Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics e-ISSN: 2656-8624
Homepage: https://ijeeemi.org/; Vol. 8, No. 4, pp. 614-632, November 2026

RESEARCH PAPER **OPEN ACCESS**

Analyzing the Contribution of Bioimpedance and Laboratory Feature Groups to Machine Learning-Based Gallstone Disease Classification

Leo Cristian Ferry and Yuni Yamasari

Department of Informatics, Universitas Negeri Surabaya, Surabaya, Indonesia

Corresponding author: Yuni Yamasari (e-mail : yuniyamasari@unesa.ac.id), **Author(s) Email**: Leo Cristian Ferry (e-mail: leo.23317@mhs.unesa.ac.id)

Abstract **Gallstone disease classification using structured patient data can benefit from different measurement sources, but the contribution of complete bioimpedance and laboratory feature groups beyond routine variables remains unclear. This study compared predefined measurement groups, rather than selecting individual predictors, to assess their incremental predictive value and consistency across models. The UCI Gallstone Dataset contained 319 participants and 38 predictors assigned to demographic, clinical, bioimpedance, and laboratory groups. Four scenarios combined a demographic and clinical baseline with bioimpedance, laboratory measurements, or both. Logistic Regression, Support Vector Machine, Random Forest, and XGBoost were tuned using GridSearch and Sand Cat Swarm Optimization (SCSO). Each optimizer used 100 fitness evaluations based on stratified five-fold cross-validation within the training data, followed by evaluation on the same held-out test set of 64 participants. Stability was examined across five repeated stratified partitions for GridSearch and five independent SCSO seeds on the original partition, while uncertainty was assessed using 4,000 stratified paired bootstrap resamples. Averaged across classifiers, adding bioimpedance improved held-out accuracy by 0.195 with GridSearch and 0.211 with SCSO; corresponding gains for laboratory measurements were 0.242 and 0.254. Mean ROC-AUC gains were 0.240 and 0.261 for bioimpedance and 0.303 and 0.327 for laboratory measurements. The highest held-out accuracy was 0.797 for Logistic Regression in Scenario C under both optimizers and Scenario D under SCSO; the best F1-score (0.817) and ROC-AUC (0.886) occurred in Scenario D with SCSO and GridSearch, respectively. Bootstrap intervals supported both additions to the baseline, but did not establish that laboratory gains were statistically greater than bioimpedance gains. The incremental value of bioimpedance after laboratory measurements remained uncertain. Neither optimizer consistently outperformed the other on held-out data. These findings require external validation because they derive from a small, single-center dataset.**

Keywords **Gallstone disease; Bioimpedance; Laboratory measurements; Machine learning; Feature group analysis.**

I. Introduction

Gallstone disease is a common gastrointestinal disorder with a substantial global burden. A systematic review and meta-analysis involving more than 32 million participants reported a pooled prevalence of approximately 6.1%, higher among women and increasing with age [1]. Another review of population-based studies published between 2000 and 2024 estimated a prevalence of 5.86% and identified substantial geographic variation [2]. Although gallstones may remain asymptomatic, symptomatic disease can cause biliary complications requiring diagnostic evaluation and treatment [1]. The prevalence of the disease and the consequences of symptomatic progression motivate investigation of patient information that may contribute to gallstone classification.

Clinical evaluation of gallstone disease integrates patient characteristics, laboratory findings, and imaging. Ultrasonography is widely used to assess gallbladder abnormalities because of its accessibility and diagnostic performance [3]. Other imaging modalities may be appropriate in selected circumstances. Demographic, clinical, and laboratory variables provide complementary information about patient characteristics and metabolic conditions. Bioelectrical impedance analysis (BIA) offers non-invasive estimates of body composition, fat distribution, and fluid compartments that are not fully described by routine anthropometric measurements. A systematic review of benign gastrointestinal diseases indicated that BIA measurements can reflect body composition and health status, although interpretation depends on standardized acquisition procedures [4]. The availability of these data sources raises the question of whether they contribute distinct predictive information when combined with commonly recorded clinical variables.

Machine learning approaches to gallbladder disease have used both imaging and structured patient data. Putra et al. [5] combined convolutional neural network feature extraction from ultrasound images with feature selection, conventional classifiers, and GridSearch optimization. For structured data, Esen et al. [6] introduced the Gallstone Dataset, containing 319 participants and 38 predictors spanning demographic characteristics, clinical conditions, laboratory measurements, and bioimpedance estimates.

Manuscript received 7 October 2026; Accepted 9 November 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.407
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Gradient Boosting achieved a reported accuracy of 85.42%, while vitamin D, C-reactive protein, total body water, and lean mass were among the influential predictors. Although this study established the dataset used in later investigations, it did not separately measure the incremental predictive contribution of complete measurement groups.

Subsequent studies have emphasized feature selection and model optimization. Sarker et al. [7] used Sand Cat Swarm Optimization (SCSO) to select 13 predictors for Random Forest and applied SHAP and DiCE for model interpretation. Shrestha et al. [8] combined Rotational Forest and Bald Eagle Search to retain 17 predictors. Das et al. [9] applied SLOA-optimized CatBoost with 19 selected features, while Tamer et al. [10] used an ensemble feature-selection approach to obtain an 11-feature Random Forest configuration. Table 1 summarizes the datasets, feature strategies, validation procedures, and reported accuracies of these investigations. The ultrasound study addresses a different classification task, whereas studies of structured gallstone data differ in feature subsets and evaluation protocols. Their reported accuracies therefore cannot be used as direct evidence of the superiority of a particular method.

Previous studies have identified influential predictors and evaluated feature selection or model optimization. Still, they have not directly quantified the added value of complete bioimpedance and laboratory measurement groups against a common demographic and clinical baseline. A predictor ranked as important within a model does not indicate whether collecting its entire measurement group improves classification beyond the information already available. This distinction is relevant when additional measurements require different resources to obtain. To address this gap, four feature groups were defined before model training: demographic, clinical, bioimpedance, and laboratory. Their composition was retained throughout the experiments. Demographic and clinical variables formed the common baseline, while bioimpedance and laboratory groups were added separately and jointly. Logistic Regression, Support Vector Machine, Random Forest, and XGBoost were evaluated using GridSearch and SCSO. Unlike earlier applications of SCSO for feature selection, SCSO was used here for hyperparameter optimization [11].

The study also investigates the reliability of feature-group comparisons. Repeated stratified train-test partitions assess sensitivity to data splitting under GridSearch, independent SCSO runs examine optimization variability on the fixed primary partition, and paired bootstrap confidence intervals quantify uncertainty in held-out performance contrasts. Supplementary analyses address predictor redundancy, preprocessing alternatives, Elastic Net regularization, laboratory feature attributions, and exploratory clinical utility. Together, these evaluations help distinguish consistent feature-group patterns from results dependent on a particular model or data partition. They provide evidence of internal stability, but not external clinical validation. The main contributions are:

1) Organizing the 38 predictors into four predefined measurement groups to examine their collective contributions without selecting individual features.

2) Comparing bioimpedance and laboratory additions separately and jointly against the same demographic–clinical baseline.

3) Evaluating the four scenarios across four classifier families and two hyperparameter optimization methods to examine consistency of the feature group patterns.

4) Assessing uncertainty and internal stability through repeated partitions, multi-seed SCSO, paired bootstrap analysis, and supplementary sensitivity experiments.

II. Materials and Method

A. Dataset and Data Collection

Table 1. Comparison of previous gallstone disease classification studies based on dataset characteristics, feature-selection approaches, machine learning classifiers, validation strategies, and reported classification accuracy.

Study	Dataset	Feature approach and classifier	Validation	Accuracy (%)
Putra et al. [5]	10,692 ultrasound images, 9 classes	VGG16, SelectKBest, SVM	80:20 split; 10-fold CV for tuning	99.35
Esen et al. [6]	UCI, 319 patients	38 features; ANOVA ranking, GB	70:30 split	85.42
Sarker et al. [7]	UCI, 319 patients	SCSO feature selection, 13 features; RF	10-fold CV	78.32
Shrestha et al. [8]	UCI, 319 patients	BES selection, 17 features; RoF	10 evaluation folds with 70:30 partitions	75.78
Das et al. [9]	UCI, 319 patients	SLOA selection, 19 features; CatBoost	5-fold CV	80.42
Tamer et al. [10]	UCI, 319 patients	Ensemble selection, 11 features; RF	70:30 split	90.62

Note: GB = Gradient Boosting; RF = Random Forest; RoF = Rotational Forest; CV = cross-validation. Reported accuracies are not directly comparable because the studies differ in datasets, training–testing partitions, feature-selection procedures, and validation protocols. The ultrasound study addresses a different classification task.

The Gallstone Dataset used in this study was introduced by Esen et al. [6] and is publicly available through the UCI

Manuscript received 7 October 2026; Accepted 9 November 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.407
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Machine Learning Repository [12]. The data were collected at the Internal Medicine Outpatient Clinic of Ankara VM Medical Park Hospital between June 2022 and June 2023 under Ankara City Hospital Ethics Committee approval E2-23-4632. The present study uses the published dataset as secondary data and involves no additional patient recruitment or direct data collection. The analyzed dataset contains 319 observations, comprising 161 individuals with gallstone disease and 158 controls, with 38 predictor variables and the binary target Gallstone Status. The predictors cover demographic, clinical, bioimpedance, and laboratory information. The released dataset contains no missing values, and categorical and binary variables are numerically coded [12]. No class balancing was applied because the class distribution was nearly balanced. In the original dataset, Gallstone Status is coded as 0 for gallstone disease and 1 for controls. For model training and performance evaluation, these labels were recoded so that gallstone disease represented the positive class (1) and controls represented the negative class (0).

B. Feature Grouping and Experimental Scenarios

The 38 predictor variables were assigned to four predefined groups according to their measurement sources and clinical characteristics: demographic, clinical, bioimpedance, and laboratory. The grouping was established before model training and remained unchanged throughout the main experiments. This approach was intended to evaluate the predictive contribution of complete measurement groups rather than identify an optimized subset of individual features. Table 2 lists the variables included in each group.

The demographic group contains five variables describing patient characteristics and routine anthropometric measurements: age, gender, height, weight, and Body Mass Index (BMI). Six variables representing recorded comorbidities and related conditions, including Hepatic Fat Accumulation (HFA), were assigned to the clinical group. The bioimpedance group comprises 14 estimates of body composition, fat distribution, and fluid compartments, whereas the laboratory group contains 13 biochemical and hematological variables. These distinctions reflect the type of patient information represented by each group rather than statistical independence among its predictors.

Particular consideration was given to the distinction between anthropometric measurements and bioimpedance-derived estimates. BMI was placed in the demographic group because it is calculated from height and weight without requiring BIA. In contrast, body fat ratio, lean mass, muscle mass, and obesity-related estimates were assigned to the bioimpedance group because they represent body composition information obtained through that measurement approach. Some variables within this group are closely related. Total body water is associated with extracellular and intracellular water, and the ECF/TBW ratio incorporates information from the fluid compartments. Lean mass, muscle mass, and body protein estimates also describe overlapping aspects of body composition, while fat-related estimates may be associated with BMI. Such relationships mean that the 14 bioimpedance variables should not be

Table 2. Definition and composition of the four predefined feature groups used for machine learning-based gallstone disease classification, including demographic, clinical, bioimpedance, and laboratory variables.

Feature group	No. of features	Variables
Demographic	5	Age, Gender, Height, Weight, Body Mass Index (BMI)
Clinical	6	Comorbidity, Coronary Artery Disease (CAD), Hypothyroidism, Hyperlipidemia, Diabetes Mellitus (DM), Hepatic Fat Accumulation (HFA)
Bioimpedance	14	Total Body Water (TBW), Extracellular Water (ECW), Intracellular Water (ICW), ECF/TBW, Total Body Fat Ratio (TBFR), Lean Mass (LM), Body Protein Content (Protein), Visceral Fat Rating (VFR), Bone Mass (BM), Muscle Mass (MM), Obesity, Total Fat Content (TFC), Visceral Fat Area (VFA), Visceral Muscle Area (VMA)
Laboratory	13	Glucose, Total Cholesterol (TC), Low Density Lipoprotein (LDL), High Density Lipoprotein (HDL), Triglyceride, Aspartate Aminotransferase (AST), Alanine Aminotransferase (ALT), Alkaline Phosphatase (ALP), Creatinine, Glomerular Filtration Rate (GFR), C-Reactive Protein (CRP), Hemoglobin (HGB), Vitamin D

Table 3. Composition of the four experimental feature scenarios used to quantify the incremental predictive contributions of bioimpedance and laboratory measurements relative to the common demographic–clinical baseline.

