to consume excessive amounts of sugar and engage in low levels of physical activity, thereby contributing to an increased risk of metabolic disorders [7], [8], [9]. These findings indicate that health risks among coffee drinkers cannot be analyzed in isolation based solely on caffeine intake; rather, they require consideration of the interactions among various lifestyle factors, such as frequency of coffee consumption, type of coffee, sugar intake, sleep duration, body mass index (BMI), physical activity, and blood pressure history [10], [11].

Advances in machine learning have been widely applied in the healthcare field for disease risk classification based on lifestyle data. Algorithms such as Random Forest, Support Vector Machine, and Gradient Boosting have proven capable of achieving high classification performance on multivariate data. The advantages of ensemble methods in handling nonlinear and complex data make them particularly relevant for lifestyle-based health risk analysis [12], [13], [14], [15]. The strengths of the ensemble approach in handling nonlinear and multivariate data make it well-suited for lifestyle-based risk analysis. Additionally, the integration of Explainable Artificial Intelligence (XAI) enables the interpretation of how variables contribute to predictive outcomes, ensuring that the system is not a black box and is easier for users to understand [10], [16], [17].

However, most previous studies have focused on the detection of specific diseases without explicitly targeting coffee drinkers as a lifestyle-based risk group. Furthermore, studies that integrate multi-algorithm comparisons within a standardized experimental framework and provide an education system based on model interpretation remain limited, particularly in the context of local Indonesian data [18], [19], [20]. This gap raises fundamental questions about how to build an intelligent system that is not only accurate in classifying the health risks of coffee drinkers but is also capable of providing personalized, data-driven educational recommendations.

As technology advances, machine learning (ML) and artificial intelligence (AI)-based approaches are increasingly being used to analyze lifestyle and health data [21]. A recent study shows that integrating lifestyle factors into machine learning algorithms can improve the accuracy of cardiovascular disease predictions [22], [23]. Predictions of heart disease risk based on wearable data and lifestyle factors also show promising results [24]. Furthermore, Holt et al [25] states that ML algorithms can identify population health patterns, predict clinical outcomes, and enable personalized interventions [26].

In the local context, the digital transformation of healthcare services in Indonesia presents significant opportunities for the adoption of AI, although it still faces infrastructure and regulatory challenges [27]. The systematic literature review also emphasizes the need for multidisciplinary collaboration to ensure that AI truly has a positive impact on healthcare services [28], [29], [30]. Another study highlights the importance of explainable AI in predicting the risk of malnutrition [31], [32] as well as lifestyle-related obesity [1], so that the prediction system is not merely a “black box” but is also understandable to users. The ensemble learning approach has also proven effective in improving the accuracy of obesity predictions based on lifestyle data [33], [34].

Given this gap, this study aims to develop a machine learning-based intelligent system to detect health risks among coffee drinkers using lifestyle analysis. This study makes the following key contributions, 1) comprehensive integration of coffee drinkers’ lifestyle variables, 2) multi-algorithm comparison within a standardized experimental framework, and 3) development of a health education system based on model interpretation for personalized lifestyle recommendations.

2. METHOD

2.1. Research Process

The research methodology consists of six main stages, ranging from problem identification to the evaluation of the machine learning model's performance. Each stage is interconnected to ensure that the developed system has a strong scientific foundation, adequate data, appropriate processing procedures, and reliable evaluation results. Fig. 1 illustrates the research stages.

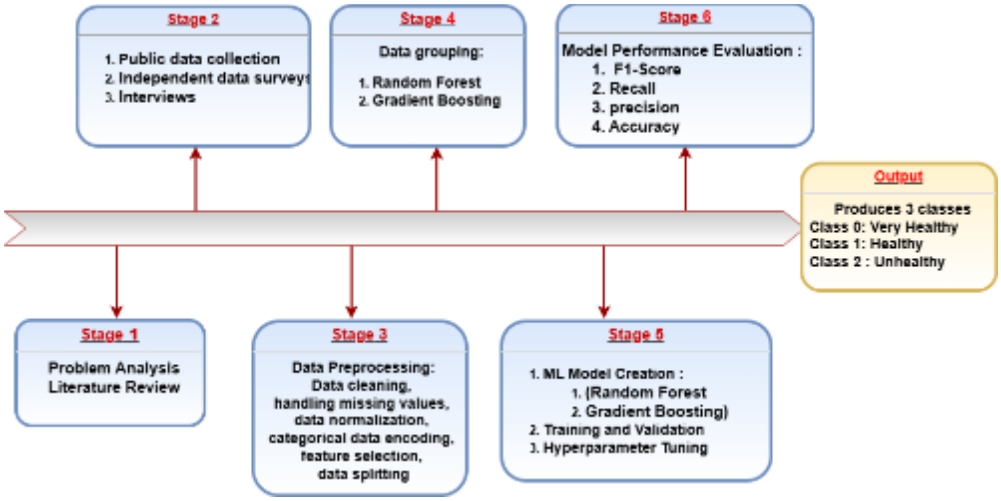

Fig. 1. Research Method

2.2. Problem Identification, Analysis, and Literature Review

To gain a deep understanding of the research problem, the main issues underlying the study were identified, such as the need to assess the health status of coffee drinkers based on lifestyle, consumption habits, and specific indicators. A problem analysis was conducted to define the research focus, objectives, benefits, and limitations. Subsequently, a literature review was performed on various relevant prior studies. The literature review aims to identify current research developments, methods that have been used, variables frequently analyzed, and research gaps that can still be explored. From this stage, the researcher obtains a theoretical foundation to determine the approach, variables, and algorithms to be used in the study.

2.3. Dataset

The dataset was obtained through a structured survey of 10,000 respondents covering nine main variables: age, gender, frequency of coffee consumption, type of coffee, sugar intake, sleep duration, physical activity, BMI, and blood pressure history. This survey was designed to capture various aspects of lifestyle and health conditions that may influence health risks, particularly among individuals who consume coffee. Nine main variables were collected: age and gender as demographic factors; coffee consumption frequency and type to describe caffeine consumption patterns; sugar intake related to metabolic risk; sleep duration and physical activity as lifestyle indicators; and BMI (Body Mass Index) and blood pressure history as indicators of physical health status.

Each respondent in the dataset was then assigned a classification label based on their health risk level, divided into three categories: very healthy, healthy, and unhealthy. These labels are likely determined based on a combination of health indicators such as BMI, blood pressure, and lifestyle habits, thereby reflecting overall health status. With this data structure, the dataset not only enables analysis of the relationship between coffee consumption and health but also supports the development of machine learning models to classify health risk levels more accurately based on individual lifestyle patterns.

