Vol. 11, No. 1, April 2026, Page. 119-132
https://journal.uin-alauddin.ac.id/index.php/instek
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

Hybrid ANOVA–SHAP Approach for Model-Agnostic Feature Selection in Gallstone Disease Prediction

Edi Jaya Kusuma[*1], Ririn Nurmandhani[2] , Yusthin Meriantti Manglapy[3], Ika Pantiawati[4], Evina Widianawati[5]

[1,2,3,4]Faculty of Health Science, Universitas Dian Nuswantoro, Semarang 50131, Indonesia
[5]Department of Biomedical Engineering, Chung Yuan Christian University, Taoyuan City, Taiwan
Email: [1*]edi.jaya.kusuma@dsn.dinus.ac.id, [2]nurmandhani@dsn.dinus.ac.id,
[3]yusthin.manglapy@dsn.dinus.ac.id, [4]ikapantia13@dsn.dinus.ac.id, [5]g11302505@cycu.edu.tw

Abstrak

Penyakit batu empedu merupakan kondisi yang cukup umum terjadi, namun sering kali tidak terdeteksi hingga muncul komplikasi, sehingga pentingnya deteksi dini menjadi semakin penting. Penelitian ini bertujuan untuk mengusulkan pendekatan seleksi fitur hibrid yang bersifat model-agnostik dengan menggabungkan metode statistik (ANOVA) dan *explainable machine learning* (*multi-model SHAP*) untuk meningkatkan kinerja prediksi batu empedu sekaligus tetap menjaga interpretabilitas model. Kerangka yang diusulkan diuji menggunakan dataset klinis yang terdiri dari 319 pasien dengan 38 variabel, yang mencakup data demografis, bioimpedansi, dan parameter laboratorium. Proses seleksi fitur dilakukan melalui beberapa konfigurasi bobot ($\alpha = 0.3, 0.5, 0.7$) guna melihat keseimbangan antara kontribusi statistik dan berbasis model. Hasil penelitian menunjukkan bahwa fitur seperti *C-Reactive Protein* (CRP), Vitamin D, *Bone Mass*, serta parameter fungsi hati secara konsisten berperan dalam prediksi, yang menggambarkan keterkaitan antara inflamasi, metabolisme, dan fungsi hati dalam pembentukan batu empedu. Performa terbaik diperoleh pada $\alpha = 0.7$ khususnya metode *Gradient Boosting* dengan akurasi 0.8646 dan *recall* 0.875, sedangkan $\alpha = 0.5$ menghasilkan nilai AUC tertinggi sebesar 0.905. Secara keseluruhan, pendekatan ini memberikan solusi yang cukup andal dan mudah dipahami untuk mendukung deteksi dini.

Kata kunci: ANOVA, Batu Empedu, Fitur Seleksi, Model Agnostik, Prediksi Klinis, SHAP

Abstract

Gallstone disease is a common condition that often remains undetected until complications arise, highlighting the importance of early prediction. This study aims to proposes a hybrid and model-agnostic feature selection approach that combines statistical methods (ANOVA) with explainable machine learning (multi-model SHAP) to improve gallstone prediction performance while maintaining interpretability. The research framework was examined using a clinical dataset of 319 patients with 38 variables, including demographic, bioimpedance, and laboratory data. Model-agnostic feature selection was performed using different weighting configurations ($\alpha = 0.3, 0.5, 0.7$) to explore the balance between statistical and model-based importance. The results show that key features such as C-Reactive Protein (CRP), Vitamin D, Bone Mass, and liver-related markers consistently contribute to prediction, reflecting the role of inflammation, metabolism, and liver function in gallstone formation.The best performance was achieved at $\alpha = 0.7$ especially Gradient Boosting method, with an accuracy of 0.8646 and recall of 0.875, while $\alpha = 0.5$ produced the highest AUC of 0.905. Overall, the proposed approach offers a robust and interpretable framework for early detection.

Keywords: *ANOVA, Clinical Prediction, Feature Selection, Gallstone, Model-Agnostic Approach, SHAP*

This work is an open access article and licensed under a Creative Commons Attribution-NonCommercial ShareAlike 4.0 International (CC BY-NC-SA 4.0)

1. INTRODUCTION

Cholecystitis is a pathological condition involving gallbladder inflammation and infection, commonly triggered by cystic duct obstruction due to gallstones. This condition may occur in acute or chronic forms, presenting with right upper abdominal pain, fever, leukocytosis, and potential liver dysfunction in severe cases. If not managed appropriately, cholecystitis can progress to serious complications such as gallbladder perforation, pericholecystic abscess, peritonitis, and systemic sepsis,

potentially increasing morbidity and mortality rates, particularly among elderly patients and those with comorbidities. It is estimated that approximately 90–95% of cholecystitis cases are caused by gallstone obstruction [1], [2].

Apart from cholecystitis, gallstones are associated with various severe complications, such as choledocholithiasis, acute biliary pancreatitis, acute cholangitis, biliary obstruction, and Mirizzi syndrome. Chronic inflammation caused by persistent gallstone irritation has further been associated with an increased risk of gallbladder carcinoma. Clinically, most individuals with gallstones remain asymptomatic, a condition referred to as asymptomatic gallstones. Consequently, the disease often remains undetected until acute or life-threatening complications develop. This silent progression significantly contributes to delayed diagnosis and intervention, thereby increasing healthcare burden and emergency surgical procedures.

Globally, the prevalence of gallstone disease is estimated to be approximately 6% [3]. The prevalence of gallstone disease was reported to have risen from 7.4% to 13.9%, based on data from the 2020 US National Health and Nutrition Examination Survey (NHANES) [4], indicating a substantial upward trend. This rise is strongly associated with lifestyle patterns commonly observed in Western countries, including high consumption of saturated fats and cholesterol-rich diets, obesity, sedentary behavior, and metabolic disorders such as diabetes mellitus and dyslipidemia.

Although traditionally the risk of gallstone formation in Indonesia has been considered relatively lower due to dietary habits and cultural patterns that favor higher fiber intake and lower fat consumption, ongoing lifestyle transitions may alter this epidemiological profile [5], [6]. Urbanization, increased intake of fast food and processed products, reduced physical activity, and the growing prevalence of obesity and metabolic syndrome [7] may contribute to a higher likelihood of gallstone development among the Indonesian population. Given the predominantly asymptomatic nature of gallstone disease and the ongoing epidemiological transition, early detection and risk prediction strategies are essential to prevent severe complications and long-term morbidity.

