Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 5, October 2025, Page. 3543-3557
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4696

Implementation of Ant Colony Optimization in Obesity Level Classification Using Random Forest

Muhammad Difha Wardana[*1], Irwan Budiman[2], Fatma Indriani[3], Dodon Turianto Nugrahadi[4], Setyo Wahyu Saputro[5], Hasri Akbar Awal Rozaq[6], Oktay Yıldız[7]

[1,2,3,4,5]Faculty of Mathematics and Natural Science, Department of Computer Science, Lambung Mangkurat University, Kalimantan, Indonesia
[6]Graduate School of Informatics, Department of Computer Science, Gazi University, Ankara, Türkiye
[7]Faculty of Engineering, Department of Computer Engineering, Gazi University, Ankara, Türkiye

Email: [1]difhawardana@gmail.com

Received : May 7, 2025; Revised : May 26, 2025; Accepted : Jul 14, 2025; Published : Oct 21, 2025

Abstract

Obesity is a pressing global health issue characterized by excessive body fat accumulation and associated risks of chronic diseases. This study investigates the integration of Ant Colony Optimization (ACO) for feature selection in obesity-level classification using Random Forests. Results demonstrate that feature selection significantly improves classification accuracy, rising from 94.49% to 96.17% when using ten features selected by ACO. Despite limitations, such as challenges in tuning parameters like alpha (α), beta (β), and evaporation rate in ACO techniques, the study provides valuable insights into developing a more efficient obesity classification system. The proposed approach outperforms other algorithms, including KNN (78.98%), CNN (82.00%), Decision Tree (94.00%), and MLP (95.06%), emphasizing the importance of feature selection methods like ACO in enhancing model performance. This research addresses a critical gap in intelligent healthcare systems by providing the first comprehensive study of ACO-based feature selection specifically for obesity classification, contributing significantly to medical informatics and computer science. The findings have immediate practical implications for developing automated diagnostic tools that can assist healthcare professionals in early obesity detection and intervention, potentially reducing healthcare costs through improved diagnostic efficiency and supporting digital health transformation in clinical settings. Furthermore, the study highlights the broader applicability of ACO in various classification tasks, suggesting that similar techniques could be used to address other complex health issues, ultimately improving diagnostic accuracy and patient outcomes.

Keywords : *Ant Colony Optimization, Feature Selection, Obesity, Random Forest.*

1. INTRODUCTION

Obesity is an increasingly severe global problem with an increasing incidence. This medical condition is characterized by excessive body fat accumulation [1]. It is also associated with a high risk of developing chronic diseases such as heart disease, stroke, type 2 diabetes, and some cancers. This problem also has enormous economic implications for society. According to data from the World Health Organization (WHO), in 2022, 2.5 billion adults aged 18 years and older were overweight, including over 890 million adults living with obesity. This corresponds to 43% of adults aged 18 years and older (43% of men and 44% of women) being overweight, a significant increase from 1990 when 25% of adults were overweight. The prevalence of overweight varied by region, from 31% in the WHO South-East Asia and African Regions to 67% in the Region of the Americas. About 16% of adults worldwide were obese in 2022, with the global prevalence of obesity more than doubling between 1990 and 2022. Additionally, 37 million children under the age of 5 years were overweight, with nearly half of these children living in Asia. In Africa, the number of overweight children under 5 years has increased by

Vol. 6, No. 5, October 2025, Page. 3543-3557
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4696

almost 23% since 2000. Over 390 million children and adolescents aged 5–19 years were overweight in 2022, with the prevalence of overweight (including obesity) rising from 8% in 1990 to 20% in 2022. The rise has occurred similarly among both boys and girls, with 19% of girls and 21% of boys being overweight in 2022. Furthermore, 8% of children and adolescents aged 5–19 years were living with obesity in 2022, up from just 2% in 1990. These statistics highlight a widespread global health issue that requires serious attention [2].

Obesity, influenced by genetic, environmental, and lifestyle factors, is now recognized as a disease with serious complications. Early detection and prevention are complex, demanding significant changes in physical activity and diet [3]. Researchers leverage data from electronic health records, insurance data, and mobile apps to develop prevention measures and early-stage treatment strategies [4]. Effective interventions typically include dietary adjustments, increased physical activity, and behavioral changes [5]. The increasing complexity of obesity-related data and the need for personalized treatment approaches have driven researchers to explore advanced computational methods for obesity assessment and classification [6]. The growing obesity rates necessitate computational methods to predict or classify obesity levels, aiding individuals in managing their diets [4].

Machine learning methods are frequently utilized in health sciences for disease complication determination, estimation, and diagnosis, aiming to enhance healthcare quality by saving time and reducing workload [7]. Recent advances in artificial intelligence have opened new possibilities for automated health monitoring and predictive analytics in obesity management [8]. Data grouping based on similar properties and patterns is crucial for analysis or classification, particularly in managing obesity from diagnosis to treatment and prevention. Accurate obesity level classification is vital for guiding medical practitioners in selecting appropriate treatment strategies [9]. The integration of machine learning with healthcare systems has shown promising results in improving diagnostic accuracy and treatment outcomes [10]. In previous research, the classification of obesity levels has been carried out using various algorithms. One of them was conducted by [11] using the KNN algorithm, which achieved an accuracy of 78.98%. Then, another study by [12] used the CNN algorithm and achieved an accuracy of 82.00%. In addition, research by [13] used Decision Tree and achieved an accuracy of up to 94.00%, and research by [14] used Multilayer Perceptron (MLP) with an accuracy of 95.06%. Of course, experimenting with other algorithms can still improve these numbers.

The Random Forest algorithm is known for superior medical data classification performance [15]. Its ensemble learning approach makes it particularly suitable for handling complex medical datasets with multiple features and potential noise [16]. This algorithm builds a decision tree with various nodes, such as root, internal, and leaf nodes, in its structure [17]. This algorithm's retrieval of attributes and data is random, according to predetermined conditions. Random Forest is widely applied in various classification cases. In research conducted by [18] with medical datasets in the form of diabetes, heart attack, and cancer, the Random Forest method produces accuracy reaching 74.03%, 83.85%, and 92.40%. The robustness of Random Forest in handling missing data and its ability to provide feature importance rankings make it an attractive choice for medical applications [19]. To improve the performance of Random Forest, it is essential to perform careful feature selection. Irrelevant features can reduce its effectiveness. Hence, an efficient approach to feature selection is required to improve the algorithm's performance.