2.4. Data Preprocessing

This stage prepares the data for use in the machine learning model training process. The preprocessing steps include: data cleaning, handling missing values, data normalization, encoding categorical data, feature selection, and data splitting to divide the dataset into an 80:20 ratio for training and testing data. This process is illustrated in Fig. 2.

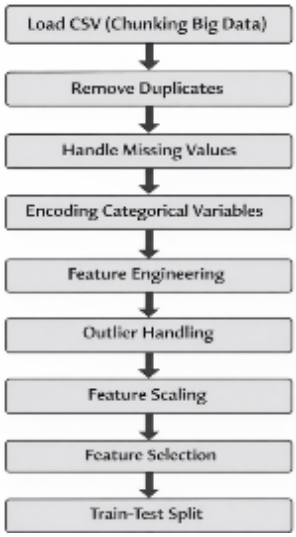

Fig. 2. Pre-processing stages of the dataset, predicting the health level of coffee drinkers

2.5. Machine Learning Architecture Models

This study employs two machine learning models: Random Forest and Gradient Boosting. The prediction results are then translated into nutrition education and healthy living recommendations. Random Forest was chosen for its ability to handle nonlinear relationships. Gradient Boosting was used to improve predictive performance through incremental learning. The machine learning models using Random Forest are shown in Fig. 3.

Fig. 3 illustrates the workflow of a machine learning-based intelligent system for predicting health risks among coffee drinkers, starting from the data collection stage through to the provision of recommendations. The process begins with raw lifestyle CSV data, which is raw data containing information on the respondents' lifestyles. Next, data preprocessing is performed, which includes removing duplicate data, handling missing values (missing value handling), encoding to convert categorical data into numerical data, and handling outliers to ensure the data is cleaner and ready for analysis. The next stage is feature engineering, where new features are created, such as coffee consumption intensity (coffee intensity), sleep quality (sleep risk), sugar consumption risk (sugar risk), and lifestyle score (lifestyle score), to enhance the quality of information in the data. Following this, feature selection is performed to choose the most relevant and influential variables for the prediction. The processed data is then fed into a Random Forest Classifier model, which consists of multiple decision trees (Tree1, Tree2, ..., TreeN) and uses a majority voting mechanism to determine the classification results. The output of this model is a health risk prediction that categorizes an individual's health risk level.

Fig. 3 illustrates the workflow for developing an ensemble-based machine learning model to predict health risks among coffee drinkers. The process begins with the Lifestyle Coffee Dataset as the primary data source, which contains lifestyle information related to coffee consumption. The data then undergoes a data preprocessing stage to clean and prepare it for analysis, followed by feature selection to identify the most relevant variables. Next, the data is split into training and test sets. (train/test split) to ensure the model can be evaluated objectively. During the modeling phase, several decision trees (Tree 1, Tree 2, Tree 3) are used, each trained separately, and then combined into an ensemble model to enhance prediction accuracy and stability. The output of this model consists of predictions that classify health risk levels, which are then used to provide health recommendations tailored to an individual's condition, ensuring the system functions not only as a predictive tool but also as a decision-making aid in maintaining a healthy lifestyle.

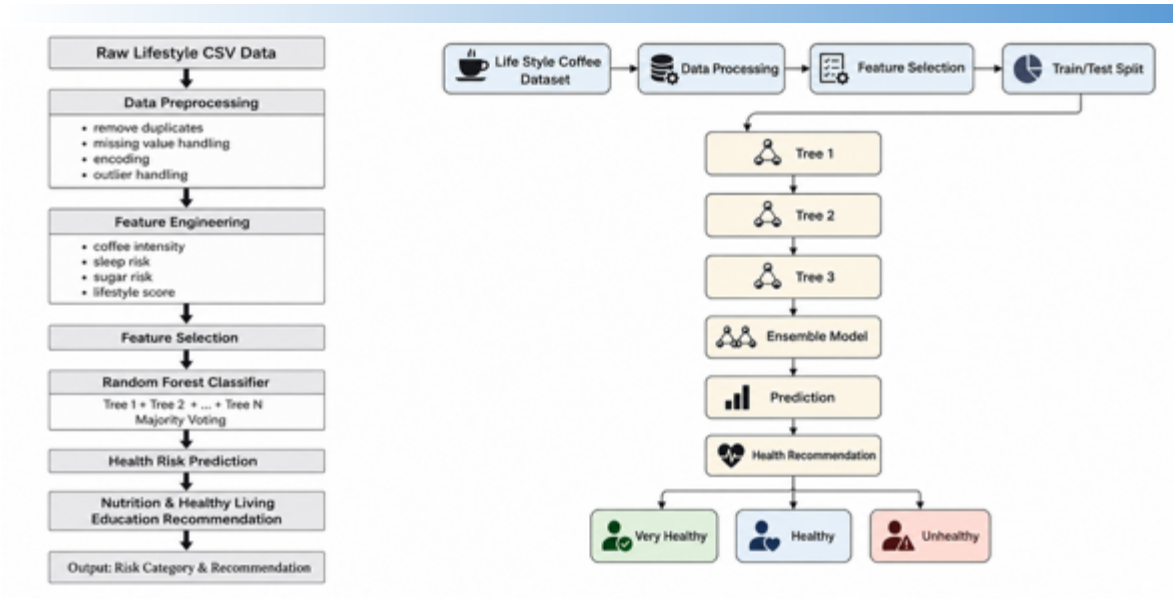

Fig. 3. Random Forest dan Gradient Boosting Architecture Model

2.6. Machine Learning Architecture Models

Parameter optimization was performed using Grid Search and Random Search with K-Fold Cross-Validation (K=5). In the Random Forest model, the optimized parameters included the number of decision trees (n_estimators) data range used+ 200.0, the maximum tree depth (max_depth) data range used: 2.00-3.00, the minimum number of samples required for a split (min_samples_split) use 80:20 and 70:30, and the minimum number of samples at a leaf node (min_samples_leaf) data range used 2.00 Meanwhile, in the Gradient Boosting model, the optimized parameters include the learning rate data range used 0.001-0,01, the number of estimators (n_estimators) data range used: 2.00-3.00, the maximum tree depth (max_depth) data range used: 2.00-3.00, and the data subsampling ratio (subsample) data range used 2.00. The parameters used for both models are presented in Table 1.