The emergence of machine learning (ML) techniques has made it possible to develop predictive models utilizing routine clinical and laboratory parameters for early screening. Several studies have reported promising results. Esen et al. [8] applied statistical feature selection (ANOVA and model-specific importance) combined with Grid Search optimization, achieving 85.42% accuracy using Gradient Boosting. Sarker et al. [9] proposed a Random Forest model optimized with Sand Cat Swarm Optimization (SCSO), reducing features from 38 to 13 and obtaining 78.32% average accuracy. Shrestha et al. [10] introduced a Rotational Forest classifier optimized with the Bald Eagle Search algorithm, achieving 78% accuracy and an AUC of 0.867. Both Sarker [9] and Shrestha [10] studies also incorporated explainable AI techniques to enhance interpretability.

Despite these encouraging findings, several methodological challenges remain. Many existing approaches rely on model-specific or statistically driven feature selection methods, which may introduce bias toward particular algorithm structures and limit generalizability across different classifiers. Metaheuristic optimization techniques, while effective in reducing dimensionality [11], [12], may also increase the risk of instability or overfitting, particularly in moderate-sized clinical datasets [13]. Furthermore, interpretability methods such as SHAP and LIME are typically applied as post-hoc analyses, rather than being systematically integrated into the feature selection process itself.

To address these limitations, this study proposes a hybrid and model-agnostic feature selection approach that integrates ANOVA and multi-model SHAP. The selection of these two methods is motivated by their complementary strengths. ANOVA provides a statistically grounded measure of feature relevance based on class separability, ensuring robustness and computational efficiency [14]. In contrast, SHAP captures non-linear relationships and interaction effects through model-based explanations [15]. Unlike existing hybrid approaches that rely on optimization techniques or single-model importance, the proposed combination reduces model dependency, improves stability through multi-model aggregation, and maintains interpretability without increasing methodological complexity.

Therefore, this study aims to propose a hybrid ANOVA–SHAP approach for model-agnostic feature selection in gallstone disease prediction, with the objective of improving predictive performance while ensuring interpretability and generalizability across different machine learning models.

2. METHODOLOGY

This study proposes a hybrid, model-agnostic feature selection framework to improve gallstone disease prediction while preserving interpretability. As presented in Figure 1, the approach first involves feature ranking, which is performed using two complementary methods: statistical analysis through ANOVA F-score and explainable machine learning via multi-model SHAP.

These two perspectives capture both linear feature relevance and complex nonlinear contributions, which are then integrated into a unified model-agnostic feature selection process. The finalized feature set is employed in both model construction and evaluation, ensuring a well-balanced trade-off between predictive capability and clinical interpretability.

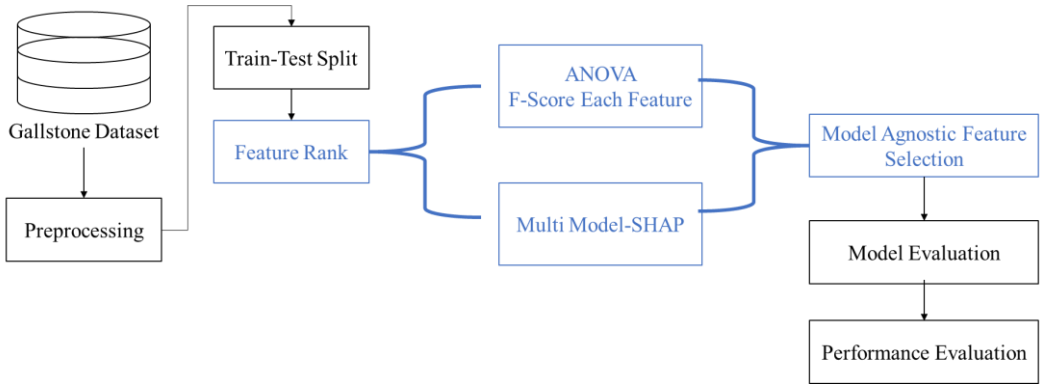


Figure 1. The proposed model-agnostic feature selection framework

2.1. Gallstone Dataset

A clinical dataset comprising 319 subjects was retrospectively collected from the Internal Medicine Outpatient Clinic of Ankara VM Medical Park Hospital between June 2022 and June 2023. Among the participants, 161 cases were clinically diagnosed with gallstone disease, indicating a relatively balanced class distribution. The data collection process was conducted following ethical approval from the Ankara City Hospital Ethics Committee (E2-23-4632) (9), ensuring conformity with established clinical research protocols [8].

The dataset consists of 38 variables representing a diverse range of patient characteristics, including demographic information (age, sex, height, weight, and body mass index), bioimpedance-derived body composition parameters (such as total body water, extracellular and intracellular water, muscle mass, body fat ratio, protein, visceral fat area, and hepatic fat accumulation), as well as routine laboratory measurements (including glucose levels, lipid profile markers, liver function tests, renal function indicators, inflammatory markers, hematological parameters, and vitamin D levels). This combination of variables reflects commonly available clinical assessments, making the dataset highly relevant for real-world screening scenarios without reliance on imaging modalities.

Importantly, the dataset is complete, with no missing values across all variables, and does not exhibit significant class imbalance. This characteristic allows the modeling process to focus on methodological robustness rather than data imputation or resampling strategies, reducing the risk of bias introduced by preprocessing steps. Furthermore, the richness and heterogeneity of features provide an ideal setting for evaluating the proposed model-agnostic feature selection approach, as it enables the identification of stable and clinically meaningful predictors across different machine learning algorithms. In line with the study objective, this dataset serves not only as a benchmark for predictive performance but also as a foundation for developing a more generalizable, interpretable, and clinically applicable early detection framework for gallstone disease.

2.2. Preprocessing

Prior to model development, a structured preprocessing procedure was applied to ensure data consistency and analytical readiness. Binary variables were treated as categorical features, while ordinal variables such as comorbidity level and hepatic fat accumulation were defined according to their inherent order, preserving their clinical interpretation.

A stratified 70:30 split was applied to divide the dataset into training and testing subsets while maintaining class balance. The target variable was converted into a numerical representation to support classification modeling.