Scenario	Included feature groups	No. of features
A – Baseline	Demographic + Clinical	11
B – Baseline + Bioimpedance	Demographic + Clinical + Bioimpedance	25
C – Baseline + Laboratory	Demographic + Clinical + Laboratory	24
D – Full	Demographic + Clinical + Bioimpedance + Laboratory	38

Manuscript received 7 October 2026; Accepted 9 November 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.407
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

interpreted as independent sources of information. The variables were nevertheless retained as a complete group to preserve the intended comparison of measurement sources. Their statistical relationships and potential multicollinearity were examined in the additional analyses.

Four experimental scenarios were constructed from these groups, as shown in Table 3. Scenario A combines demographic and clinical variables and serves as the common baseline. Scenario B adds bioimpedance variables to the baseline, while Scenario C adds laboratory variables. Scenario D includes all four groups. Comparisons of B with A and C with A assess the incremental contribution of bioimpedance and laboratory information beyond the same baseline. The D−B and D−C comparisons examine the additional contribution of each group when the other is already included. All scenarios were evaluated using the same classifiers and optimization procedures, with hyperparameters selected separately for each classifier–scenario combination. No feature selection was applied in the main experiments, allowing changes in predictive performance to be assessed without altering the predefined composition of the measurement groups.

C. Data Partitioning and Preprocessing

The dataset was divided into training and test sets using a stratified 80:20 partition with random seed 42, resulting in 255 training observations and 64 held-out observations. Stratification preserved the relative proportions of gallstone cases and controls [13]. The training set was used for model fitting and hyperparameter selection through stratified five-fold cross-validation, with shuffled folds generated using the same random seed [14]. The primary experiment used identical training and test indices, as well as the same cross-validation folds, across all four feature scenarios, classifiers, and optimization methods. The held-out test set was excluded from model fitting and hyperparameter selection.

The released dataset contained no missing values, so imputation was not required. All predictors were numerically represented, including categorical and binary variables [12]. In the main experiments, binary predictors retained their original 0/1 coding before preprocessing, while Comorbidity and Hepatic Fat Accumulation (HFA) were used with their existing numerical codes. Comorbidity ranged from 0 to 3 and HFA from 0 to 4; neither variable was converted into dummy indicators in the primary pipeline.

StandardScaler was applied to all predictors, including binary variables, within the classification pipeline. For each cross-validation fold, the scaler was fitted using only the training subset and then applied to the corresponding validation subset to prevent information leakage [14], [15]. After hyperparameter selection, the complete pipeline was refitted on the 255 training observations before evaluating the held-out test set. Although feature scaling is particularly relevant to Logistic Regression and Support Vector Machine, the same preprocessing procedure was retained for Random Forest and XGBoost to maintain consistency across classifiers [16]. Alternative scaling and encoding configurations were evaluated separately through sensitivity analysis, as described in Section II-G.

D. Classification Models

Four supervised classifiers were selected to represent different learning mechanisms commonly used with structured medical data: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and XGBoost [17], [18], [19], [20]. LR provides a linear probabilistic model, SVM supports nonlinear classification through kernel functions, RF combines decision trees using bagging, and XGBoost applies gradient boosting. All classifiers were evaluated using the same feature scenarios, preprocessing procedure, and validation partitions to examine the consistency of feature-group contributions across model families.

1) Logistic Regression: LR was used as the linear reference classifier. For a predictor vector x, the probability of gallstone disease is expressed using the logistic function in Eq. (1) [14], [20]:

$$P(y = 1 \mid x) = \frac{1}{1 + \exp\left[-\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right)\right]} \quad (1)$$

where β_0 is the intercept, β_j is the coefficient of the predictor x_j, and p is the number of predictors. L2 regularization was applied using the LBFGS solver, with a maximum of 2,000 iterations. The inverse regularization strength C was optimized; smaller values impose stronger regularization.

2) Support Vector Machine: SVM with the Radial Basis Function (RBF) kernel was used to model nonlinear decision boundaries [18], [20]. The soft-margin optimization problem and its constraints are given in Eq. (2) and Eq. (3) [18], [20]:

$$\min_{w,b,\xi} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} \xi_i \quad (2)$$

$$y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i,\ \xi_i \geq 0 \quad (3)$$

where w is the weight vector, b is the bias, ξ_i is the slack variable, and C controls the penalty for margin violations. The function ϕ represents the feature mapping, and $y_i \in \{-1, +1\}$ denotes the class label in the SVM formulation. The RBF kernel is defined in Eq. (4) [18]:

$$K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2) \quad (4)$$

where γ controls the influence of distances between observations. Both C and γ were optimized, while the RBF kernel remained fixed.

3) Random Forest: RF combines decision trees trained using bootstrap samples and randomly selected predictor subsets [18], [20]. Gini impurity was used as the node-splitting criterion and is calculated using Eq. (5) [18], [20]:

$$Gini(t) = 1 - \sum_{c=1}^{C} p_{c,t}^2 \quad (5)$$

where $p_{c,t}$ represents the proportion of observations belonging to class c at node t. The final classification is obtained through majority voting, as expressed in Eq. (6) [18], [20]:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \ldots, h_M(x)\} \quad (6)$$

where $h_m(x)$ is the prediction of tree m and M is the number of trees. The number of trees, maximum depth, minimum split size, and minimum leaf size were

Manuscript received 7 October 2026; Accepted 9 November 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.407
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

optimized, while bootstrap sampling and the Gini criterion remained fixed.

4) XGBoost: XGBoost constructs an ensemble of trees sequentially, with each additional tree improving the model's predictions [18], [20]. The additive update at iteration t is given in Eq. (7) [14], [20]:

$$F_i^{(t)} = F_i^{(t-1)} + f_t(\mathrm{x}_i) \tag{7}$$

Here, $F_i^{(t)}$ is the ensemble score for observation i, and f_t is the newly added tree. For binary classification, the score is transformed into a probability using the logistic function. The regularized training objective combines prediction loss and tree complexity, as shown in Eq. (8) [14], [20]:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} \ell\left(y_i, \sigma\left(F_i^{(t)}\right)\right) + \sum_{k=1}^{t} \Omega\left(f_k\right). \tag{8}$$

Here, ℓ represents binary classification loss, σ is the logistic function, and Ω penalizes tree complexity. The model used a binary logistic objective, while the number of estimators, maximum depth, and learning rate were optimized.

E. Hyperparameter Optimization

Hyperparameter optimization was performed independently for each classifier and feature scenario. GridSearch and Sand Cat Swarm Optimization (SCSO) used identical stratified five-fold cross-validation partitions and mean validation accuracy as the optimization objective [21], [22]. Preprocessing was fitted within each training fold to prevent information leakage. For a hyperparameter configuration θ, the fitness function is defined in Eq. (9) [21]:

$$f(\theta) = \frac{1}{K}\sum_{k=1}^{K} \mathrm{Accuracy}_k(\theta), \qquad K = 5. \tag{9}$$

where $Accuracy_k(\theta)$ represents the classification accuracy obtained on the validation fold k. Accuracy was selected as the optimization objective because the dataset was nearly balanced.

1) GridSearch: GridSearch evaluated 100 predefined hyperparameter configurations for each classifier–scenario combination. The search spaces covered different regularization strengths and model complexities within a fixed evaluation budget. For LR and SVM, the regularization parameter C and RBF kernel coefficient γ were examined on logarithmic scales [21]. RF candidates varied ensemble size, tree depth, and minimum sample requirements, while XGBoost candidates covered the number of estimators, tree depth, and learning rate. The candidate values are listed in Table 4. Each configuration was evaluated using Eq. (9), and the one with the highest mean cross-validation accuracy was selected. The procedure is summarized in Algorithm 1.

Algorithm 1. GridSearch-based hyperparameter optimization using stratified five-fold cross-validation to maximize mean classification accuracy across predefined parameter combinations.

Input: Training data D_{train}, classifier M, predefined feature scenario S, parameter grid G, and five fixed stratified CV folds

Output: Optimal hyperparameter configuration θ^* and corresponding mean five-fold cross-validation accuracy f^*

```
Initialize best fitness f* = -∞
Generate 100 configurations from G (Table 4)
for each configuration θ in G do
    for k = 1 to 5 do
        Fit the preprocessing and classifier pipeline on training fold k using θ
        Predict labels on validation fold k
        Calculate validation accuracy
    end for
    Compute mean CV accuracy f(θ) using Eq. (9)
    if f(θ) > f* then
        Set f* = f(θ) and θ* = θ
    end if
end for
Return θ* and f*
```

2) Sand Cat Swarm Optimization: SCSO is a population-based metaheuristic inspired by the searching and attacking behavior of sand cats [11], [23]. Each agent represented a candidate hyperparameter configuration rather than a feature subset. The algorithm updated candidate positions through exploration and exploitation while retaining the predefined feature composition of each scenario. At update iteration t, the global sensitivity parameter $r_G^{(t)}$, agent sensitivity $r_i^{(t)}$, and search-phase control variable $R_i^{(t)}$ were calculated using Eq. (10) [11]:

$$\begin{aligned} r_G^{(t)} &= S_M\left(1 - \frac{t}{T}\right), \\ r_i^{(t)} &= r_G^{(t)} u_i^{(t)}, \\ R_i^{(t)} &= r_G^{(t)}\left(2v_i^{(t)} - 1\right). \end{aligned} \tag{10}$$

Here, $S_M = 2$ is the maximum initial sensitivity, $T = 9$ is the number of population update iterations, and $t = 0, \ldots, T-1$. The variables $u_i^{(t)}$ and $v_i^{(t)}$ are independently sampled from a uniform distribution over $[0,1]$. The magnitude of $R_i^{(t)}$ determines whether the agent explores or exploits the search space. When $\left|R_i^{(t)}\right| > 1$, exploration updates each coordinate relative to a randomly selected population member, as expressed in Eq. (11) [11]:

$$X_{i,j}^{(t+1)} = r_i^{(t)}\left(X_{q,j}^{(t)} - w_{i,j}^{(t)} X_{i,j}^{(t)}\right) \tag{11}$$

where $X_{i,j}^{(t)}$ represents coordinate j of agent i, $X_{q,j}^{(t)}$ is the corresponding coordinate of a randomly selected population member, and $w_{i,j}^{(t)}$ is a random value drawn uniformly from $[0,1]$.

When $|R| \le 1$, exploitation uses the distance in Eq. (12) and position update in Eq. (13) [11]:

$$D_{i,j}^{(t)} = \left|a_{i,j}^{(t)} X_{\mathrm{best},j}^{(t)} - X_{i,j}^{(t)}\right| \tag{12}$$

$$X_{i,j}^{(t+1)} = X_{\mathrm{best},j}^{(t)} - r_i^{(t)} D_{i,j}^{(t)} \cos\left(\alpha_{i,j}^{(t)}\right) \tag{13}$$

Here, $X_{\mathrm{best},j}^{(t)}$ denotes coordinate j of the best solution found up to iteration t, and $a_{i,j}^{(t)}$ is sampled uniformly from $[0,1]$. The direction parameter $\alpha_{i,j}^{(t)}$ was selected through roulette-wheel sampling from 360 integer labels, with

Manuscript received 7 October 2026; Accepted 9 November 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.407
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Table 4. Hyperparameter search spaces and optimization domains for Logistic Regression, Support Vector Machine, Random Forest, and XGBoost under GridSearch and Sand Cat Swarm Optimization (SCSO).