Table 1. Hyperparameter tuning is used in the optimization

No	Hyperparameter for Random Forest	Data Value	Hyperparameter for Gradient Boosting	Data Value
1	n_estimators	200.0	learning_rate	0.001-0.05
2	max_depth	3.00	n_estimators	200-800
3	min_samples_split	80:20	max_depth	3.00
4	min_samples_leaf	2.00	subsample	2.00

2.7. Machine Learning Architecture Models

The next step is to evaluate the performance of the classification model that has been built. The evaluation is conducted using several evaluation metrics, namely the Confusion Matrix, Precision, Recall, F1-Score, and Accuracy. The results of this evaluation are used to assess the model's ability to accurately classify the health levels of coffee drinkers based on lifestyle. The formulas for Precision, Recall, F1-Score, and Accuracy are explained as follows [35], [36], [37]:

a) Accuracy

Accuracy is an effective evaluation metric when the class distribution in a dataset is relatively balanced. However, accuracy has limitations on imbalanced datasets, where the number of data points in one class is significantly greater than in the others. In such cases, a model can achieve a high accuracy score simply by predicting the majority class, even though its performance on the minority class is poor. This can lead to misleading interpretations, especially when classification of the minority class is critical. This phenomenon is known as the accuracy paradox, which demonstrates that accuracy alone is insufficient for evaluating the performance of a classification model [38], [39]. The formula for calculating the accuracy value is shown in Equation (1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \ldots (1)$$

Where, TP (True Positive) represents the number of correctly predicted positive instances, TN (True Negative) represents the number of correctly predicted negative instances, FP (False Positive) represents the number of negative instances incorrectly classified as positive, and FN (False Negative) represents the number of positive instances incorrectly classified as negative.

b) Precision

In classification problems with more than two classes, accuracy alone is not sufficient to indicate the accuracy of predictions for each class. Therefore, the precision metric is used, which indicates the proportion of correct positive predictions. The formula for precision is shown in Equation (2).

$$Precision = \frac{TP}{(TP + FP)} \quad \dots (2)$$

c) F1-Score

The F1 score is an evaluation metric that combines precision and recall using a harmonic mean. This metric accounts for both false positives and false negatives, making it more effective for use on data with class imbalance compared to accuracy. The F1 score is more heavily influenced by the lower of the precision and recall values. The F1 score is the appropriate metric when false positive and false negative classification errors have a balanced impact on model performance [40].

$$F1 - Score = \frac{2*Precision*Recall}{Precision+Recall} \quad \dots (3)$$

d) Recall

Recall represents the model's success rate in detecting all data that truly belong to the positive class. The higher the recall value, the better the model's ability to capture positive data, as the number of false negatives is relatively low. Conversely, a low recall indicates that the model is still missing many positive instances during the classification process. The formula for recall is shown in Equation (4).

$$Recall = \frac{TP}{(TP + FN)} \quad \dots (4)$$

2.8. Model Output

The final stage of this study is to generate predictions of coffee drinkers' health levels based on lifestyle factors classified into three main categories: very healthy, healthy, and unhealthy. These classification results can serve as a basis for early detection among the general public, enabling the implementation of appropriate intervention measures. In this study, the output consists of a coffee drinker's health level, divided into three classes: very healthy = 0, healthy = 1, and unhealthy = 2.

3. RESULT AND DISCUSSION

3.1. Experimental Results

To assess the performance of each model, training and validation were conducted. The experimental results obtained from applying hyperparameter tuning to the Random Forest and Gradient Boosting models were then analyzed and presented in Table 2.

Table 2. Comparison of Random Forest with Gradient Boosting

Random Forest	Gradient Boosting
Parallel trees	Sequential trees
Reducing variance	Reducing bias
Faster training	More accurate
Robust to noise	Susceptible to overfitting
No dependencies between trees	Trees correct each other's errors

The hyperparameter tuning methods used in this study are Grid Search and Random Search. Grid Search performs an exhaustive search of all predefined parameter combinations, enabling it to find the optimal configuration more systematically. In contrast, Random Search performs a random search of parameters within a predefined parameter space, making it more efficient in terms of computational time.

The analysis demonstrates the performance of each model based on the best parameter combinations generated by both tuning methods. Table 3 presents the experimental results of the Random Forest and Gradient Boosting methods with various parameters and cross-validation.

Thus, a comparison between the two architectural models in detecting the health of coffee drinkers can be observed, as shown in Table 3.

Table 3. Results of Random Forest and Gradient Boosting Analysis with Hyperparameter Tuning and Cross Validation (CV)

Model	Split	Tunning	n_estimator	CV3 = 3	CV3 = 4	CV3 = 5
Random Forest	80: 20	Grid	200.0	0.9473	**0.9492**	0.9482
Random Forest	80: 20	Random	200.0	0.9473	**0.9492**	0.9482
Random Forest	70: 30	Grid	200.0	0.9453	0.9478	0.9491
Random Forest	70: 30	Random	200.0	0.9453	0.9478	0.9491
Random Forest	90: 10	Grid	200.0	0.9448	0.9479	0.9486
Random Forest	90: 10	Random	200.0	0.9448	0.9479	0.9486
Random Forest	60: 40	Grid	200.0	0.9477	0.9484	0.9498
Random Forest	60: 40	Random	200.0	0.9477	0.9484	0.9498
GradientBoosting	80: 20	Grid	200.0	0.05	0.9314	0.9311
GradientBoosting	80: 20	Random	200.0	0.05	0.9314	0.9311
GradientBoosting	70: 30	Grid	200.0	0.05	0.9275	0.9271
GradientBoosting	70: 30	Random	200.0	0.05	0.9275	0.9271
GradientBoosting	90: 10	Grid	200.0	0.05	0.9276	0.9299
GradientBoosting	90: 10	Random	200.0	0.05	0.9276	0.9299
GradientBoosting	60: 40	Grid	200.0	0.05	0.9283	0.9286
GradientBoosting	60: 40	Random	200.0	0.05	0.9283	0.9286
GradientBoosting	80: 20	Grid	600.0	0.05	0.9309	0.9180
GradientBoosting	80: 20	Grid	800.0	0.05	0.9397	0.9349
GradientBoosting	80: 20	Grid	800.0	0.05	**0.9417**	**0.9422**

To assess the performance of both models, an accuracy test was conducted using the confusion matrix method. The evaluation metrics used include recall, precision, F1-Score, and accuracy. Table 4 shows the confusion matrix evaluation results for the Random Forest model.

Based on the test results in Table 4 using the Random Forest method, an overall accuracy of 0.91 was obtained, indicating that the model is capable of classifying data with a high degree of accuracy. Further analysis was conducted using the confusion matrix to examine the classification performance for each class in greater detail. The experimental results show that Random Forest performs best compared to the other methods. Tables 4 and 5 show the performance of each model based on the confusion matrix evaluation.

