



Clinical Impact of an Artificial Intelligence–Based Decision Support System in Routine Hospital Practice

S.T. Gopukumar

University: Nanobioinformatics Unit, Department of General Surgery, Saveetha Medical College and Hospital, Saveetha Institute of Medical and Technical Sciences (SIMATS), Saveetha University, Chennai, Tamilnadu, India, Email: gopukumar@live.com, ORCID: <https://orcid.org/0000-0001-8160-2414>

***Corresponding Author's: S.T. Gopukumar**

Received:07/06/2026; Revised: 12/07/2026; Accepted: 20/07/2026; Published: 28/07/2026

ABSTRACT

Gallstone disease remains a common biliary disorder with substantial diagnostic and clinical consequences, creating a need for practical tools that can support early risk identification. This study developed and internally validated an artificial intelligence-based clinical decision-support model for gallstone disease using a publicly available de-identified dataset of 319 patients and 38 candidate predictors. Demographic, comorbidity, anthropometric, laboratory, and body-composition variables were evaluated through elastic-net logistic regression, random forest, gradient boosting, and radial-basis-function support vector machine models using repeated stratified five-fold cross-validation. Model performance was assessed through receiver operating characteristic area under the curve, accuracy, sensitivity, specificity, precision, F1 score, Brier score, calibration measures, risk stratification, permutation importance, and decision-curve analysis. The extended elastic-net model was selected as the final model because it combined interpretability with good discrimination and calibration, achieving an area under the curve of 0.848, accuracy of 78.1%, sensitivity of 72.8%, specificity of 83.2%, and Brier score of 0.161. At an optimized threshold of 0.393, sensitivity increased to 85.4% while specificity remained 72.0%. Vitamin D, C-reactive protein, bone mass, extracellular fluid distribution, and visceral fat area were influential predictors. The model showed meaningful risk separation and potential net clinical benefit, but external multicenter and prospective validation remain necessary before routine implementation across diverse clinical settings and patient populations.

Keywords: Artificial intelligence, Clinical decision support, Gallstone disease, Machine learning, Risk prediction

1. Introduction

Gallstone disease is one of the most prevalent diseases of the biliary system and is a significant clinical problem due to its high prevalence, recurrent symptoms, and potential for complications. It is found in different populations and is related to demographic, metabolic, dietary, and lifestyle factors. Age and sex, obesity, diabetes, and other metabolic features of the individual have been found to play a critical role in gallstone development by population studies; and many people with gallstones will not experience symptoms until complications arise (Song et al., 2020). These observations have also highlighted that gallstone disease is linked to both modifiable and non-modifiable risk factors and that identification of patients at high risk for gallstones before any clinically significant biliary event (CSBE) is warranted (Gu et al., 2020).

The pathogenesis of gallstones is especially related to disturbances in lipid metabolism. When there are abnormal levels of total cholesterol, triglycerides, LDL, and HDL, this can predispose to the formation of stones. At the population level, serum lipid profiles are significantly related to the presence of gallbladder stones, suggesting that biochemical parameters that are often obtained routinely can yield useful information to predict gallstones (Wang et al., 2020). Other factors such as lifestyle can also impact the risk of disease. A healthy nutritional pattern, regular exercise, normal body weight and abstinence from smoking have been linked to reduced risk of developing symptomatic gallstone disease, but it is a complex relationship between metabolic health and biliary disease (Wirth et al., 2020).

Gallstone disease is not only a problem with uncomplicated biliary colic. Acute cholecystitis, choledocholithiasis, cholangitis, biliary pancreatitis and other systemic complications can occur, which may necessitate urgent investigation and treatment. Some clinical and biochemical parameters can help classify patients who are more likely to have these adverse effects, but risk estimation by using single parameters and/or clinical judgment is not consistent (Cao et al., 2020). The main modalities for diagnosis are currently clinical, lab testing, and imaging (abdominal ultrasonography). Though early diagnosis and timely treatment for acute calculous cholecystitis is supported by evidence, the clinical presentation may vary, and in resource-constrained environments, imaging may not be available (Pisano et al., 2020).

AI and machine learning techniques provide a chance to enhance gallstone risk assessment by simultaneously considering a combination of various patient-level parameters. These methods can detect nonlinear relationships, interactions, and patterns that are not apparent with traditional statistical methods. A recent study has validated the ability to build clinically relevant models based on a combination of bioimpedance and laboratory data and machine learning techniques, a critical proof-of-concept of the potential for predictors of gallstone disease to be developed from commonly available measurements (Esen et al., 2024). Machine-learning methods that rely on imaging have also been promising for the detection of cholelithiasis, but inconsistency in the datasets, algorithms, and validation methods has restricted their direct clinical use (Ahmed et al., 2024).

