Journal of Statistics, Computing and Interdisciplinary Research
ISSN (P): 2707-7101

J. Stat. Comp. & Interdis. Rese.
Volume 7(2) (2025), 157-176
http://scir.wum.edu.pk/index.php/ojs

Automated Detection of Gallstone Disease Using Machine Learning Techniques: A Data-Driven Approach

Khadija Kanwal[1*], Iqra Batool[1], Ishmal Naz[1], Aiza Shabbir[1], Meherwar Fatima[1], Humera Batool Gill[1]

[1] Institute of Computer Science and Information Technology, The Women University Multan, Pakistan.

***Corresponding Author:** Khadija Kanwal, khadijakanwal.6022@wum.edu.pk

Abstract

Gallstone disease is a very general gastrointestinal condition characterized by the origination of stones within the gallbladder, which can lead to inflammation, abdominal pain, and serious complications. Early detection of this disease is helpful for effective treatment. This research presents a comprehensive framework to evaluate the performance of stacked ensemble classifiers in supervised learning tasks. The presented approach employs three heterogeneous model ensembles: (i) XGBoost, LightGBM, and CatBoost; (ii) Random Forest, Extra Trees, and Gradient Boosting; and (iii) K-Nearest Neighbors, Support Vector Classifier, and Decision Tree. These models were integrated with a meta-learning strategy using Logistic Regression within a Stacking-Classifier architecture, with pass through enabled to enrich the meta-model with original input features alongside base learner predictions. Stratified five-fold cross-validation was applied to evaluate the performance and ensure the robustness of the presented model across varying class distributions. This multi-level validation approach enabled a rigorous comparison across diverse classifier families. The proposed model achieves 97.0\% accuracy, demonstrating significant improvement over existing research. The experimental results demonstrate that the proposed method has significant potential to support decision-making, reduce diagnostic errors, and improve patient outcomes.

Keywords: *Computed tomography (CT); Gallstone disease (GSD); Recursive Feature Elimination (RFE); K-Nearest Neighbors (KNN); Decision Tree (DT).*

1. Introduction

Gallstones are small, but hard stones form inside the gallbladder, which is a tiny organ just below the liver that helps you digest fats. These stones form due to the presence of cholesterol or other substances in bile, a fluid produced by the body to break down food. Gallstones do not cause symptoms, but can cause sudden stomach pain, bloating, or nausea, especially after eating fatty foods. Major causes such as being overweight, losing weight too quickly, or having a family history of gallstones can increase the risk. In addition, complications range from acute conditions such as cholecystitis, cholangitis, and pancreatitis to long-term risk of gallbladder cancer. For this purpose, ultrasound serves as the primary diagnostic imaging modality due to its cost-effectiveness and high sensitivity; however, CT and MRCP provide additional precision when needed [1]. The rate of gallstone disease is increasing due to a combination of biological differences between sexes, unhealthy habits, and inherited traits. Although some people may not feel any symptoms, others can have serious health problems, such as gallbladder infections or blocked bile flow. Recent research studies have shown that the gallbladder is not just a storage organ—it also plays a role in how the body manages fat and sugar [2]. Portincasa and colleagues linked gallstone disease to broader metabolic disorders, explaining how these conditions are connected and why early prevention is so important [3].

In another study [4] described the key factors that increase gallstones, including older age, obesity, elevated BMI, diabetes mellitus, hypercholesterolemia, and menopause in women. This higher risk among women is closely related to metabolic health. In support of this, Bastard and Fève point out that inflammation in fat tissue, both throughout the body and around organs, plays a major role in the health problems caused by obesity. They suggest that targeting this inflammation could help treat conditions such as obesity and type 2 diabetes [5]. Abdominal ultrasound is still the first choice for detecting gallstones, with around 84\% detection ability and 99\% accuracy [6–8]. However, it becomes less reliable for the detection of very small stones, especially those less than 3 mm in size [9,10]. In addition, tools such as bioelectrical impedance methods, including BIA and the more advanced BIVA, help assess body composition. These methods are useful for creating personalized nutrition plans and evaluating a person's risk of metabolic disorders.