Model	Hyperparameter	GridSearch	SCSO search domain
Logistic Regression	C	100 log-spaced values from 0.001 to 100	10^{-3}–10^{2}, log scale
SVM (RBF)	C	10 log-spaced values from 0.01 to 100	10^{-2}–10^{2}, log scale
	γ	10 log-spaced values from 0.0001 to 1	10^{-4}–10^{0}, log scale
Random Forest	n_estimators	50, 100, 150, 225, 300	Integer, 50–300
	max_depth	5, 10, 15, 20, 25	Integer, 5–25
	min_samples_split	2, 5	Integer, 2–5
	min_samples_leaf	1, 2	Integer, 1–2
XGBoost	n_estimators	50, 100, 150, 225, 300	Integer, 50–300
	max_depth	2, 3, 4, 6, 8	Integer, 2–8
	learning_rate	0.01, 0.05, 0.10, 0.20	Continuous, 0.01–0.20

Note: GridSearch evaluates predefined parameter combinations, whereas SCSO explores bounded parameter domains. Both optimization methods use a budget of 100 fitness evaluations per classifier–scenario combination.

probabilities proportional to their values. The selected label was used directly in the cosine calculation, following the implemented procedure. Candidate positions were clipped to the predefined search bounds. Integer-valued parameters were rounded to valid integers, while C and γ were optimized in logarithmic coordinates and transformed back before model evaluation. The search terminated after nine population updates, corresponding to exactly 100 fitness evaluations, including the initial population.

Algorithm 2. Sand Cat Swarm Optimization (SCSO) for classifier hyperparameter tuning using population-based exploration and exploitation with stratified five-fold cross-validation accuracy as the fitness function.

Input: Training data, classifier, predefined feature scenario, parameter bounds (Table 4), five fixed stratified CV folds, population size of 10 agents, and a maximum budget of 100 fitness evaluations.

Output: Best decoded hyperparameter configuration θ^* and corresponding mean five-fold cross-validation accuracy f^*

```
Initialize 10 agents uniformly within the encoded parameter bounds (seed = 42)
Decode and evaluate each agent using Eq. (9)
Store the best position Xbest and fitness f*
Set evaluation count E = 10 and maximum population updates T = 9 (100 fitness evaluations in total)
for t = 0 to T - 1 do
    Compute global sensitivity rG using Eq. (10)
    Copy the current population
    for each agent i = 1 to 10 do
        Sample agent sensitivity ri and phase variable Ri using Eq. (10)
        for each parameter coordinate j do
            Select α by roulette-wheel sampling
            if |Ri| > 1 then
                Update position using Eq. (11)
            else
                Compute distance using Eq. (12)
                Update position using Eq. (13)
            end if
        end for
        Clip the updated position to its bounds
    end for
    Replace the population with all updated agents
    Decode and evaluate all 10 agents using mean five-fold CV accuracy
    Update Xbest and f* if a better solution is found
    Set E = E + 10
end for
Return decoded Xbest as θ* and fitness f*
```

The SCSO procedure is summarized in Algorithm 2. GridSearch and SCSO used the same parameter bounds, cross-validation folds, and budget of 100 fitness evaluations per classifier–scenario combination. However, GridSearch evaluated a predefined discrete grid, whereas SCSO explored numerical parameter domains and could revisit configurations after parameter conversion. Thus, equal evaluation budgets did not necessarily imply identical search coverage or computational time. The primary experiment comprised 32 combinations of four classifiers, four feature scenarios, and two optimizers using random seed 42. For each combination, the optimal hyperparameters were selected by mean cross-validation accuracy. The corresponding preprocessing and classification pipeline was then refitted on all 255 training observations and evaluated on the same 64 held-out observations, as illustrated in Fig. 1. Variability across independent SCSO seeds was assessed separately in the robustness analysis.

F. Evaluation and Comparative Analysis

Model performance was evaluated at two stages. During hyperparameter selection, accuracy was assessed through stratified five-fold cross-validation using only the 255 training observations and reported as the mean ± standard deviation across folds. After selecting the optimal hyperparameters, each preprocessing and classification pipeline was refitted on the complete training set and evaluated on the 64 held-out observations [14], [24], [25]. The test set was not used to select models or hyperparameters.

Manuscript received 7 October 2026; Accepted 9 November 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.407
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Five primary performance metrics were used: Accuracy, Precision, Recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (ROC-AUC), with gallstone disease treated as the positive class [20], [25]. Accuracy was used as the hyperparameter optimization objective because the dataset was nearly balanced. The threshold-dependent metrics are defined in Eq. (14), Eq. (15), Eq. (16), Eq. (17) [24], [25]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (14)$$

$$Precision = \frac{TP}{TP + FP} \quad (15)$$

$$Recall = \frac{TP}{TP + FN} \quad (16)$$

$$F1 = 2 \cdot \frac{Precision \times Recall}{Precision + Recall} \quad (17)$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. ROC curves were generated from continuous model outputs, using predicted probabilities for LR, RF, and XGBoost and decision scores for SVM. ROC-AUC was calculated to summarize discrimination across classification thresholds [24], [25].

Feature-group contributions were evaluated within the same classifier and optimization method. For a performance metric M, the changes associated with adding bioimpedance and laboratory variables to the baseline are defined in this study in Eq. (18) and Eq. (19):

$$\Delta_{\mathrm{BIA}} = M_B - M_A \quad (18)$$

$$\Delta_{\mathrm{Lab}} = M_C - M_A \quad (19)$$

where M_A, M_B, and M_C are the corresponding metric values for Scenarios A, B, and C. The additional contributions obtained in the Full scenario relative to the Bioimpedance and Laboratory scenarios are defined in Eq. (20) and Eq. (21):

$$\Delta_{\mathrm{Full|BIA}} = M_D - M_B \quad (20)$$

$$\Delta_{\mathrm{Full|Lab}} = M_D - M_C \quad (21)$$

where M_D represents the metric value for Scenario D. Positive differences indicate improved performance relative to the reference scenario, whereas negative differences indicate decreased performance. These comparisons correspond to B−A, C−A, D−B, and D−C, respectively. The contrasts were calculated separately for each classifier and optimization method, and their averages across classifiers were used to summarize the overall feature-group patterns. The difference between

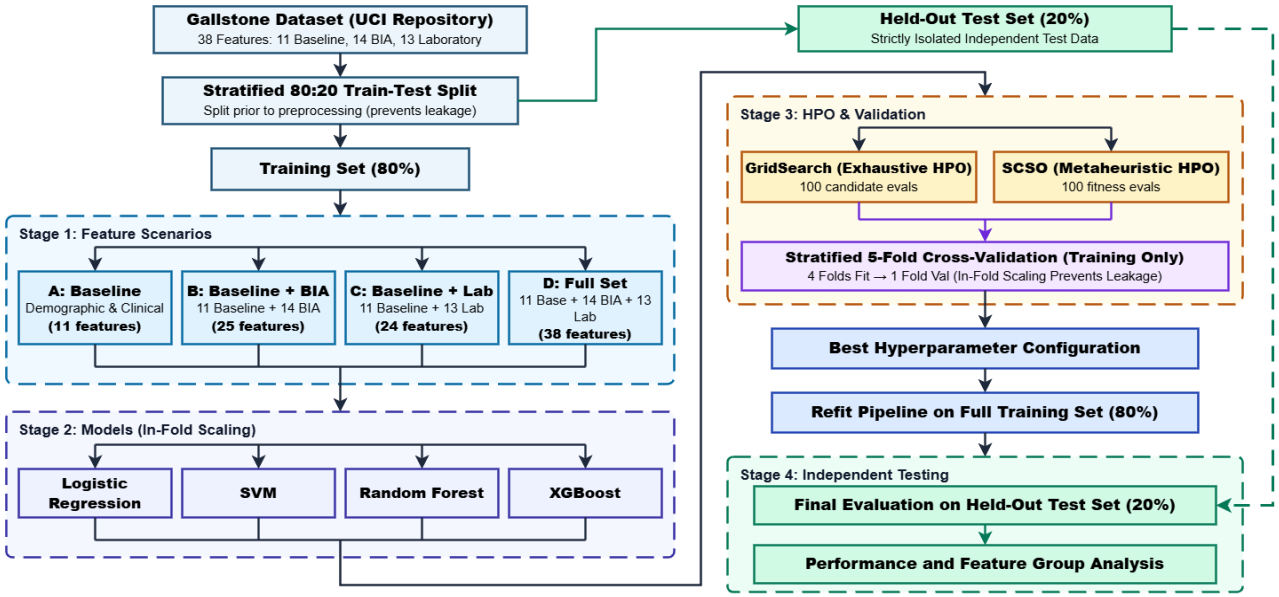


Fig. 1. Workflow of the proposed machine learning framework for gallstone disease classification and predefined feature-group comparison. The framework includes stratified training–testing partitioning, four feature scenarios, four classifiers, five-fold cross-validation, hyperparameter optimization using GridSearch or Sand Cat Swarm Optimization (SCSO), and evaluation on an independent held-out test partition.

Additional clinical classification measures included sensitivity, specificity, balanced accuracy, and the Matthews Correlation Coefficient (MCC), together with confusion matrices for selected configurations. Sensitivity was equivalent to recall, while specificity represented the proportion of correctly classified controls, calculated as $TN/(TN + FP)$. Balanced accuracy was calculated as the mean of sensitivity and specificity. MCC incorporated all four confusion-matrix outcomes to assess classification performance while accounting for both positive and negative predictions. These measures were included to examine the balance between identifying gallstone cases and avoiding false-positive classifications.

SCSO and GridSearch for the same classifier, scenario, and performance metric is defined in Eq. (22):

$$\Delta_{\mathrm{HPO}} = M_{\mathrm{SCSO}} - M_{\mathrm{GridSearch}} \quad (22)$$

This comparison was used to assess the influence of hyperparameter optimization on predictive performance and the consistency of feature-group conclusions. Performance differences were examined descriptively across classifier families and optimization methods. Their uncertainty was further assessed using paired bootstrap confidence intervals, while repeated data partitions were used to evaluate the stability of the scenario comparisons.

Manuscript received 7 October 2026; Accepted 9 November 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.407
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

The procedures for these additional analyses are described in Section II-G.

G. Robustness and Additional Analyses

To assess the stability of feature-group contributions, GridSearch experiments were repeated across five stratified 80:20 train-test partitions using random seeds 42, 52, 62, 72, and 82. Five partitions provided a computationally manageable internal sensitivity assessment, although they were not intended to capture the full range of sampling variability. Hyperparameters were independently optimized within each training partition using stratified five-fold cross-validation. Identical partitions and validation folds were used across classifiers and feature scenarios, resulting in 80 classifier–scenario evaluations. SCSO variability was examined separately using five optimization seeds (42, 52, 62, 72, and 82), while keeping the original data partition and cross-validation folds fixed. Each SCSO run used 100 fitness evaluations, with held-out performance, best-so-far fitness history, and computation time recorded. These analyses distinguished partition-dependent performance changes from stochastic optimization variability. However, the repeated partitions drew from the same 319 participants and could share test observations, providing limited evidence of internal stability rather than independent external validation.

Statistical uncertainty was assessed using 4,000 stratified paired bootstrap resamples of the original 64 held-out observations. Cases and controls were resampled with replacement within their respective classes, and the same patient indices were used for each paired comparison. Percentile-based 95% confidence intervals were calculated for differences in accuracy, F1-score, and ROC-AUC across the four feature-scenario contrasts. Intervals were obtained for Logistic Regression and for contrasts averaged across the four classifiers under each optimization method. ROC-AUC confidence intervals were also estimated for Logistic Regression in all four scenarios. The bootstrap procedure used the predictions of the fitted models without repeating model training or hyperparameter optimization. Consequently, these intervals describe uncertainty conditional on the original models and test partition rather than variation across independently trained models or external populations.