Feature transformation was conducted based on variable type. Continuous variables were standardized to address differences in scale across heterogeneous clinical and laboratory measurements using z-score normalization (1) [16]:

$$x' = \frac{(x - \mu)}{\sigma} \tag{1}$$

where x represents the original value, μ is the mean, and σ is the standard deviation of each feature. Categorical variables were encoded into numerical representations to enable effective model learning. Although no missing values were observed, simple imputation strategies were included as a precaution to enhance robustness.

Overall, this preprocessing framework ensures consistent feature representation while preserving clinical relevance, providing a stable and unbiased foundation for subsequent model-agnostic feature selection and multi-model evaluation.

2.3. Hybrid Feature Rank

To identify the most relevant predictors for gallstone classification, this study proposes a hybrid feature ranking approach that combines statistical significance and model-based interpretability. Specifically, Analysis of Variance (ANOVA) is used to capture the linear discriminatory power of each feature, while SHapley Additive exPlanations (SHAP) are employed to quantify feature contributions across multiple machine learning models. This combination aims to reduce bias associated with single-method feature selection and improve generalizability.

Initially, ANOVA F-scores are calculated for each feature to assess their discriminative capability between gallstone and non-gallstone classes. Higher F-scores indicate stronger statistical separation. In parallel, SHAP values are calculated using several tree-based and linear models to capture both linear and non-linear feature effects. A global importance score is obtained by computing the mean of absolute SHAP values across all samples for each model. The use of absolute values ensures that the magnitude of feature contribution is captured consistently, enabling fair comparison across features and models. While this approach does not explicitly preserve the direction of influence (positive or negative), it is appropriate for feature ranking, where the primary objective is to identify the most influential variables rather than their directional effects. These importance scores are then aggregated across models to produce a unified SHAP-based ranking, thereby reducing model-specific bias. Although feature importance may vary across individual models, the aggregation process helps mitigate model-specific bias by emphasizing features that consistently contribute across different learning algorithms. This ensures that the selected features are not dominated by the behavior of a particular model, but instead reflect more generalizable patterns within the data. The calculation of unified SHAP-based ranking can be performed using Equation (2) as shown below [15]:

$$I_j^{SHAP} = \left(\frac{1}{M}\right) \sum_{m=1}^{M} \left(\left(\frac{1}{n}\right) \sum_{i=1}^{n} |\varphi_{i,j}^{m}| \right) \tag{2}$$

where:

- I_j^{SHAP} : Aggregated SHAP importance of feature j across multiple models (m)
- M : Total number of models
- n : Total number of samples
- $\phi_{i,j}^{(m)}$: SHAP value of feature $j$for the i-th sample in model m

To integrate both perspectives, ANOVA and SHAP scores are normalized and combined into a single composite score in Equation (3):

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

$$I_j^{final} = \alpha \cdot I_j^{ANOVA} + (1 - \alpha) \cdot I_j^{SHAP} \quad (3)$$

I_j^{final} represents the final combined importance score for feature j, integrating both statistical significance and model-based interpretability. I_j^{ANOVA} corresponds to the normalized ANOVA F-score, reflecting the feature's ability to discriminate between classes, while I_j^{SHAP} captures its contribution across multiple machine learning models. The parameter $\alpha \in [0,1]$ controls the relative weighting between these two components. In this study, equal weighting ($\alpha = 0.5$) is initially applied to balance statistical and model-driven perspectives, ensuring that both statistical relevance and model-based importance are equally considered. To further examine the effect of different weighting schemes, two additional configurations are introduced: $\alpha = 0.3$ and $\alpha = 0.7$, representing SHAP-dominant and ANOVA-dominant settings, respectively.

These values are selected as representative points to capture the transition between model-driven and statistically driven feature importance without introducing unnecessary complexity. Features are subsequently ranked based on this combined score, forming the basis for the next stage of model-agnostic feature selection.

2.4. Model Agnostic Feature Selection

Following the hybrid feature ranking stage, a model-agnostic feature selection strategy was employed to determine the most robust subset of predictors. Instead of relying on a single classifier, this approach evaluates feature subsets across multiple machine learning algorithms to ensure stability and generalizability [17].

Features were first ranked based on the combined importance score. A progressive selection scheme was then applied by incrementally selecting the top-k features. For each subset, performance was evaluated using a diverse group of classifiers representing different learning paradigms, including Logistic Regression (linear), Random Forest and Gradient Boosting (ensemble-based), Decision Tree (rule-based), AdaBoost (boosting), and Gaussian Naïve Bayes (probabilistic). This diversity ensures that the selected features are not biased toward a specific model structure[18].

For each candidate subset of size k, model performance was assessed using 5-fold cross-validation on the training data. The mean performance across all models was computed to obtain a model-agnostic evaluation score. The optimal number of features was determined by maximizing this aggregated performance:

$$k^* = argmax_k \left(\left(\frac{1}{L}\right) \sum_{l=1}^{L} CVScore_{l(k)} \right) \quad (4)$$

k^* represents the optimal number of selected features, L is the number of evaluated models, and $\text{CVScore}_l(k)$ denotes the cross-validated performance of model l using the top-k ranked features. The evaluation is conducted consistently across all models to ensure comparability.

This selection strategy ensures that the chosen feature subset demonstrates stable performance across multiple classifiers rather than being optimized for a single model. By prioritizing cross-model consistency, the approach enhances generalizability and reduces the risk of overfitting, making it particularly suitable for clinical prediction tasks such as early detection of gallstone disease.

2.5. Model Performance Evaluation

Following the selection of the optimal feature subset, model performance was assessed using a diverse range of classifiers, including Logistic Regression, Random Forest, Gradient Boosting, AdaBoost, Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Gaussian Naïve Bayes, and Multi-Layer Perceptron (MLP). These models represent diverse learning paradigms to ensure a comprehensive assessment. All models were trained using the selected features without hyperparameter tuning, allowing a fair comparison and isolating the impact of the proposed feature selection method. Performance was evaluated on the test set using Accuracy, Precision, Recall, F1-score, and AUC to capture overall classification quality and clinical relevance.