Several factors can increase the risk of developing gallstones, such as high BMI, larger waist, diabetes, and smoking. On the other hand, drinking coffee regularly can help reduce risk, while alcohol does not seem to have much effect [11,12] . Bioelectrical impedance analysis (BIA) is also used in cancer care to measure body composition and help predict patient outcomes, although more studies are needed to make it more reliable in clinical practice. Diagnosis of malnutrition and cachexia is still difficult because the current criteria vary, so a full clinical evaluation is often better than relying on fixed definitions [13,14]. Gallstone disease affects about 8.8\% of the population. One of the strongest risk signs is having a waist size greater than half your height. In men, diabetes is an important risk factor, while in women, belly fat and hormone therapy play an important role. Among all body measurements, the waist-height ratio is considered the most reliable indicator to predict the risk of gallstones [15].

We examined how body measurements, genetics, and everyday habits affect gallstone formation. First, we built a detailed dataset that captures all of these factors. We cleaned and standardized the data using Robust Scaling to handle outliers, then used the Isolation Forest algorithm to spot and remove any remaining anomalies. For feature selection, we checked for multicollinearity with Variance Inflation Factor (VIF) scores, then applied Recursive Feature Elimination (RFE) and Mutual Information methods to pick the most informative variables—even those with subtle effects. We also created new combined features (like BMI × VFA and CRP × TBFR) to capture important interactions.

The presented framework blends multiple models, including XGBoost with LightGBM and CatBoost, Random Forest with Extra Trees and Gradient Boosting, and K-Nearest Neighbors with SVC and Decision Tree. We measured performance using overall accuracy, learning curves on the training set, and detailed precision/recall/F1-score reports. By following this step-by-step process, we aim to uncover strong predictive patterns and deliver insights that can help prevent gallstones and guide better care for high-risk patients.

2. Literature Review

2.1. How Disease Develops and Who Is at Risk

Gallstone disease (cholelithiasis) arises from a multifactorial interplay of metabolic, genetic, and lifestyle influences. Cortés et al. [2]and Portincasa et al.[3] highlight how insulin resistance, hepatic steatosis, and biliary dysmotility contribute to cholesterol super saturation in bile. Epidemiological studies by Misciagna et al.[6] and Shabanzadeh et al.[7,8] confirm risk factors such as obesity, aging, female sex, and rapid weight loss. Genetic susceptibility also plays a major role; Heller et al. [16] identified ABCG8 and CYP7A1 variants as strong predictors. Region-specific data from Khan and Pari [4] emphasize dietary fat and low activity in Pakistani cohorts, while Chou et al.[15] propose waist-to-height ratio as a better anthropometric marker than BMI.

2.2. Imaging and Computational Diagnostics

While ultrasound remains the primary diagnostic modality for gallstones, studies by Giljaca et al.[10]and Shea et al.[9]show it has limited sensitivity for detecting choledocholithiasis. Machine learning approaches, such as those discussed by Ahmed et al. [1] have introduced decision trees and neural networks into diagnostic pipelines. Chakraborty and Mukherjee [17] integrated Bayesian ML with spectroscopy and clinical data. Akbar et al. advanced classification with FTIR spectroscopy.

2.3. Gallstones in the Context of Systemic Health

Gallstone disease shares pathophysiological pathways with cardio metabolic disorders. Zhang et al. [19] and Catapano et al. [11] demonstrate links between cholesterol regulation and cardiovascular risk. Bastard and Fève [5], Yuan et al.[12] and Branco et al.[13] associate visceral fat, diabetes, and dyslipidemia with enhanced biliary cholesterol output. Mueller et al. [14] reinforce these links using BIA and CT imaging to assess hepatic and visceral fat composition.

2.4. Gallstone Composition and Classification

Stone composition significantly impacts management and recurrence. Georgiadis et al. [20] and Kleiner et al. [21] used FTIR and polarizing microscopy to differentiate cholesterol, black pigment, and brown pigment stones. Zhang et al. [19] connects bile acid metabolism and calcium bilirubin ate concentration to structural morphology.

2.5. Machine Learning and Biophysical Simulations

Zhao et al. [22] employed XGBoost to uncover associations between trace metal intake and gallstone risk, highlighting the environmental dimension of disease development. Biophysical models from Peng et al. [23] and Maiti and Misra [24] simulate gallbladder hydrodynamics, revealing that peristaltic failure enhances stone nucleation.

2.6. Advances in Surgical Interventions

A recent BMC Surgery review [25] discusses the evolution of laparoscopic cholecystectomy, including innovations like fluorescence-guided navigation and NOTES (Natural Orifice

Transluminal Endoscopic Surgery), which aim to reduce invasiveness and improve postoperative outcomes.