Potential redundancy among predictors was examined using pairwise Spearman rank correlations within and between the predefined feature groups. Correlations were calculated descriptively from the 319 observations without using outcome labels to select or remove variables. Variance inflation factors were computed for continuous predictors, excluding binary variables and the numerically coded Comorbidity and Hepatic Fat Accumulation (HFA) variables. The stability of L2-regularized Logistic Regression coefficients was also examined across the five cross-validation training folds using the original GridSearch-selected hyperparameters. Preprocessing sensitivity was assessed by comparing the original standardization of all predictors with four alternatives: standardizing only nonbinary predictors, applying robust scaling to nonbinary predictors, omitting scaling, and one-hot encoding Comorbidity and HFA while scaling other nonbinary predictors and retaining binary indicators. These experiments used the original data partition and fixed GridSearch-selected hyperparameters, without additional tuning.

Elastic Net Logistic Regression was included as a supplementary regularized baseline. The model was evaluated across the same five stratified train-test partitions used in the repeated GridSearch experiments. Its hyperparameters were optimized independently within each training partition using stratified five-fold cross-validation. The search included 20 logarithmically spaced values of the inverse regularization strength C, ranging from 0.001 to 100, and five L1 mixing ratios (0.1, 0.3, 0.5, 0.7, and 0.9), giving 100 candidate configurations per scenario. StandardScaler was fitted within the training folds, and Elastic Net was implemented using the SAGA solver. Its performance and feature-scenario contrasts were compared with those of the original L2-regularized Logistic Regression on matching data partitions.

Additional model interpretation and exploratory clinical evaluations focused on GridSearch-optimized Logistic Regression. Laboratory predictor contributions were examined in Scenarios C and D using linear-model SHAP attributions calculated on the log-odds scale relative to the training-data mean reference. Mean absolute attribution values were used to summarize predictor importance. Because this attribution procedure uses an independent-feature reference and the predictors may be correlated, the results were interpreted as contributions to fitted predictions rather than independent physiological effects. Calibration was examined for all four scenarios using Brier scores and expected calibration error calculated from five fixed-width probability bins. Decision-curve analysis was performed over probability thresholds from 0.05 to 0.95 in increments of 0.05, comparing model net benefit with treat-all and treat-none strategies. These analyses used predictions from the original held-out test set and were considered exploratory because of the limited sample size and the absence of clinically established decision thresholds.

III. Results

A. Feature Scenario Performance

Table 5 summarizes cross-validation accuracy and held-out performance for the four feature scenarios under GridSearch and SCSO. Scenario A, comprising demographic and clinical variables, yielded mean cross-validation accuracies of 0.584–0.631 across classifiers and optimizers. Held-out accuracy ranged from 0.422 to 0.531 and ROC-AUC from 0.450 to 0.515, indicating limited discrimination using the baseline alone. Adding bioimpedance variables in Scenario B increased mean cross-validation accuracy to 0.678–0.729 and held-out accuracy to 0.656–0.719. Held-out F1-scores ranged from 0.686 to 0.735, with ROC-AUC values of 0.729–0.786. Each classifier-optimizer combination improved on its corresponding Scenario A configuration in all three held-out metrics, although the magnitude of improvement varied. Scenario C, which added laboratory measurements to the baseline, achieved mean cross-validation accuracies of 0.725–0.780 and held-out accuracies of 0.672–0.797. Held-out ROC-AUC ranged

Manuscript received 7 October 2026; Accepted 9 November 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.407
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Table 5. Comparison of five-fold cross-validation accuracy and held-out classification performance across four feature scenarios and four machine learning classifiers optimized using GridSearch and Sand Cat Swarm Optimization (SCSO).

Scenario	Model	GridSearch				SCSO			
		CV Acc.	Acc.	F1	AUC	CV Acc.	Acc.	F1	AUC
A	LR	0.620±0.063	0.484	0.492	0.497	0.620±0.063	0.484	0.492	0.497
A	SVM	0.624±0.053	0.531	0.559	0.503	0.631±0.061	0.484	0.522	0.450
A	RF	0.600±0.064	0.484	0.522	0.515	0.620±0.064	0.422	0.464	0.510
A	XGB	0.584±0.068	0.453	0.493	0.479	0.596±0.061	0.500	0.515	0.486
B	LR	0.729±0.066	0.688	0.706	0.741	0.729±0.066	0.688	0.706	0.739
B	SVM	0.718±0.074	0.719	0.735	0.740	0.725±0.055	0.703	0.725	0.729
B	RF	0.678±0.113	0.656	0.686	0.734	0.694±0.104	0.656	0.686	0.733
B	XGB	0.678±0.112	0.672	0.687	0.738	0.698±0.072	0.688	0.697	0.786
C	LR	0.725±0.062	**0.797**	0.794	0.844	0.725±0.062	**0.797**	0.794	0.846
C	SVM	0.769±0.061	0.688	0.697	0.773	0.780±0.056	0.734	0.738	0.811
C	RF	0.773±0.033	0.719	0.735	0.773	0.769±0.038	0.703	0.747	0.809
C	XGB	0.765±0.039	0.719	0.743	0.814	0.769±0.043	0.672	0.720	0.786
D	LR	0.773±0.038	0.781	**0.806**	**0.886**	0.773±0.038	**0.797**	**0.817**	**0.885**
D	SVM	**0.788±0.078**	0.766	0.762	0.855	**0.804±0.089**	0.766	0.762	0.857
D	RF	0.780±0.061	0.703	0.716	0.834	0.788±0.073	0.719	0.735	0.847
D	XGB	0.784±0.031	0.672	0.696	0.766	0.784±0.024	0.672	0.720	0.800

Note: CV accuracy is reported as mean ± standard deviation across five stratified folds. Accuracy, F1-score, and ROC-AUC were calculated on the original held-out test set. Bold values indicate the highest value within each performance column, ties are highlighted equally.

from 0.773 to 0.846. Logistic Regression obtained an accuracy of 0.797 and an F1-score of 0.794 under both optimizers, with ROC-AUC values of 0.844 for GridSearch and 0.846 for SCSO. Scenario C generally performed better than Scenario B, although not in every classifier accuracy of 0.797 was achieved in Scenario C under both optimizers and Scenario D under SCSO. However, the full feature set did not improve every metric. For GridSearch-optimized Logistic Regression, accuracy decreased from 0.797 in Scenario C to 0.781 in Scenario D.

Table 6. Mean differences in held-out accuracy, F1-score, and ROC-AUC between predefined feature scenarios under GridSearch and Sand Cat Swarm Optimization (SCSO), with 95% confidence intervals obtained from 4,000 stratified paired bootstrap resamples.

Optimizer	Contrast	ΔAccuracy (95% CI)	ΔF1-Score (95% CI)	ΔROC-AUC (95% CI)
GridSearch	B−A	+0.195 [0.070, 0.320]	+0.187 [0.072, 0.312]	+0.240 [0.117, 0.363]
	C−A	+0.242 [0.121, 0.359]	+0.226 [0.101, 0.359]	+0.303 [0.164, 0.437]
	D−B	+0.047 [−0.035, 0.125]	+0.042 [−0.039, 0.123]	+0.097 [0.011, 0.181]
	D−C	0.000 [−0.051, 0.051]	+0.003 [−0.049, 0.053]	+0.034 [−0.018, 0.085]
SCSO	B−A	+0.211 [0.086, 0.332]	+0.205 [0.089, 0.330]	+0.261 [0.139, 0.384]
	C−A	+0.254 [0.129, 0.379]	+0.251 [0.121, 0.391]	+0.327 [0.192, 0.460]
	D−B	+0.055 [−0.031, 0.145]	+0.055 [−0.030, 0.145]	+0.100 [0.012, 0.188]
	D−C	+0.012 [−0.031, 0.059]	+0.009 [−0.031, 0.049]	+0.034 [−0.015, 0.085]

Note: Values are mean differences across four classifiers, with percentile-based 95% confidence intervals from 4,000 stratified paired bootstrap resamples of the original 64 held-out observations. Intervals are conditional on the fitted models and test partition. Positive values indicate improved performance in the first scenario. Intervals crossing zero indicate uncertainty in the direction of the performance difference.

optimizer configuration.

Scenario D, comprising all 38 predictors, achieved cross-validation accuracies of 0.773–0.804, with the highest value of 0.804 obtained by SVM with SCSO. On the held-out test set, GridSearch-optimized Logistic Regression achieved a ROC-AUC of 0.886, compared with 0.497 in Scenario A. This absolute increase of 0.389 indicates improved discrimination between gallstone cases and controls within the test sample. Logistic Regression obtained the highest held-out F1-score of 0.817 with SCSO in Scenario D, while the maximum

B. Feature Group Contribution

Table 6 presents the held-out performance differences for four feature-scenario contrasts, averaged across classifiers for each optimization method, together with their 95% paired bootstrap confidence intervals. Adding bioimpedance to the baseline (B−A) improved accuracy, F1-score, and ROC-AUC for every classifier under both optimization methods. The respective mean gains with GridSearch were 0.195, 0.187, and 0.240, compared with 0.211, 0.205, and 0.261 with SCSO. Adding laboratory measurements (C−A) produced larger mean gains of

Manuscript received 7 October 2026; Accepted 9 November 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.407
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

0.242, 0.226, and 0.303 under GridSearch and 0.254, 0.251, and 0.327 under SCSO. All classifier–optimizer combinations showed positive changes, although the larger C−A averages do not establish statistical superiority of laboratory measurements over bioimpedance.

Adding laboratory variables to the Bioimpedance scenario (D−B) increased mean held-out accuracy by 0.047 with GridSearch and 0.055 with SCSO. The corresponding F1-score gains were 0.042 and 0.055, and ROC-AUC gains were 0.097 and 0.100. F1-score and ROC-AUC improved in every classifier–optimizer combination, whereas accuracy improved in three of four classifiers under each optimizer. Conversely, adding bioimpedance to the Laboratory scenario (D−C) produced mean accuracy changes of 0.000 and 0.012 and F1-score changes of 0.003 and 0.009, respectively. ROC-AUC increased by 0.034 under both methods. GridSearch accuracy decreased for three classifiers, while SCSO showed two increases and two unchanged results. The additional contribution of bioimpedance after laboratory measurements was small and varied across classifiers.

Paired bootstrap analysis using 4,000 stratified resamples showed positive mean B−A and C−A differences, with all 95% confidence intervals for accuracy, F1-score, and ROC-AUC excluding zero under both optimizers. With GridSearch, the mean accuracy gain was 0.195 (95% CI: 0.070–0.320) for B−A and 0.242 (0.121–0.359) for C−A. For D−B, the ROC-AUC improvement was 0.097 (95% CI: 0.011–0.181) with GridSearch and 0.100 (0.012–0.188) with SCSO, while the accuracy and F1-score intervals included zero. The D−C accuracy difference was 0.000 (95% CI: −0.051 to 0.051) with GridSearch and 0.012 (−0.031 to 0.059) with SCSO. All D−C confidence intervals included zero, indicating an uncertain incremental benefit from bioimpedance after laboratory measurements. Fig. 2 illustrates the accuracy contrasts and their confidence intervals. These estimates are conditional on the fitted models and original held-out test set, rather than variability across independently trained models.

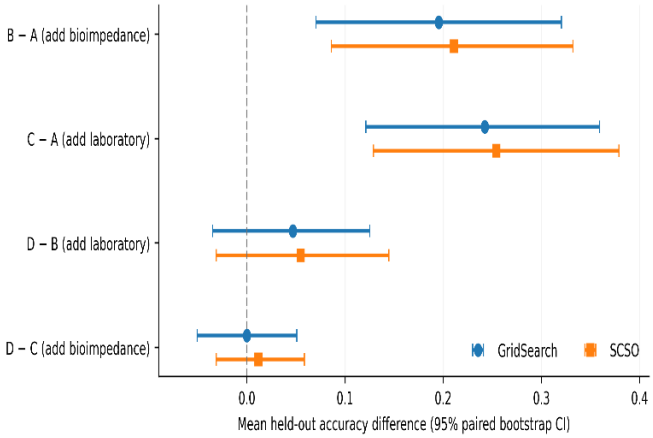


Fig. 2. Mean differences in held-out classification accuracy between predefined feature scenarios using GridSearch and Sand Cat Swarm Optimization (SCSO). Points represent mean accuracy differences across four classifiers, and error bars indicate 95% confidence intervals obtained from 4,000 stratified paired bootstrap resamples. B−A and C−A represent bioimpedance and laboratory additions to the baseline, respectively; D−B and D−C represent additional laboratory and bioimpedance contributions.