AI has been studied for its ability to detect biliary stones and abnormalities, with variable findings and often using retrospective imaging data and not pooled clinical data (Blum et al., 2025). Explainable deep-learning systems enable better understanding of how computer-aided cholelithiasis diagnosis has become more transparent, including by highlighting regions of the image that influence the model's predictions, but with the requirement for high-quality imaging and infrastructure (Kumar et al., 2025). Incorporating interpretability methods like SHAP analysis and counterfactual explanations into other machine-learning models has contributed to a better understanding of the contributions of predictors and to explaining the more understandable classification results (Sarker et al., 2025).

Being a successful predictor is not enough for clinical decision support. Interpretability, calibration, and the ability to generate information that can be used by clinicians without losing the professional judgment of the clinician are also needed. Explainable AI can enhance clinical decision-support systems by pinpointing key factors, boosting transparency, and enabling end-users to assess the clinical realism of the predictions (Liu et al., 2024). Notwithstanding these developments, there are only a few studies that have investigated the discrimination and calibration performance, threshold performance, predictor importance, risk stratification, and potential for clinical use of multiple algorithms for the prediction of gallstones using the same analytical framework. The aim of this study was thus to put together and internally validate a clinical decision-support model based on AI using routine clinical data for gallstone disease. It also tested the routine and extended predictors, determined the most influential predictors, and compared the model to determine whether it is suitable for individual risk stratification for confirmatory assessment.

Objectives of the study:

- 1.To develop an AI-based model for predicting gallstone disease using routine clinical data.
- 2.To internally validate and compare the performance of multiple machine-learning algorithms.
- 3.To identify key predictors and assess the model's calibration, risk stratification, and clinical utility.

2. Methodology

2.1 Study Design and Data Source

A secondary analysis of a publicly available, de-identified clinical dataset comprising 319 patients and 39 variables, including 38 predictors and one binary outcome, was conducted to develop and internally validate an AI-based clinical decision-support model for gallstone disease (Esen et al., 2024). No patient contact or intervention was involved, and the study adhered to principles of transparent prediction-model reporting.

2.2 Study Population and Outcome Definition

All records available were reviewed and checked for eligibility; those that were included had complete gallstone status records and predictor data, and those that were excluded were duplicates, had indeterminate outcomes, and clinically impossible values. The final sample consisted of 319 patients (161 without gallstones and 158 with gallstones) with outcome coded as 0 (absence) and 1 (presence) of gallstone disease.

2.3 Candidate Predictor Variables

A total of 38 different predictors were tested, among which were demographic, comorbidities, anthropometrics, body composition, metabolic variables, lipid variables, liver variables, renal variables, inflammatory, hematological, and nutritional variables. Two predictor sets were developed; one included the routine clinical measures and the other included the measures of body composition, visceral fat, and hepatic fat.

2.4 Data Quality Assessment and Preprocessing

No records were found to be missing, duplicate, or inconsistent, and there were no distribution problems with the data. The extreme obesity-percentage value ($>300\%$) was kept for the primary analysis and subsequently conducted sensitivity analysis; categorical variables were coded, continuous variables were standardized, and all preprocessing was done within cross-validation folds to prevent data leakage.

2.5 Descriptive and Comparative Statistical Analysis

All statistical analyses and machine-learning procedures were performed using Python. Data handling and numerical computation used pandas and NumPy, inferential analyses used SciPy, model development and validation used scikit-learn, and publication-quality figures were generated using Matplotlib.

2.6 Model Development and Internal Validation

Routine and extended predictor sets were used for leakage-safe feature selection, and four algorithms (elastic-net logistic regression, random forest, gradient boosting, and SVM) were trained. Several stratified five-fold cross validations were performed, and the final model was chosen based on a balance of performance, calibration, interpretability, and parsimony (not accuracy) instead of on accuracy alone.

2.7 Model Performance Assessment

AUC with bootstrap confidence intervals as well as accuracy, sensitivity, specificity, precision, and F1 score were used to assess the performance. The Brier score, calibration intercept, and slope were used to assess calibration, and accuracy was measured with default (0.50) and optimized probability thresholds with a classification performance assessment.

2.8 Model Interpretation and Predictor Importance

The final model was interpreted with the help of permutation-based feature importance, with which, if the performance decreased after disrupting the feature, the predictor was contributing to the model. Importance was related to the predictive value, but not necessarily to causality, and was correlated with features that were kept after internal feature selection.

2.9 Risk Stratification and Clinical Utility

Predicted probabilities were divided into three categories: low (<0.30), intermediate ($0.30-0.69$), and high (≥ 0.70), and prevalence was computed for each category. Net clinical benefit was estimated and compared with treat-all and treat-none strategies for diagnostic prioritization using decision-curve analysis.

2.10 Sensitivity Analysis

An extreme obesity-percentage outlier was removed from the set to run a sensitivity analysis and evaluate robustness of findings to removal of this outlier by assessing model performance using AUC and Brier score.

