Open Access and Peer Reviewed Journal ISSN 2394-2231 https://www.ijctjournal.org/

RESEARCH ARTICLE OPEN ACCESS

Evolutionary Hyperparameter Optimization of Machine Learning Classifiers for Gallstone Disease Prediction

Mohit Kharbanda[1], Arshi Husain[2], Virendra P. Vishwakarma[3]

[1,2,3]*USIC&T, GGSIPU, Delhi, India*

[1]mohit93kharbanda@gmail.com

Abstract – This study investigate whether evolutionary hyperparameter search can meaningfully improve classifier performance for gallstone disease (GSD) diagnosis. Working from a cohort of 319 patients described by 38 clinical and demographic variables, we split the data into training (80%) and testing (20%) partitions using stratified sampling, and standardized features using statistics computed solely from the training partition to avoid information leakage. A Genetic Algorithm (GA) then searched the hyperparameter space of six classifiers – Random Forest (RF), Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbors (KNN), and Logistic Regression (LR) with each candidate configuration scored by 5-fold stratified cross-validation restricted to the training partition. Tuned models were refit on the full training set and evaluated once, on the held-out test partition, using accuracy, precision, recall, F1-score, and confusion-matrix analysis. XGBoost led every metric at 90.62%, with a ROC-AUC of 93.46%, RF and SVM tied at 79.69% accuracy, LR reached 78.12%, KNN 67.19%, and DT 62.50%. The margin between XGBoost and the remaining classifiers indicates that, for this dataset, pairing gradient boosting with evolutionary tuning offers a genuine advantage over the alternatives tested.

Keywords – Gallstone disease, Machine learning, Genetic Algorithm, Synthetic resampling, XGBoost, Ensembles, Decision support systems

1. Introduction

Cholelithiasis (gallstone disease) or GSD is one of the more common conditions affecting the hepatobiliary system, arising when stones form within the gallbladder or the bile ducts. Patients can stay symptom-free for years or progress toward serious complications. cholecystitis, obstruction of the biliary tract or pancreatitis triggered by a lodged stone. Catching the condition early therefore has real clinical weight. Standard diagnostic pathways history and physical exam, blood work, ultrasonography, CT work well in principle, but their accuracy in practice depends heavily on case complexity and on the interpreting clinician's experience. That dependency is part of why computer-aided diagnostic tools for gallstone detection have drawn sustained research interest.

Machine learning (ML) and, increasingly, deep learning (DL) have found a natural niche in this space. Classifiers built on Logistic Regression (LR), SVM, RF, gradient boosting variants (including XGBoost), KNN and neural network architectures now appear routinely in gallstone risk-assessment studies.

One representative comparison pitted LR against Gaussian Naive Bayes, a Multilayer Perceptron, and SVM for gallstone risk prediction. LR was ahead on AUC [1]. A separate strand of research built purely on bioelectrical impedance measurements found that BMI, body fat percentage, waist circumference and gender alone carried enough signal to be clinically useful [5]. Still other work has folded in biochemical markers Vitamin D level, C-reactive protein, total body water, lean mass as additional predictors [8, 9].

Since no individual variable or model tends to dominate on its own, ensemble strategies have become common. Pipelines stacking RF, XGBoost, LR, Gradient Boosting, Extra Trees, LightGBM, and SVM together have consistently beaten their individual component models on gallstone classification benchmarks[16, 17]. A parallel thread applies metaheuristic search rather than fixed ensembling, tuning feature subsets and model settings jointly and reporting meaningful gains from doing so [18]. A result that reinforces the broader point that both *which* features are used and *how* a model is configured matter for GSD prediction.

Imaging-based approaches form a second, largely separate research track, with ultrasound as the default non-invasive modality for spotting gallstones. Bozdag et al. built a content-based retrieval system on CNN-derived features and reported 0.92711 average precision for gallstone cases [4]; a related pipeline combined CNN feature extraction with a conventional ML classifier on top and reported comparable gains [15]. Working instead from handcrafted texture and histogram features, Hong et al. reached 96.33% accuracy with an RF classifier on ultrasound regions of interest [3]. DL methods have since been extended to cine-image sequences for spotting acute gallbladder pathology [14], and a scoping review of this literature frames AI as an increasingly standard second opinion for radiologists reading gallbladder ultrasounds [13].