Feature selection is an essential process in data analysis [20]. In the context of medical data, effective feature selection not only improves model performance but also enhances interpretability, which is crucial for clinical decision-making [21]. One exciting approach in feature selection optimization is to use Ant Colony Optimization (ACO). ACO is a metaheuristic algorithm inspired by the behavior of ants. The basic concept of this algorithm is to follow the trail of ants in finding the optimal path through the feature graph [22]. The adaptive nature of ACO makes it particularly suitable

Vol. 6, No. 5, October 2025, Page. 3543-3557
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4696

for complex optimization problems where traditional methods may fail to find optimal solutions [23], [24]. This will help select the best features to achieve the desired goal in data analysis. Previous research conducted by [25] proved that ACO with Naive Bayes provides promising model results; in that study, ACO as a feature selection was able to increase the accuracy results, which in the first test were 86.45%, and by using feature selection, to 95.84%.

Despite numerous studies on obesity classification using machine learning, there remains a significant research gap in optimizing feature selection specifically for obesity datasets. Previous studies have not adequately explored the potential of nature-inspired algorithms like ACO for feature selection in obesity classification. Most existing research focuses on algorithmic improvements without systematically addressing the feature optimization challenges inherent in obesity-related datasets. This research addresses this gap by implementing ACO as a feature selection method combined with Random Forest, which has not been extensively investigated in the obesity classification domain. The novelty of this study lies in the systematic integration of bio-inspired optimization with ensemble learning specifically tailored for obesity level prediction, providing a novel contribution to both medical informatics and computational intelligence fields. Therefore, this research uses other algorithms to improve the model's quality for the problems raised.

2. METHOD

The research methodology is illustrated in Figure 1, which shows the complete workflow from data collection to evaluation. The implementation was conducted using Python 3.10 with the scikit-learn library for the Random Forest algorithm and a custom ACO implementation. Data preprocessing was performed using the pandas and numpy libraries.

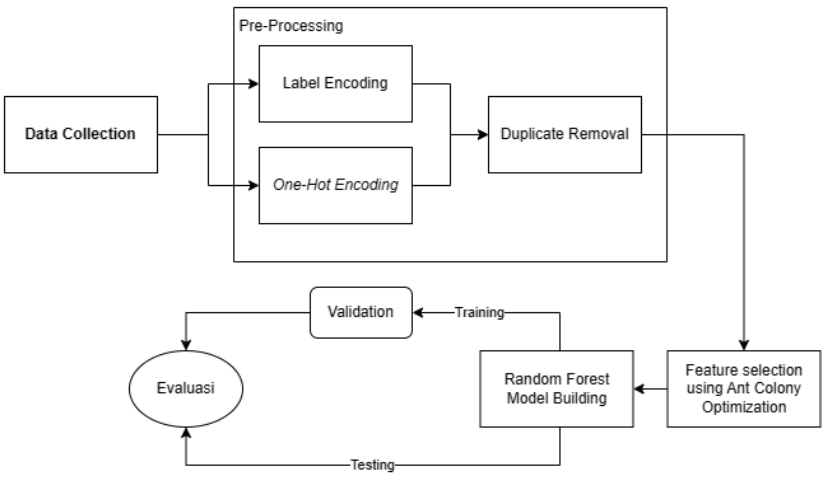

Figure 1. Flow of Research

2.1. Data Collection

The dataset used in this study is among the top relevant public datasets on obesity rates obtained from the Kaggle Website. This data is available through the link https://www.kaggle.com/datasets/aravindpcoder/obesity-or-cvd-risk-classifyregressorcluster [26]. The data contains numerical and continuous data for analysis based on classification algorithms. The data consists of obesity level estimations from individuals in Mexico, Peru, and Colombia, aged between 14 and 61, with diverse eating habits and physical conditions. Data was collected using a web-based platform where anonymous users completed a survey. The dataset collected amounted to 2111 data points, has 17 features, and has 7 classes. The attributes related to eating habits include. Each attribute can be seen in Table 1.

Vol. 6, No. 5, October 2025, Page. 3543-3557
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4696

Table 1. Features of the obesity risk dataset

No	Variables	Features	Description
1	X1	Gender	Gender with Male & Female categories.
2	X2	Age	Age in years.
3	X3	Height	Height (M).
4	X4	Weight	Body weight (Kg).
5	X5	family_history_with_overweight	Has a family with a high body weight.
6	X6	FAVC	Frequency of Eating High-Calorie Foods.
7	X7	FCVC	Frequency of Vegetable Consumption for one week.
8	X8	NCP	Number of main meals in a day.
9	X19	CAEC	Eat a snack.
10	X10	SMOKE	Smoking or not.
11	X11	CH20	Daily water consumption.
12	X12	SCC	Monitoring of calorie consumption.
13	X13	FAF	Frequency of physical activity.
14	X14	TUE	Time spent using technology devices.
15	X15	CALC	Alcohol Consumption.
16	X16	MTRANS	Transportation is used daily.
17	Y	NObeyesdad	Obesity Level.

After data collection, initial data cleaning steps were undertaken to ensure the quality and usability of the dataset. These steps included removing duplicate records to ensure each entry was unique, encoding categorical variables into numerical formats suitable for classification algorithms, and splitting the dataset into training and testing sets with an 80:20 ratio. This approach ensured balanced and reliable model evaluation, with cross-validation further dividing the training data to provide robust and stable performance through multiple iterations.

2.2. Data Preprocessing

One-hot encoding was used to convert categorical variables into binary vectors, enabling the machine-learning models to process the data effectively. Specifically, categorical variables such as gender were transformed, resulting in binary columns like "gender_male" and "gender_female," with each category represented by a binary value [27]. One-hot encoding benefits machine learning models by allowing them to understand and process categorical data more effectively [28].