Table 4. Confusion matrix test results for the Random Forest model

Actual \ Predicted	Class 0	Class 1	Class 2	Total	Accuracy
Class 0	2270	106	0	2376	0.9492
Class 1	129	1195	108	1432	0.9000
Class 2	1	26	165	192	0.6000
Total	**2400**	**1327**	**273**	**4000**	**4000**

Table 5. Confusion matrix test results for the Gradient Boosting model

Actual \ Predicted	Class 0	Class 1	Class 2	Total	Accuracy
Class 0	1120	68	0	1188	0.9422
Class 1	59	599	58	716	0.8700

Class 2	1	17	78	96	0.5700
Total	**1180**	**684**	**136**	**2000**	**2000**

3.2. Analysis of Model Performance Evaluation Results

To assess the performance of the two models, an accuracy test was conducted using the confusion matrix method. The evaluation metrics included accuracy, precision, recall, and F1-score. Table 5 presents the overall confusion matrix evaluation results for the Random Forest and Gradient Boosting models. Based on the results, the Random Forest model achieved an overall accuracy of 0.91, while the Gradient Boosting model achieved an accuracy of 0.90. These results indicate that both models were able to classify the health risk levels of coffee drinkers with a high degree of accuracy. However, a more detailed analysis was required to examine the classification performance for each health risk class, particularly the Very Healthy, Healthy, and Unhealthy categories.

Table 6. Confusion matrix test results for the Random Forest model and Gradient Boosting

Model	Accuracy	Precision	Recall	F1
Random Forest	0.91	0.91	0.91	0.91
Gradient Boosting	0.90	0.90	0.90	0.90

Based on the evaluation results in Table 6, Random Forest showed slightly better overall performance than Gradient Boosting. The Random Forest model achieved more stable classification results across the three health risk classes because it combines multiple decision trees through a bagging mechanism, which helps reduce variance and improves robustness against data imbalance and overlapping feature patterns. This characteristic enables Random Forest to distinguish the Very Healthy, Healthy, and Unhealthy classes more consistently.

In contrast, Gradient Boosting is more susceptible to classification errors in determining the Very Healthy class. This is because Gradient Boosting builds decision trees sequentially, where each new tree focuses on correcting the errors made by the previous trees. Although this approach can improve predictive performance, it also makes the model more sensitive to noisy data, outliers, and small differences between adjacent classes. In this study, the Very Healthy and Healthy classes may share similar lifestyle characteristics, such as moderate coffee consumption, adequate sleep duration, normal BMI, and relatively balanced physical activity. As a result, Gradient Boosting may produce a stricter decision boundary and misclassify some Very Healthy respondents as Healthy when their lifestyle indicators are close to the threshold between the two classes.

Therefore, although both models produced high classification performance, Random Forest demonstrated better stability and generalization in classifying lifestyle-based health risk levels. The class-level comparison shown in Table 7 further confirms that Random Forest performed better in the Healthy class, achieving 90%, while Gradient Boosting achieved 87%. Meanwhile, both models achieved 95% accuracy in the Very Healthy class, indicating that both algorithms were effective at identifying respondents with low health risk, although Gradient Boosting remained more sensitive to borderline cases between the Very Healthy and Healthy categories.

Table 7. Comparative Results of Random Forest and Gradient Boosting Model Performance Tests

Evaluation Metrics	Random Forest	Gradient Boosting
Accuracy	**0.91**	0.90
Precision (Macro Avg)	0.82	0.80
Recall (Macro Avg)	**0.88**	0.86
F1-Score (Macro Avg)	**0.84**	0.82
Precision (Weighted Avg)	**0.91**	0.90
Recall (Weighted Avg)	**0.91**	0.90
F1-Score (Weighted Avg)	**0.91**	0.90

Table 8. Comparative Results of the Performance Test of the Random Forest and Gradient Boosting Models based on Classes

Class	Precision RF	Precision GB	Recall RF	Recall GB	F1 RF	F1 GB
Class 0	0.95	0.95	**0.96**	0.94	**0.95**	0.95
Class 1	**0.90**	0.88	0.83	**0.84**	**0.87**	0.86

Class 2	**0.60**	0.57	**0.86**	0.81	**0.71**	0.67

Overall, the macro average of 0.84 indicates that the model's performance is fairly balanced across all classes, although there is a performance imbalance in the minority class. The weighted average of 0.91 indicates that the model performs very well when considering the distribution of data counts across each class. These results indicate that the Random Forest method provides stable and accurate classification performance, particularly for the majority class; however, there is still room for improvement in the minority class to increase precision and reduce classification errors. The confusion matrix plots for the performance of the Random Forest and Gradient Boosting models are shown in Fig. 4 and 5.

Fig. 4 shows the confusion matrix for the Random Forest method in classifying three classes, where the diagonal values represent correct predictions and the off-diagonal values indicate errors. The model performed best on class 0 (very healthy) with 2,270 correct predictions and minimal errors, indicating excellent ability to recognize this class. For class 1 (healthy), there were 1,195 correct predictions, though misclassifications into classes 0 and 2 were still observed, indicating similarities in characteristics among the classes. Meanwhile, class 2 (unhealthy) had 165 correct predictions with some errors, likely influenced by the smaller dataset size. Overall, the model's performance is considered good with 91% accuracy, as shown by the dominance of values on the main diagonal, although there is still room for improvement, particularly for the minority class, through data balancing techniques or model optimization.

Fig. 5 shows the confusion matrix for the Gradient Boosting method, indicating that the model performs quite well in classifying the three classes, with the diagonal values serving as an indicator of correct predictions. For class 0, there were 1,120 correct predictions with a few errors in class 1, indicating that the model is quite accurate in recognizing this class. For class 1, 599 data points were classified correctly, but there were still errors in classes 0 and 2, indicating similarities in characteristics between classes. Meanwhile, in class 2, there were 78 correct predictions with relatively few errors, although some data were still predicted as class 1. Overall, the Gradient Boosting model demonstrated good performance with a dominance of values on the main diagonal, although classification errors still occurred, particularly in classes with closely related characteristics.