3. Results

3.1 Study Population and Data Quality

In total, 319 patient records were analysed. Of this group, 158 patients (49.5%) had gallstones while 161 patients (50.5%) did not have gallstones. There were a total of 38 predictor variables that were candidates. There were no missing observations or duplicate records found. A single extreme value was identified in the obesity-percentage variable and was kept in the main analysis and was analyzed separately using sensitivity analysis. Table 1 shows the analytical sample and data-quality characteristics.

Table 1. Analytical sample and data-quality characteristics

Parameter	Value
Total records assessed, n	319
Records included, n	319
Gallstone-positive patients, n (%)	158 (49.5)
Gallstone-negative patients, n (%)	161 (50.5)
Candidate predictors, n	38
Missing observations, n (%)	0 (0.0)
Duplicate records, n (%)	0 (0.0)
Extreme obesity values >300%, n (%)	1 (0.3)

3.2 Clinical and Laboratory Characteristics According to Gallstone Status

The median C-reactive protein, total body fat content, and visceral fat area were found to be significantly increased in the patients with gallstone disease when compared with the patients without gallstones. The gallstone-positive group had lower levels of vitamin D, lean mass, and bone mass as well as hemoglobin and aspartate aminotransferase. Vitamin D, C-reactive protein, lean mass, total body fat ratio, and bone mass were the greatest Standardized Group Difference measures. The differences in clinical, laboratory, and body composition findings, according to gallstone status, are detailed in Table 2.

Table 2. Selected clinical, laboratory, and body-composition characteristics according to gallstone status

Variable	Gallstone absent (n=161)	Gallstone present (n=158)	Standardized difference	mean	p-value
C-reactive protein	0.00 (0.00–0.31)	1.29 (0.20–3.45)	0.584		<0.001
Vitamin D	25.27 (19.62–30.20)	17.00 (9.95–23.77)	−0.757		<0.001
Aspartate aminotransferase	20.00 (16.00–27.00)	16.50 (14.00–21.00)	−0.272		<0.001
Lean mass	73.52 ± 8.38	69.72 ± 8.08	−0.462		<0.001
Total body fat ratio	26.39 ± 8.41	30.19 ± 8.06	0.462		<0.001
Bone mass	2.90 (2.50–3.30)	2.60 (2.30–3.00)	−0.442		<0.001
Hemoglobin	14.76 ± 1.74	14.07 ± 1.75	−0.400		<0.001
Total fat content	20.10 (15.10–26.70)	24.75 (18.62–29.75)	0.344		<0.001
Visceral fat area	10.60 (7.90–14.50)	12.48 (9.10–15.79)	0.283		0.005
Gender category 0, n (%)	94 (58.4)	68 (43.0)	−0.311		0.009

Values are presented as mean ± standard deviation, median (interquartile range), or n (%).

3.3 Internally Validated Model Performance

Four machine-learning algorithms were tested with both a basic and an expanded set of predictors. The extended models that performed best numerically (on a receiver operating characteristic area under the curve of 0.851) were gradient boosting. The area under the curve was the same (0.851) for the random forest, with lower sensitivity. The extended elastic-net logistic regression model had an area under the curve of 0.848, accuracy of 78.1%, sensitivity of 72.8%, specificity of 83.2%, and Brier score of 0.161. An extended logistic regression with elastic-net regularization was chosen as the final clinical decision-support model due to its high discrimination that remains comparable to the best nonlinear models, and its interpretability and good calibration. Random forest had the best model performance in the routine clinical model (AUC 0.833). The use of predictors of body composition did therefore lead to about 1.5–1.8 percent more discrimination. The performance of the extended prediction models is shown in Figure 1. The performance of the candidate prediction models is presented in Table 3.

Table 3. Repeated stratified cross-validated performance of candidate models

Predictor set	Model	ROC-AUC	95% CI	Accuracy	Sensitivity	Specificity	Precision	F1 score	Brier score
Extended	Gradient boosting	0.851	0.808–0.893	0.781	0.759	0.801	NA	NA	0.155
Extended	Random forest	0.851	0.807–0.892	0.777	0.728	0.826	NA	NA	0.160

Extended	Elastic-net logistic regression	0.848	0.803–0.891	0.781	0.728	0.832	0.810	0.767	0.161
Extended	RBF support vector machine	0.840	0.795–0.881	0.774	0.734	0.814	NA	NA	0.160
Routine	Random forest	0.833	0.788–0.875	0.774	0.722	0.826	NA	NA	0.169
Routine	Gradient boosting	0.825	0.780–0.866	0.749	0.741	0.758	NA	NA	0.176
Routine	Elastic-net logistic regression	0.822	0.775–0.868	0.762	0.741	0.783	NA	NA	0.175
Routine	RBF support vector machine	0.797	0.751–0.845	0.737	0.741	0.733	NA	NA	0.183