Data splitting is fundamental in data preprocessing and machine learning model development. This process refers to separating a dataset into distinct subsets used in various stages of data analysis, model development, and performance evaluation [29]. The strategy for performing data splitting can vary depending on the purpose and size of the dataset. In this research, data splitting is performed with a data division of 80:20, which provides 80% of training and 20% of testing data, creating a balance between the training and test data for stable and reliable model evaluation. Cross-validation is a process of dividing data into training data and testing data [30]. Fold the definition for the K value, which is determined according to the research needs [31]. One set of data is used as training data, the rest as testing data, and so on, sequentially and alternately for each set. The cross-validation technique can estimate the value of unknown tuning parameters and the prediction error rate in the final model [32]. In this process, the model is trained using 80% of the entire data as training data, while the remaining 20% is kept for the final model evaluation process. The model is then tested on the remaining 1 part of

Vol. 6, No. 5, October 2025, Page. 3543-3557
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4696

the data as test data. This process is repeated ten times, so each piece of data is used as test data once, and the training data covers the entire data in the desired proportion.

2.3. Ant Colony Optimization

Ant Colony Optimization (ACO) is one of the methods for solving complex optimization problems using a metaheuristic approach. This method is inspired by the behavior of ant colonies in hunting, where ants communicate and cooperate in various ways [33]. ACO belongs to the nature-inspired intelligence techniques, namely swarm intelligence, which refers to collective behavior capable of self-regulation [34]. Although ACO has attracted the interest of many researchers due to its ability to handle various complex problems and provide optimal solutions, it is more effective in handling numerical data than text data [35]. ACO was first applied to the Traveling Salesman Problem (TSP) [36]. The probability of moving from point i to point j at time t is calculated using the following formula.

$$P_{i,j}^{k}(t) = \frac{[\tau_{i,j}(t)]^{\alpha}[\eta_{i,j}(t)]^{\beta}}{\sum_{l \epsilon N_i^k}[\tau_{i,l}(t)]^{\alpha}[\eta_{i,l}(t)]^{\beta}} \qquad (1)$$

Where is N_i^k the feasible neighborhood of the node i for ants k, $(\tau_{i,j}(t)$is the pheromone value on edge (i,j) at time t, aIs the weight of the pheromone, $n_{i,j}(t)$Is the previously available heuristic information on the edge (i,j) at time, t and β is the weight of the heuristic information. The pheromone value $(\tau_(i,j)(t)$ is updated according to the following equation.

$$\tau_{i,j}(t) = \rho.\tau_{i,j}(t-1) + \sum_{k=1}^{n} \Delta\tau_{i,j}^{k}(t) \qquad (2)$$

In the context of feature selection, we use ACO to help find the most relevant subset of features from the dataset's features. Initially, the pheromone value for each feature is initialized randomly. Next, the ants will build a subset of features based on probabilities influenced by the pheromone level and heuristic information. Each iteration of the ACO algorithm involves the process of feature subset construction by each ant, evaluation of the quality of the feature subset using the Random Forest model, and updating the pheromone value based on the quality of the found feature subset. This process repeats for several iterations in the hope of finding the best feature subset that improves the performance of the classification model.

2.4. Random Forest

Random Forest was first officially published by Leo Breiman in 2001. It is one of the methods used for classification and regression. This method is an ensemble of learning methods that use decision trees as basic classifiers, which are built and combined, as we can see in Figure 2 [37]. Random Forest has several advantages, and it can produce high accuracy. In addition, it is considered quite strong in handling outliers and noise. Random Forest is also simple and easily parallelized [38].

A Random Forest is an extension of a Decision Tree that uses multiple Decision Trees. Each Decision Tree is trained using individual samples, and each attribute is randomly selected from a subset of attributes in a decision tree [39]. The Random Forest method is used in building a decision tree consisting of root, internal, and leaf nodes to randomly select attributes and data according to the imposed conditions [40]. To better understand how a Decision Tree functions, it is helpful to understand its basic components: the root node, internal nodes, and leaf nodes. The root node is the topmost node representing the entire dataset, internal nodes (decision nodes) represent decision points with at least two branches, and leaf nodes are the final outputs holding class labels or continuous values. The diagram below illustrates these components.

Vol. 6, No. 5, October 2025, Page. 3543-3557
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4696

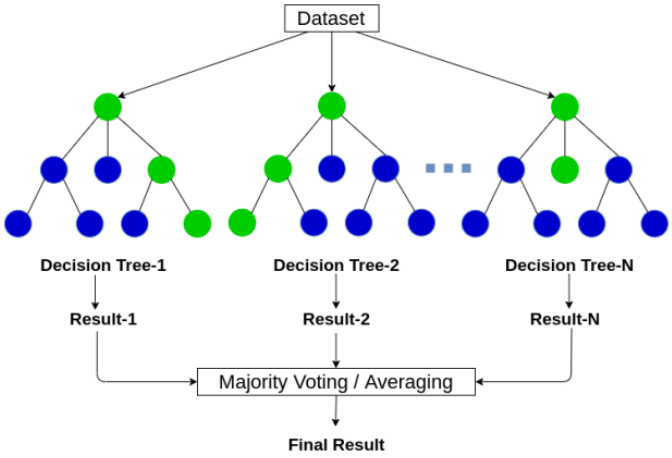

Figure 2. Random Forest Example

2.5. Performance Evaluation

A confusion matrix is one of the measurement tools in the form of a 2x2 matrix, which is used to obtain the classification accuracy of datasets against active and inactive classes in the algorithm used [41]. The evaluation stage aims to determine the level of accuracy of the results of using the algorithm or method [42]. The testing process is known as the Confusion matrix, which represents the correctness of a classification.

The confusion matrix helps us calculate several model performance evaluation metrics, including accuracy, precision, recall, F1 score, and so on. It contains actual and predicted information on the classification system. Table 2 displays the Confusion matrix and its calculation formula.