Cross-validation showed more uniform changes: B−A, C−A, and D−B improved accuracy, F1-score, and ROC-AUC for every classifier under both optimizers. For D−C, accuracy and F1-score increased in all eight classifier–optimizer combinations and ROC-AUC in seven. The held-out results were less uniform, especially for D−C, indicating that validation improvements were not always

Table 7. Optimal hyperparameter configurations selected by GridSearch and Sand Cat Swarm Optimization (SCSO), with held-out performance and optimization-time variability across five independent SCSO seeds.

Scenario–Model	GridSearch best	SCSO best	SCSO Accuracy	SCSO ROC-AUC	SCSO Time (s)
A–LR	C=0.850	C=0.981	0.484 ± 0.000	0.496 ± 0.002	7.3 ± 5.4
A–SVM	C=12.92, γ=0.129	C=11.29, γ=0.0950	0.512 ± 0.017	0.477 ± 0.017	4.1 ± 0.1
A–RF	50/5/2/1	68/7/2/1	0.456 ± 0.040	0.497 ± 0.018	56.6 ± 6.4
A–XGB	50/3/0.100	61/3/0.140	0.509 ± 0.014	0.497 ± 0.019	15.5 ± 4.5
B–LR	C=2.420	C=3.068	0.688 ± 0.000	0.740 ± 0.001	9.0 ± 2.3
B–SVM	C=12.92, γ=0.0167	C=10.55, γ=0.0161	0.703 ± 0.011	0.741 ± 0.019	4.0 ± 0.1
B–RF	225/5/2/2	171/8/4/2	0.637 ± 0.017	0.725 ± 0.011	126.8 ± 46.2
B–XGB	100/3/0.0500	146/2/0.133	0.684 ± 0.036	0.770 ± 0.022	25.6 ± 6.5
C–LR	C=1.918	C=3.068	0.787 ± 0.021	0.845 ± 0.001	5.9 ± 0.4
C–SVM	C=4.642, γ=0.0167	C=100.00, γ=0.0038	0.719 ± 0.016	0.806 ± 0.007	4.2 ± 0.2
C–RF	100/15/5/1	141/5/5/1	0.706 ± 0.007	0.787 ± 0.016	111.8 ± 41.5
C–XGB	50/3/0.100	50/2/0.0309	0.672 ± 0.000	0.785 ± 0.001	16.0 ± 3.3
D–LR	C=2.154	C=3.068	0.787 ± 0.009	0.886 ± 0.001	6.8 ± 0.4
D–SVM	C=100.00, γ=0.0060	C=100.00, γ=0.0047	0.753 ± 0.020	0.854 ± 0.011	4.4 ± 0.2
D–RF	225/10/2/2	80/14/3/2	0.725 ± 0.009	0.841 ± 0.007	124.3 ± 37.6
D–XGB	150/3/0.0100	201/2/0.0100	0.741 ± 0.039	0.836 ± 0.024	38.8 ± 14.6

Note: Optimal parameters correspond to the primary experiment with seed 42. RF parameters are listed as n_estimators/max_depth/min_samples_split/min_samples_leaf, and XGBoost parameters as n_estimators/max_depth/learning_rate. LR and SVM use C as the regularization parameter, with γ denoting the RBF kernel coefficient. SCSO performance and time are reported as mean ± standard deviation over five independent optimizer seeds (42, 52, 62, 72, and 82) using the same train-test partition. Parameter values are rounded for presentation.

Manuscript received 7 October 2026; Accepted 9 November 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.407
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

retained during testing.

The repeated GridSearch experiment assessed stability across five stratified train-test partitions. Mean accuracy differences were 0.150 ± 0.036 for B−A, 0.210 ± 0.080 for C−A, 0.077 ± 0.069 for D−B and 0.017 ± 0.020 for D−C, calculated by averaging classifier differences within each partition and then summarizing the five partition-level means. B−A and C−A improved accuracy in all 20 classifier–partition comparisons. D−B improved in 16 comparisons, was unchanged in one, and decreased in three; D−C improved in seven, was unchanged in four, and decreased in nine. Across the five evaluated partitions, both additions to the baseline consistently improved accuracy, whereas the additional contribution of bioimpedance after laboratory measurements remained uncertain.

C. GridSearch and SCSO Comparison

The performance differences between GridSearch and SCSO were generally smaller than those between feature scenarios. During cross-validation, SCSO achieved higher accuracy in 10 of the 16 classifier–scenario combinations, matched GridSearch in five, and performed worse in one. The mean cross-validation differences (SCSO minus GridSearch) were +0.007 for accuracy, +0.007 for F1-score, and +0.005 for ROC-AUC. On the held-out test set, SCSO achieved higher accuracy in five combinations, matched GridSearch in six, and performed worse in five. The corresponding mean differences were −0.003 for accuracy, +0.001 for F1-score, and +0.005 for ROC-AUC. These small differences indicate that the cross-validation advantage of SCSO did not consistently translate into better held-out performance.

Table 7 presents the optimal hyperparameters from the primary experiment and the mean ± standard deviation of SCSO performance and optimization time across five independent random seeds. For Scenario D, SCSO increased SVM cross-validation accuracy from 0.788 to 0.804, but both methods yielded held-out accuracy of 0.766 and F1-score of 0.762; ROC-AUC differed only slightly (0.855 versus 0.857). For Random Forest in the same scenario, GridSearch selected 225 trees, and SCSO selected 80 with different splitting parameters, yielding held-out accuracies of 0.703 and 0.719. These examples illustrate that different parameter settings can produce similar results despite differences in cross-validation fitness, rather than establishing an advantage for either search strategy.

The optimizers also largely agreed on the direction of feature group contributions. Across four classifiers and five metrics, directional agreement for B−A and C−A was 20/20 (100%) in both cross-validation and held-out testing. For D−B, agreement was 19/20 (95%) and 18/20 (90%), respectively. For D−C, it decreased to 17/20 (85%) and 11/20 (55%). This agreement indicates that the main baseline additions were less sensitive to optimizer choice, whereas small incremental differences in the Full scenario were more dependent on the selected configurations.

Five independent SCSO seeds revealed classifier dependent variability: held-out accuracy standard deviations across the 16 classifier scenario combinations ranged from 0.000 to 0.040. Logistic Regression was generally stable, whereas Random Forest and XGBoost showed greater variation in some configurations. Mean SCSO optimization times ranged from approximately 4.0–4.4 seconds for SVM to 56.6–126.8 seconds for Random Forest across scenarios. In the primary experiment, mean search times were 56.5 seconds for GridSearch and 50.6 seconds for SCSO. The multi-seed standard deviations reflect stochastic optimization on a fixed data partition, not variation across test partitions. Likewise, these implementation-specific timings do not establish general computational superiority.

Fig. 3 shows best-so-far cross-validation accuracy over the 100 SCSO fitness evaluations. Across four classifiers and five seeds, mean improvement from the initial 10 evaluations to the final evaluation was approximately 0.013–0.017, depending on the scenario. Improvements were generally larger early in the search and smaller near the end. The resulting plateaus indicate limited additional improvement under this budget; they neither establish convergence to a global optimum nor guarantee the same behavior with a larger search budget.

D. Feature Redundancy and Sensitivity Analyses

Descriptive Spearman correlation analysis across the 319 observations identified substantial overlap among bioimpedance measurements: 18 of the 91 within-group predictor pairs had absolute correlation coefficients of at least 0.80. The strongest correlations involved total body fat ratio and lean mass ($\rho = -0.992$), intracellular water and muscle mass ($\rho = 0.975$), and total body water and muscle mass ($\rho = 0.967$). Cross group overlap was also observed between BMI and the bioimpedance-derived obesity measure ($\rho = 0.913$). Five of 70 demographic–bioimpedance pairs exceeded the same correlation threshold. These relationships indicate that several body-composition estimates represent overlapping information. Among continuous predictors, variance inflation factors

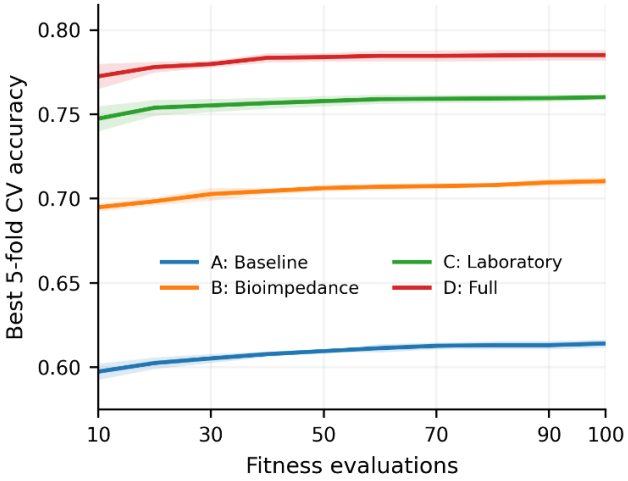


Fig. 3. Convergence behavior of Sand Cat Swarm Optimization (SCSO) across four classifiers and five independent optimization seeds for each feature scenario. Curves represent mean best-so-far five-fold cross-validation accuracy over 100 fitness evaluations, while shaded regions indicate one standard deviation across seed-wise classifier averages.

Manuscript received 7 October 2026; Accepted 9 November 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.407
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

(VIFs) were particularly high for weight (207.45), lean mass (112.05), extracellular water (109.47), and BMI (74.19). Logistic Regression coefficients also showed sign reversals across the five cross-validation folds for 37 of the 98 feature scenario coefficients examined, indicating sensitivity to the training samples and correlated predictors. However, redundancy and coefficient instability do not establish that the variables lack predictive value. No predictors were removed from the primary experiments because the objective was to compare intact measurement groups.

Preprocessing sensitivity was evaluated on the original held-out split using fixed GridSearch-selected hyperparameters. Compared with standardizing all predictors, scaling only nonbinary variables changed mean accuracy by +0.001, robust scaling by +0.006, and one-hot encoding Comorbidity and HFA by +0.020 across the 16 classifier–scenario combinations, although effects varied among configurations. Omitting scaling reduced mean accuracy by 0.047, driven mainly by a 0.176 decrease for SVM. Random Forest and XGBoost were unaffected by changes to scaling alone. Because hyperparameters were not retuned, these results describe sensitivity to preprocessing choices rather than comparisons between independently optimized pipelines.

Elastic Net Logistic Regression was evaluated across five stratified train-test partitions. Table 8 shows mean held-out accuracies of 0.563 ± 0.040, 0.697 ± 0.014, 0.750 ± 0.102, and 0.791 ± 0.045 for Scenarios A–D, respectively. Mean ROC-AUC increased from 0.597 ± 0.062 in Scenario A to 0.845 ± 0.050 in Scenario D. The mean accuracy differences were +0.134 for B−A, +0.188 for C−A, +0.094 for D−B, and +0.041 for D−C. The D−C difference was positive in three partitions and negative in two, indicating that the additional accuracy gain from bioimpedance was not consistent when laboratory measurements were available. Elastic Net and L2-regularized Logistic Regression achieved the same mean Scenario D accuracy (0.791) on matched partitions.

E. Clinical Performance and Model Interpretation

Clinical performance was assessed using Logistic Regression, which achieved the highest held-out ROC-AUC in the primary experiment. Table 9 presents confusion matrices and additional classification metrics for the four GridSearch scenarios and SCSO-optimized Scenario D. The held-out test set contained 32 gallstone cases and 32 controls. Under GridSearch, Scenario C correctly identified 25 cases and 26 controls, yielding sensitivity of 0.781, specificity of 0.813, balanced accuracy of 0.797, and MCC of 0.594. Scenario D correctly identified 29 cases but only 21 controls. Its sensitivity increased to 0.906, while specificity decreased to 0.656. Compared with Scenario C, this reduced false negatives from seven to three but increased false positives from six to eleven. The change indicates that Scenario D classified more observations as gallstone cases, improving case detection at the expense of additional false alarms. Its balanced accuracy and MCC decreased to 0.781 and 0.581, respectively. With SCSO, Scenario D retained sensitivity of 0.906 and achieved specificity of 0.688, balanced accuracy of 0.797, and MCC of 0.608.