Vol. 11, No. 1, April 2026, Page. 119-132
https://journal.uin-alauddin.ac.id/index.php/instek
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

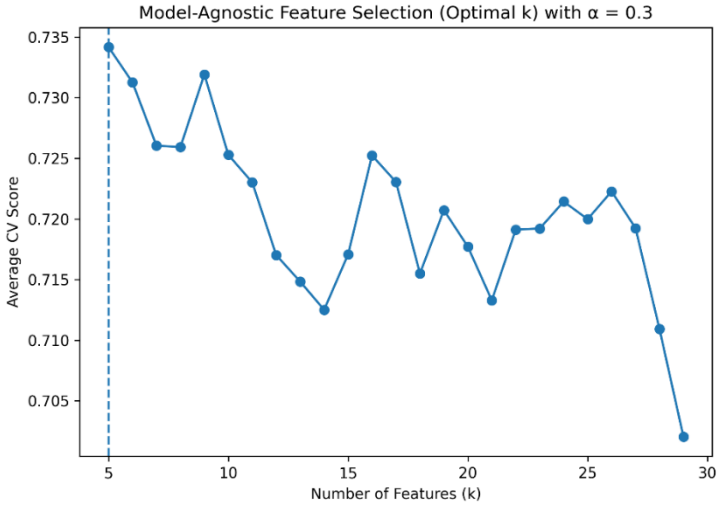
3. RESULT AND DISCUSSION

This section presents the results of the proposed framework for gallstone prediction. In addition to evaluating feature selection and model performance, this study also investigates the effect of the weighting parameter $\alpha = \{0.3, 0.5, 0.7\}$, which controls the balance between ANOVA and multi-model SHAP in the hybrid feature ranking process.

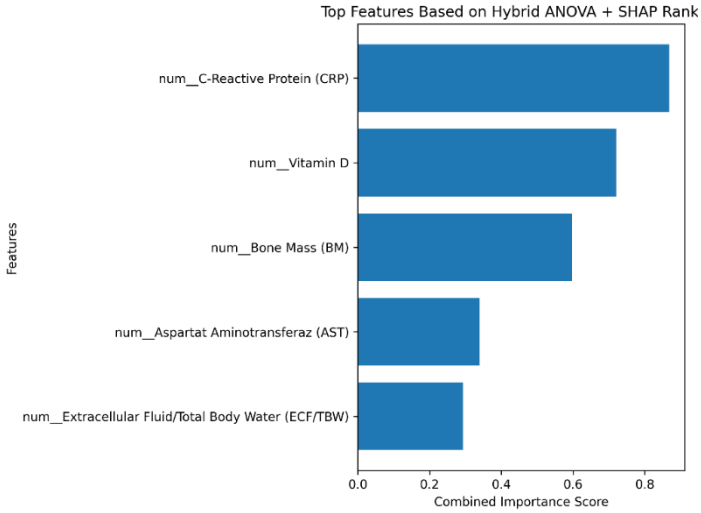
Experiments were conducted under different weighting configurations to assess how the contribution of statistical and model-based importance influences feature selection and predictive performance. The analysis is divided into two parts: feature selection results and model performance evaluation.

3.1. Feature Selection Results

To better understand how different perspectives of feature importance influence the selection process, the proposed hybrid ranking method was evaluated under three weighting configurations: $\alpha = 0.3$, $\alpha = 0.5$, and $\alpha = 0.7$.



(a) Model-agnostic feature selection performance curve under α=0.3



(b) Top 5 feature selected

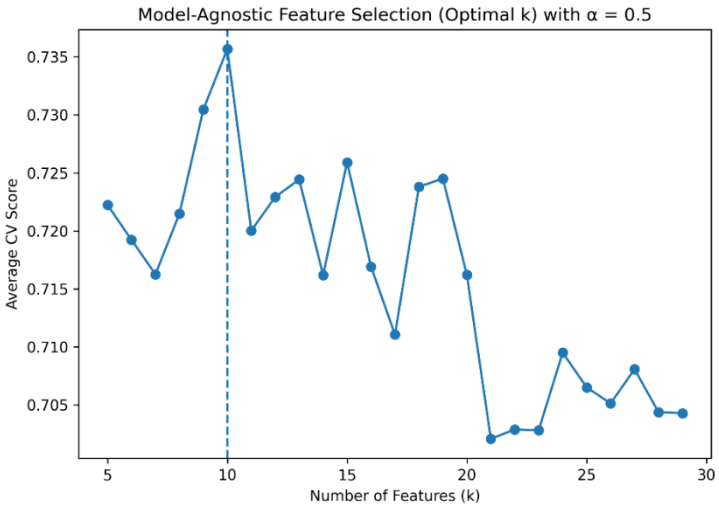
Figure 2. (a) Model-agnostic feature selection performance curve under $\alpha = 0.3$, showing the relationship between the number of selected features (k) and the average cross-validation score; (b) Top 5 feature selected based on ANOVA and SHAP with 0.3:0.7 ratio

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

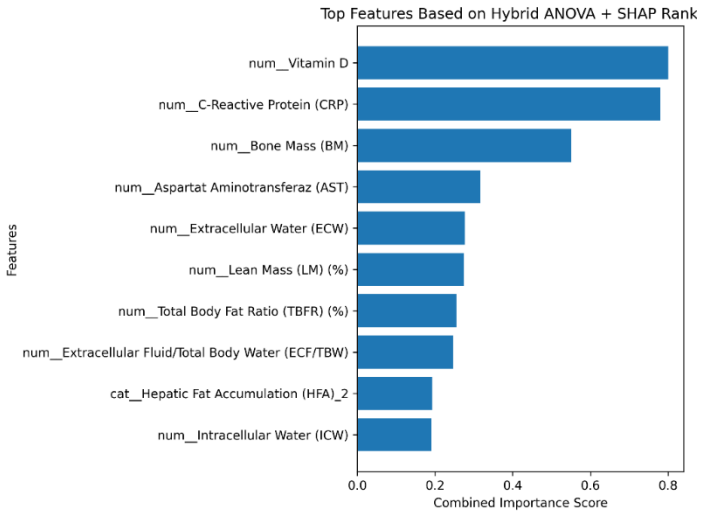
These settings represent SHAP-dominant, balanced, and ANOVA-dominant scenarios, respectively. Rather than exploring a wide range of values, these three configurations were intentionally chosen to capture the transition from model-driven to statistically driven feature selection in a clear and interpretable manner.

Across all configurations, several features consistently appeared among the top-ranked predictors, most notably C-Reactive Protein (CRP), Vitamin D, Bone Mass, and Aspartate Aminotransferase (AST). From a clinical perspective, these variables are closely related to inflammatory responses, metabolic conditions, and liver function, all of which are known to be associated with gallstone formation. Their consistent presence across different weighting schemes suggests that the proposed approach is able to identify not only statistically relevant features, but also clinically meaningful ones.