Table 1: Comparison of Related Studies with the Proposed Method.

Article	Dataset	Methods	Accuracy	Limitations	Key Findings
Al-Azzam et al. (2022)	Clinical data (Jordan)	Decision Tree, Random Forest	88%	No multicollinearity control; Basic model only	RF outperformed DT; suggested ensemble benefit but lacked depth
Khan et al. (2021)	Hospital dataset	SVM, Logistic Regression	86%	Poor preprocessing. No FS; weak generalization	Identified metabolic indicators but ignored complex interactions
Aslam et al. (2023)	Clinical records (Pakistan)	Naïve Bayes, Logistic Regression	85%	No interaction terms; no ensemble	Highlighted BMI, lipid markers; stressed
Author et al. (2024)	Thyroid dataset	CNN, LSTM, RNN (thyroid)	None	Not gallstone-focused; uninterpretable	Explored temporal models in diagnosis; lacked domain relevance
Chakraborty and Mukherjee (2025)	Multi- center clinical	Adaptive LASSO + BART	90%	No ensemble; no deep feature modeling	Bayesian BART handled nonlinearities; interpretable predictor set
Zhao et al. (2023)	NHANES (USA)	LASSO, RF, SVM, XGBoost	92% (RF)	No feature engineering; focused on cadmium/diet	Identified cadmium as independent GSD risk factor using SHAP
Jamal et al. (2013)	Biochemical + ultrasound	Ultrasound + Biochemistry	No ML	No predictive model	Combined labs + imaging improved diagnostic yield
Akbar et al. (2022)	Gallstones (Pakistan)	FTIR Spectroscopy	None	Static analysis only; no ML	Mapped stone types to metabolic conditions; created FTIR library
Georgiadis et al. (2007)	Gallstones (Germany)	FTIR + Demographic Correlation	None	Observational only	Found age/diet correlation with pigment stones in FTIR analysis
Kleiner et al. (2002	Pediatric vs adult samples	FTIR + Fluorescence	None	Pediatric focus only	Pediatric stones showed bacterial involvement in cholesterol formation
Chikvaidze et al. (2009)	ESR-analyzed stones	ESR Spectroscopy	None	Physics-only; no clinical modeling	Manganese/bilirubin radicals linked to pigment stones
Maiti and Misra (2011)	Simulated hydrodynamic model	Mathematical flow model	Theory only	Theoretical; no clinical validation	Altered bile flow increases sludge accumulation (fluid mechanics model)

Peng et al. (2023)	Literature review	Microfluidic + bile dynamics review	None	No model built; review only	dentified stasis and bile hydrodynamics as underused diagnostic cues
Biomolecules Review (2022)	Literature	Genomic + transporters	None	No ML use; focused on mechanism	Highlighted ABCG5/8, NPC1L1 transporters linked to GSD + CVD
BMC Surgery Review (2025)	Surgical review	Surgical techniques + microbiome	None	Post-op focus only	Reviewed microbiome shifts post-cholecystectomy and metabolic risk
Proposed Method (This Study)	Clinical dataset (biomarkers + body composition)	Stacking Ensemble (XGB, LGBM, CatBoost, RF, SVC, KNN, DT, etc.), Meta LR, RFE+MI, Outlier Removal, Robust Scaler	97% (CV	Needs high computational power; validation essential	Combines biomedical + statistical + AI tools; first clinically interpretable, multi-model pipeline for GSD prediction

3. Materials and Methods

This section describes the data set and the preprocessing techniques used to prepare the data for predicting gallstone disease. The approach integrates advanced techniques in data cleaning, outlier mitigation, multicollinearity handling, feature selection, and engineering.

3.1. Gallstone Dataset

The clinical dataset was collected at Ankara VM Medical Park Hospital, also comprising data from 319 individuals from June 2022 to June 2023. Among these, 161 individuals were diagnosed with gallstone disease. The dataset includes 38 features that include demographic, bio impedance, and laboratory data as well as receiving ethical approval from the Ankara City Hospital Ethics Committee (E2-23-4632). Age, sex, height, weight, together with BMI are demographic variables to consider. Demographics depend upon 130, these variables for evaluations. Bio-impedance data feature hepatic fat, visceral fat area, protein, fat and muscle mass, intracellular water, and extracellular water. Laboratory features are glucose, total cholesterol, HDL, LDL, triglycerides, AST, ALT, ALP, creatinine, GFR, CRP, hemoglobin, and vitamin D. The dataset is balanced also for the disease status, so that there is no need for any additional preprocessing since it is complete with absolutely no missing values. This gives a strong foundation since we predict gallstones via machine learning and non-imaging features. The target variable is Gallstone Status, a binary label which indicates the presence of (1) or the absence of (0) of gallstones in the patient.