Fig. 4 presents the held-out ROC curves for GridSearch Logistic Regression. ROC-AUC increased from 0.497 (95% CI: 0.359–0.644) in Scenario A to 0.741 (0.611–0.863) in B, 0.844 (0.741–0.936) in C, and 0.886 (0.797–0.953) in D. Paired bootstrap ROC-AUC differences were +0.244 (95% CI: 0.101–0.388) for B−A and +0.347 (0.190–0.503) for C−A. However, the D−C difference of +0.042 (−0.036–0.121) remained uncertain because its interval included zero. The confidence intervals support discrimination gains from adding either measurement group to the baseline, while the additional gain from bioimpedance after laboratory measurements remains uncertain.

Among laboratory predictors, C-reactive protein (CRP) and vitamin D had the largest mean absolute Logistic Regression SHAP attributions in Scenario C (0.691 and 0.669 log-odds units), followed by alanine aminotransferase (ALT, 0.394), total cholesterol (0.381), and aspartate aminotransferase (AST, 0.379). In Scenario D, vitamin D (0.622) and CRP (0.598) remained the leading laboratory predictors, followed by hemoglobin (0.417), high-density lipoprotein (0.357), and AST (0.335). These values are mean absolute attributions on the log-odds scale and indicate how laboratory measurements contributed to fitted predictions, not the direction or independent physiological effects of individual variables. Correlations among predictors further limit their interpretation.

Calibration and exploratory clinical utility were assessed for GridSearch-optimized Logistic Regression. Brier scores decreased from 0.271 in Scenario A to 0.211, 0.165, and 0.136 in Scenarios B, C, and D, respectively. Expected calibration error (five probability bins) was slightly lower for Scenario C than D (0.056 versus 0.063). Decision curve analysis showed threshold-dependent net benefit: at a threshold of 0.40, Scenario D achieved 0.354 compared with 0.312 for C and 0.167 for treat-all; at 0.50, Scenario C achieved 0.297 compared with 0.281 for D.

Table 8. Held-out classification performance of Elastic Net-regularized Logistic Regression across five repeated stratified training–testing partitions for the four predefined feature scenarios.

Scenario	Accuracy	F1-score	ROC-AUC
A — Baseline	0.563 ± 0.040	0.548 ± 0.042	0.597 ± 0.062
B — Bioimpedance	0.697 ± 0.014	0.725 ± 0.020	0.719 ± 0.034
C — Laboratory	0.750 ± 0.102	0.760 ± 0.088	0.823 ± 0.062
D — Full	0.791 ± 0.045	0.800 ± 0.044	0.845 ± 0.050

Note: Values represent mean ± standard deviation across five stratified 80:20 train-test partitions with independently optimized Elastic Net hyperparameters. Standard deviations reflect variation across data partitions, not cross-validation folds. ROC-AUC denotes the area under the receiver operating characteristic curve.

Manuscript received 7 October 2026; Accepted 9 November 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.407
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Table 9. Confusion-matrix outcomes and clinical classification metrics for selected Logistic Regression configurations evaluated on the held-out test set.

Config.	TN / FP	FN / TP	Sens.	Spec.	Prec.	Bal. Acc.	MCC	ROC-AUC (95% CI)
GridSearch A	15 / 17	16 / 16	0.500	0.469	0.485	0.484	−0.031	0.497 [0.359, 0.644]
GridSearch B	20 / 12	8 / 24	0.750	0.625	0.667	0.688	0.378	0.741 [0.611, 0.863]
GridSearch C	26 / 6	7 / 25	0.781	0.813	0.806	0.797	0.594	0.844 [0.741, 0.936]
GridSearch D	21 / 11	3 / 29	0.906	0.656	0.725	0.781	0.581	0.886 [0.797, 0.953]
SCSO D	22 / 10	3 / 29	0.906	0.688	0.744	0.797	0.608	0.885 [0.794, 0.952]

Note: TN, FP, FN, and TP denote true negatives, false positives, false negatives, and true positives, respectively. The two paired-count columns represent the rows of each 2 × 2 confusion matrix, with actual controls in the first row (TN, FP) and actual gallstone cases in the second (FN, TP). Sens.: sensitivity; Spec.: specificity; Prec.: precision; Bal. Acc.: balanced accuracy; MCC: Matthews Correlation Coefficient. ROC-AUC confidence intervals were obtained from 4,000 stratified bootstrap resamples.

These evaluations remain exploratory because they used only 64 held-out observations with a balanced class distribution and lacked a clinically established decision threshold.

IV. Discussion

A. Feature Group Contributions

The main finding is that complete bioimpedance and laboratory feature groups each improved classification when added to the demographic and clinical baseline. On the original held-out test set, mean accuracy gains for B−A were 0.195 with GridSearch and 0.211 with SCSO, compared with 0.242 and 0.254 for C−A. These gains were observed across all four classifiers under both optimizers, and the 95% paired bootstrap confidence intervals for mean changes in accuracy, F1-score, and ROC-AUC were above zero. The repeated GridSearch experiment also found positive B−A and C−A accuracy differences across all five data partitions. Laboratory measurements produced larger average gains, but the study did not directly test whether these gains were significantly greater than those from bioimpedance. The results support the predictive contribution of the predefined feature groups within this dataset, rather than independent effects of individual variables.

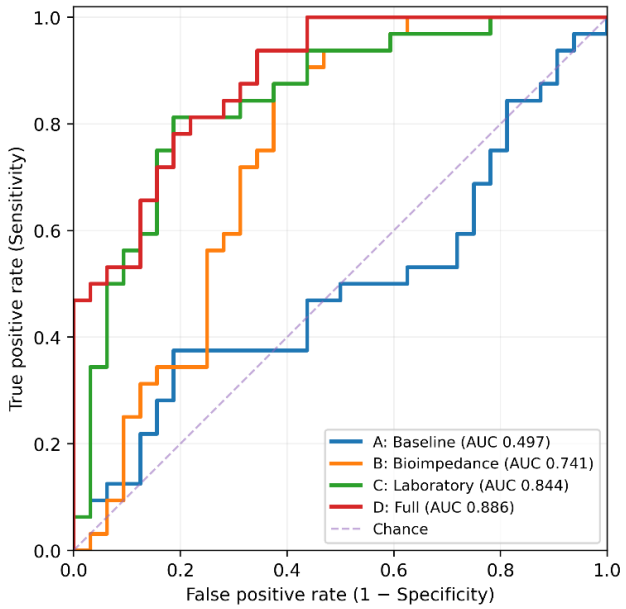


Fig. 4. Receiver operating characteristic (ROC) curves of GridSearch-optimized Logistic Regression for gallstone disease classification on the held-out test set. Scenarios A, B, C, and D represent the demographic–clinical baseline, baseline with bioimpedance, baseline with laboratory measurements, and complete feature set, respectively. The diagonal reference line represents random classification, and the area under the ROC curve (AUC) quantifies discrimination.

The bioimpedance contribution is notable because weight and Body Mass Index (BMI) were already included in the baseline. Bostan and Yılmaz [26] reported differences in body fat percentage, trunk fat, and visceral fat between participants with and without gallstones, while Zhang et al. [27] identified associations between central adiposity and gallstone disease after accounting for BMI. These findings provide a plausible explanation for Scenario B: measures of fat distribution and body composition may convey information that weight and BMI alone do not capture. Nevertheless, the correlation analysis identified substantial overlap among bioimpedance estimates of water compartments, lean tissue, and fat-related characteristics. High variance inflation factors and variation in Logistic Regression coefficients further complicate individual feature interpretation. The B−A improvement therefore reflects the collective contribution of the 14 bioimpedance variables, not 14 independent predictive effects.

Laboratory measurements cover metabolic, inflammatory, and hepatobiliary processes that are incompletely represented by baseline characteristics. Prior studies associated gallstone disease with metabolic syndrome, abdominal obesity, and low HDL cholesterol [28], insulin-resistance indicators based on triglycerides, glucose, and anthropometric measures [29], liver-related markers including AST, ALT, and ALP [30], and elevated high-sensitivity C-reactive protein [31]. The SHAP analysis likewise indicated that the fitted Logistic Regression models used information from multiple laboratory domains. CRP and vitamin D had the largest mean absolute attributions among laboratory predictors in Scenarios C and D, while liver enzymes and lipid-related variables also contributed. The original Gallstone Dataset study likewise identified vitamin D and inflammatory indicators as influential [6]. However, SHAP values explain fitted predictions and do not establish the direction, independence, or causal role of the underlying physiological relationships.

The incremental benefit of combining measurement groups depended on which group was already included.

Manuscript received 7 October 2026; Accepted 9 November 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.407
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Adding laboratory measurements to Scenario B increased mean held-out ROC-AUC by 0.097 with GridSearch and 0.100 with SCSO; both 95% paired bootstrap confidence intervals were above zero. The corresponding accuracy and F1-score intervals included zero. Adding bioimpedance to Scenario C produced a smaller mean ROC-AUC gain of 0.034 under either optimizer, with confidence intervals that included zero. D−C accuracy changes were also inconsistent across the five repeated GridSearch partitions. The modest D−C difference could reflect overlapping information from body composition and metabolic measurements, although the present analysis cannot quantify that overlap. These findings do not establish that bioimpedance is redundant; its additional predictive value may depend on the measurements already available.

B. Model and Hyperparameter Optimization Stability

The principal feature-group patterns were broadly consistent across Logistic Regression, SVM, Random Forest, and XGBoost despite their different learning mechanisms. Adding bioimpedance or laboratory measurements to the common baseline improved the main held-out metrics under both optimization methods. Repeated GridSearch evaluations across five stratified partitions and the supplementary Elastic Net analysis generally preserved these patterns, indicating that they were not restricted to the initial split or L2-regularized Logistic Regression. However, absolute predictive performance differed among classifiers, and the smaller gains obtained by combining measurement groups were less consistent. Agreement in the direction of a feature group contrast should therefore not be interpreted as equivalence among classifiers or configurations.

The optimization method had a smaller influence on held-out performance than feature composition. Across the 16 classifier–scenario combinations, the mean accuracy difference between SCSO and GridSearch was +0.007 during cross-validation but −0.003 on the held-out test set. The modest cross-validation gain did not consistently translate into better test performance, particularly with model selection based on a limited dataset [21], [32]. Both methods used the same validation folds and 100 fitness evaluations per combination, although GridSearch evaluated predefined configurations while SCSO explored bounded numerical domains. Equal evaluation budgets did not ensure identical search coverage or computational time. The observed differences describe the tested implementations rather than a general advantage of either optimizer.

Across five independent SCSO seeds, the standard deviation of held-out accuracy ranged from 0.000 to 0.040 among the 16 classifier–scenario combinations. Logistic Regression was generally stable, while Random Forest and XGBoost showed greater variation in some configurations. Best-so-far cross-validation fitness improved mainly during earlier evaluations and changed less toward the end of the budget. Such leveling indicates limited additional improvement within the observed runs, not convergence to a global optimum. Runtime also varied substantially among classifiers because evaluating a candidate required repeated model fitting. Reviews of metaheuristic applications in healthcare identify search-space characteristics, computational requirements, and data limitations as relevant challenges [33], [34]. Accordingly, fitness curves, variability across seeds, and runtime should be considered alongside final predictive performance.

Directional agreement between GridSearch and SCSO on held-out results was 100% (20/20) for B−A and C−A, 90% (18/20) for D−B, and 55% (11/20) for D−C across four classifiers and five metrics. When two scenarios yield similar results, different selected hyperparameters can alter the direction of a small performance difference without substantially changing classification quality. Repeated partitions tested sensitivity to the training and test samples, whereas independent SCSO seeds assessed stochastic optimization on a fixed partition. These analyses provide complementary evidence for internal stability, but neither establishes external reproducibility. The conclusions remain limited to the examined dataset, predefined measurement groups, classifier families, search spaces, and evaluation procedures.