At the same time, the composition of the selected features shows important variations as the weighting changes. When SHAP is given more influence ($\alpha = 0.3$), the Model-Agnostic feature selection becomes more selective, resulting in a reduced subset of only 5 features, as illustrated in Fig. 2(a).



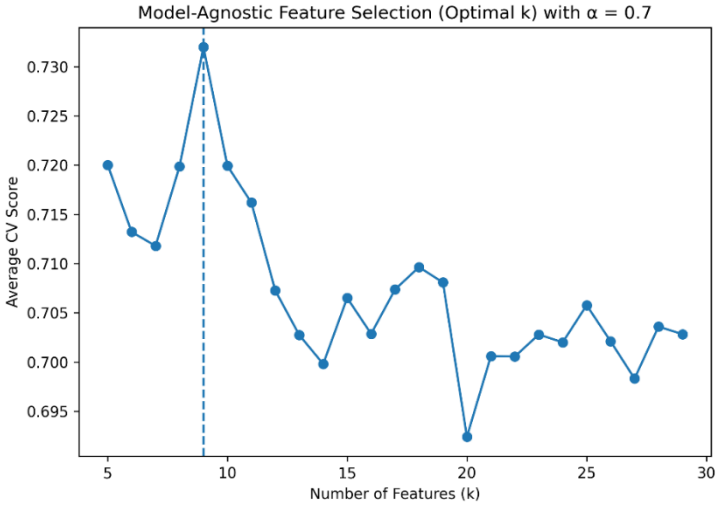
(a) Model-agnostic feature selection performance curve under $\alpha = 0.5$



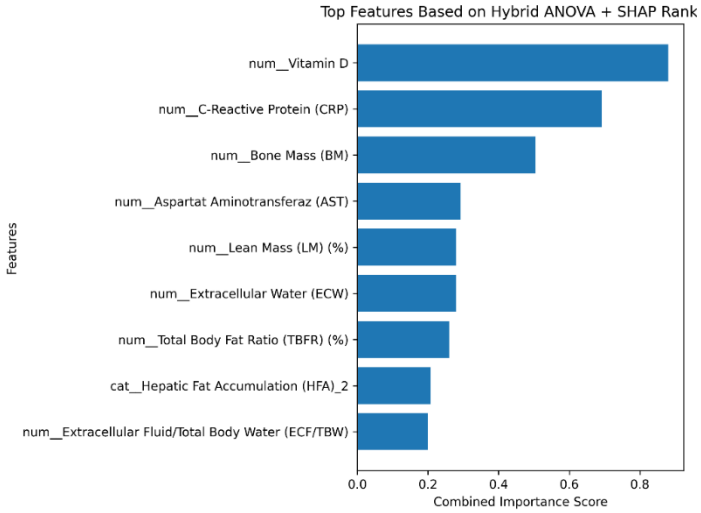
(b) Top 10 feature selected

Figure 3. Model-agnostic feature selection performance curve under $\alpha = 0.5$, showing the relationship between the number of selected features (k) and the average cross-validation score; (b) Top 10 feature selected based on ANOVA and SHAP with 0.5:0.5 ratio

In Fig. 3(a), under the balanced setting ($\alpha = 0.5$), the number of selected features increases to 10, reflecting a more comprehensive representation that combines statistically significant variables with interaction-driven features. This configuration retains key predictors while also incorporating additional variables such as intracellular water (ICW), indicating the contribution of both SHAP and ANOVA in capturing complex relationships.



(a) Model-agnostic feature selection performance curve under $\alpha = 0.7$



(b) Top 9 feature selected

Figure 4. Model-agnostic feature selection performance curve under $\alpha = 0.7$, showing the relationship between the number of selected features (k) and the average cross-validation score; Top 9 feature selected based on ANOVA and SHAP with 0.7:0.3 ratio

In contrast, when the weighting shifts toward ANOVA ($\alpha = 0.7$), the feature subset becomes slightly more compact, with 9 selected features which illustrated in Fig 4(a). In this scenario, Vitamin D becomes the most dominant feature, followed by CRP and Bone Mass, reflecting stronger statistical separability between classes. Additionally, features such as HDL are introduced, while some interaction-driven variables are excluded, indicating a shift toward features with higher discriminative power based on statistical testing.

The decision to focus on these three values of α was motivated by the need to balance analytical clarity and experimental efficiency. By examining representative points along the weighting spectrum, it becomes easier to observe how the feature selection process evolves without introducing unnecessary complexity.

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

3.2. Feature Interpretation and Comparative Analysis

The effect of different weighting configurations can be observed across Figures 2–4. As shown in Figure 2 when the $\alpha = 0.3$, under the SHAP-dominant setting, the model benefits from a more selective feature subset, where only the most influential variables contribute to prediction. In this configuration, CRP emerges as the most dominant feature, followed by Vitamin D, Bone Mass, AST, and fluid balance indicators such as ECF/TBW (Fig.2(b)). This finding further supports the observation that SHAP-based importance tends to prioritize a compact set of highly impactful features.

Under the balanced configuration, as illustrated in Figure 3 ($\alpha = 0.5$), the number of selected features increases to 10, indicating a more comprehensive representation. This setting combines statistically significant variables with interaction-driven features, retaining key predictors such as CRP, Vitamin D, and Bone Mass while also incorporating additional variables like intracellular water (ICW). This suggests that both ANOVA and SHAP contribute to capturing more complex relationships across physiological domains.

In contrast, when the weighting shifts toward ANOVA, as shown in Figure 4 ($\alpha = 0.7$), the feature subset becomes slightly more compact, consisting of 9 features. In this configuration, Vitamin D emerges as the most dominant feature, followed by CRP and Bone Mass, reflecting stronger statistical separability between classes. The inclusion of features such as HDL, alongside the exclusion of some interaction-driven variables, indicates a shift toward features with higher discriminative power based on statistical testing.

