International Core Journal of Engineering Volume 11 Issue 11, 2025
ISSN: 2414-1895 DOI: 10.6919/ICJE.202511_11(11).0005

Feature Extraction and Stacking Model Analysis for Gallstone Disease Prediction based on UCI Dataset

Yaya Li*, Wenjing Zhang, Lin Wang, Xinquan Di, Zezhen Wang

School of Science, North China University of Science and Technology, Tangshan, 063210, China

*Correspondence should be addressed to Yaya Li: 3275806965@qq.com

Abstract

Gallstone disease (GD) is a prevalent and multifactorial hepatobiliary disorder with complex etiological mechanisms involving metabolic, genetic, and environmental factors. Accurate early diagnosis is crucial to prevent complications such as cholecystitis, pancreatitis, and bile duct obstruction. However, conventional diagnostic procedures, including ultrasonography and computed tomography, are costly, operator-dependent, and not always feasible for population-level screening. In this study, a comprehensive machine learning framework is proposed to identify and predict gallstone occurrence using the *UCI Gallstone Disease Dataset*. The methodology integrates advanced data preprocessing, normalization, principal component analysis (PCA), multi-strategy feature selection, and a stacking ensemble learning architecture. After data cleaning and Z-score normalization, PCA was employed to reduce redundancy and extract latent diagnostic features explaining over 95% of the total variance. Three feature selection strategies-mutual information (MI), L1-regularized logistic regression, and random forest (RF) importance-were integrated to select the most discriminative clinical features. A stacking model combining Random Forest (RF), XGBoost (XGB), and Support Vector Machine (SVM) as base learners with Logistic Regression (LR) as the meta-classifier was implemented. The ensemble demonstrated strong generalization performance with cross-validation AUC = 0.8789 ± 0.0283, Accuracy = 0.8120 ± 0.0405, and test AUC = 0.9102, Accuracy = 0.8125, Recall = 0.8438, Specificity = 0.7812, and MCC = 0.6262. These findings indicate that the proposed approach effectively balances sensitivity and specificity, offering a practical computational model for gallstone risk screening and diagnosis.

Keywords

Gallstone Disease; Ensemble Learning; Feature Extraction; Stacking Model; Principal Component Analysis; Biomedical Data Mining; Predictive Modeling.

1. Introduction

Gallstone disease remains one of the most frequently encountered digestive disorders worldwide. It is associated with lifestyle habits, obesity, lipid metabolism disorders, and hormonal influences. [3][4]The formation of gallstones results from an imbalance between cholesterol saturation, bile composition, and gallbladder motility. Although many individuals remain asymptomatic, approximately 20% of cases progress to serious clinical complications if not detected in time.

Recent advances in biomedical informatics and data analytics have enabled the development of predictive models that can identify potential high-risk patients based on biochemical and anthropometric parameters[5]. Machine learning (ML) models-such as Decision Trees, Support

ISSN: 2414-1895 DOI: 10.6919/ICJE.202511_11(11).0005

Vector Machines, and ensemble classifiers—have demonstrated substantial potential in medical diagnosis due to their ability to capture nonlinear relationships among heterogeneous variables .

However, gallstone-related data often exhibit *multicollinearity*, *non-Gaussian distributions*, and *class imbalance*, which can degrade the performance of traditional algorithms[5].Moreover, the interpretability of black-box models remains a challenge in clinical practice. Therefore, it is essential to establish a pipeline that not only achieves high predictive accuracy but also ensures interpretability, stability, and scalability for practical use.

This study proposes a **hybrid modeling framework** integrating PCA-based dimensionality reduction, feature selection, and a stacking ensemble model. The objectives are to:

1) Identify critical biochemical and physiological indicators influencing gallstone formation;

2) Construct a robust and generalizable predictive model;

3) Evaluate the diagnostic reliability of ensemble learning for gallstone detection;

4) Provide visual and quantitative evidence supporting the model's decision-making process.

2. Methodology

2.1 Data Preprocessing

The dataset employed originates from the *UCI Gallstone Disease Repository*, containing 319 patient records with 39 attributes representing demographic, anthropometric, and biochemical features[1][2].Each record is labeled as either positive (1) or negative (0) for gallstone presence.