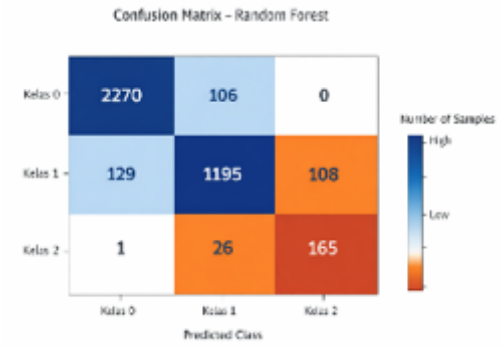

Fig. 4. Grafik confusion Matrix model Random Forest

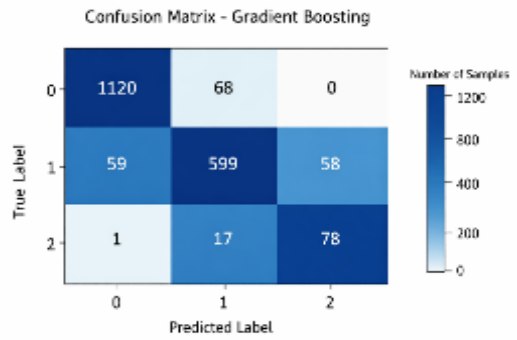

Fig. 5. Grafik confusion Matrix model Gradient Boosting

3.3. Feature Importance Analysis

The machine learning model used in this study identified five lifestyle variables that have the greatest influence on determining health risk levels among coffee drinkers. These variables are the frequency of coffee consumption, sleep duration, sugar intake, body mass index (BMI), and physical activity. Among these, the frequency of coffee consumption emerged as the most influential factor, as it is directly associated with the amount of caffeine entering the body. Excessive coffee consumption may increase the risk of health problems such as hypertension, sleep disturbances, and metabolic disorders. Sleep duration was identified as the second most important variable because inadequate sleep can intensify the negative effects of caffeine while reducing the body's ability to recover physiologically. In addition, sugar intake, BMI, and physical activity also play significant roles in shaping health outcomes, highlighting the importance of maintaining a balanced lifestyle. Overall,

these findings demonstrate that the machine learning model is capable of effectively identifying the lifestyle factors that contribute most significantly to predicting health risk levels among coffee drinkers.

Sugar intake is the next significant factor, particularly because coffee consumption is often accompanied by added sugar, which risks raising blood glucose levels and triggering metabolic diseases such as diabetes. BMI (Body Mass Index) reflects overall body condition, where an unhealthy BMI can increase susceptibility to various diseases. Meanwhile, physical activity serves as a protective factor that can mitigate the negative effects of unhealthy consumption patterns.

Overall, these findings confirm that health risks among coffee drinkers are not solely influenced by coffee itself, but also result from the complex interaction of various lifestyle factors. Thus, a lifestyle analysis approach is highly relevant in supporting early detection and data-driven decision-making.

3.4. Discussion

The results showed that Random Forest provided the best performance due to its ability to handle nonlinear relationships between lifestyle variables. The ensemble method was also more robust against noise and overfitting than other methods. This finding is consistent with previous research that found Random Forest to be effective for lifestyle-based health risk classification.

Furthermore, the integration of lifestyle variables into a single model enhances the ability to predict health risks more comprehensively. The developed system not only provides predictions but also educational recommendations that can be used as a data-driven early detection tool.

Based on Tables 5 and 6, the Random Forest model performed slightly better than Gradient Boosting across almost all evaluation metrics. Random Forest achieved an accuracy of 0.91, higher than Gradient Boosting's 0.90. Furthermore, the macro-average value for Random Forest was also higher, indicating a more balanced classification across classes.

In the per-class evaluation, both models performed almost the same in class 0. However, Random Forest had an advantage in minority classes, especially in class 2, with a recall value of 0.86 compared to Gradient Boosting of 0.81. This indicates that Random Forest is better able to detect minority classes compared to Gradient Boosting. Overall, Random Forest provided more stable performance and was superior in handling data imbalance compared to Gradient Boosting. A comparison of the performance of the two models is shown in Fig. 6.

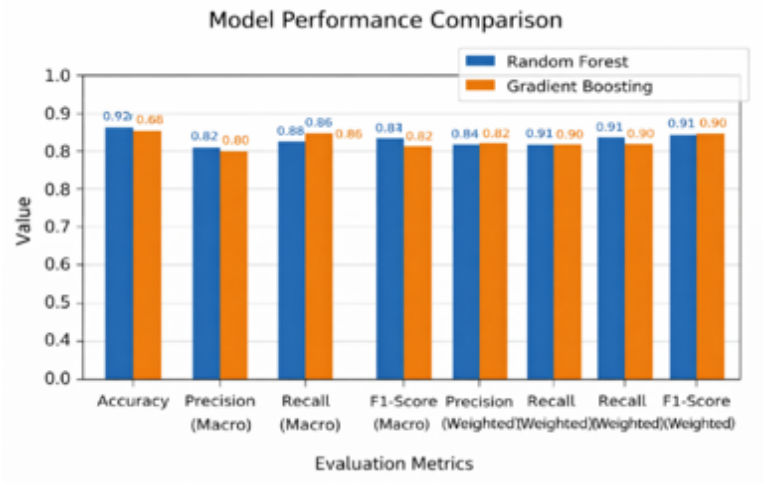

Fig. 6. Comparison graph of the performance of the Random Forest and Gradient Boosting models

The model performance comparison chart shows that Random Forest consistently outperforms Gradient Boosting on nearly all evaluation metrics. In terms of accuracy, Random Forest achieves a score of 0.92, higher than Gradient Boosting's 0.88, indicating better overall predictive capability. For the precision, recall, and F1-score metrics, both on a macro scale and weighted, both models demonstrate relatively balanced performance; however, Random Forest remains slightly superior, particularly in weighted recall and F1-score, which demonstrate the model's stability in handling imbalanced data distributions.

Nevertheless, Gradient Boosting also demonstrates competitive performance with a margin that is not too significant, particularly in macro recall, which is nearly equivalent. This indicates that both models possess good generalization capabilities, but Random Forest is more stable in classifying all classes evenly. This difference in performance is likely influenced by the characteristics of the algorithm, with Random Forest being more robust to noise and overfitting.

For future research, several steps can be taken to improve the model's performance. First, conducting a more in-depth hyperparameter optimization, such as tuning the number of trees (n_estimators), tree depth, and learning rate in Gradient Boosting. Second, applying data balancing techniques such as SMOTE or under sampling to address class imbalance, particularly for the minority class. Third, adding new features or performing feature engineering to enable the model to capture more complex patterns. Fourth, exploring other methods such as ensemble stacking or deep learning models to improve accuracy. With these steps, it is hoped that future research can produce models that are more accurate, stable, and capable of generalizing data better.

The confusion matrix results show that the Gradient Boosting method can deliver good classification performance, though there are still some limitations that need to be noted. The best performance is observed in class 0, which has the highest number of correct predictions, indicating that the model can clearly recognize patterns in this class. In class 1, although the number of correct predictions is quite high, misclassifications into classes 0 and 2 still occur, indicating an overlap of characteristics between classes, causing the model to struggle in distinguishing decision boundaries optimally. Meanwhile, in class 2, the model's performance is relatively lower compared to other classes, likely due to the smaller amount of data (class imbalance), resulting in the model being less capable of learning patterns comprehensively.