Table 1. Comparison of Selected Features Across α Configurations and Previous Studies

Method / Study	Approach Type	Feature Count	Characteristic
Esen et al. [8]	Statistical (ANOVA-based)	31	Linear, statistically driven
Sarker et al. [9]	RF + SCSO (optimization-based)	13	Model-dependent, optimized subset
Shrestha et al. [10]	RoF + BES	17	Metaheuristic-driven selection
Proposed (α = 0.3)	SHAP-dominant	5	Compact and interaction-focused
Proposed (α = 0.5)	Hybrid (balanced)	10	Balanced representation
Proposed (α = 0.7)	ANOVA-dominant	9	Compact, statistically discriminative

When compared to previous studies in Table 1, the proposed approach shows clear advantages in terms of stability and flexibility. Earlier work, such as Esen et al. [8], relies primarily on statistical methods, which are effective in identifying linearly separable features but may overlook more complex relationships. On the other hand, approaches proposed by Sarker et al. [9] and Shrestha et al. [10] focus on model-driven or optimization-based strategies, which can capture non-linear patterns but are often tied to specific models and may produce less consistent feature sets. By combining both perspectives in a model-agnostic framework, the proposed method is able to consistently identify key predictors while allowing controlled adjustments through the weighting parameter.

Overall, these findings highlight the flexibility of the proposed framework. By tuning the balance between statistical and model-based importance, the method can adapt to the underlying characteristics of the data, producing a feature set that is not only compact and stable, but also clinically meaningful.

3.3. Model Performance Evaluation

This section evaluates the predictive performance of multiple machine learning models using the selected feature subsets obtained under different weighting configurations ($\alpha = 0.3$, $\alpha = 0.5$, and $\alpha = 0.7$). The evaluation focuses on standard classification metrics, including accuracy, precision, recall, F1-score, and area under the curve (AUC), to facilitate a comprehensive evaluation of predictive performance.

Table 2. Comprehensive Model Performance Across All Weighting Configurations

Model	α	Accuracy	Precision	Recall	F1 Score	AUC
Gradient Boosting	0.3	0.8021	0.8222	0.7708	0.7957	0.8607
Gradient Boosting	0.5	0.8542	0.8400	**0.8750**	0.8571	**0.9045**
Gradient Boosting	0.7	**0.8646**	**0.8571**	**0.8750**	**0.8660**	0.9028
AdaBoost	0.3	0.8125	0.8000	0.8333	0.8163	0.8639
AdaBoost	0.5	0.8125	0.8000	0.8333	0.8163	0.9034
AdaBoost	0.7	0.8021	0.7636	**0.8750**	0.8155	0.8898
Random Forest	0.3	0.7917	0.7917	0.7917	0.7917	0.8707
Random Forest	0.5	0.8438	0.8511	0.8333	0.8421	0.8939
Random Forest	0.7	0.8021	0.8085	0.7917	0.8000	0.8937
Logistic Regression	0.3	0.7604	0.7907	0.7083	0.7473	0.8377
Logistic Regression	0.5	0.8229	0.8444	0.7917	0.8172	0.8824
Logistic Regression	0.7	0.7917	0.8043	0.7708	0.7872	0.8750
SVM	0.3	0.8021	0.8222	0.7708	0.7957	0.8685
SVM	0.5	0.8021	0.7959	0.8125	0.8041	0.8906
SVM	0.7	0.7917	0.7917	0.7917	0.7917	0.8889
MLP	0.3	0.7917	0.7917	0.7917	0.7917	0.8763
MLP	0.5	0.8542	0.8542	0.8542	0.8542	0.8976
MLP	0.7	0.7917	0.8182	0.7500	0.7826	0.8932
KNN	0.3	0.7604	0.7660	0.7500	0.7579	0.8201
KNN	0.5	0.7708	0.7955	0.7292	0.7609	0.8268
KNN	0.7	0.7813	0.7872	0.7708	0.7789	0.8340
Naive Bayes	0.3	0.7396	0.7805	0.6667	0.7191	0.8129
Naive Bayes	0.5	0.7396	0.7674	0.6875	0.7253	0.8051
Naive Bayes	0.7	0.7604	0.7778	0.7292	0.7527	0.8103
Decision Tree	0.3	0.6979	0.7021	0.6875	0.6947	0.6979
Decision Tree	0.5	0.7188	0.7143	0.7292	0.7216	0.7188
Decision Tree	0.7	0.7292	0.7037	0.7917	0.7451	0.7292

From Table 2, ensemble-based methods consistently demonstrate superior performance compared to other models. In particular, Gradient Boosting shows the most stable and competitive results, achieving the highest performance under the ANOVA-dominant setting ($\alpha = 0.7$), with an accuracy of 86.46% and an AUC exceeding 0.90. This indicates that a feature subset emphasizing statistical separability is highly effective in improving classification performance for this dataset.

Under the balanced configuration ($\alpha = 0.5$), both Gradient Boosting and MLP achieve comparable performance, with accuracy reaching 85.42%. This suggests that combining statistical and model-based feature importance results in a robust and well-generalized feature representation. The presence of multiple models with similar performance also indicates that the selected features are stable and transferable across different learning algorithms.

In contrast, the SHAP-dominant configuration ($\alpha = 0.3$) produces a more compact feature subset, which leads to slightly lower performance compared to other settings. The best result in this configuration is achieved by AdaBoost, with an accuracy of 81.25%. While this setting effectively captures highly influential features, the reduced number of variables may limit the model's ability to capture complementary information necessary for optimal classification.

Overall, the comparison across different α values highlights an important trade-off between feature compactness and predictive performance. While SHAP-dominant selection produces a smaller and more selective feature set, ANOVA-dominant selection yields better performance by leveraging stronger

statistical separability. The balanced configuration provides a middle ground, offering a comprehensive feature representation with competitive performance.

These findings demonstrate that the proposed hybrid framework is not only flexible in feature selection but also effective in supporting predictive modeling. By adjusting the weighting parameter, the approach can be tailored to prioritize either interpretability, compactness, or performance, depending on the application requirements. This adaptability is particularly valuable in clinical decision-support systems, where both accuracy and interpretability are essential.

Table 3 presents a comparative analysis between the proposed hybrid feature selection framework and previous studies in gallstone prediction. The comparison highlights differences in methodology, feature selection strategies, and predictive performance.

From a performance perspective, the proposed method achieves competitive and, in some cases, superior results compared to prior studies.