CT imaging adds a third data source. RF models built on combined clinical and CT features have exceeded 0.91 AUC when predicting the severity of acute gallstone pancreatitis [2], and dedicated deep-learning pipelines for CT-based stone detection are an active area [19] – together confirming that AI-assisted diagnosis is not confined to any single imaging modality.

What is comparatively rare in this literature is work that (a) relies purely on structured clinical and biochemical data rather than imaging, and (b) tunes model hyperparameters systematically rather than using library defaults or a handful of manually chosen settings. Several studies touch feature optimization or explainability in isolation, but the two rarely appear together inside one pipeline [9, 10]. This problem is important from a practical point of view, because a diagnostic model that cannot demonstrate which variables determine its predictions will be difficult for practitioners to accept.

This paper aims to directly address this gap. We propose a GSD classification pipeline that is purely based on structured clinical and biomedical variables, employs an evolutionary search strategy for exploring the hyperparameters of each classifier rather than manually tuning them, and isolates the test partition during this search. Specifically, our contributions are:

1. An ML-based classification framework for gallstone disease built on structured clinical and biomedical parameters.
2. A Genetic Algorithm-driven hyperparameter search applied independently to six different classifier families.
3. A head-to-head comparison of those six classifiers under matched optimization conditions.
4. A leakage-controlled experimental design, with GA search and cross-validation confined strictly to the training partition.
5. A quantitative evaluation against standard classification metrics to gauge how well the approach generalizes.

The rest of the paper proceeds as follows. Section II surveys related work. Section III lays out the dataset and methodology. Section IV reports and discusses results. Section V concludes and points to future directions.

2. Literature Review

Gallbladder disease study GSD study in particular involves today a large variety of types of data: clinical, biochemical, bioimpedance, ultrasound, CT, and metabolomic [1, 5, 8, 4]. The subsections that follow categorize earlier studies according to the types of data collected instead of their order of time of publication.

Structured clinical and body-composition data

LR, GNB, MLP, and SVM were compared in terms of their performance on NHANES survey data and clinical cohort through SHAP analysis, which revealed Relative Fat Mass, Weight-to-Waist Index, Waist Circumference, Waist-to-Height Ratio, and BMI to be the best predictors, LR performed the best with AUC scores of 0.699 on training and 0.727 on validation sets [1]. Lu et al. employed bioelectrical impedance dataset without any additional clinical data to build a nomogram based on gender, BMI, body fat percentage, and waist circumference reaching 0.770 AUC value, which can be regarded as a benchmark in terms of how much information non-invasive body composition analysis provides [5]. This was supplemented by Esen et al., who added laboratory measurements to clinical and bioimpedance data [8]. Another method of tuning was used by Sarker et al., who optimized RF classifier using Sand Cat Swarm Optimization and then added SHAP/DiCE explanation [9].

Ensembles and metaheuristic search

Ensemble approaches feature prominently in this area. Indirani et al. discovered that stacking up clinical and biochemical factors in an ensemble outperformed each of their model-only benchmarks [16], while Akben and Yumrutaş found a similar trend with a non-invasive risk-prediction ensemble for gallstones [17]. Conversely, Zeqo and Naka went the route of metaheuristics and used them to find optimal features and parameters, discovering some tangible benefits as well [18]. Lastly, Salem et al. chose yet another approach by using metabolomics together with clinical factors in machine learning and deep learning models [7].

Ultrasound-based imaging

For the imaging methods, the CNN-feature retrieval approach of Bozdag et al. differentiated between multiple gallbladder diseases with very high precision in the retrieval task [4], while a later approach combining CNN features with a traditional classifier achieved a similar result [15], showing good transferability of pre-trained CNNs as a feature extraction approach even without using a full end-to-end classifier architecture. Hong et al. took a different path by employing handcrafted binary, histogram, and texture features, which were used to fine-tune the classifier (random forest), achieving the best results [3]. A recent scoping review of this sub-field by Wang et al. highlights the fact that AI approaches are increasingly becoming integrated into the standard workflow of ultrasound for gallbladder examination [13].