These findings indicate that data distribution and feature similarity across classes significantly influence model performance. Compared to other methods, such as Random Forest, Gradient Boosting still yields competitive results, but requires further adjustments, such as hyperparameter optimization, feature quality improvement, or the application of data balancing techniques (e.g., SMOTE) to improve accuracy, particularly for the minority class. Thus, although the model is already quite reliable, there is still room for development to improve classification performance more evenly across all classes.

4. CONCLUSION

Based on the test results and performance comparison between the Random Forest and Gradient Boosting models, it can be concluded that both models are capable of providing good classification performance in detecting the health risk levels of coffee drinkers. However, the Random Forest model demonstrated slightly better performance than Gradient Boosting, with an accuracy of 0.91, compared to 0.90 achieved by Gradient Boosting. The precision, recall, and F1-score values also indicate that Random Forest produced more stable classification results across the three health risk categories, namely Very Healthy, Healthy, and Unhealthy. In addition, Random Forest showed better capability in handling minority classes, particularly Class 2, as reflected by its higher recall value compared to Gradient Boosting. This indicates that Random Forest is more effective in identifying respondents from classes with smaller sample sizes and is better able to reduce classification errors.

Overall, the results of the confusion matrix and classification report evaluations show that Random Forest can be considered the best model in this study because it provides more accurate, stable, and reliable performance in classifying lifestyle-based health risk levels among coffee drinkers. Beyond its technical performance, the developed model also has practical value as an educational tool. By using lifestyle indicators such as coffee consumption frequency, sugar intake, sleep duration, physical activity, BMI, and blood pressure, the system can help coffee drinkers understand their health status more easily. The classification results can provide early information about whether a person is categorized as Very Healthy, Healthy, or Unhealthy, allowing users to reflect on their daily habits and make better lifestyle decisions. Therefore, this system can support health awareness by encouraging coffee drinkers to balance coffee consumption with adequate sleep, controlled sugar intake, regular physical activity, and healthier daily routines.

For future research, several improvements can be made to enhance model performance and practical implementation. First, data balancing techniques such as SMOTE or class weighting can be applied

to improve the classification performance of minority classes. Second, more advanced ensemble methods such as stacking or blending can be explored by combining Random Forest and Gradient Boosting to achieve more optimal results. Third, broader hyperparameter optimization using Bayesian Optimization or Optuna can be applied to obtain the best parameter configuration more efficiently. In addition, future studies may explore other algorithms, such as XGBoost, LightGBM, or deep learning-based models, to compare classification performance. Future development may also integrate the model into a web-based or mobile-based application so that coffee drinkers can independently input their lifestyle data and receive simple educational feedback regarding their health risk level.

Acknowledgment

We would like to express our gratitude to the Center for Research and Community Service at Ahmad Dalam University for guiding and assisting us throughout the implementation of this research. We would also like to thank the Muhammadiyah Center for Research and Community Service for assisting and guiding us in conducting our research activities and for providing financial support.

Declarations

Sri Winiarti: Contributed to this article by conducting analysis and model development, training and evaluating the model, and drafting the article in Indonesian according to the publisher's template.
Ahmad Azhari: Contributed to this article by proofreading and translating the manuscript into English, as well as editing the article.
Nathaniela Isya Nur Rofiah: Contributed to data collection and data preprocessing to prepare the data for analysis in this article.
Conflict of interest: There is no conflict of interest among the authors in the writing of this article, as all authors contributed to its preparation.

Data and Software Availability Statements

The data was obtained from the kaggle.com platform, which can be accessed at https://www.kaggle.com/datasets/uom190346a/global-coffee-health-dataset.

REFERENCES

[1] T. Khater, H. Tawfik, and B. Singh, "Explainable artificial intelligence for investigating the effect of lifestyle factors on obesity," *Intell. Syst. Appl.*, vol. 23, p. 200427, Sep. 2024, doi: 10.1016/j.iswa.2024.200427.

[2] J. Schlichtiger, S. Brunner, A. Strüven, J. M. Hoppe, and C. Stremmel, "Effects of Caffeine Intake on Self-Administered Sleeping Quality and Wearable Monitoring of Sleep in a Cohort of Young Healthy Adults," *Nutrients*, vol. 17, no. 9, p. 1503, Apr. 2025, doi: 10.3390/nu17091503.

[3] Z. Gaeini, Z. Bahadoran, P. Mirmiran, and F. Azizi, "Tea, coffee, caffeine intake and the risk of cardio-metabolic outcomes: findings from a population with low coffee and high tea consumption," *Nutr. Metab.*, vol. 16, no. 1, p. 28, Dec. 2019, doi: 10.1186/s12986-019-0355-6.

[4] X.-X. Wang, C.-Y. Cao, X.-F. Wang, X.-Y. Wang, H.-B. Tong, and Y.-F. Liu, "Investigating the genetic and causal relationship between coffee/caffeine consumption and stroke: genome-wide association and bidirectional Mendelian randomization study," *Cereb. Cortex*, vol. 35, no. 9, p. bhaf265, Sep. 2025, doi: 10.1093/cercor/bhaf265.

[5] M. F. Mendoza, R. M. Sulague, T. Posas-Mendoza, and C. J. Lavie, "Impact of Coffee Consumption on Cardiovascular Health," *Ochsner J.*, vol. 23, no. 2, pp. 152–158, 2023, doi: 10.31486/toj.22.0073.

[6] A. D. Ahmadabad, L. Jahangiry, N. Gilani, M. A. Farhangi, E. Mohammadi, and K. Ponnet, "Lifestyle patterns, nutritional, and metabolic syndrome determinants in a sample of the older Iranian population," *BMC Geriatr.*, vol. 24, no. 1, p. 36, Jan. 2024, doi: 10.1186/s12877-024-04659-1.

[7] A. Azwana and R. S. Siregar, "Analisis Kualitatif Preferensi dan Nilai Konsumsi Kopi pada Generasi Urban di Kecamatan Medan Petisah," *J. Educ. Hum. Soc. Sci. JEHSS*, vol. 8, no. 3, pp. 1430–1438, Feb. 2026, doi: 10.34007/jehss.v8i3.2945.

[8] A. F. Ramadhani and M. N. Dwiputri, "Karakteristik Pola Konsumsi Produk Kopi dan Estimasi Paparan Kafein Harian Pada Kelompok Dewasa Awal: Studi Korelasi Terhadap Indikator Stres," *J. Sains Dan Teknol. Pangan*, vol. 11, no. 1, Feb. 2026, doi: 10.63071/cgkvp824.