Table 3. Comparison of Proposed Method with Previous Studies

Study	Method	Feature Selection Approach	Model	*k*	Accuracy	Recall	AUC
Esen et al. [8]	Machine Learning + Grid Search	Statistical (ANOVA / feature importance)	Gradient Boosting	31	0.8542	0.8700	0.850
Sarker et al. [9]	RF + SCSO	Optimization-based (metaheuristic)	Random Forest	13	0.7832	0.7501	–
Shrestha et al. [10]	RoF + BESA	Optimization-based	Rotational Forest	17	0.7579	0.7470	0.867
Proposed (α = 0.3)	Hybrid (ANOVA + SHAP)	SHAP-dominant	AdaBoost	5	0.8125	0.8333	0.864
Proposed (α = 0.5)	Hybrid (ANOVA + SHAP)	Balanced	Gradient Boosting	10	0.8542	**0.8750**	**0.905**
Proposed (α = 0.7)	Hybrid (ANOVA + SHAP)	ANOVA-dominant	Gradient Boosting	9	**0.8646**	**0.8750**	0.903

k : Feature count

In particular, the ANOVA-dominant configuration ($\alpha = 0.7$) yields the highest accuracy of 0.8646, slightly surpassing the performance reported by Esen et al. [8], which achieved 0.8542 using a statistically driven approach. Meanwhile, the balanced configuration ($\alpha = 0.5$) achieves the highest AUC (0.905), indicating strong overall discriminative ability.

A key observation lies in the recall values of the proposed method, which reach 0.875 under both $\alpha = 0.5$and $\alpha = 0.7$. This is crucial in healthcare, where high recall signifies the model's ability to accurately detect gallstone cases while minimizing false negatives. In contrast, previous studies do not explicitly report recall, limiting the assessment of their clinical reliability.

In terms of feature efficiency, the proposed approach demonstrates a clear advantage. While Sarker et al. [9] and Shrestha et al. [10] rely on 13 and 17 features, respectively, the proposed method achieves comparable or better performance using fewer features, particularly under $\alpha = 0.7$(9 features) and $\alpha = 0.3$(as few as 5 features). This indicates that the hybrid selection strategy is capable of identifying a more compact yet informative feature subset. Another important distinction is the methodological complexity. Previous studies employ optimization-based techniques such as SCSO and BES, which are often computationally intensive and tightly coupled with specific models. In contrast, the proposed method utilizes a model-agnostic hybrid approach that integrates statistical and explainable machine learning techniques without requiring metaheuristic optimization. This makes the framework more flexible and easier to generalize across different models.

Overall, the results indicate that the proposed approach achieves strong and consistent predictive performance across different weighting configurations. The superior results obtained under the ANOVA-dominant setting suggest that statistical separability plays a key role in this dataset, while the balanced configuration provides competitive performance with a more comprehensive feature representation [19]. Importantly, the consistently high recall values demonstrate the model's

effectiveness in identifying gallstone cases, which is critical for minimizing missed diagnoses in clinical settings [20].

From a clinical perspective, the selected features highlight key mechanisms involved in gallstone formation. C-Reactive Protein (CRP) reflects systemic inflammation, which may alter bile composition and promote cholesterol crystallization [21]. Vitamin D is closely linked to lipid metabolism, and its imbalance may contribute to dyslipidemia and disrupted bile acid regulation. Bone Mass also emerges as an important predictor, suggesting a potential link between mineral metabolism and gallstone formation. As Bone Mass is associated with calcium homeostasis, disturbances in calcium balance may influence biliary processes, including calcium precipitation within bile. In addition, it may reflect broader metabolic conditions such as obesity or hormonal imbalance, which are known risk factors for gallstones [22]. Liver-related variables such as AST and Hepatic Fat Accumulation (HFA) further indicate the role of hepatic function in bile production [23], while fluid balance indicators (ECW, ICW, ECF/TBW) suggest that overall physiological status may also contribute. Overall, these findings show that the model captures clinically meaningful patterns, supporting its interpretability and relevance for early detection of gallstone disease.

4. CONCLUSION

This study proposes a hybrid and model-agnostic feature selection approach for gallstone prediction by integrating statistical significance (ANOVA) and explainable machine learning (multi-model SHAP). The approach is designed to address limitations of previous methods that rely solely on either statistical or model-specific feature selection, which may introduce bias and limit generalizability.

The findings demonstrate that the proposed approach is capable of identifying a compact yet clinically meaningful set of features. Key predictors such as C-Reactive Protein (CRP), Vitamin D, Bone Mass, and liver-related markers consistently emerge across different configurations, highlighting the roles of inflammation, metabolism, and hepatic function in gallstone formation. By combining statistical and model-based perspectives, the method provides a balanced feature selection strategy that supports both interpretability and robustness.

From a predictive perspective, the results indicate that emphasizing statistical and model-driven importance can improve reliability while maintaining a manageable number of features. This supports the practical applicability of the approach in real-world clinical settings.

Overall, this study contributes to the field of clinical machine learning by introducing a simple yet robust model-agnostic feature selection approach that bridges predictive performance and clinical interpretability. The ability to capture clinically relevant patterns while maintaining transparency makes the proposed method particularly suitable for early detection and decision-support systems.

Despite these contributions, several limitations remain. The study is based on a single-center clinical dataset, which may limit generalizability across different populations. In addition, the current approach focuses on tabular clinical data without incorporating imaging or longitudinal information.

Future work may explore the integration of multi-modal data, such as ultrasound imaging and temporal patient records, as well as validation on larger and more diverse populations. Further development toward real-time deployment and integration into healthcare systems also represents an important direction for practical implementation.

5. ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the Institute for Research and Community Service (LPPM) of Universitas Dian Nuswantoro for providing financial support through the Internal Research Grant scheme under Grant ID 032/F.9/UDN-09/II/2026. The support has been instrumental in facilitating the completion of this research.

REFERENCES

[1] M. K. Azzami, A. Rastono, and B. Adhayati, “A Case Control Study on The Relationship Between Bodey Mass Index as a Risk Factors for Incidence of Cholelithiasis at RSUD Banten,” *Int. J. Med. Public Health*, vol. 2, no. 1, pp. 31–42, 2025.

[2] J. R. Gallaher and A. Charles, “Acute Cholecystitis A Review,” *JAMA*, vol. 327, no. 10, p. 965, Mar. 2022, doi: 10.1001/jama.2022.2350.

[3] X. Wang *et al.*, "Global Epidemiology of Gallstones in the 21st Century: A Systematic Review and Meta-Analysis," *Clinical Gastroenterology and Hepatology*, vol. 22, no. 8, pp. 1586–1595, Aug. 2024, doi: 10.1016/j.cgh.2024.01.051.