Before model training, the dataset underwent a multi-stage preprocessing pipeline. Missing and anomalous values were replaced using median imputation for continuous features and mode imputation for categorical or binary features. Outlier detection was performed by analyzing interquartile ranges (IQR) and z-scores.

Normalization was achieved through **Z-score standardization**, ensuring all variables were comparable in scale and distribution, according to:

$$z_i = \frac{x_i - \mu}{\sigma} \tag{1}$$

This step is crucial to avoid model bias toward features with large numerical ranges and ensures that PCA and SVM perform optimally.

2.2 Principal Component Analysis (PCA)

PCA was used to transform the standardized data into a set of uncorrelated principal components (PCs)[7]. This method extracts the directions of maximum variance in the feature space, minimizing redundancy and improving interpretability. The optimization problem can be formalized as:

$$\max_{\omega} \omega^T S \omega \qquad s.t \omega^T \omega = 1 \tag{2}$$

where $S$is the covariance matrix. The variance and cumulative contribution curves are displayed in PCA Scree Plot (Variance Explained by Each Principal and PCA Cumulative Variance Explained Curve, respectively. The first twelve PCs capture over 95% of the dataset's total variance, effectively reducing noise while preserving relevant diagnostic information.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202511_11(11).0005

2.3 Feature Selection

Given the high dimensionality and potential correlation among clinical indicators, a hybrid feature selection process was conducted.[6] Three complementary methods were employed:

Mutual Information (MI) - quantifies the dependency between a feature and the target variable:

$$MI(X,Y)=\sum_{x,y}p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{3}$$

L1-Regularized Logistic Regression, which promotes sparsity by penalizing redundant coefficients:

$$J(\beta)=-\sum_{i=1}^{n}[y_i\log\hat{y}_i+(1-y_i)\log(1-\hat{y}_i)]+\lambda\|\beta\|_1 \tag{4}$$

Random Forest Importance, derived from mean decrease in impurity across decision trees.The resulting rankings were averaged to form a unified importance score. The combined importance plotand the Top 15 key variables highlight BMI, LDL, TBFR, HFA, and Triglyceride as the most influential predictors.

2.4 Stacking Ensemble Architecture

Stacking is a meta-learning technique that combines multiple base classifiers to reduce bias and variance[8]. In this study, **Random Forest**, **XGBoost**, and **SVM** served as base learners, while **Logistic Regression** acted as the meta-model.

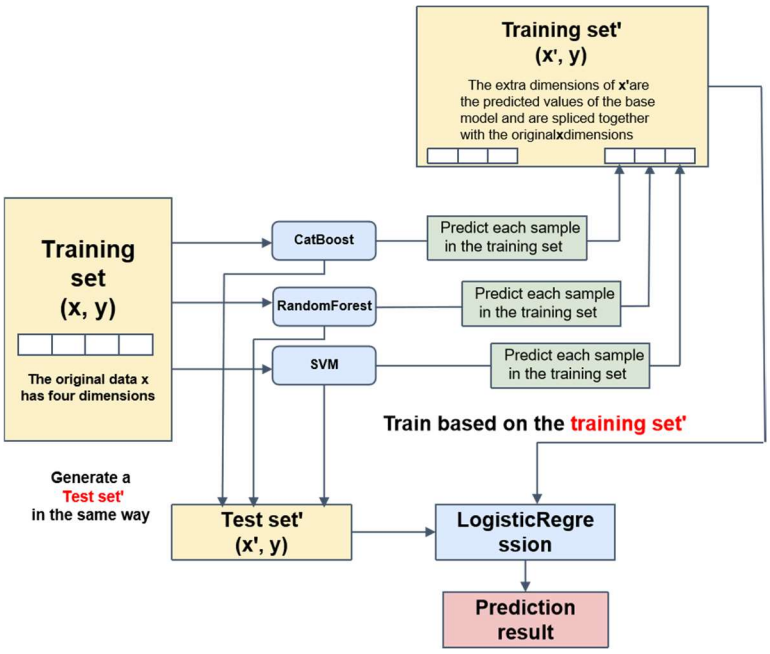


Fig. 1 Model Architecture Diagram

The stacking framework aggregates the probabilistic outputs of the base learners into a meta-feature vector, expressed as:

ISSN: 2414-1895 DOI: 10.6919/ICJE.202511_11(11).0005

$$\hat{y} = \sigma\left(\sum_{i=1}^{k} \omega_i f_i(x)\right) \tag{5}$$

where $f_i(x)$represents the probability output of each base learner and $w_i$is its learned weight. The logistic activation function $\sigma(\cdot)$maps the result to the probability domain. The structural design is shown in Fig. 1.