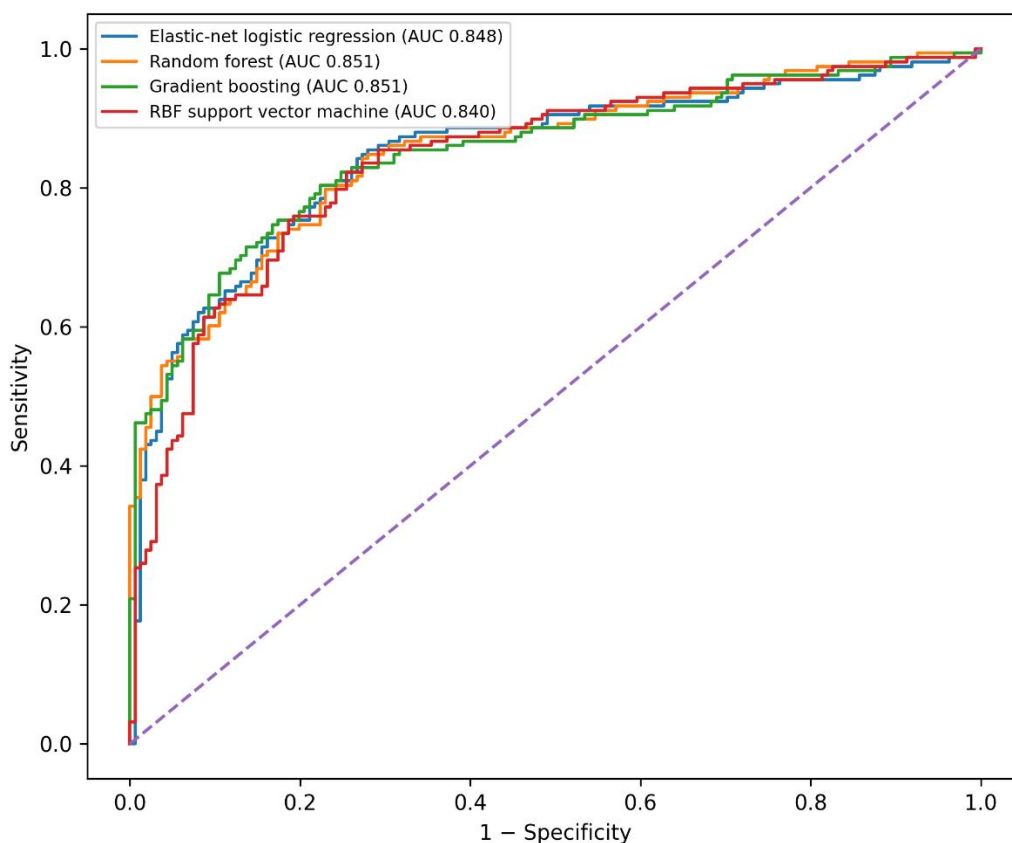


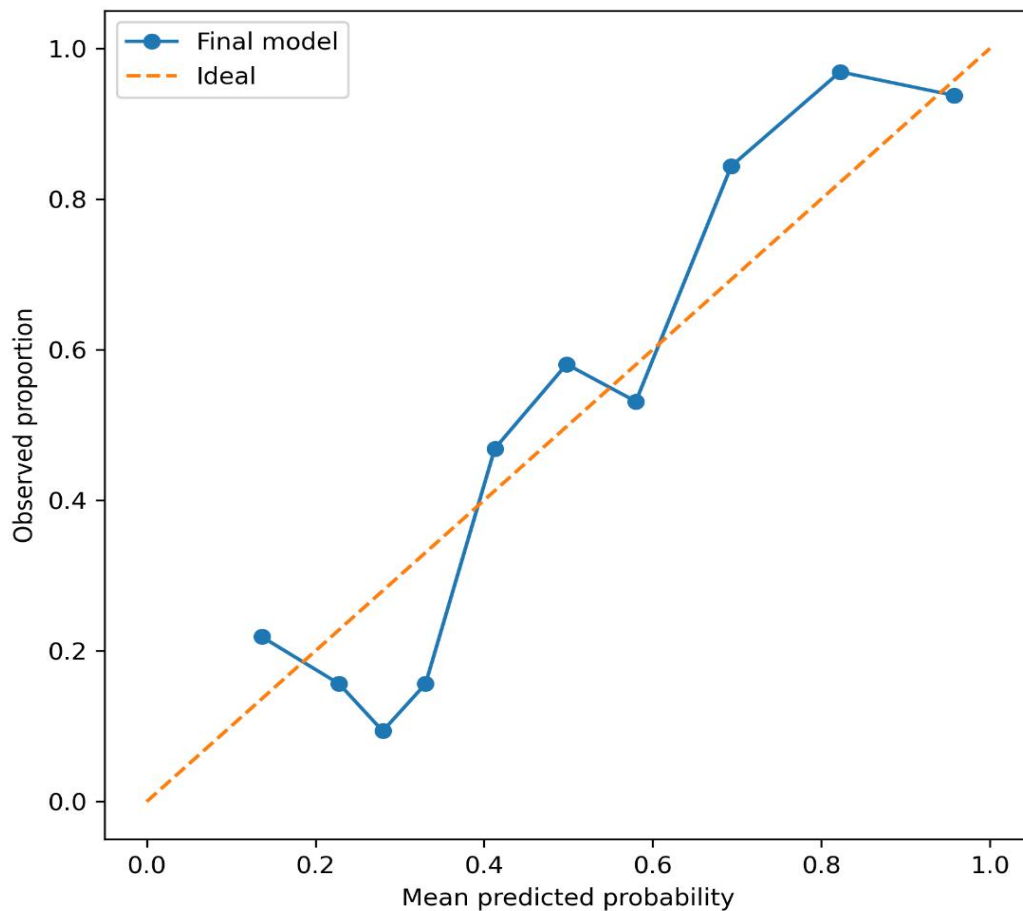
Figure 1. Cross-Validated Receiver Operating Characteristic Curves of the Extended Prediction Models