[4] A. Unalp-Arida and C. E. Ruhl, "Increasing gallstone disease prevalence and associations with gallbladder and biliary tract mortality in the US," *Hepatology*, vol. 77, no. 6, pp. 1882–1895, Jun. 2023, doi: 10.1097/HEP.0000000000000264.

[5] H. Hendarto, F. N. Akbar, J. B. Muzakki, R. A. Amri, S. N. A. Nugraha, and H. Adlani, "Obesity, dyslipidemia, and diabetes mellitus as risk factors in cholelithiasis," *Electronic Journal of General Medicine*, vol. 20, no. 6, p. em549, Nov. 2023, doi: 10.29333/ejgm/13814.

[6] C. W. Candra, I. K. Sudartana, I. W. G. A. E. Putra, K. P. Yasa, I. M. Mahayasa, and K. D. Ariyanta, "Risk Factors Affecting the Occurrence of Gallstones at Prof. Dr. IGNG Visits General Hospital Denpasar Bali," *International Journal Of Scientific Advances*, vol. 5, no. 3, 2024, doi: 10.51542/ijscia.v5i3.12.

[7] Alchamdani and A. S. Anas, "Urban Obesity in Transition: Socioeconomic, Lifestyle, and Environmental Drivers in Jakarta, Indonesia," *Medicor : Journal of Health Informatics and Health Policy*, vol. 2, no. 2, pp. 113–124, Apr. 2024, doi: 10.61978/medicor.v2i2.748.

[8] 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*, vol. 103, no. 8, p. e37258, Feb. 2024, doi: 10.1097/MD.0000000000037258.

[9] 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.

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

[11] E. J. Kusuma, I. Pantiawati, and S. Handayani, "Melanoma Classification based on Simulated Annealing Optimization in Neural Network," *Knowledge Engineering and Data Science*, vol. 4, no. 2, p. 97, Mar. 2022, doi: 10.17977/um018v4i22021p97-104.

[12] A. Das, S. Mollick, H. Mondal, T. J. Bala, A. Nag, and A. Guho, "Meta-heuristic Algorithms for High-Dimensional Feature Selection: A Systematic Review of Methodologies, Applications, and Emerging Challenges with Future Research Directions," in *Feature Fusion for Next-Generation AI*, 1st ed., Md. M. Anindya Nag and A. K. B. Hassan, Eds., Springer Cham, 2026, pp. 51–61. doi: 10.1007/978-3-031-94386-7_5.

[13] E. J. Kusuma, R. Nurmandhani, I. Pantiawati, Y. M. Manglapy, and E. Widianawati, "A Random Forest and SMOTE-Based Machine Learning Model for Predicting Recurrence in Papillary Thyroid Carcinoma," *Jurnal Teknik Informatika (Jutif)*, vol. 6, no. 4, pp. 2019–2034, Aug. 2025, doi: 10.52436/1.jutif.2025.6.4.4854.

[14] V. Bhaskar P, V. R, B. Rao K, E. L. Devi, S. Saripalle, and R. Dhumpati, "An Interpretable Stacking Ensemble Model AB-CBLC for High-Accuracy Lung Cancer Detection Using Clinical and Behavioral Data," in *2025 8th International Conference on Computing Methodologies and Communication (ICCMC)*, IEEE, Jul. 2025, pp. 1580–1587. doi: 10.1109/ICCMC65190.2025.11140773.

[15] D. V. Artamonov *et al.*, "Interpretable Machine Learning with SHAP Identifies Key Biomarkers in a Multi-Factorial Spectrum of Age-Related Neurological and Metabolic Conditions," *Int. J. Mol. Sci.*, vol. 27, no. 4, p. 1805, Feb. 2026, doi: 10.3390/ijms27041805.

[16] E. J. Kusuma, R. Nurmandhani, L. Aryani, I. Pantiawati, and G. F. Shidik, "Optimasi Model Extreme Gradient Boosting Dalam Upaya Penentuan Tingkat Risiko Pada Ibu Hamil Berbasis Bayesian Optimization (BOXGB)," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 12, no. 1, pp. 111–120, Feb. 2025, doi: 10.25126/jtiik.20251219001.

[17] J. Lötsch and A. Ultsch, "A model-agnostic framework for dataset-specific selection of missing value imputation methods in pain-related numerical data," *Canadian Journal of Pain*, vol. 10, no. 1, Dec. 2026, doi: 10.1080/24740527.2025.2595160.

[18] S. S. Ayon *et al.*, "Explainable AI framework for improved Thalassemia mental health classification and feature selection," *PLoS One*, vol. 21, no. 1, p. e0341168, Jan. 2026, doi: 10.1371/journal.pone.0341168.

[19] I. Maruotto, F. K. Ciliberti, P. Gargiulo, and M. Recenti, "Feature Selection in Healthcare Datasets: Towards a Generalizable Solution," *Comput. Biol. Med.*, vol. 196, p. 110812, Sep. 2025, doi: 10.1016/j.compbiomed.2025.110812.

[20] K. Zhang *et al.*, "Machine learning model for differentiating xanthogranulomatous cholecystitis and gallbladder cancer in multicenter largescale study," *NPJ Digit. Med.*, vol. 8, no. 1, p. 590, Oct. 2025, doi: 10.1038/s41746-025-01991-7.

[21] J. Chen, Z.-T. Liu, J.-T. Lyu, and G.-P. Jiang, "Impact of metabolic disorders on gallstone disease and perioperative recovery after laparoscopic cholecystectomy," *Hepatobiliary & Pancreatic Diseases International*, vol. 23, no. 6, pp. 604–612, Dec. 2024, doi: 10.1016/j.hbpd.2024.08.001.

[22] I. F. Amri, F. H. N. Rohim, M. I. Nurul Azka, and M. S. Rakhmawati, "Analysis of Gallstone Incidence Factors Using a Logistic Regression Model.," *EKSAKTA: Journal of Sciences and Data Analysis*, Oct. 2025, doi: 10.20885/EKSAKTA.vol6.iss2.art3.

[23] X. Feng *et al.*, "MAFLD mediates the association between CHR and gallstones in the U.S. adults: evidence from NHANES 2021–2023," *BMC Gastroenterol.*, vol. 25, no. 1, p. 268, Apr. 2025, doi: 10.1186/s12876-025-03805-2.