C. Comparison with Gallstone Studies

Previous studies using the Gallstone Dataset have primarily examined feature selection and model optimization, with performance reported under different evaluation procedures. Esen et al. [6] reported 85.42% accuracy for Gradient Boosting using all 38 predictors and identified influential laboratory and body-composition measurements, without testing the incremental contribution of intact measurement groups. Sarker et al. [7] obtained mean ten-fold cross-validation accuracies of 78.32% with 13 SCSO-selected features and 78.42% with the complete feature set. Shrestha et al. [8] reported 78.00% accuracy and a ROC-AUC of 0.867 using all features, compared with 75.78% and 0.860 after Bald Eagle Search reduced the input to 17 features. These results show that feature reduction can preserve performance without necessarily improving it.

Das et al. [9] reported a mean five-fold cross-validation accuracy of 79.58% for CatBoost with 38 predictors and 80.42% after SLOA-based selection and optimization with 19 predictors. Tamer et al. [10] achieved the highest accuracy of 90.62% using Random Forest with 11 selected features. Together, these studies illustrate the emphasis on obtaining compact, predictive feature subsets. However, differences in the classifiers, selected variables, and validation procedures prevent a direct ranking of the reported accuracies.

The reported results were obtained using different evaluation procedures. Sarker et al. [7] and Das et al. [9] reported mean accuracies from ten-fold and five-fold cross-validation, respectively. Shrestha et al. [8] used ten evaluation partitions with a 70:30 training–testing ratio rather than conventional nonoverlapping ten-fold test sets. Esen et al. [6] and Tamer et al. [10] reported results using 70:30 train-test partitions. In the present study, five-fold cross-validation was restricted to hyperparameter selection within the training data, followed by evaluation on a separate stratified 20% test set. The highest held-out accuracy was 79.7%, while the repeated-partition and bootstrap analyses provided additional assessments of

Manuscript received 7 October 2026; Accepted 9 November 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.407
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

internal variability. Differences in test partitions, validation procedures, and model selection prevent direct performance rankings, even when studies use the same original dataset.

Indirani et al. [35] examined several ensemble configurations, including a seven-classifier voting model and variants incorporating SHAP-based feature selection and probability calibration. The seven-classifier model achieved 79.69% accuracy and a ROC-AUC of 0.9150 at the default threshold. A refined SHAP-based ensemble achieved 79.69% accuracy and a ROC-AUC of 0.9238, while adjusting its decision threshold to 0.40 increased accuracy to 87.50%, with approximately 94% recall for gallstone-positive cases. In the present study, the highest held-out accuracy, F1-score, and ROC-AUC were 0.797, 0.817, and 0.886, respectively, obtained from different configurations. Classification thresholds were not optimized in the primary experiment. Differences in model design, feature selection, and threshold adjustment therefore limit direct comparisons of predictive performance.

The contribution of this study lies in comparing complete measurement sources rather than seeking the smallest or most accurate selected subset. Unlike approaches focused on influential individual predictors or feature selection [6], [7], [8], [9], [10], [35], the bioimpedance and laboratory groups were kept fixed and added separately and jointly to a common baseline. Their incremental contributions were examined across classifier families and optimization methods, with additional analyses of uncertainty and stability. This complementary design does not demonstrate that fixed groups outperform feature selection or that the observed differences would generalize to independent clinical populations.

D. Implications and Limitations

The predictive gains from additional measurements must be considered alongside their acquisition requirements. Laboratory variables provided larger average improvements than bioimpedance when added to the same baseline, but collecting a complete laboratory panel may require blood sampling, additional costs, and tests not routinely available, such as vitamin D. Bioimpedance offers non-invasive body composition assessment but requires suitable equipment and standardized procedures. Some laboratory tests may already be available from routine clinical assessments, whereas others would need to be obtained specifically. The small and uncertain D−C gains also raise the question of whether additional bioimpedance measurements would justify their acquisition costs when laboratory data are already available. Because measurement costs, acquisition time, and availability were not evaluated, the observed gains cannot establish whether obtaining either complete measurement group would be justified in routine care.

The clinical analyses also showed that improved discrimination does not guarantee better classification at a given threshold. For GridSearch-optimized Logistic Regression, moving from the Laboratory to the Full scenario increased sensitivity from 0.781 to 0.906 but reduced specificity from 0.813 to 0.656. The Full scenario had a lower Brier score, whereas the Laboratory scenario had slightly lower expected calibration error. Decision-curve analysis favored the Full scenario at a probability threshold of 0.40, but the Laboratory scenario at 0.50. The balance between missed gallstone cases and false-positive classifications would depend on downstream diagnostic decisions, which were not specified in this dataset. Without a clinically established threshold or a representative evaluation population, these findings remain exploratory and do not support patient-level use [36], [37], [38].

The small, single-center dataset limits the precision and generalizability of these findings. It contained 319 participants, with 64 observations in the primary held-out test set, making performance estimates sensitive to individual classification errors. The nearly balanced case-control distribution may also differ from the prevalence in intended clinical populations, affecting positive predictive value, calibration, and decision-curve net benefit. Accuracy gains from B−A and C−A remained positive across all five repeated stratified partitions, but these partitions reused observations from the same cohort. Paired bootstrap intervals quantified uncertainty on the original held-out test set conditional on the fitted models, without accounting for variability from retraining. Neither analysis substitutes for validation in an independent clinical cohort. Larger independent cohorts representing intended clinical populations are needed to assess model transportability and the precision of performance estimates [24], [36], [39], [40], [41]. Such evaluations should consider discrimination, calibration, and clinical usefulness rather than accuracy alone.

The fixed feature groups permit assessment of their collective predictive contribution but cannot identify indispensable individual variables or determine whether smaller subsets would perform similarly. The groups were defined by their measurement sources rather than statistical independence, so potentially redundant predictors were intentionally retained. Strong correlations among bioimpedance measurements, high variance inflation factors, and coefficient instability complicate individual-feature interpretation. Bioimpedance estimates may also depend on hydration, device characteristics, acquisition procedures, and body composition prediction equations [42], [43], which could not be examined independently using the released dataset. Likewise, SHAP attributions describe fitted model predictions, not independent physiological or causal effects.

Finally, GridSearch and SCSO were compared within predefined search spaces and a 100-fitness-evaluation budget. Independent SCSO seeds characterized stochastic variation on a fixed data partition but did not demonstrate global optimality or general optimizer superiority. Similar held-out results should therefore be interpreted within the specified implementation, search domains, and evaluation budget. Preprocessing sensitivity and Elastic Net analyses examined alternative modeling choices but remained internal to the same dataset. Future studies should combine multicenter external validation with standardized bioimpedance acquisition, prospective assessment of calibration and clinically justified decision thresholds, and evaluation of

Manuscript received 7 October 2026; Accepted 9 November 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.407
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

measurement costs and net benefit in realistic care settings.

V. Conclusion

This study assessed the contribution of complete bioimpedance and laboratory feature groups to gallstone classification across four predefined scenarios, four classifiers, and two hyperparameter optimizers. Adding either group to the demographic and clinical baseline improved held-out performance. Averaged across classifiers, the accuracy gains from bioimpedance were 0.195 with GridSearch and 0.211 with SCSO, compared with 0.242 and 0.254 from laboratory variables. The highest held-out accuracy was 0.797, achieved by Logistic Regression in Scenario C under both optimizers and Scenario D under SCSO. The highest F1-score (0.817) came from Scenario D with SCSO, while the highest ROC-AUC (0.886) came from Scenario D with GridSearch. These maximum values were not obtained from a single model configuration. Across five repeated data partitions, both additions to the baseline showed consistent accuracy gains. The paired bootstrap intervals for B−A and C−A excluded zero across all three primary metrics, whereas the D−B results varied by metric and all D−C intervals included zero. Neither optimizer consistently outperformed the other on the held-out test set. The findings are limited by the single-center dataset of 319 participants, the 64 held-out observations, and correlations among predictors. Sensitivity–specificity trade-offs and exploratory calibration and decision-curve results do not establish clinical usefulness. Larger independent multicenter cohorts are needed to assess generalizability, standardized bioimpedance acquisition, clinically relevant decision thresholds, and whether performance gains justify additional measurement costs.

Acknowledgments

The authors would like to thank the Department of Informatics, Universitas Negeri Surabaya, for the academic support provided during the preparation of this study.

Funding

This research received no external funding.

Author Contributions

Leo Cristian Ferry contributed to conceptualization, methodology, software implementation, validation, formal analysis, data curation, visualization, and preparation of the original manuscript. Yuni Yamasari contributed to supervision, methodology, validation, and review and editing of the manuscript. Both authors reviewed and approved the final version of the manuscript.

Data Availability

The dataset analyzed in this study is publicly available in the UCI Machine Learning Repository as the “Gallstone” dataset. It can be accessed at https://archive.ics.uci.edu/dataset/1150/gallstone-1. No new patient data were collected in the present study.

Code Availability

The Python notebooks and supporting scripts for data preprocessing, model training, hyperparameter optimization, feature-group comparisons, and supplementary analyses are publicly available at https://github.com/Leo-Cristiann/gallstone-feature-group-analysis. The repository includes experimental settings, random seeds, software dependencies, and instructions for reproducing the analyses.

Institutional Review Board Statement

This study involved secondary analysis of a publicly available dataset and did not include new participant recruitment or data collection. Ethical approval for the original dataset collection was granted by the Ankara City Hospital Ethics Committee (Approval No. E2-23-4632), as reported by the dataset creators.

Informed Consent Statement

No new informed consent was obtained for this secondary analysis of publicly available, de-identified data. Consent procedures for the original data collection were the responsibility of the original investigators.

Consent for Publication

Not applicable. This study analyzed a publicly available secondary dataset and does not report personally identifiable participant information.

Conflict of Interest

The authors declare that they have no competing financial or non-financial interests that could have influenced the research.

Declaration of Generative AI and AI-Assisted Technologies

During the preparation of this manuscript, the authors used ChatGPT (OpenAI) to assist with language refinement, manuscript organization, and restructuring text in response to reviewer comments. AI-assisted suggestions were reviewed by the authors, and the authors take full responsibility for the final content of the manuscript.

References

[1] X. Wang *et al.*, “Global Epidemiology of Gallstones in the 21st Century: A Systematic Review and Meta-Analysis,” *Clin. Gastroenterol. Hepatol.*, vol. 22, no. 8, pp. 1586–1595, 2024, doi: 10.1016/j.cgh.2024.01.051.

[2] Q. Ning, F. Liu, Y. Fang, X. Zhu, J. Liu, and Z. Li, “Estimating global prevalence of gallbladder stones in general population from 2000 to 2024: systematic review and meta-analysis,” *Ann. Med.*, vol. 57, no. 1, p. 2570795, Dec. 2025, doi: 10.1080/07853890.2025.2570795.

[3] L. Mencarini, A. Vestito, R. M. Zagari, and M. Montagnani, “New Developments in the

Manuscript received 7 October 2026; Accepted 9 November 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.407
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Ultrasonography Diagnosis of Gallbladder Diseases," *Gastroenterol. Insights*, vol. 15, no. 1, pp. 42–68, 2024, doi: 10.3390/gastroent15010004.

[4] J. R. Quist, C. L. Rud, S. Brantlov, L. C. Ward, S. M. Dahl Baunwall, and C. L. Hvas, "Bioelectrical impedance analysis as a clinical marker of health status in adult patients with benign gastrointestinal disease: A systematic review," *Clin. Nutr. ESPEN*, vol. 59, pp. 387–397, 2024, doi: 10.1016/j.clnesp.2023.12.145.

[5] R. A. Putra, G. A. Pradipta, and P. D. W. Ayu, "Gallbladder Disease Classification from Ultrasound Images Using CNN Feature Extraction and Machine Learning Optimization," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 7, no. 4, pp. 1089–1111, Oct. 2025, doi: 10.35882/jeeemi.v7i4.1030.

[6] İ. Esen, H. Arslan, S. Aktürk Esen, M. Gülşen, N. Kültekin, and O. Özdemir, "Early prediction of gallstone disease with a machine learning-based method from bioimpedance and laboratory data," *Medicine (Baltimore).*, vol. 103, no. 8, p. e37258, Feb. 2024, doi: 10.1097/MD.0000000000037258.