3.4 Calibration and Classification Thresholds

Finally, the last elastic-net model exhibited generally satisfactory calibration. The intercept and the slope of the calibration were 0.026 and 1.111, respectively. The model was successful at the conventional probability threshold of 0.50 with 78.1% accuracy and was more specific than sensitive. The optimum probability level was 0.393. At this cut-off point, sensitivity rose by 12.6% (from 72.8% to 85.4%) while specificity fell by 11.2% (from 83.2% to 72.0%). The accuracy remained almost the same (78.1% to 78.7%) and the F1 score rose from 76.7% to 79.9%. The predictions and observations of gallstone risk are plotted against one another in Figure 2. The final model's performance for default and optimal thresholds is presented in Table 4.

Table 4. Performance of the final model at alternative classification thresholds

Metric	Threshold 0.500	Threshold 0.393
Accuracy	0.781	0.787
Sensitivity	0.728	0.854
Specificity	0.832	0.720
Precision	0.810	0.750
F1 score	0.767	0.799
ROC-AUC	0.848	0.848
Brier score	0.161	0.161
Calibration intercept	0.026	0.026
Calibration slope	1.111	1.111

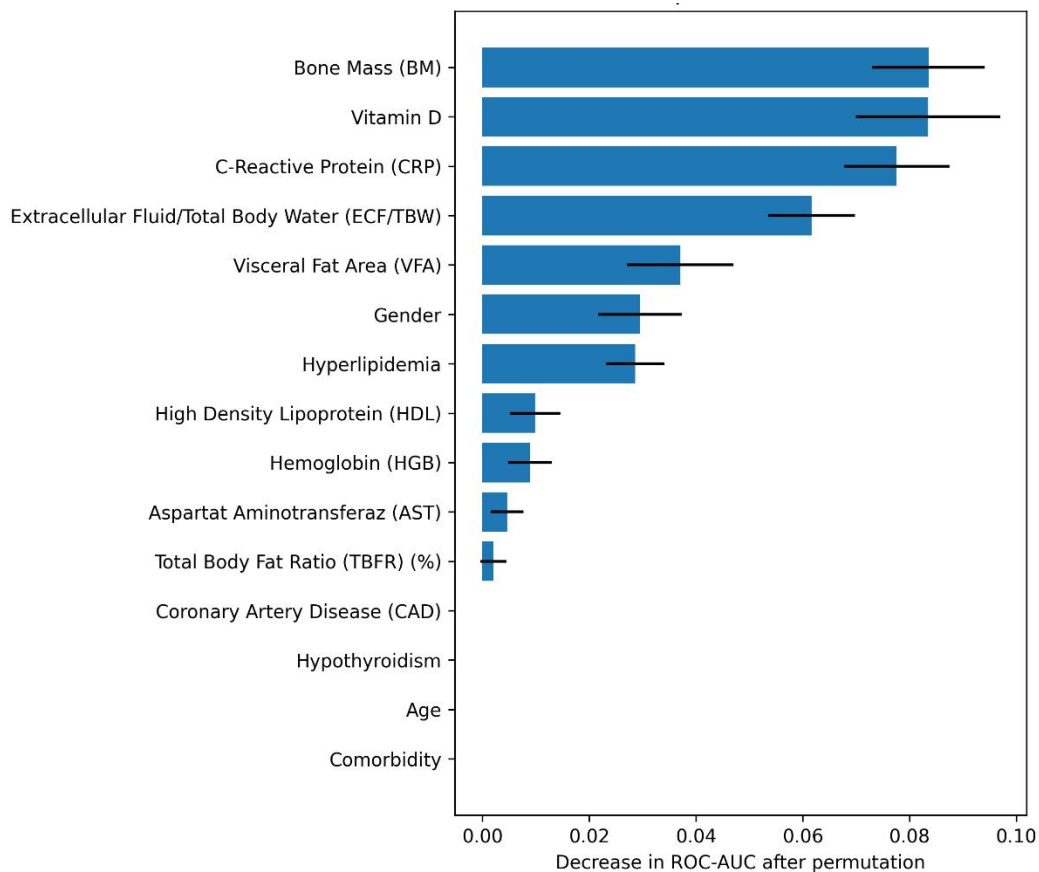
**Figure 2. Calibration Plot of the Final Elastic-Net Logistic Regression Model**