CT-based approaches

Ma et al. utilized CT scan characteristics along with clinical factors in an RF model to classify the severity of acute gallstone pancreatitis, achieving 0.91 AUC [2]. Veena and Gowrishankar tested SSD, Faster R-CNN, and Mask R-CNN for the purpose of automating the process of stone localization from CT [6], while Meyer et al. used a specially developed deep learning model for CT scan stone localization in their recent study [19].

Explainability

There is, however, a relatively small yet emerging strand that takes interpretability as a primary objective. Kumar et al. created an explainable CNN so that clinicians could understand its reasoning on cases of cholelithiasis [10], while the SCSO-optimized RF model created by Sarker et al. used SHAP and DiCE explanation tools [9].

Where this leaves open ground

In the literature, two themes emerge. First, research focused on images [4, 3, 14, 19] significantly outweighs research based solely on structured data and/or biochemical information. Second, very few approaches involve a systematic search for the optimal set of hyperparameters because most pipelines rely either on one algorithm or an ensemble of algorithms without optimization; very few papers discuss how optimization and evaluation are done on completely disjoint sets of data. The framework introduced in this paper addresses all of the above aspects at once.

3. Materials and Methods

3.1. Dataset Description

The dataset used in this work has the following properties:

- **Records:** 319 patients.
- **Attributes:** 39 total – 38 predictor variables plus the target label, Gallstone Status.
- **Class balance:** 161 records labeled class 0, 158 labeled class 1 – close to an even split.
- **Feature retention:** all 38 predictors kept throughout; no feature selection or dimensionality reduction was applied at any stage.
- **Missing-value handling:** entries that could not be parsed as numeric were treated as missing and filled with that feature's median value.
- **Target encoding:** the categorical label was mapped to binary integers via label encoding.

The end result of this preprocessing is a single feature matrix retaining all 38 original attributes, used unmodified for every model trained afterward.

3.2. Training and Testing Strategy

Evaluation followed a strict train/test separation:

- An 80:20 stratified split preserved class proportions in both partitions.
- 255 records went into training and hyperparameter search; the remaining 64 were held out exclusively for final testing.
- The 64-record test partition never entered the GA optimization loop or any cross-validation fold.
- Feature scaling used `StandardScaler` fitted only on the training partition, then applied unchanged to both partitions – so no test-set statistics influenced the scaling parameters.

This separation is what keeps the reported test performance a fair estimate of generalization rather than an artifact of the tuning process.

3.3. Genetic Algorithm-Based Hyperparameter Optimization

Hyperparameter search – and only hyperparameter search – was delegated to a Genetic Algorithm (GA), implemented with the DEAP evolutionary computation library. Feature selection was explicitly out of scope for the GA; all 38 predictors remained in play for every classifier.

Each chromosome encoded one candidate hyperparameter configuration. An initial population was sampled randomly within pre-set parameter bounds, and each chromosome's fitness was its mean accuracy across 5-fold stratified cross-validation on the training partition, with a fixed random seed controlling the fold shuffling for reproducibility.

The evolutionary settings were:

- **Crossover:** two-point, probability 0.7
- **Mutation:** uniform integer, probability 0.3
- **Selection:** tournament, tournament size 3
- **Generations:** 10
- **Population size:** 15

The chromosome with the best mean cross-validation accuracy at the end of the run became that classifier's final hyperparameter configuration. This process ran independently, from a fresh population, for each of the six target classifiers: RF, XGBoost, SVM, DT, KNN, and LR.

3.4. Evaluated Machine Learning Models

Six classifiers were trained and compared under this framework.