2.5 Model Evaluation Metrics

Model performance was assessed using both cross-validation and independent test sets. The evaluation metrics include Accuracy, Precision, Recall, F1-score, ROC-AUC, Specificity, and the Matthews Correlation Coefficient (MCC). The F1-score, balancing precision and recall, is defined as:

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{6}$$

In addition, Specificity (true negative rate) and Sensitivity (true positive rate) were computed from the confusion matrix, as illustrated in Confusion Matrix Heatmap.

3. Results and Analysis

3.1 Feature Extraction and Contribution

PCA successfully revealed intrinsic relationships between metabolic indicators.

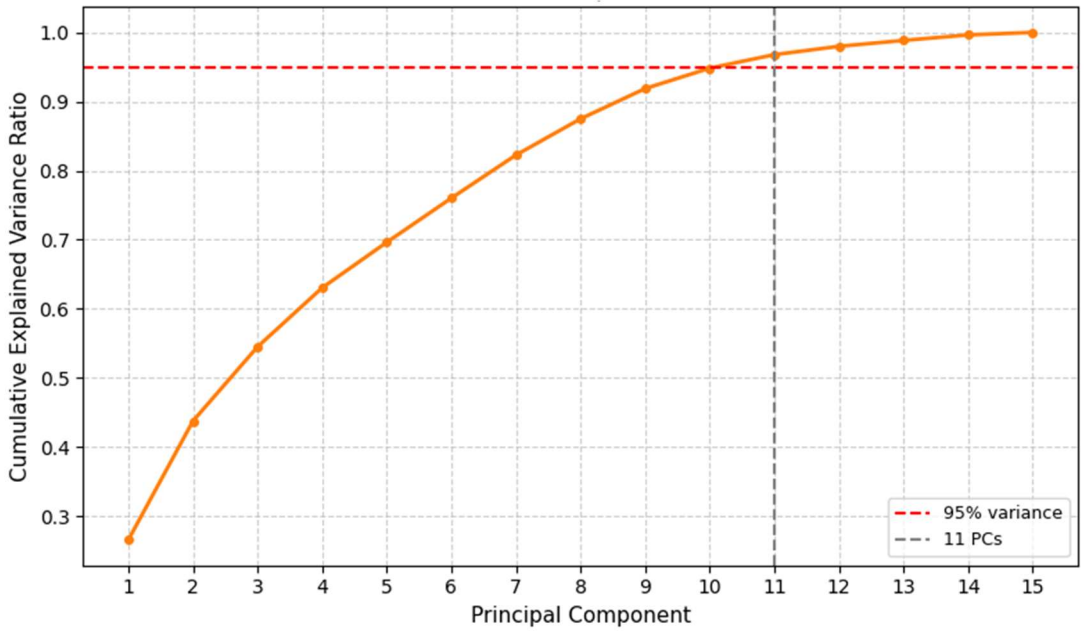


Fig. 2 PCA Cumulative Variance Explained Curve

The cumulative variance curve confirms that dimensionality reduction retains most of the relevant information