3.5 Predictor Importance and Risk Stratification

The last feature-selection process left 16 predictors. The highest permutation-importance values were obtained for bone mass and vitamin D; the other factors were in the following descending order: ECF/TBW, visceral fat area, sex, and hyperlipidemia. The model was able to differentiate patients into clinically distinct risk categories. In the low-risk group, 15/95 patients (15.8%) had gallstones. Of children with intermediate risk, 70 had gallstone disease, and of those with high-risk, 73 had gallstone disease. The high-risk group had a 92.4% prevalence of gallstones. The predictors' importance in the final model is shown in Figure 3 by their permutation. Predictor importance, risk stratification, and sensitivity analysis results are shown in Table 5.

Table 5. Predictor importance, risk stratification, and sensitivity analysis

Analysis	Numerical result
Bone mass permutation importance	0.0835
Vitamin D permutation importance	0.0835
C-reactive protein permutation importance	0.0776
Extracellular fluid/total body water ratio permutation importance	0.0617
Visceral fat area permutation importance	0.0371
Gender permutation importance	0.0295
Hyperlipidemia permutation importance	0.0286
High-density lipoprotein permutation importance	0.0098
Hemoglobin permutation importance	0.0089
Aspartate aminotransferase permutation importance	0.0046
Low-risk group, gallstone cases/total	15/95
Low-risk group, observed prevalence (%)	15.8
Intermediate-risk group, gallstone cases/total	70/145
Intermediate-risk group, observed prevalence (%)	48.3
High-risk group, gallstone cases/total	73/79
High-risk group, observed prevalence (%)	92.4
Primary-analysis ROC-AUC, n=319	0.848
Sensitivity-analysis ROC-AUC, n=318	0.817
Primary-analysis Brier score, n=319	0.161
Sensitivity-analysis Brier score, n=318	0.174

**Figure 3. Permutation-Based Feature Importance of the Final Clinical Decision-Support Model**

3.6 Clinical Utility and Sensitivity Analysis

The decision-curve analysis showed that the final model had the greatest calculated net benefit compared to all patients or no patients at threshold probabilities of around 0.21–0.80. The model thus had the potential to be used to identify patients that may benefit from prioritized confirmatory assessment. One record was removed from the data because it had an obesity-percentage value over 300%, and a sensitivity analysis was done. The area under the curve in the

primary analysis was 0.848, while in the reduced sample it was 0.817, and the Brier score was 0.161 vs 0.174 in the two analyses, respectively. The results showed moderate sensitivity to observations in the relatively small data set, suggesting the need for more observations. The findings showed moderate sensitivity of model performance to individual observations in the relatively small data set, suggesting that observations are needed.

4. Discussion

In this study, an artificial intelligence–based clinical decision-support model was developed and internally validated, and several routine and extended clinical data were used for predicting gallstone disease. The final elastic-net logistic regression model showed good discrimination (receiver operating characteristic area under the curve (AUC), 0.848), a good accuracy (78.1%), a sensitivity of 72.8%, and a specificity of 83.2%, at the default probability threshold. Gradient boosting and random forest performed slightly better in terms of numerical discrimination, but the elastic-net model was chosen as it had a better predictive performance–calibration–interpretability–clinical practicality balance. This was aligned with current guidance that prediction models should be evaluated using clearly reported data and clinically relevant performance metrics and a clear definition of the purpose of the model, rather than discrimination (Collins et al., 2024a).

The performance observed suggested that the disease of gallstones could be reasonably predicted from the demographic, biochemical, and anthropometric/body composition parameters. There was little difference between the ROC-AUCs of the best nonlinear model and the final model with elastic-net regularization; adding more complexity didn't offer a great leap in discrimination. A simpler and more interpretable model has a lot of clinical relevance as it is more likely to be evaluated, communicated, and used in decision support workflows. Repeated stratified cross-validation and feature selection that does not allow leakage despite a relatively small sample size and a high number of candidate predictors still presented a risk of overfitting. The quality of the prediction model should also be evaluated based on the model design, analysis, and applicability to the target clinical population (Moons et al., 2025).