3.4.1. K-Nearest Neighbors (KNN). KNN makes no assumptions about the underlying data distribution and performs no explicit training pass; at inference time, it looks up the k closest training points to a query and votes among their labels. GA search here covered the neighbor count, the vote-weighting scheme, and the distance metric.

3.4.2. Support Vector Machine (SVM). SVM looks for the widest possible margin between classes, choosing the hyperplane that separates them with the most breathing room on either side; a kernel trick lets that boundary bend to fit non-linear patterns without explicitly mapping into higher-dimensional space. GA search covered the regularization strength C, kernel coefficient γ, and kernel choice.

3.4.3. Random Forest (RF). RF trains many decision trees on bootstrapped resamples of the training data and averages their votes, which damps the overfitting tendency any single tree would show on its own. GA search here covered tree count, depth, split thresholds, leaf size, and the feature-sampling strategy used at each split.

3.4.4. Extreme Gradient Boosting (XGBoost). XGBoost grows trees one at a time, and each addition is fit specifically to the errors the ensemble-so-far still gets wrong, rather than being

Table 1. Classifier Hyperparameters Tuned via Genetic Algorithm

Classifier	Optimized Hyperparameter Space
Random Forest (RF)	Number of Estimators (`n_estimators`), Maximum Tree Depth (`max_depth`), Minimum Samples Split (`min_samples_split`), Minimum Samples Leaf (`min_samples_leaf`), Maximum Features (`max_features`)
XGBoost (XGB)	Number of Estimators (`n_estimators`), Maximum Tree Depth (`max_depth`), Learning Rate (η), Minimum Child Weight (`min_child_weight`), Subsample Ratio (`subsample`), Column Subsampling Ratio (`colsample_bytree`)
Support Vector Machine (SVM)	Regularization Parameter (C), Kernel Coefficient (γ), Kernel Type
Decision Tree (DT)	Maximum Tree Depth (`max_depth`), Minimum Samples Split (`min_samples_split`), Minimum Samples Leaf (`min_samples_leaf`), Splitting Criterion
K-Nearest Neighbors (KNN)	Number of Neighbors (`n_neighbors`), Weighting Strategy (`weights`), Distance Metric
Logistic Regression (LR)	Regularization Parameter (C), Solver

trained independently like an RF tree. GA search covered the estimator count, tree depth, learning rate, minimum child weight, and the two subsampling ratios (row-wise and column-wise).

3.4.5. Decision Tree (DT). A single DT recursively splits the feature space along threshold rules chosen to best separate the classes at each node, producing a decision path a clinician could read directly. GA search covered maximum depth, split and leaf sample thresholds, and the splitting criterion itself.

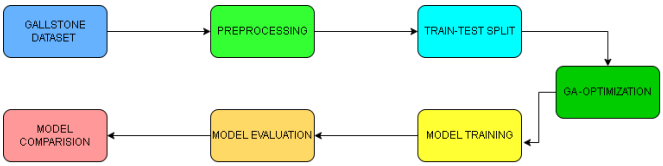

Figure 1. Work flow of our methodology

3.4.6. Logistic Regression (LR). LR fits a linear decision boundary and passes it through a logistic function to produce a class-membership probability rather than a hard label. GA search here was limited to the regularization strength C and the solver algorithm.

3.5. Experimental Workflow

The full pipeline runs in sequence: preprocess the raw data, apply the 80:20 stratified split, fit the scaler on the training partition only, run GA-based hyperparameter search per classifier under 5-fold stratified cross-validation, refit each tuned model on the complete training set, and score once on the untouched test partition. Accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices are then compared across classifiers to identify the strongest performer.

3.6. Performance Metrics

Four confusion-matrix quantities – true/false positives and negatives (TP, FP, TN, FN) – feed all reported metrics:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{3}$$

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Accuracy (Eq. 1) captures overall correctness; precision (Eq. 2) captures how trustworthy a positive prediction is; recall (Eq. 3) captures how many true positive cases the model actually catches; and F1 (Eq. 4) balances the latter two into a single number. ROC-AUC is reported alongside these as a threshold-independent measure of separability between the two classes.