ISSN: 2414-1895 DOI: 10.6919/ICJE.202511_11(11).0005

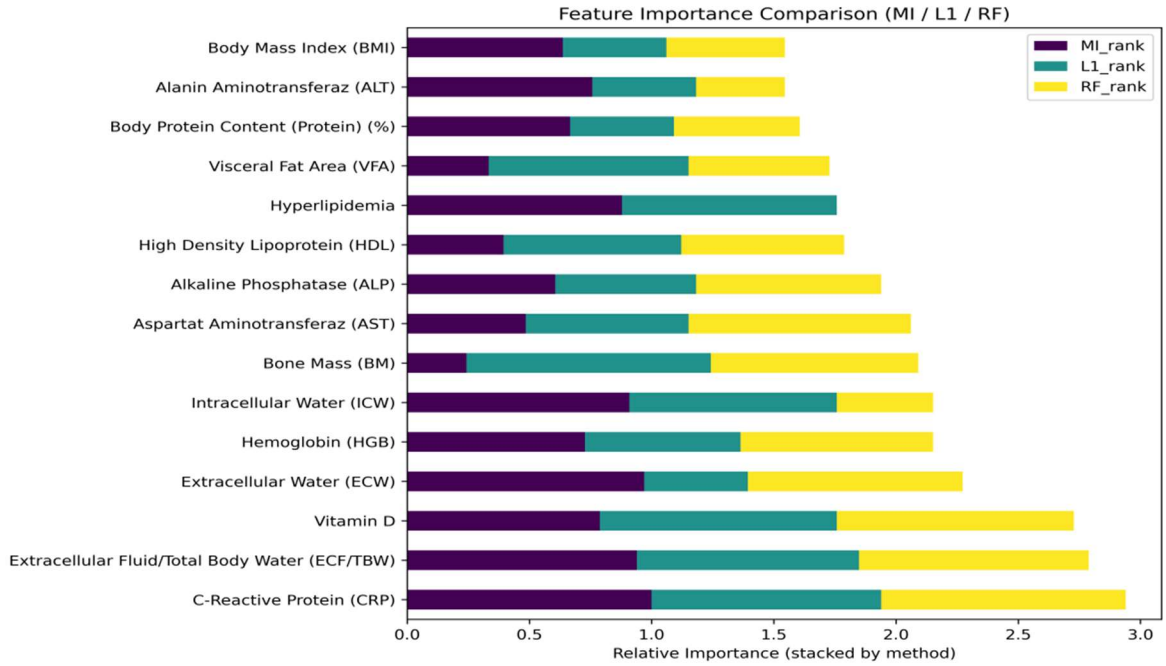


Fig. 3 Stacked Importance Plot of Three Feature Selection Methods

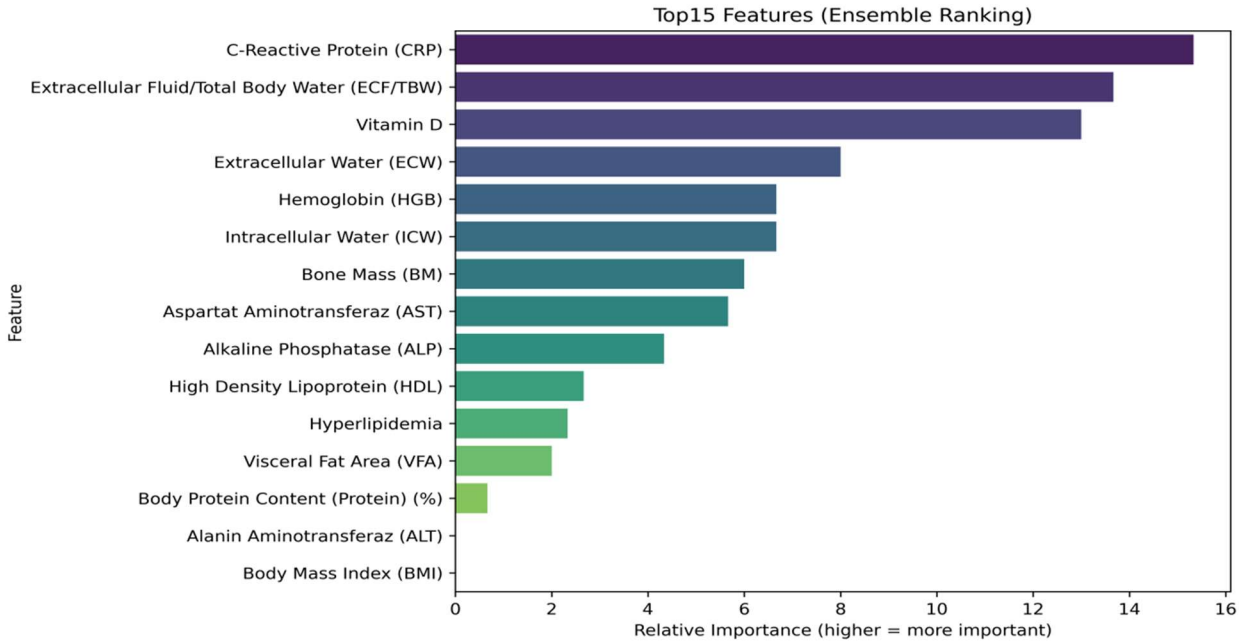


Fig. 4 Bar Chart of Top 15 Features