The final model had generally acceptable calibration values, which were an intercept of 0.026 and a slope of 1.111. The values showed that the predicted probabilities were fairly similar to the actual ones, though the slope of > 1.0 showed there was some underfitting of the distribution of predicted risks or that there was a lack of diversity. The use of the model was especially important in the process of calibration, as a model can discriminate well and at the same time underestimate or overestimate the individual risk. When model outputs are employed to prioritize diagnostics, reliable estimates of probability become particularly important when the patients could be divided into low-, intermediate-, and high-risk groups (Huang et al., 2020). With these prevalence rates of gallstones observed in these groups, the model was able to differentiate between clinical groups.

The optimized probability threshold of 0.393 resulted in an increase of sensitivity from 72.8% to 85.4%, and a decrease of specificity from 83.2% to 72.0%. This trade-off can be acceptable if the model is employed as a screening/triage tool to look for those patients who might warrant confirmatory ultrasonography. A lower threshold will increase the number of patients who are referred who are positive for gallstones, but will also increase the number of negative patients being referred. Therefore, the selection of threshold should be based on the clinical use, resources available for imaging, the risk of delayed diagnosis, and the level of acceptable burden of further investigation. Evaluation of the model should be carried out beyond the development phase and include both intended-use and external validation and implementation before considering (Collins et al., 2024b).

Finally, the decision-curve analysis showed that the final model had the highest estimated net benefit when threshold probabilities were between about 0.21 and 0.80, suggesting that it would be beneficial to evaluate all patients or no patients at these levels. This result indicated the possibility of clinical utility in a wide spectrum of clinical preferences. Because it takes into account the consequences of true positive and false positive classifications—rather than just statistical accuracy measures—it is called decision-curve analysis (Vickers & Holland, 2021). This should be viewed with caution, as the benefits of the calculated net benefit were projections based on past data and not actual improvements in workflow, patient outcomes, or healthcare utilization. The model thus had potential clinical usefulness, but lacked clinical impact. This is relevant to the clinical consequences of using models with similar AUC; some have extremely different consequences based on the threshold chosen and the context of use (Sadatsafavi et al., 2021).

The bone mass, vitamin D, C-reactive protein, extracellular fluid distribution, visceral fat area, gender and hyperlipidaemia were identified as influential variables in the permutation interpretation. These variables were in the domains of nutrition, inflammation, metabolism and body composition, which could all contribute to the susceptibility to gallstone disease. Their significance should not be interpreted as that of a causal effect, as they are measures of contribution to prediction, not biological effect. Explainable outputs can enhance clinician trust by highlighting the most significant variables in a prediction, but simplified output tools can also be misleading, hard to interpret, and misapplied without the context of the clinical situation (Abbas et al., 2025).

The extended predictor set only slightly outperforms the routine clinical predictor set. This indicated that body-composition measurements provide extra information and were not necessary for clinically useful prediction. It may be more practical to use a reduced model that uses variables that are readily available in areas without bioimpedance devices. Performance must now be externally validated to be sure it is consistent within hospitals, populations, laboratory systems, and disease-prevalence environments. Validating such should maintain the original model, evaluate calibration and discrimination separately, and analyse the performance of subgroups as well as determine if there is a

need for recalibration before use (Riley et al., 2024).

The study had several limitations, including being single-source with a small sample, retrospective in design, using internal validation only, not being completely certain of the coding of some of the variables, one implausible measurement of obesity, and lack of symptomatic, imaging, treatment, and follow-up data. There was no comparator of clinician judgment, and the model was not rolled out in clinical practice. The results thus show the viability of a clinical decision support model for research, rather than for immediate clinical use.

5. Conclusion

Gallstone disease prediction using routine and extended clinical data was feasible through an internally validated artificial intelligence-based clinical decision-support model. The final elastic-net logistic regression model was more clinically meaningful in terms of risk stratification, acceptable in terms of calibration, and better discriminatory than the top nonlinear models, and was also more interpretable. The optimized probability threshold increased the model's sensitivity, while maintaining adequate specificity, suggesting the model's potential to be used to identify patients who could benefit from priority ultrasonography or further evaluation. The results of the decision curve analysis also indicated a net benefit over a wide range of clinically relevant thresholds. Vitamin D, C-reactive protein, bone mass, visceral fat area, extracellular fluid distribution, gender, and hyperlipidemia were important contributors to predicting these outcomes, but not interpreted causally. The modest improvement in performance from body-composition variables suggested that a simpler model of the routine data might be plausible in non-equipment-based settings. The results corroborate the model for use as a research stage decision support tool and not as a substitute for clinical judgment. External multicenter validation, prospective implementation, workflow testing, and comparing with clinician-only decision-making are needed before routine clinical adoption.