4. Experimental Results and Discussion

Table 2 lays out how the six GA-tuned classifiers compare once evaluated on the held-out test partition. The spread between the top and bottom performers is wide – close to 30 accuracy points – which is itself informative: it suggests the choice of algorithm matters more here than any remaining tuning headroom within a given algorithm family.

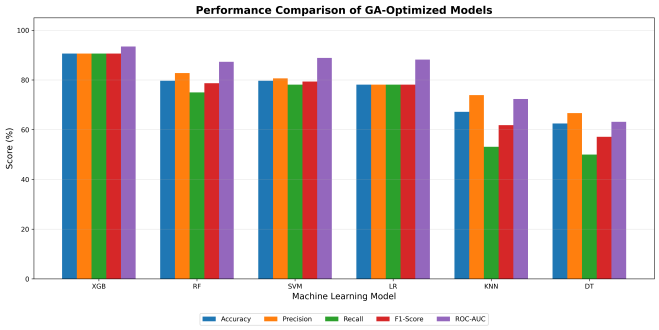

Figure 2. Performance Comparison of all models

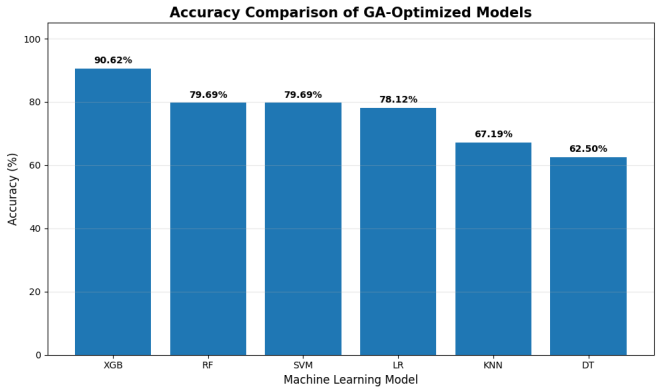

Figure 3. Accuracy Comparison of all models

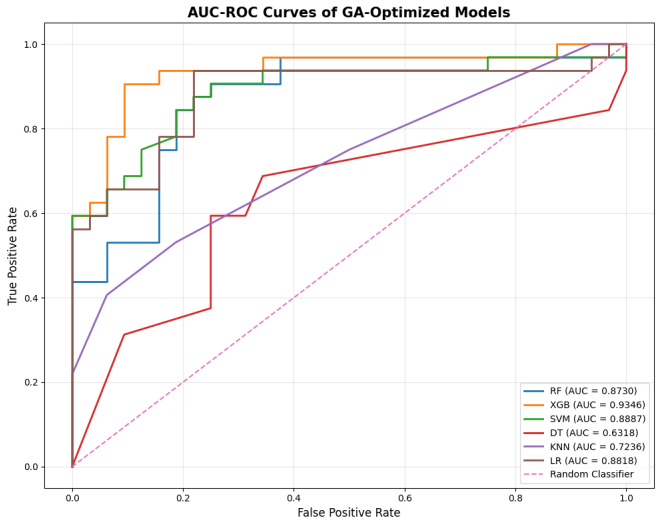

Figure 4. AUC Curves of all models

XGBoost takes a clear lead from the rest, tying up with 90.62in terms of accuracy, precision, recall, and F1-Score, along with ROC-AUC of 93.46%. The interesting thing to remember here is that the model's precision and recall have come out to be equal values, which implies that there is no compromise being made between the two.

RF and SVM are quite close competitors, each of them having an accuracy of 79.69% but reaching that threshold differently: the accuracy of RF of 82.76% compared to the recall of 75.00% implies that minimizing the number of false positives is preferred, while for SVM, recall is 78.12%, compared to the precision of 80.65%. This can be confirmed by their ROC-AUC indicators (87.30% and 88.87%, respectively).