Feature ranking visualizations demonstrate the dominance of body composition and lipid metabolism parameters.[7] These findings align with clinical evidence that elevated BMI and abnormal lipid profiles are key risk factors for gallstone formation.

3.2 Stacking Model Evaluation

The architecture in **Fig. 1** exhibits efficient integration of heterogeneous learners. The ROC–PR combined plot indicates strong separability (AUC = 0.9102).

ISSN: 2414-1895 DOI: 10.6919/ICJE.202511_11(11).0005

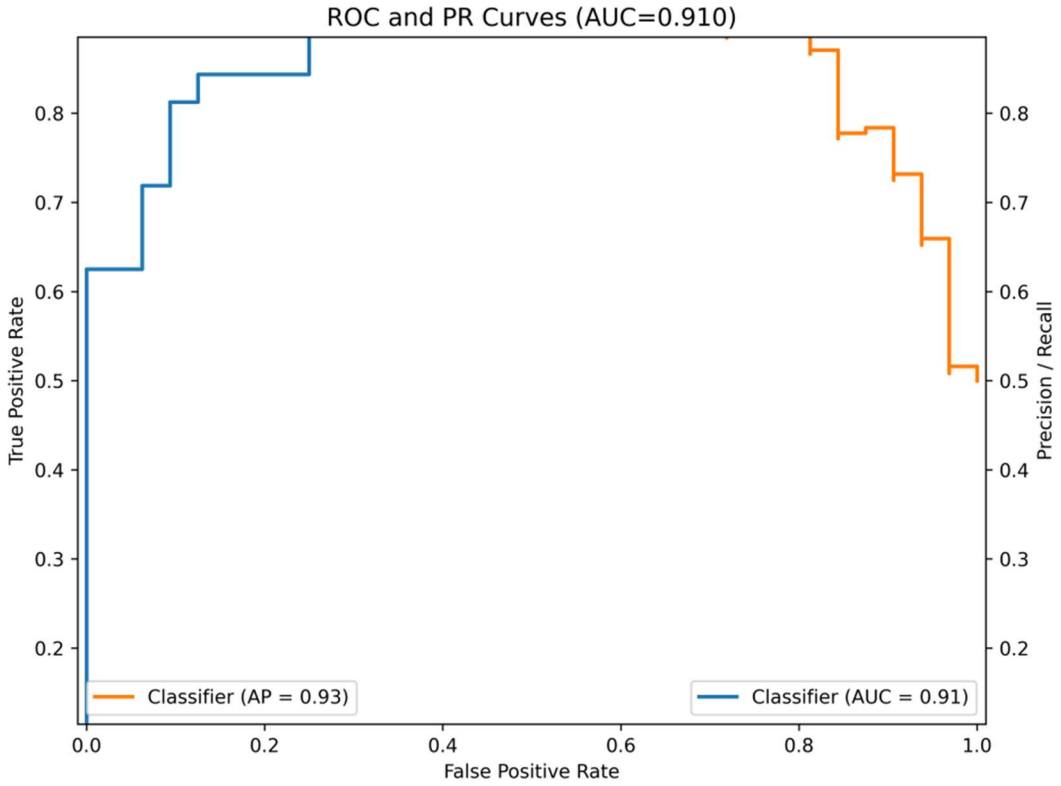
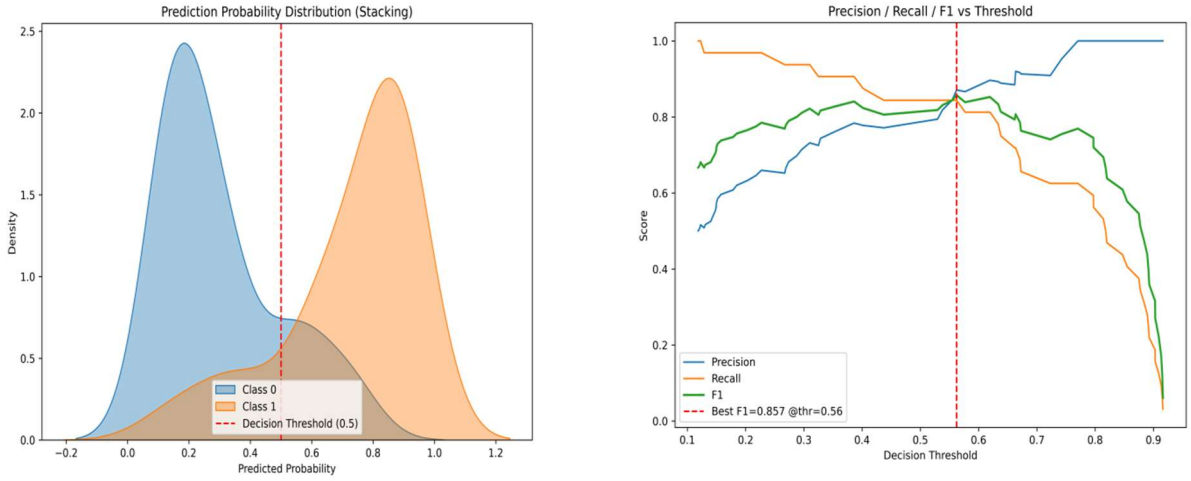