Table 2. Confusion Matrix

Classification		Class Prediction	
		Class = Yes	Class = No
Class Observation	Class = Yes	TP (True Positive)	FN (False Negative)
	Class = No	FP (False Positive)	TN (True Negative)

A confusion matrix is divided into four main components, namely True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN). True Positive (TP) is the number of cases where the model predicts positive and the true value is also positive. For example, in the case of disease detection, if the model predicts a person is sick (positive) and the person is actually ill, this is TP. False Positive (FP) is the number of cases where the model predicts positive, but the actual value is negative. For example, if the model predicts a person is sick, but the person is actually healthy, this is FP. False Positive is also known as Type I Error. False Negative (FN) is the number of cases where the model predicts negative, but the actual value is positive. For example, if the model predicts a person is healthy (negative), but the person is actually sick, this is an FN. False Negative is also known as Type II Error. True Negative (TN) is the number of cases where the model predicts negative, and the true value is negative. For example, if the model predicts a person is healthy and the person is actually healthy, this is a TN [43], [44]. From the TP, FP, FN, and TN calculations, the proportion of correct predictions out of the total predictions, the proportion of positive predictions that are actually positive, the proportion of positive values successfully identified by the model, and the proportion of negative values successfully identified by the model will be calculated [45], [46].

Vol. 6, No. 5, October 2025, Page. 3543-3557
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4696

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

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

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (6)$$

3. RESULT

This study aims to determine the potential of Ant Colony Optimization (ACO) in feature selection and provide the best performance results on obesity-level classification. Two tests were conducted, namely the application of Random Forest with all features and the application of Random Forest using features selected by ACO. Data preprocessing is carried out before testing. Cross-validation with a value of K-10 was then performed to evaluate the performance of the model objectively. Feature selection is performed to identify the most relevant features, with ACO parameters optimized to achieve the best performance.

3.1. Preprocessing

In this study, preprocessing in the form of Label Encoding is carried out on several binary categorical features such as 'family_history_with_overweight,' 'FAVC,' 'SMOKE,' and 'SCC,' and converted into binary numbers, namely 0 and 1. One-hot encoding is carried out on categorical features in the form of nominals such as 'Gender,' 'CAEC,' 'CALC,' and 'MTRANS.' The obesity level dataset has 17 features, 16 independent features, and one feature for labels. After one-hot encoding, the additional features became 23 independent features. The features obtained can be seen in Figure 3.

```
[5 rows x 24 columns]
Index(['Age', 'Height', 'Weight', 'family_history_with_overweight', 'FAVC',
       'FCVC', 'NCP', 'SMOKE', 'CH2O', 'SCC', 'FAF', 'TUE', 'Gender_Male',
       'CAEC_Frequently', 'CAEC_Sometimes', 'CAEC_no', 'CALC_Frequently',
       'CALC_Sometimes', 'CALC_no', 'MTRANS_Bike', 'MTRANS_Motorbike',
       'MTRANS_Public_Transportation', 'MTRANS_Walking', 'NObeyesdad_encoded'],
      dtype='object')
```

Figure 3. Data Information

Furthermore, after exploring the dataset, we found 24 duplicate data points, so we removed them. Then, the dataset was divided into two parts: 80% of the data became training data, and the remaining 20% became test data.

3.2. Feature Selection

Selecting the correct parameter values is crucial to run the ACO algorithm efficiently. After conducting various parameter experiments, the best parameter selection in feature selection and classification accuracy was identified as 1 for the alpha value, 1 for the beta value, and 0.6 for the evaporation rate. Therefore, these values are adopted for the algorithm. In addition, other parameters have also been adjusted. Various combinations of alpha, beta, and evaporation rates were tested to observe their impact on the algorithm's performance. The experiments involved running the ACO algorithm multiple times with different parameter sets and evaluating the classification accuracy of the resulting models. The combination of $\alpha = 1$, $\beta = 1$, and evaporation rate = 0.6 consistently produced the best results in terms of feature selection efficiency and model accuracy, as can be seen in Table 3.

Vol. 6, No. 5, October 2025, Page. 3543-3557
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4696

Table 3. Selected Parameter Values for Ant Colony Optimization

Parameter	Value
Alpha (α)	1
Beta (β)	1
Evaporation Coefficient (ρ)	0.6
Number of Ants	16
Number of Iterations	5

Not all features have a reasonable correlation in this classification. As described earlier, this research applies the Ant Colony Optimization (ACO) approach to determine the choice of features.

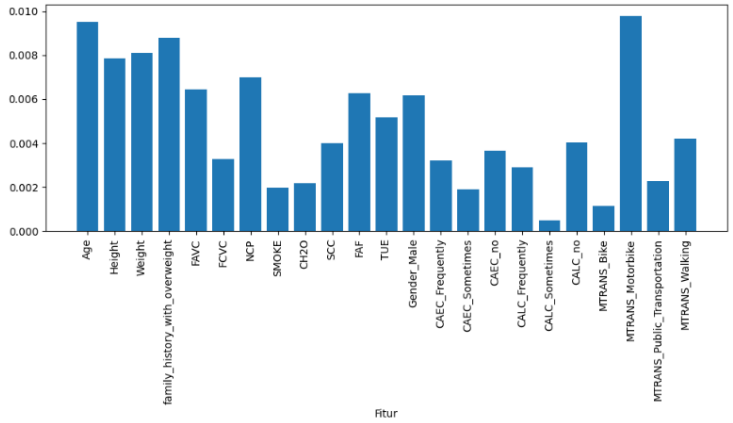

Figure 4. Pheromone value of features selected by ACO

In Figure 4, the pheromone value generated in each feature can be seen. The average pheromone value generated is 0.0048. In the ACO algorithm, a higher pheromone value indicates that ants have often selected a feature during iterations, which can be interpreted as a "better" feature in the context of feature selection.

Table 4. Features Selected by ACO

Features	Pheromone Value
MTRANS_Motorbike	0.0097
Age	0.0095
family_history_with_overweight	0.0087
Weight	0.0080
Height	0.0078
NCP	0.0069
FAVC	0.0064
FAF	0.0062
Gender_Male	0.0061
TUE	0.0051

Table 4. shows the features ACO selected with pheromone values above average. The pheromone value indicates how important the feature is in determining the data classification. In the table, it can be seen that Age, family_history_with_overweight, MTRANS_Motorbike, and Weight have the highest pheromone values. This shows that these features are most likely crucial in determining the data classification. It shows that the pheromone values for all the selected features are relatively close.

Vol. 6, No. 5, October 2025, Page. 3543-3557
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4696

3.3. Cross-validation

The ACO selected features from a total of 23 to 10 features. The model was trained using 80% of the data, with the remaining 20% reserved for final evaluation. Cross-validation with K=10 was employed, a common choice in machine learning for balancing training and testing data. This method ensures robust model evaluation by dividing the data into ten equal parts.