[9] M. Maulidah and N. Hidayati, "Prediksi Kesehatan Tidur dan Gaya Hidup Menggunakan Machine Learning," *CONTEN Comput. Netw. Technol.*, vol. 4, no. 1, pp. 81–86, Dec. 2024, doi: 10.31294/conten.v4i1.4918.

[10] K. Wairimu H., "Integrating Lifestyle Data into Personalized Health Solutions," *Res. Output J. Public Health Med.*, vol. 5, no. 1, pp. 100–106, 2025, doi: 10.59298/ROJPHM/2025/51100106.

[11] M. G. Sghaireen *et al.*, "Machine Learning Approach for Metabolic Syndrome Diagnosis Using Explainable Data-Augmentation-Based Classification," *Diagnostics*, vol. 12, no. 12, p. 3117, Dec. 2022, doi: 10.3390/diagnostics12123117.

[12] S. Patidar, D. Kumar, and D. Rukwal, "Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction," in *Advances in Transdisciplinary Engineering*, R. M. Singari and P. K. Kankar, Eds., IOS Press, 2022. doi: 10.3233/ATDE220723.

[13] S. M. Abdulrahman, M. Rashid, and F. Al-Hashimi, "Optimized deep learning architectures for the classification of colorectal cancer diagnosis using whole slide images," *PeerJ Comput. Sci.*, vol. 11, p. e3241, Nov. 2025, doi: 10.7717/peerj-cs.3241.

[14] G. Airlangga, "Comparative Analysis of Machine Learning Models for Chronic Disease Indicator Classification Using U.S. Chronic Disease Indicators Dataset," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 3, pp. 1034–1042, Jun. 2024, doi: 10.57152/malcom.v4i3.1403.

[15] T. Paul, M. Assaduzzaman, N. Fahad, and M. J. Hossen, "Enhancing risk prediction for diabetes, hypertension, and heart disease using SMOTE-ENN balancing with PCA and gradient boosting in healthcare AI," *Intell.-Based Med.*, vol. 13, p. 100339, Mar. 2026, doi: 10.1016/j.ibmed.2025.100339.

[16] M. Salimparsa, K. Sedig, D. J. Lizotte, S. S. Abdullah, N. Chalabianloo, and F. T. Muanda, "Explainable AI for Clinical Decision Support Systems: Literature Review, Key Gaps, and Research Synthesis," *Informatics*, vol. 12, no. 4, p. 119, Oct. 2025, doi: 10.3390/informatics12040119.

[17] A. Jovic, N. Frid, K. Brkic, and M. Cifrek, "Interpretability and accuracy of machine learning algorithms for biomedical time series analysis – a scoping review," *Biomed. Signal Process. Control*, vol. 110, p. 108153, Dec. 2025, doi: 10.1016/j.bspc.2025.108153.

[18] M. R. Khadiki and V. A. Fitria, "Klasifikasi Penyakit Ginjal Kronis Menggunakan K-Nearest Neighbors dengan Feature selection Pearson Correlation Coefficient," *J. Inf. Syst. Res. JOSH*, vol. 6, no. 3, pp. 1767–1776, Apr. 2025, doi: 10.47065/josh.v6i3.7131.

[19] S. M. S. Islam *et al.*, "Machine Learning Approaches for Predicting Hypertension and Its Associated Factors Using Population-Level Data From Three South Asian Countries," *Front. Cardiovasc. Med.*, vol. 9, p. 839379, Mar. 2022, doi: 10.3389/fcvm.2022.839379.

[20] M. J. Corbi-Cobo-Losey *et al.*, "Coffee Consumption and the Risk of Metabolic Syndrome in the 'Seguimiento Universidad de Navarra' Project," *Antioxidants*, vol. 12, no. 3, p. 686, Mar. 2023, doi: 10.3390/antiox12030686.

[21] A. Higuera-Gómez *et al.*, "Computational algorithm based on health and lifestyle traits to categorize lifemetabotypes in the NUTRiMDEA cohort," *Sci. Rep.*, vol. 14, no. 1, p. 24835, Oct. 2024, doi: 10.1038/s41598-024-75110-z.

[22] H.-J. Kim *et al.*, "Machine Learning–Based Analysis of Lifestyle Risk Factors for Atherosclerotic Cardiovascular Disease: Retrospective Case-Control Study," *JMIR Med. Inform.*, vol. 13, pp. e74415–e74415, Aug. 2025, doi: 10.2196/74415.

[23] Z. Ungvari and S. K. Kunutsor, "Coffee consumption and cardiometabolic health: a comprehensive review of the evidence," *GeroScience*, vol. 46, no. 6, pp. 6473–6510, Jul. 2024, doi: 10.1007/s11357-024-01262-5.

[24] S. Tuly, S. Majumder, A. Chowdhury, A. Habib, and S. A. Suha, "Predicting Heart Disease Risk with Lifestyle Factors: Uncovering Vital Predictors via Feature Selection and Machine Learning Techniques," in *Proceedings of the 3rd International Conference on Computing Advancements*, Dhaka Bangladesh: ACM, Oct. 2024, pp. 466–473. doi: 10.1145/3723178.3723240.

[25] J. R. Holt *et al.*, "Enhancing Health Research with Machine Learning: Practical Case Studies Using the All of Us Researcher Workbench," *Data Sci. Sci.*, vol. 4, no. 1, p. 2523871, Dec. 2025, doi: 10.1080/26941899.2025.2523871.

[26] D. Dixon *et al.*, "Unveiling the Influence of AI Predictive Analytics on Patient Outcomes: A Comprehensive Narrative Review," *Cureus*, May 2024, doi: 10.7759/cureus.59954.

[27] S. Hanifa and K. E. Wicaksono, "Digital Transformation of Health Services in Indonesia through the Utilization of Artificial Intelligence, Big Data, and Telemedicine: Systematic Literature Review-VOSviewer," *Proceeding Int. Conf. Inov. Sci. Technol. Educ. Child. Health*, vol. 5, no. 1, pp. 181–192, Jun. 2025, doi: 10.62951/icistech.v5i1.280.

[28] M. F. Wibowo, A. Pyle, E. Lim, J. W. Ohde, N. Liu, and J. Karlström, "Insights Into the Current and Future State of AI Adoption Within Health Systems in Southeast Asia: Cross-Sectional Qualitative Study," *J. Med. Internet Res.*, vol. 27, p. e71591, Jun. 2025, doi: 10.2196/71591.