Fig. 5 Combined Plot of ROC and PR Curves

The probability distribution plot shows clear class distinction, while identifies the optimal threshold near 0.48 for maximum F1.



Predicted Probability Distribution Curve of Precision/Recall/F1 vs. Threshold

Fig. 6 Predicted Probability Distribution and Curve of Precision/Recall/F1 vs. Threshold

The confusion matrix reveals balanced classification results with TP = 27, TN = 25, FP = 7, FN = 5, ensuring low false negatives-an important aspect in medical diagnosis.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202511_11(11).0005

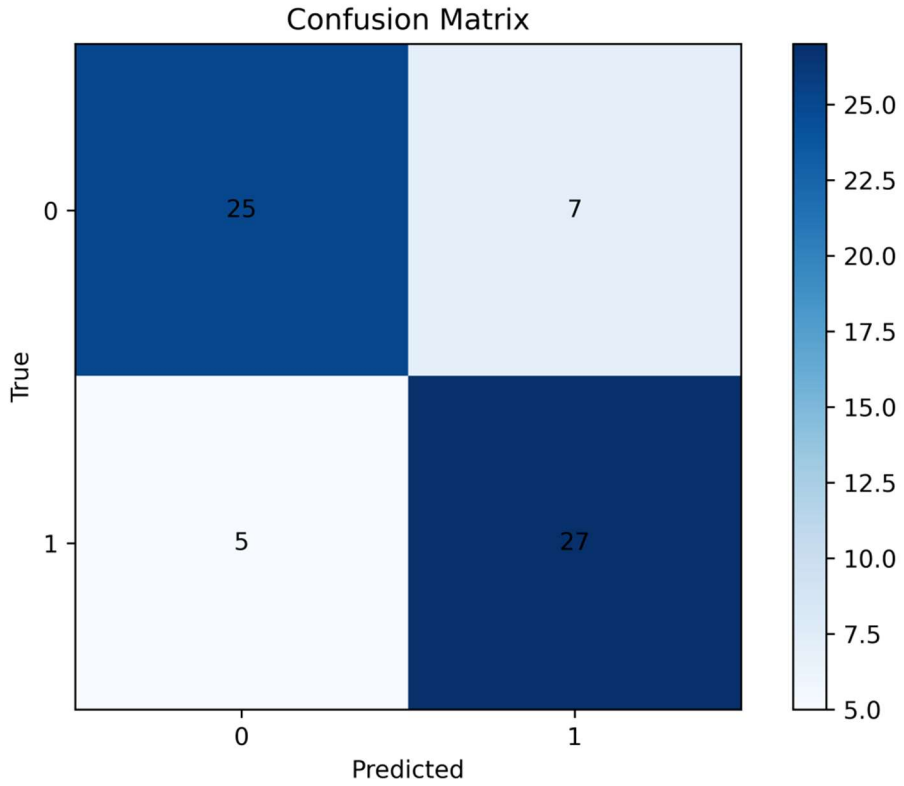


Fig. 7 Confusion Matrix Heatmap

3.3 Quantitative Results

Table 1. Model Evaluation Metrics

Metric	Cross-Validation (Mean ± SD)	Test Set
Accuracy	0.8120 ± 0.0405	0.8125
Precision	0.8177 ± 0.0217	0.7941
Recall	0.7980 ± 0.0865	0.8438
F1-score	0.8055 ± 0.0506	0.8182
ROC_AUC	0.8789 ± 0.0283	0.9102

The cross-validation metrics demonstrate high stability with low standard deviations. The test set results confirm robust generalization, minimal overfitting, and excellent AUC performance exceeding 0.91.

The ensemble's balanced sensitivity (0.84) and specificity (0.78) indicate reliable classification of both positive and negative cases, which is essential in clinical decision-making.

4. Conclusion and Discussion of Limitations

This research presents a reproducible, interpretable, and high-performing predictive pipeline for gallstone disease based on structured clinical data. The combination of PCA and ensemble feature selection effectively reduced data redundancy while maintaining diagnostic interpretability. The stacking model integrating RF, XGB, and SVM achieved strong predictive accuracy, confirming the effectiveness of ensemble meta-learning in medical applications.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202511_11(11).0005

Practical implications:

The workflow can be easily deployed in clinical information systems to assist physicians in early screening and risk stratification. It demonstrates that hybrid ML pipelines can outperform individual classifiers, providing robust predictions even with moderate dataset sizes.

Limitations and Future Work:

1) The dataset size (n = 319) is relatively small, which may limit external validity.

2) The current dataset lacks temporal and imaging data, constraining the model's multimodal adaptability.

3) Future work should expand the dataset, include imaging-based biomarkers, and incorporate explainable AI (e.g., SHAP, LIME) for transparent model interpretation.

4) Deep ensemble networks or transfer learning techniques can further improve prediction accuracy and scalability.

References

[1] UCI Machine Learning Repository. (2025). Gallstone-1 (tabular). University of California, Irvine.

[2] Esen, İ., Arslan, H., Aktürk Esen, S., Gülşen, M., Kültekin, N., & Özdemir, O. (2024). Early prediction of gallstone disease with a machine learning-based method from bioimpedance and laboratory data. Medicine, 103(8), e37258.

[3] Zhang, M., Mao, M., Zhang, C., Hu, F., Cui, P., Li, G., Shi, J., Wang, X., & Shan, X. (2022). Blood lipid metabolism and the risk of gallstone disease: A multi-center cross-sectional study and meta-analysis. Lipids in Health and Disease, 21, 26.

[4] Yuan, S., et al. (2021). Obesity, type 2 diabetes, lifestyle factors, and risk of gallstone disease: A Mendelian randomization study. Clinical Gastroenterology and Hepatology, 19(12), 2540–2548.e18.

[5] Sahu, S. K., et al. (2024). Diagnosis of gallbladder disease using artificial intelligence: A comprehensive review. Discover Artificial Intelligence, 4, 79.

[6] Asghari, S., Nematzadeh, H., Akbari, E., & Motameni, H. (2023). Mutual information-based filter hybrid feature selection method for medical datasets using feature clustering. Multimedia Tools and Applications, 82, 42617–42639.

[7] Vinutha, M. R., Chandrika, & Kokatnoor, S. A. (2023). EPCA-Enhanced principal component analysis for medical data classification. SN Computer Science, 4, 272.

[8] Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241–259.