Table 5. Data Validation Comparison

RF	RF+ACO
95.20%	91.79%
93.11%	92.53%
95.20%	91.79%
96.40%	95.52%
93.11%	95.52%
94.61%	92.48%
94.01%	91.72%
93.71%	97.74%
94.01%	95.48%
92.21%	94.73%
94.16%	93.93%

Table 5 shows the cross-validation results using Random Forest (RF) and Random Forest combined with the Ant Colony Optimization Algorithm (RF+ACO). The RF model achieved an average accuracy of 94.16%, with performance ranging from 92.21% to 96.40% across folds, demonstrating consistent performance. The RF+ACO model had an average accuracy of 93.93%, with some folds showing high accuracy, such as 95.52% in the 4th fold and 97.74% in the 8th fold, but overall, its average performance was slightly lower than the RF model.

3.4. Random Forest Model Building and Evaluation

Furthermore, testing is performed by applying a Random Forest with all features. The results are shown in Table 6.

Table 6. Performance Results of the Random Forest Classification Algorithm

Class	Precision	Recall	F1-Score	Overall Accuracy
Insufficient_Weight	0.95	0.90	0.92	
Normal_Weight	0.80	0.92	0.85	
Obesity_Type_I	0.99	0.97	0.98	
Obesity_Type_II	1.00	1.00	1.00	94,49%
Obesity_Type_III	1.00	1.00	1.00	
Overweight_Level_I	0.96	0.87	0.91	
Overweight_Level_II	0.94	0.94	0.94	

Table 6 shows that the Random Forest classification algorithm performs satisfactorily in classifying obesity levels, as evidenced by the high F1-score and overall accuracy values. Fl-Score analysis, the harmonic mean of Precision and Recall, shows that the algorithm can classify all classes well, with F1-Score values above 0.90 for all classes. This indicates that the algorithm can identify and

Vol. 6, No. 5, October 2025, Page. 3543-3557
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4696

classify data with high precision and accuracy. Overall Accuracy reached 94.49%, marking the algorithm's ability to classify most of the data correctly. Performance analysis for each class shows that the algorithm has good Precision and Recall for each class. This indicates a high ability to classify and identify the data correctly. A prominent example is the Obesity Type II and Obesity Type III classes, where the algorithm achieved perfect Precision and Recall of 1.00. This shows the outstanding ability to classify the data for both classes.

Further testing will be done by applying the subset of features Ant Colony Optimization (ACO) selected to the Random Forest classification. The results can be seen in Table 7.

Table 7. Performance Results of the Random Forest Classification Algorithm with Feature Subsets Selected by Ant Colony Optimization (ACO)

Class	Precision	Recall	F1-Score	Overall Accuracy
Insufficient_Weight	0.97	0.95	0.96	
Normal_Weight	0.88	0.93	0.90	
Obesity_Type_I	1.00	0.96	0.98	
Obesity_Type_II	1.00	1.00	1.00	96,17%
Obesity_Type_III	1.00	1.00	1.00	
Overweight_Level_I	0.96	0.89	0.92	
Overweight_Level_II	0.92	1.00	0.96	

Table 7 shows that the combination of Random Forest and ACO algorithms for feature selection produces impressive performance in classifying obesity levels. This is evidenced by the overall accuracy value, which is 96.17%. High Precision and Recall values reinforce the algorithm's ability to classify each class well for most obesity classes. The best results were achieved in the Obesity_Type_II (Class 3) and Obesity Type III (Class 4) classes, where the algorithm achieved perfect Precision and Recall (1.00).

In addition, the combination of the Random Forest and ACO algorithm also showed good performance in Normal Weight (Class 1) and Overweight_Level_II (Class 6) classes, although there were some misclassifications. On the other hand, the performance of the algorithm tends to be lower in the Obesity_Type_I (Class 2) and Overweight Level_I (Class 5) classes, especially in terms of lower Recall, indicating that there are still some that are not correctly identified in these classes. The combination of the Random Forest and Ant Colony Optimization (ACO) algorithm has a higher Overall Accuracy than the Random Forest algorithm.

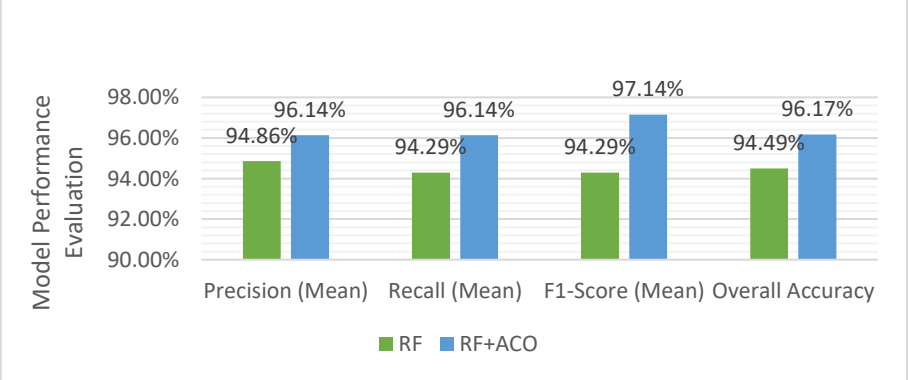

Figure 5. Performance Results of Classification Algorithms

Vol. 6, No. 5, October 2025, Page. 3543-3557
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4696

In Figure 5, performance increases after model evaluation using a combination of Random Forest with features selected by Ant Colony Optimization feature selection.

4. DISCUSSIONS

This research explores the implementation of the ACO algorithm for feature selection in obesity-level classification using the Random Forests algorithm. The results showed that the proposed approach selected a smaller subset of features (10) from 23 independent features in the initial dataset. Testing the Random Forest classification model with all features resulted in an accuracy of 94.49% while using ten features selected by ACO increased the accuracy to 96.17%. This shows that proper feature selection can improve the performance of obesity classification models, resulting in a significant improvement in prediction.