[29] E. Wianto, H. Toba, C.-H. Chen, and M. Malinda, "Sensor-based Data Acquisition via Ubiquitous Device to Detect Muscle Strength Training Activities," presented at the AHFE 2023 Hawaii Edition, 2023. doi: 10.54941/ahfe1004213.

[30] N. Crane, C. Hagerman, O. Horgan, and M. Butryn, "Patterns and Predictors of Engagement With Digital Self-Monitoring During the Maintenance Phase of a Behavioral Weight Loss Program: Quantitative Study," *JMIR MHealth UHealth*, vol. 11, p. e45057, Jul. 2023, doi: 10.2196/45057.

[31] F. Di Martino, F. Delmastro, and C. Dolciotti, "Explainable AI for malnutrition risk prediction from m-Health and clinical data," *Smart Health*, vol. 30, p. 100429, Dec. 2023, doi: 10.1016/j.smhl.2023.100429.

[32] B. Zhou *et al.*, "Clustering of lifestyle behaviours and analysis of their associations with MAFLD: a cross-sectional study of 196,515 individuals in China," *BMC Public Health*, vol. 23, no. 1, p. 2303, Nov. 2023, doi: 10.1186/s12889-023-17177-3.

[33] S. Ahmadi and N. Wening, "The Role of Artificial Intelligence for Medical Professionals in Indonesia: A Systematic Literature Review," *J. Soc. Sci. JoSS*, vol. 4, no. 4, pp. 218–227, May 2025, doi: 10.57185/joss.v4i4.437.

[34] J.-H. Jeong, I.-G. Lee, S.-K. Kim, T.-E. Kam, S.-W. Lee, and E. Lee, "DeepHealthNet: Adolescent Obesity Prediction System Based on a Deep Learning Framework," *IEEE J. Biomed. Health Inform.*, vol. 28, no. 4, pp. 2282–2293, Apr. 2024, doi: 10.1109/JBHI.2024.3356580.

[35] M. H. Kurniawan, H. Handiyani, T. Nuraini, and R. T. S. Hariyati, "Artificial Intelligence (AI) in Nursing Services: A Literature Review," *Faletehan Health J.*, vol. 10, no. 01, pp. 77–84, Apr. 2023, doi: 10.33746/fhj.v10i01.556.

[36] P. H. Trenggono and A. Bachtiar, "Peran Artificial Intelligence Dalam Pelayanan Kesehatan : A Systematic Review," *J. Ners*, vol. 7, no. 1, pp. 444–451, Apr. 2023, doi: 10.31004/jn.v7i1.13612.

[37] K. Riehl, M. Neunteufel, and M. Hemberg, "Hierarchical confusion matrix for classification performance evaluation," *J. R. Stat. Soc. Ser. C Appl. Stat.*, vol. 72, no. 5, pp. 1394–1412, Dec. 2023, doi: 10.1093/jrsssc/qlad057.

[38] I. Markoulidakis and G. Markoulidakis, "Probabilistic Confusion Matrix: A Novel Method for Machine Learning Algorithm Generalized Performance Analysis," *Technologies*, vol. 12, no. 7, p. 113, Jul. 2024, doi: 10.3390/technologies12070113.

[39] V. Sinap, "A novel hyperparameter tuning method for enhanced intrusion detection in network security," *Turk. J. Eng.*, vol. 9, no. 3, pp. 519–534, Jul. 2025, doi: 10.31127/tuje.1624366.

[40] S. Sathyanarayanan, "Confusion Matrix-Based Performance Evaluation Metrics," *Afr. J. Biomed. Res.*, pp. 4023–4031, Nov. 2024, doi: 10.53555/AJBR.v27i4S.4345.

AUTHOR BIOGRAPHY

Sri Winiarti is a lecturer at Ahmad Dahlan University with a passion and expertise in artificial intelligence. She was born in Indonesia and has long been active as an academic, involved in teaching, research, and knowledge development in information technology.

Since 2001, Sri Winiarti's research focus has continued to evolve in line with technological advances. In the early stages (2001–2010), her research focused primarily on data mining, decision support systems, and database processing to support information analysis. From 2011–2018, her research focus shifted to machine learning, including the application of algorithms such as Support Vector Machines (SVMs), Decision Trees, and other classification methods across domains such as education and healthcare.

From 2019 to the present, her research has increasingly focused on the development of deep learning-based intelligent systems, image analysis, and the integration of AI in digital health, technology-based education, and lifestyle analysis. Furthermore, she has begun to emphasize the ethical aspects of AI and the responsible use of technology. Her competencies include data processing, predictive model development, AI-based pattern analysis, and the development of applicable intelligent systems. She is also active in scientific publications and contributes to the development of AI-based technological innovations relevant to societal needs.

Ahmad Azhari is a lecturer in Informatics at Ahmad Dahlan University. He is an active academic involved in teaching, research, and development of information technology, particularly in the fields of intelligent systems and software engineering.
From 2016 to the present, Ahmad Azhari's research focus has been adaptive to technological advances. In the initial period (2016–2019), his research focused heavily on information systems development, software engineering, and web-based and mobile applications to support the needs of organizations and public services. Entering the 2020–2022 period, his research focus shifted to the application of machine learning and artificial intelligence (AI), including the use of classification and prediction algorithms in data processing. From 2023 to the present, his research has increasingly shifted toward the development of integrated intelligent systems, the utilization of data analytics, and the application of AI in various fields such as education, digital services, and decision support systems. Furthermore, he has begun to emphasize the importance of system efficiency and the development of technology-based solutions that adapt to the needs of Industry 4.0. His competencies include programming, system analysis and design, software development, data processing, and the implementation of AI technology in various real-world situations. Ahmad Azhari is also active in scientific research and publications, contributing to the development of applicable and impactful technological innovations. With his experience and expertise, Ir. Ahmad Azhari is committed to fostering the development of competent, innovative human resources in the informatics field, ready to face the challenges of the digital era.

Nathaniela Isya Nur Rofiah is a student in the Informatics Study Program at Ahmad Dahlan University, graduating in 2023. She has a strong interest in information technology, particularly artificial intelligence and machine learning. Since the beginning of her studies, she has been actively developing her competencies in programming, data processing, and predictive model analysis. Currently, she is conducting research in machine learning, focusing on developing classification and data analysis models to solve real-world problems, such as pattern detection, prediction, and data-driven decision-making.
In her research, she utilizes various algorithms, such as Decision Trees, Support Vector Machines (SVMs), and ensemble methods, to improve model performance. She is also interested in data preprocessing, data exploration, and model evaluation using metrics such as accuracy, precision, recall, and F1-score. With a strong passion for learning, she is committed to continuously developing her skills in machine learning and contributing to the development of innovative, technology-based solutions that benefit society.