Table 2. Performance Comparison Across GA-Optimized Classification Models

Metric	XGBoost	RF	SVM	LR	KNN	DT
Accuracy (%)	**90.62**	79.69	79.69	78.12	67.19	62.50
Precision (%)	**90.62**	82.76	80.65	78.12	73.91	66.67
Recall (%)	**90.62**	75.00	78.12	78.12	53.12	50.00
F1-score (%)	**90.62**	78.69	79.37	78.12	61.82	57.14
ROC-AUC (%)	**93.46**	87.30	88.87	88.18	72.36	63.18

LR came in at 78.12% accuracy and 88.18% ROC-AUC notably, its ROC-AUC is close to RF and SVM's despite a lower raw accuracy, suggesting the default 0.5 threshold may not be optimal for this particular model on this dataset. KNN and DT lagged well behind both, at 67.19% and 62.50% accuracy; their recall figures (53.12% and 50.00%) are the real story here – both models are missing roughly half of the true positive cases, which would be a serious limitation in any clinical deployment context.

Why XGBoost wins here is at least partly explainable by what it is structurally: an additive ensemble that keeps correcting its own residual errors round by round, rather than fitting once and stopping. Combined with a hyperparameter set the GA searched specifically for this dataset – rather than defaults borrowed from another problem – that iterative error-correction seems to translate into a real accuracy advantage rather than a marginal one. Both KNN and DT, by contrast, have comparatively little capacity to model interactions among 38 correlated clinical variables, which likely explains why they trail so far behind the ensemble and margin-based methods.

5. Conclusion

We built a GA-optimized classification pipeline for gallstone disease using 38 clinical and demographic variables from 319 patients, with hyperparameter search, cross-validation, and model comparison all kept strictly within the training partition until final evaluation.

Across the six classifiers tested, XGBoost was the clear winner: 90.62% on accuracy, precision, recall, and F1-score, correctly classifying 58 of the 64 held-out test cases. RF and SVM followed at 79.69% accuracy, LR at 78.12%, and KNN and DT trailed at 67.19% and 62.50% respectively.

The implication from this analysis can be straightforward: the use of evolutionary tuning combined with a gradient boosting classifier appears to be a powerful combination for this type of structured-data GSD classification problem. However, the dataset under consideration is small in size 319 samples which means that the following natural question is how this advantage would manifest itself in larger, more diverse cohorts of patients.

6. Future Work

There are several directions that can be pursued based on the work done. Applying the pipeline to a larger dataset could provide insight into how generalizable the results obtained are. Alternative optimization strategies swarm based methods, Bayesian optimization are worth benchmarking against the GA approach used here to see whether the choice of search algorithm itself matters much once each is given a comparable search budget.

Layering explainable AI techniques (SHAP, feature-importance analysis) on top of the tuned XGBoost model would address the interpretability gap noted in Section II, and would likely be a prerequisite for any real clinical adoption. External validation on cohorts drawn from different populations or care settings is the other obvious gap, since a single 319-patient dataset cannot speak to how the model behaves outside the population it was trained on. Longer term, the best-performing model here could form the core of a clinically oriented decision-support tool, provided it goes through appropriate calibration, validation, and interpretability work first.

Conflict of Interest

The authors declare no competing interests related to this work.

Acknowledgment

This research was supported by a fellowship under the Visvesvaraya PhD Scheme, which the authors gratefully acknowledge.

References

[1] L. Deng, S. Wang, D. Wan, Q. Zhang, W. Shen, X. Liu, and Y. Zhang, "Relative fat mass and physical indices as predictors of gallstone formation: Insights from machine learning and logistic regression," *International Journal of General Medicine*, pp. 509–527, 2025.

[2] Y. Ma, P. Yue, J. Zhang, J. Yuan, Z. Liu, Z. Chen, H. Zhang, C. Zhang, Y. Zhang, C. Dong et al., "Early prediction of acute gallstone pancreatitis severity: A novel machine learning model based on CT features and open access online prediction platform," *Annals of Medicine*, vol. 56, no. 1, p. 2357354, 2024.