The superior performance of the proposed RF+ACO method (96.17% accuracy) compared to existing approaches demonstrates a significant advancement in obesity classification. Unlike previous studies by Dewi & Dwidasmara [8] (78.98% with KNN) and KIVRAK [9] (82.00% with CNN), our approach leverages bio-inspired optimization to intelligently select the most relevant features. This contributes to the field of medical informatics by providing a more efficient and interpretable model for healthcare decision support systems. The ACO algorithm's ability to identify optimal feature subsets (reducing from 23 to 10 features) addresses the curse of dimensionality problem common in medical datasets, making it particularly valuable for real-time clinical applications where computational efficiency is crucial.

Nonetheless, this study has limitations, namely, using ACO techniques in feature selection requires careful parameter settings to achieve optimal results. The computational complexity of ACO parameter tuning (α, β, and evaporation rate) may present challenges in clinical implementation, requiring domain expertise for optimal configuration. Despite these limitations, this study provides important implications for developing a more efficient and accurate obesity classification system. The proposed method significantly outperforms traditional approaches, with accuracy improvements of 17.19% over KNN, 14.17% over CNN, 2.17% over Decision Tree, and 1.11% over MLP, demonstrating its superiority in obesity classification tasks. Using feature selection methods such as ACO can help reduce the dimension of irrelevant data, thereby speeding up the model-building process and improving the interpretability of the results. This has immediate practical implications for healthcare systems, as reduced feature sets enable faster diagnosis processing and lower computational costs in clinical environments. Overall, this study makes an essential contribution to obesity classification by introducing a practical approach to feature selection. Table 8 compares the results of this study with those of previous studies.

Table 8. Comparative Study

Study	Algorithm	Accuracy
Dewi & Dwidasmara [11]	KNN	78,98%.
KIVRAK [12]	CNN	82,00%
Devi et al. [13]	Decision Tree	94,00%
Alqahtani [14]	Multilayer Perceptron (MLP)	95,06%
The proposed model	RF+ACO	96,17%

The comparison table reveals that the algorithm used in this study (Random Forest with Ant Colony Optimization) achieved the highest accuracy (96.17%) compared to other algorithms used in different studies. This indicates that the proposed approach yields superior performance in classifying the obesity level of patients. Applying the ACO algorithm as feature selection significantly improved

Vol. 6, No. 5, October 2025, Page. 3543-3557
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4696

prediction accuracy. The findings showed that the use of 10 features selected by ACO, including MTRANS Motorbike, Age, family_history_with_overweight, Weight, Height, NCP, FAVC, FAF, Gender_Male, TUE, was able to provide sufficient information to classify the level of obesity with an accuracy reaching 96.917%. This confirms that not all features in the dataset contribute equally to accurate prediction. The proposed approach improves the prediction accuracy and increases the model's generalizability to new data.

5. CONCLUSION

Based on this research, several significant conclusions can be drawn regarding the implementation of Ant Colony Optimization for obesity classification. The integration of Ant Colony Optimization with Random Forest successfully improved obesity classification accuracy from 94.49% to 96.17%, representing a significant 1.68% improvement and achieving the highest accuracy compared to existing methods including KNN (78.98%), CNN (82.00%), Decision Tree (94.00%), and MLP (95.06%). ACO effectively reduced feature dimensionality from 23 to 10 features (56.5% reduction) while maintaining superior classification performance, demonstrating its efficiency in feature selection optimization and addressing the curse of dimensionality problem common in medical datasets. This research significantly contributes to developing a more efficient and accurate obesity classification system, where the ACO feature selection method can effectively deal with large and complex data dimensions, enabling more concise and efficient model building.

The proposed method demonstrates excellent performance across all obesity classes with precision values ranging from 0.88 to 1.00, indicating robust and reliable classification capabilities suitable for clinical applications. This research contributes significantly to medical informatics and computer science by providing an efficient, interpretable model that can support healthcare professionals in obesity diagnosis and intervention planning, with immediate applications in digital health transformation. The findings have substantial implications for developing automated diagnostic tools in healthcare systems, potentially reducing diagnostic time by 56.5% through feature reduction, improving computational efficiency, and supporting real-time clinical decision-making processes.

Future research should focus on conducting external validation using independent datasets across different populations and healthcare settings to ensure cross-cultural applicability and test the generalizability of the resulting model. Additional investigations should explore other swarm intelligence algorithms, such as Particle Swarm Optimization and Genetic Algorithm, for comparative analysis and potential hybrid optimization approaches. The development of real-time implementation frameworks for integration into existing clinical decision support systems and electronic health record platforms represents a crucial next step for practical deployment. Furthermore, extending the methodology to multi-class health condition classification beyond obesity, including diabetes, cardiovascular diseases, and metabolic syndrome prediction, would broaden the impact of this research. Finally, implementing automated parameter tuning mechanisms to reduce the complexity of ACO configuration for non-expert users in clinical environments would enhance the accessibility and adoption of this approach. This research is expected to make an ongoing contribution to the field of obesity classification and the development of more sophisticated data analysis methods in intelligent healthcare systems, ultimately improving patient outcomes and healthcare delivery efficiency.

ACKNOWLEDGEMENT

This work was supported by Lambung Mangkurat University and Gazi University, which provided essential resources and assistance.

Vol. 6, No. 5, October 2025, Page. 3543-3557
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4696

REFERENCES

[1] I. M. Putra, I. Tahyudin, H. A. A. Rozaq, A. Y. Syafa'At, R. Wahyudi, and E. Winarto, "Classification analysis of COVID19 patient data at government hospital of banyumas using machine learning," in *2021 2nd International Conference on Smart Computing and Electronic Enterprise: Ubiquitous, Adaptive, and Sustainable Computing Solutions for New Normal, ICSCEE 2021*, Jun. 2021, pp. 271–274, doi: 10.1109/ICSCEE50312.2021.9498020.

[2] W. H. Organization, "Obesity and Overweight," *WHO*, 2024. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight (accessed Apr. 07, 2024).

[3] M. Safaei, E. A. Sundararajan, M. Driss, W. Boulila, and A. Shapi'i, "A systematic literature review on obesity: Understanding the causes & consequences of obesity and reviewing various machine learning approaches used to predict obesity," *Comput. Biol. Med.*, vol. 136, p. 104754, Sep. 2021, doi: 10.1016/j.compbiomed.2021.104754.