[7] P. Sarker, J.-J. Tiang, and A.-A. Nahid, "Gallstone Classification Using Random Forest Optimized by Sand Cat Swarm Optimization Algorithm with SHAP and DiCE-Based Interpretability," *Sensors*, vol. 25, no. 17, p. 5489, Sep. 2025, doi: 10.3390/s25175489.

[8] K. Shrestha, P. Sarker, J.-J. Tiang, and A.-A. Nahid, "Metaheuristic-based gallstone classification using rotational forest explained with SHAP," *Front. Digit. Heal.*, vol. 7, p. 1727559, 2026, doi: 10.3389/fdgth.2025.1727559.

[9] P. Das, M. A. U. Khan, P. Sarker, and A.-A. Nahid, "Gallstone disease classification using SLOA-optimized CatBoost classifier with explainable AI," *PLOS ONE*, vol. 21, no. 6, p. e0342945, Jun. 2026, doi: 10.1371/journal.pone.0342945.

[10] C. Ç. Tamer, H. Arslan, and T. Çağlıkantar, "A novel method for predicting gallstones based on ensemble feature selection method," *Sci. Rep.*, Jul. 2026, doi: 10.1038/s41598-026-53394-7.

[11] A. Seyyedabbasi and F. Kiani, "Sand Cat swarm optimization: a nature-inspired algorithm to solve global optimization problems," *Eng. Comput.*, vol. 39, no. 4, pp. 2627–2651, 2023, doi: 10.1007/s00366-022-01604-x.

[12] İ. Esen, H. Arslan, S. Aktürk, M. Gülşen, N. Kültekin, and O. Özdemir, "Gallstone," *UCI Machine Learning Repository*. 2024. Accessed: Nov. 08, 2026. [Online]. Available: https://archive.ics.uci.edu/dataset/1150/gallstone-1

[13] S. Szeghalmy and A. Fazekas, "A Comparative Study of the Use of Stratified Cross-Validation and Distribution-Balanced Stratified Cross-Validation in Imbalanced Learning," *Sensors*, vol. 23, no. 4, p. 2333, 2023, doi: 10.3390/s23042333.

[14] A. Pfob, S.-C. Lu, and C. Sidey-Gibbons, "Machine learning in medicine: a practical introduction to techniques for data pre-processing, hyperparameter tuning, and model comparison," *BMC Med. Res. Methodol.*, vol. 22, no. 1, p. 282, Nov. 2022, doi: 10.1186/s12874-022-01758-8.

[15] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, p. 100804, 2023, doi: 10.1016/j.patter.2023.100804.

[16] J. Herrera Pinheiro *et al.*, "The Impact of Feature Scaling in Machine Learning: Effects on Regression and Classification Tasks," *IEEE Access*, vol. 13, pp. 199903–199931, Jan. 2025, doi: 10.1109/ACCESS.2025.3635541.

[17] N. Rabbani, G. Y. E. Kim, C. J. Suarez, and J. H. Chen, "Applications of machine learning in routine laboratory medicine: Current state and future directions," *Clin. Biochem.*, vol. 103, pp. 1–7, 2022, doi: 10.1016/j.clinbiochem.2022.02.011.

[18] M. Shehab *et al.*, "Machine learning in medical applications: A review of state-of-the-art methods," *Comput. Biol. Med.*, vol. 145, p. 105458, 2022, doi: 10.1016/j.compbiomed.2022.105458.

[19] B. He and Z. Qiu, "Development and validation of an interpretable machine learning for mortality prediction in patients with sepsis," *Front. Artif. Intell.*, vol. 7, p. 1348907, 2024, doi: 10.3389/frai.2024.1348907.

[20] A. Al-Nafjan, A. Aljuhani, A. Alshebel, A. Alharbi, and A. Alshehri, "Artificial Intelligence in Predictive Healthcare: A Systematic Review," *J. Clin. Med.*, vol. 14, no. 19, p. 6752, 2025, doi: 10.3390/jcm14196752.

[21] B. Bischl *et al.*, "Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges," *WIREs Data Min. Knowl. Discov.*, vol. 13, no. 2, p. e1484, Mar. 2023, doi: 10.1002/widm.1484.

[22] N. I. S. Lu'o, D. F. Sengkey, and V. F. F. Joseph, "Combinations of Optimization Method and Balancing Technique in Hypertension Classification with Machine Learning," *Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 7, no. 2, pp. 387–400, May 2025, doi: 10.35882/ijeeemi.v7i2.86.

[23] F. Anka and N. Aghayev, "Advances in Sand Cat Swarm Optimization: A Comprehensive Study," *Arch. Comput. Methods Eng.*, vol. 32, no. 5, pp. 2669–2712, 2025, doi: 10.1007/s11831-024-10217-0.

[24] G. S. Collins *et al.*, "Evaluation of clinical prediction models (part 1): from development to external validation," *BMJ*, vol. 384, p. e074819, Jan. 2024, doi: 10.1136/bmj-2023-074819.

[25] S. R. Master, T. C. Badrick, A. Bietenbeck, and S. Haymond, "Machine Learning in Laboratory Medicine: Recommendations of the IFCC Working Group," *Clin. Chem.*, vol. 69, no. 7, pp. 690–698,

Manuscript received 7 October 2026; Accepted 9 November 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.407
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Jul. 2023, doi: 10.1093/clinchem/hvad055.

[26] M. S. Bostan and S. Yılmaz, "Prospective evolution of body compositions based on bioelectrical impedance analysis and water intake on patients with gallstone," *Ulus. Travma Acil Cerrahi Derg.*, vol. 29, no. 8, pp. 850–857, Aug. 2023, doi: 10.14744/tjtes.2023.94034.

[27] M. Zhang *et al.*, "Independent association of general and central adiposity with risk of gallstone disease: observational and genetic analyses," *Front. Endocrinol. (Lausanne).*, vol. 15, p. 1367229, Mar. 2024, doi: 10.3389/fendo.2024.1367229.

[28] H. Zakeri Fardi, K. Mojtahedi, S. Maroufizadeh, F. Joukar, and F. Mansour-Ghanaei, "The association between metabolic syndrome and gallstone disease: A cross-sectional study from the PERSIAN Guilan cohort study," *Endocr. Metab. Sci.*, vol. 17, p. 100221, 2025, doi: 10.1016/j.endmts.2025.100221.

[29] J. Liu, Z. Hu, D. Bo, X. Zhang, Z. Zhang, and X. Liu, "Predictive role of insulin resistance surrogates in gallstone disease," *Medicine (Baltimore).*, vol. 104, no. 14, p. e41478, Apr. 2025, doi: 10.1097/MD.0000000000041478.

[30] S. Li *et al.*, "Liver Function-Related Indicators and Risk of Gallstone Diseases—A Multicenter Study and a Systematic Review and Meta-Analysis," *Gastroenterol. Res. Pract.*, vol. 2024, no. 1, p. 9097892, Jan. 2024, doi: 10.1155/2024/9097892.

[31] Z. Jiang, H. Jiang, X. Zhu, D. Zhao, and F. Su, "The relationship between high-sensitivity C-reactive protein and gallstones: a cross-sectional analysis," *Front. Med.*, vol. 11, p. 1453129, Nov. 2024, doi: 10.3389/fmed.2024.1453129.

[32] T. J. Bradshaw, Z. Huemann, J. Hu, and A. Rahmim, "A Guide to Cross-Validation for Artificial Intelligence in Medical Imaging," *Radiology: Artificial Intelligence*, vol. 5, no. 4, p. e220232, Jul. 2023, doi: 10.1148/ryai.220232.

[33] E. H. Houssein, E. Saber, A. A. Ali, and Y. M. Wazery, "Integrating metaheuristics and artificial intelligence for healthcare: basics, challenging and future directions," *Artif. Intell. Rev.*, vol. 57, no. 8, p. 205, 2024, doi: 10.1007/s10462-024-10822-2.

[34] A. Lameesa, M. Hoque, M. S. Bin Alam, S. F. Ahmed, and A. H. Gandomi, "Role of metaheuristic algorithms in healthcare: a comprehensive investigation across clinical diagnosis, medical imaging, operations management, and public health," *J. Comput. Des. Eng.*, vol. 11, no. 3, pp. 223–247, Jun. 2024, doi: 10.1093/jcde/qwae046.

[35] M. Indirani, S. Sudheer, R. Mahaveerakannan, and P. Ruba, "Gallstone Disease Prediction Using Clinical and Biochemical Features Through Ensemble Learning Techniques," *Int. J. Comput. Intell. Syst.*, vol. 19, no. 1, p. 19, 2026, doi: 10.1007/s44196-025-01083-0.

[36] R. D. Riley *et al.*, "Evaluation of clinical prediction models (part 2): how to undertake an external validation study," *BMJ*, vol. 384, p. e074820, Jan. 2024, doi: 10.1136/bmj-2023-074820.

[37] L. Zhao *et al.*, "Understanding decision curve analysis in clinical prediction model research," *Postgrad. Med. J.*, vol. 100, no. 1185, pp. 512–515, Jul. 2024, doi: 10.1093/postmj/qgae027.

[38] R. Strandberg, P. Jepsen, and H. Hagström, "Developing and validating clinical prediction models in hepatology – An overview for clinicians," *J. Hepatol.*, vol. 81, no. 1, pp. 149–162, 2024, doi: 10.1016/j.jhep.2024.03.030.

[39] K. G. M. Moons *et al.*, "PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods," *BMJ*, vol. 388, p. e082505, Mar. 2025, doi: 10.1136/bmj-2024-082505.

[40] O. Efthimiou, M. Seo, K. Chalkou, T. Debray, M. Egger, and G. Salanti, "Developing clinical prediction models: a step-by-step guide," *BMJ*, vol. 386, p. e078276, Sep. 2024, doi: 10.1136/bmj-2023-078276.

[41] R. D. Riley *et al.*, "Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study," *BMJ*, vol. 384, p. e074821, Jan. 2024, doi: 10.1136/bmj-2023-074821.

[42] F. Campa *et al.*, "High-standard predictive equations for estimating body composition using bioelectrical impedance analysis: a systematic review," *J. Transl. Med.*, vol. 22, no. 1, p. 515, May 2024, doi: 10.1186/s12967-024-05272-x.

[43] S. Brantlov *et al.*, "Metrology in Bioelectrical Impedance Analysis (BIA): From Measurement Science to Clinical and Research Applications," *Sensors*, vol. 26, no. 13, p. 4017, Jun. 2026, doi: 10.3390/s26134017.

AUTHOR BIOGRAPHY



Leo Cristian Ferry is an undergraduate student in the Department of Informatics Engineering at Universitas Negeri Surabaya, Indonesia. His academic interests include machine learning, data mining, artificial intelligence, and their applications in healthcare and structured data analysis. During his undergraduate study, he worked on projects involving data processing, classification, and machine learning model evaluation. His current research focuses on analyzing the contribution of different feature groups to machine learning-based gallstone disease classification, with particular attention to bioimpedance and laboratory information, classifier comparison, and hyperparameter optimization. He is also interested in the practical application of computational methods for solving real-world problems and in developing reliable data-driven systems for healthcare and other applied domains.

Manuscript received 7 October 2026; Accepted 9 November 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.407
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).



Yuni Yamasari is a lecturer in the Informatics Engineering department at Universitas Negeri Surabaya (UNESA). She completed her academic degrees entirely at Institut Teknologi Sepuluh Nopember (ITS), earning her bachelor's degree in 2001, her master's in 2008, and her doctorate in Electrical Engineering in 2020. Her research interests and expertise span software engineering, data mining, artificial intelligence, and applied informatics. Alongside teaching and mentoring student research, she actively contributes to scientific publications, academic books, community service initiatives, and intellectual property development, with a strong focus on advancing technology applications in Indonesia's education and health sectors. Her current work also reflects a continuing interest in data-driven methods and practical artificial intelligence applications.

Manuscript received 7 October 2026; Accepted 9 November 2026; Available online 20 November 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i4.407
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).