[3] C. Hong, I. Zafar, M. M. Ayaz, R. Kanwal, F. Kanwal, M. Dauelbait, M. Bourhia, and Y. A. B. Jardan, "Analysis of machine learning algorithms for real-time gallbladder stone identification from ultrasound images in clinical decision support systems," *International Journal of Computational Intelligence Systems*, vol. 18, no. 1, pp. 1–28, 2025.

[4] A. Bozdag, M. Yildirim, M. Karaduman, H. B. Mutlu, G. Karaduman, and A. Aksoy, "Detection of gallbladder disease types using a feature engineering-based developed CBIR system," *Diagnostics*, vol. 15, no. 5, p. 552, 2025.

[5] J.-H. Lu, G.-X. Tong, X.-Y. Hu, R.-F. Guo, and S. Wang, "Construction and evaluation of a nomogram to predict gallstone disease based on body composition," *International Journal of General Medicine*, pp. 5947–5956, 2022.

[6] A. Veena and S. Gowrishankar, "Context based healthcare informatics system to detect gallstones using deep learning methods," *International Journal of Advanced Technology and Engineering Exploration*, vol. 9, no. 96, p. 1661, 2022.

[7] N. M. Salem, K. M. Jack, H. Gu, A. Kumar, M. Garcia, P. Yang, and V. Dinu, "Machine and deep learning identified metabolites and clinical features associated with gallstone disease," *Computer Methods and Programs in Biomedicine Update*, vol. 3, p. 100106, 2023.

[8] I. 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, 2024.

[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, 2025.

[10] D. Kumar, M. A. Mehta, K. Kotecha, and A. Kulkarni, "Computer-aided cholelithiasis diagnosis using explainable convolutional neural network," *Scientific Reports*, vol. 15, p. 4249, 2025.

[11] A. M. Obaid, A. Turki, H. Bellaaj, and M. Ksantini, "Diagnosis of gallbladder disease using artificial intelligence: A comparative study," *International Journal of Computer Intelligence Systems*, vol. 17, p. 46, 2024.

[12] K. Kanwal, I. Batool, I. Naz, A. Shabbir, M. Fatima, and H. Batool Gill, "Automated detection of gallstone disease using machine learning techniques: A data-driven approach," *JSCIR*, vol. 7, no. 2, pp. 157–176, 2025.

[13] X. Wang, H. Zhang, Z. Bai, X. Xie, and Y. Feng, "Current status of artificial intelligence analysis for the diagnosis of gallbladder diseases using ultrasonography: A scoping review," *Translational Gastroenterology and Hepatology*, vol. 10, 2024.

[14] C. Ge, J. Jang, P. Svrcek, V. Fleming, and Y. H. Kim, "Exploring deep learning applications using ultrasound single view cines in acute gallbladder pathologies: Preliminary results," *Academic Radiology*, vol. 32, no. 2, pp. 770–775, 2025.

[15] R. A. Putra, G. A. Pradipta, and P. D. W. Ayu, "Gallbladder disease classification from ultrasound images using CNN feature extraction and machine learning optimization," *Journal of Electronic, Electromedical, and Medical Informatics*, vol. 7, no. 4, pp. 1089–1111, 2025.

[16] M. Indirani, S. Sudheer, R. Mahaveerakannan et al., "Gallstone disease prediction using clinical and biochemical features through ensemble learning techniques," *International Journal of Computational Intelligence Systems*, vol. 19, p. 19, 2026.

[17] S. B. Akben and H. Yumrutaş, "A new ensemble learning approach for non-invasive gallstone risk prediction," in *2025 16th International Conference on Electrical and Electronics Engineering (ELECO)*, Bursa, Türkiye, 2025, pp. 1–6.

[18] E. Zeqo and E. Naka, "A metaheuristic and machine learning-based algorithm for predicting gallstone disease," in *2025 International Conference on Big Data, Knowledge and Control Systems Engineering (BdKCSE)*, Bankya, Bulgaria, 2025, pp. 1–8.

[19] M. T. Meyer, J. Wasserthal, J. Cyriac et al., "Automated detection of gallbladder stones using a deep learning algorithm on computed tomography scans," *Abdominal Radiology*, 2026.