[4] R. Kaur, R. Kumar, and M. Gupta, "Predicting risk of obesity and meal planning to reduce the obese in adulthood using artificial intelligence," *Endocrine*, vol. 78, no. 3, pp. 458–469, Oct. 2022, doi: 10.1007/s12020-022-03215-4.

[5] S. Pfeifflé *et al.*, "Current Recommendations for Nutritional Management of Overweight and Obesity in Children and Adolescents: A Structured Framework," *Nutrients*, vol. 11, no. 2, p. 362, Feb. 2019, doi: 10.3390/nu11020362.

[6] Z. Helforoush and H. Sayyad, "Prediction and classification of obesity risk based on a hybrid metaheuristic machine learning approach," *Front. Big Data*, vol. 7, Sep. 2024, doi: 10.3389/fdata.2024.1469981.

[7] M. S. Sirsat, E. Fermé, and J. Câmara, "Machine Learning for Brain Stroke: A Review," *J. Stroke Cerebrovasc. Dis.*, vol. 29, no. 10, 2020, doi: 10.1016/j.jstrokecerebrovasdis.2020.105162.

[8] L. Huang *et al.*, "The Role of Artificial Intelligence in Obesity Risk Prediction and Management: Approaches, Insights, and Recommendations," *Medicina (B. Aires).*, vol. 61, no. 2, p. 358, Feb. 2025, doi: 10.3390/medicina61020358.

[9] L. M. Dang, K. Min, H. Wang, M. J. Piran, C. H. Lee, and H. Moon, "Sensor-based and vision-based human activity recognition: A comprehensive survey," *Pattern Recognit.*, vol. 108, p. 107561, Dec. 2020, doi: 10.1016/j.patcog.2020.107561.

[10] S. A. Alowais *et al.*, "Revolutionizing healthcare: the role of artificial intelligence in clinical practice," *BMC Med. Educ.*, vol. 23, no. 1, p. 689, Sep. 2023, doi: 10.1186/s12909-023-04698-z.

[11] A. M. S. I. Dewi and I. B. G. Dwidasmara, "Implementation of the K-Nearest Neighbor (KNN) Algorithm for Classification of Obesity Levels," *JELIKU (Jurnal Elektron. Ilmu Komput. Udayana)*, vol. 9, no. 2, p. 277, 2020, doi: 10.24843/jlk.2020.v09.i02.p15.

[12] M. Kıvrak, "Deep Learning-Based Prediction of Obesity Levels According to Eeating Habits and Physical Condition," *J. Cogn. Syst.*, vol. 6, no. 1, pp. 24–27, Jun. 2021, doi: 10.52876/jcs.939875.

[13] M. S. Devi, P. S. Ramesh, A. Joshi, K. Maithili, and A. P. Chand, "Probable Deviation Outlier-Based Classification of Obesity with Eating Habits and Physical Condition," in *Intelligent Manufacturing and Energy Sustainability*, Springer Singapore, 2023, pp. 81–93.

[14] A. Alqahtani, F. Albuainin, R. Alrayes, N. Al Muhanna, E. Alyahyan, and E. Aldahasi, "Obesity Level Prediction Based on Data Mining Techniques," *IJCSNS Int. J. Comput. Sci. Netw. Secur.*, vol. 21, no. 3, p. 103, 2021, doi: 10.22937/IJCSNS.2021.21.3.14.

[15] M. Z. Alam, M. S. Rahman, and M. S. Rahman, "A Random Forest based predictor for medical data classification using feature ranking," *Informatics Med. Unlocked*, vol. 15, p. 100180, 2019, doi: 10.1016/j.imu.2019.100180.

[16] A. Sarica, A. Cerasa, and A. Quattrone, "Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review," *Front. Aging Neurosci.*, vol. 9, Oct. 2017, doi: 10.3389/fnagi.2017.00329.

[17] P. A. Balaji and V. Sugumaran, "Fault detection of automobile suspension system using decision tree algorithms: A machine learning approach," *Proc. Inst. Mech. Eng. Part E J. Process Mech. Eng.*, p. 095440892311526, Jan. 2023, doi: 10.1177/09544089231152698.

[18] V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, "AI-based smart prediction of clinical

Vol. 6, No. 5, October 2025, Page. 3543-3557
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4696

disease using random forest classifier and Naive Bayes," *J. Supercomput.*, vol. 77, no. 5, pp. 5198–5219, May 2021, doi: 10.1007/s11227-020-03481-x.

[19] E. Battistella, D. Ghiassian, and A.-L. Barabási, "Improving the performance and interpretability on medical datasets using graphical ensemble feature selection," *Bioinformatics*, vol. 40, no. 6, Jun. 2024, doi: 10.1093/bioinformatics/btae341.

[20] D. S. Khafaga, E.-S. M. El-kenawy, F. Alrowais, S. Kumar, A. Ibrahim, and A. A. Abdelhamid, "Novel Optimized Feature Selection Using Metaheuristics Applied to Physical Benchmark Datasets," *Comput. Mater. Contin.*, vol. 74, no. 2, pp. 4027–4041, 2023, doi: 10.32604/cmc.2023.033039.

[21] S.-C. Lu, C. L. Swisher, C. Chung, D. Jaffray, and C. Sidey-Gibbons, "On the importance of interpretable machine learning predictions to inform clinical decision making in oncology," *Front. Oncol.*, vol. 13, Feb. 2023, doi: 10.3389/fonc.2023.1129380.

[22] C. Liu *et al.*, "An improved heuristic mechanism ant colony optimization algorithm for solving path planning," *Knowledge-Based Syst.*, vol. 271, p. 110540, Jul. 2023, doi: 10.1016/j.knosys.2023.110540.

[23] D. Yilmaz Eroglu and U. Akcan, "An Adapted Ant Colony Optimization for Feature Selection," *Appl. Artif. Intell.*, vol. 38, no. 1, Dec. 2024, doi: 10.1080/08839514.2024.2335098.

[24] M. Dorigo, V. Maniezzo, and A. Colorni, "Ant system: optimization by a colony of cooperating agents," *IEEE Trans. Syst. Man, Cybern. Part B*, vol. 26, no. 1, pp. 29–41, Feb. 1996, doi: 10.1109/3477.484436.