References

1. Abbas, Q., Jeong, W., & Lee, S. W. (2025, August). Explainable AI in clinical decision support systems: a meta-analysis of methods, applications, and usability challenges. In *Healthcare* (Vol. 13, No. 17, p. 2154). MDPI.
2. Ahmed, A. S., Ahmed, S. S., Mohamed, S., Salman, N. E., Humidan, A. A. M., Ibrahim, R. F., ... & Elamir, A. (2024). Advancements in cholelithiasis diagnosis: a systematic review of machine learning applications in imaging analysis. *Cureus*, 16(8).
3. Blum, J., Wood, L., & Turner, R. (2025). Artificial intelligence in the detection of choledocholithiasis: a systematic review. *HPB*, 27(1), 1-9.
4. Cao, Z., Wei, J., Zhang, N., Liu, W., Hong, T., He, X., & Qu, Q. (2020). Risk factors of systematic biliary complications in patients with gallbladder stones. *Irish Journal of Medical Science* (1971-), 189(3), 943-947.
5. Collins, G. S., Dhiman, P., Ma, J., Schlusser, M. M., Archer, L., Van Calster, B., ... & Riley, R. D. (2024). Evaluation of clinical prediction models (part 1): from development to external validation. *bmj*, 384.
6. Collins, G. S., Moons, K. G., Dhiman, P., Riley, R. D., Beam, A. L., Calster, B. V., ... & Logullo, P. (2024). TRIPOD+ AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *bmj*.
7. Esen, I., Arslan, H., Aktürk, S., Gülşen, M., Kültekin, N., & Özdemir, O. (2024). Gallstone [Dataset]. UCI Machine Learning Repository. <https://archive.ics.uci.edu/dataset/1150/gallstone1>
8. Esen, İ., Arslan, H., Esen, S. A., 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.
9. Gu, Q., Zhou, G., & Xu, T. (2020). Risk factors for gallstone disease in Shanghai: an observational study. *Medicine*, 99(3), e18754.
10. Huang, Y., Li, W., Macheret, F., Gabriel, R. A., & Ohno-Machado, L. (2020). A tutorial on calibration measurements and calibration models for clinical prediction models. *Journal of the American Medical Informatics Association*, 27(4), 621-633.
11. Kumar, D., Mehta, M. A., Kotecha, K., & Kulkarni, A. (2025). Computer-aided cholelithiasis diagnosis using explainable convolutional neural network. *Scientific Reports*, 15(1), 4249.
12. Liu, S., McCoy, A. B., Peterson, J. F., Lasko, T. A., Sittig, D. F., Nelson, S. D., ... & Wright, A. (2024). Leveraging explainable artificial intelligence to optimize clinical decision support. *Journal of the American Medical Informatics Association*, 31(4), 968-974.
13. Moons, K. G., Damen, J. A., Kaul, T., Hooft, L., Navarro, C. A., Dhiman, P., ... & van Smeden, M. (2025). PROBAST+ AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *bmj*, 388.
14. Pisano, M., Allievi, N., Gurusamy, K., Borzellino, G., Cimbanassi, S., Boerna, D., ... & Ansaloni, L. (2020). 2020 World Society of Emergency Surgery updated guidelines for the diagnosis and treatment of acute calculus cholecystitis. *World journal of emergency surgery*, 15(1), 61.
15. Riley, R. D., Archer, L., Snell, K. I., Ensor, J., Dhiman, P., Martin, G. P., ... & Collins, G. S. (2024). Evaluation of clinical prediction models (part 2): how to undertake an external validation study. *bmj*, 384.
16. Sadatsafavi, M., Adibi, A., Puhani, M., Gershon, A., Aaron, S. D., & Sin, D. D. (2021). Moving beyond AUC: decision curve analysis for quantifying net benefit of risk prediction models. *European Respiratory Journal*.

17. Sarker, P., Tiang, J. J., & Nahid, A. A. (2025). Gallstone Classification Using Random Forest Optimized by Sand Cat Swarm Optimization Algorithm with SHAP and DiCE-Based Interpretability. *Sensors*, 25(17), 5489.
18. Song, S. T., Shi, J., Wang, X. H., Guo, Y. B., Hu, P. F., Zhu, F., ... & Xie, W. F. (2020). Prevalence and risk factors for gallstone disease: a population-based cross-sectional study. *Journal of digestive diseases*, 21(4), 237-245.
19. Vickers, A. J., & Holland, F. (2021). Decision curve analysis to evaluate the clinical benefit of prediction models. *The Spine Journal*, 21(10), 1643-1648.
20. Wang, J., Shen, S., Wang, B., Ni, X., Liu, H., Ni, X., ... & Liu, H. (2020). Serum lipid levels are the risk factors of gallbladder stones: a population-based study in China. *Lipids in health and disease*, 19(1), 50.
21. Wirth, J., Joshi, A. D., Song, M., Lee, D. H., Tabung, F. K., Fung, T. T., ... & Wu, K. (2020). A healthy lifestyle pattern and the risk of symptomatic gallstone disease: results from 2 prospective cohort studies. *The American journal of clinical nutrition*, 112(3), 586-594.