[25] K. Minari and D. P. Rini, "Optimasi Algoritma Naive Bayes Menggunakan Ant Colony Optimization untuk Klasifikasi Data Penderita Penyakit Stroke," *Repository UNSRI*, 2023. https://repository.unsri.ac.id/137557/ (accessed Apr. 07, 2024).

[26] F. M. Palechor and A. de la H. Manotas, "Obesity or CVD Risk Dataset," *Kaggle*, 2023. https://www.kaggle.com/datasets/aravindpcoder/obesity-or-cvd-risk-classifyregressorcluster (accessed Apr. 07, 2024).

[27] K. P. N. V Satya Sree, J. Karthik, C. Niharika, P. V. V. S. Srinivas, N. Ravinder, and C. Prasad, "Optimized Conversion of Categorical and Numerical Features in Machine Learning Models," in *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, Nov. 2021, pp. 294–299, doi: 10.1109/I-SMAC52330.2021.9640967.

[28] M. K. Dahouda and I. Joe, "A Deep-Learned Embedding Technique for Categorical Features Encoding," *IEEE Access*, vol. 9, pp. 114381–114391, 2021, doi: 10.1109/ACCESS.2021.3104357.

[29] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," *Glob. Transitions Proc.*, vol. 3, no. 1, pp. 91–99, 2022, doi: 10.1016/j.gltp.2022.04.020.

[30] D. Rastogi, P. Johri, V. Tiwari, and A. A. Elngar, "Multi-class classification of brain tumour magnetic resonance images using multi-branch network with inception block and five-fold cross validation deep learning framework," *Biomed. Signal Process. Control*, vol. 88, p. 105602, Feb. 2024, doi: 10.1016/j.bspc.2023.105602.

[31] M. Gholizadeh, R. Saeedi, A. Bagheri, and M. Paeezi, "Machine learning-based prediction of effluent total suspended solids in a wastewater treatment plant using different feature selection approaches: A comparative study," *Environ. Res.*, vol. 246, p. 118146, Apr. 2024, doi: 10.1016/j.envres.2024.118146.

[32] D. S. Pandey, H. Raza, and S. Bhattacharyya, "Development of explainable AI-based predictive models for bubbling fluidised bed gasification process," *Fuel*, vol. 351, p. 128971, Nov. 2023, doi: 10.1016/j.fuel.2023.128971.

[33] M. Scianna, "The AddACO: A bio-inspired modified version of the ant colony optimization algorithm to solve travel salesman problems," *Math. Comput. Simul.*, vol. 218, pp. 357–382, Apr. 2024, doi: 10.1016/j.matcom.2023.12.003.

[34] N. Kumari and D. P. Acharjya, "Data classification using rough set and bioinspired computing in healthcare applications - an extensive review," *Multimed. Tools Appl.*, vol. 82, no. 9, pp. 13479–13505, Apr. 2023, doi: 10.1007/s11042-022-13776-1.

[35] S. Gite *et al.*, "Textual Feature Extraction Using Ant Colony Optimization for Hate Speech

Vol. 6, No. 5, October 2025, Page. 3543-3557
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4696

Classification," *Big Data Cogn. Comput.*, vol. 7, no. 1, p. 45, Mar. 2023, doi: 10.3390/bdcc7010045.

[36] B. Lammel, K. Gryzlak, R. Dornberger, and T. Hanne, "An ant colony system solving the travelling salesman region problem," in *2016 4th International Symposium on Computational and Business Intelligence (ISCBI)*, Sep. 2016, pp. 125–131, doi: 10.1109/ISCBI.2016.7743270.

[37] E. K. Sahin, "Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest," *SN Appl. Sci.*, vol. 2, no. 7, p. 1308, Jul. 2020, doi: 10.1007/s42452-020-3060-1.

[38] J. S. Rhodes, A. Cutler, and K. R. Moon, "Geometry- and Accuracy-Preserving Random Forest Proximities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10947–10959, Sep. 2023, doi: 10.1109/TPAMI.2023.3263774.

[39] X. Zhou, P. Lu, Z. Zheng, D. Tolliver, and A. Keramati, "Accident Prediction Accuracy Assessment for Highway-Rail Grade Crossings Using Random Forest Algorithm Compared with Decision Tree," *Reliab. Eng. Syst. Saf.*, vol. 200, p. 106931, Aug. 2020, doi: 10.1016/j.ress.2020.106931.

[40] P. H. Putra, A. Azanuddin, B. Purba, and Y. A. Dalimunthe, "Random forest and decision tree algorithms for car price prediction," *J. Mat. Dan Ilmu Pengetah. Alam LLDikti Wil. 1*, vol. 3, no. 2, pp. 81–89, Apr. 2023, doi: 10.54076/jumpa.v3i2.305.

[41] E. A. Afify, A. Sharaf, and A. E., "Facebook Profile Credibility Detection using Machine and Deep Learning Techniques based on User's Sentiment Response on Status Message," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 12, 2020, doi: 10.14569/IJACSA.2020.0111273.

[42] M. Mahmud *et al.*, "Implementation of C5.0 Algorithm using Chi-Square Feature Selection for Early Detection of Hepatitis C Disease," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 6, no. 2, pp. 116–124, Mar. 2024, doi: 10.35882/jeeemi.v6i2.384.

[43] S. A. Hicks *et al.*, "On evaluation metrics for medical applications of artificial intelligence," *Sci. Rep.*, vol. 12, no. 1, p. 5979, Apr. 2022, doi: 10.1038/s41598-022-09954-8.

[44] R. Kundu, R. Das, Z. W. Geem, G.-T. Han, and R. Sarkar, "Pneumonia detection in chest X-ray images using an ensemble of deep learning models," *PLoS One*, vol. 16, no. 9, p. e0256630, Sep. 2021, doi: 10.1371/journal.pone.0256630.

[45] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, Dec. 2020, doi: 10.1186/s12864-019-6413-7.

[46] A. A. Salih and A. M. Abdulazeez, "Evaluation of Classification Algorithms for Intrusion Detection System: A Review," *J. Soft Comput. Data Min.*, vol. 02, no. 01, Apr. 2021, doi: 10.30880/jscdm.2021.02.01.004.