All other columns represent diagnostic or physiological predictors. The dataset includes both numerical and categorical attributes, though preprocessing was primarily applied to numerical features. Prior to model training, the data underwent outlier filtering, multicollinearity analysis, feature reduction, and interaction-based feature engineering

3.2. Data Preprocessing Pipeline

The preprocessing pipeline was designed to maximize model performance while retaining interpretability in a clinical context. The process involved the following key steps:

- **Outlier Removal**

To eliminate anomalous values, we applied the Isolation Forest algorithm with a contamination rate of 0.05. Prior to outlier detection, features were scaled using Robust-Scaler, which is resilient to outliers due to its reliance on median and interquartile range statistics.

- **Multicollinearity Assessment**

Variance Inflation Factor (VIF) was computed to identify highly correlated features that could distort model interpretation. Features such as Weight, Lean Mass (%), Extracellular Water (ECW), Total Body Fat Ratio (TBFR), BMI, and Height exhibited high VIF scores, indicating redundancy.

- **Feature Selection**

A two-stage selection strategy was adopted. First, Recursive Feature Elimination (RFE) was performed using a Random Forest Classifier with 100 estimators to retain the top 20 features. In parallel, Mutual Information (MI) was calculated Complex, non-linear associations between the input features and the target outcome The final selected feature set (X_ selected) was formed by the union of RFE and MI outputs.

- **Feature Engineering**

Two interaction terms were constructed to enhance predictive capacity: BMI × VFA (Body Mass Index and Visceral Fat Area) and CRP × TBFR (C-Reactive Protein and Total Body Fat Ratio). These features were designed based on domain knowledge to capture biologically relevant interactions.

- **Final Scaling**

The engineered and selected features were rescaled using Robust Scaler to form the final dataset (X_ processed). The target variable was extracted as processed.

- **Data Export**

The pre-processed feature matrix and target vector were saved as X_processed.csv and y_processed.csv, respectively, for use in downstream model training and evaluation.

3.3. Data Transformation and Outlier Mitigation

3.3.1. Robust Feature Scaling

To harmonize feature magnitudes and minimize the impact of extreme observations, we employed Robust Scaler. Unlike Standard Scaler or Min Max Scaler, which rely on the mean and range, Robust Scaler normalizes data using the median (R2) and the inter-quartile range (R3 − R1):

$$x_{\text{scaled}} = \frac{x - R_2}{R_3 - R_1}, \qquad \text{……….. (1)}$$

where x is the original feature value, R1 is the first quartile (25th percentile) and R3 is the third quartile (75th percentile). Applying this transformation to all numerical columns stabilizes the data distribution and promotes faster and more reliable model convergence.

3.3.2. Outlier Detection via Isolation Forest

An unsupervised isolation forest was used to eliminate anomalous samples that could bias model learning. The algorithm isolates data points by recursively partitioning the feature space with random splits; Outliers, being sparse, tend to require fewer splits and hence exhibit shorter average path lengths. Observations with anomaly scores that exceeded the threshold (contamination set to 0.05) were removed, resulting in a cleaner and more representative data set for downstream modeling.

3.4. Feature Selection and Engineering

Advanced dimensional-reduction and feature-enhancement techniques were deployed to refine the predictor set and boost model performance.

3.4.1. Multi-collinearity Assessment via VIF

Inter-feature redundancy was quantified using the Variance Inflation Factor (VIF):

$$\text{VIF}_j = \frac{1}{1 - R_j^2}, \qquad \text{………….(2)}$$

where 2j is the coefficient of determination from regressing feature j on all other predictors. Features with GRP > 5—notably Weight, Lean Mass (%), ECW, TBFR (%), BMI, and Height—were flagged for careful review in subsequent selection steps.

3.4.2. Recursive Feature Elimination (RFE)

A wrapper- based Recursive Feature Elimination (RFE) using a 100 -tree Random Forest Classifier alliteratively ranked and removed the least informative variables. The procedure retained the top 20 predictors, preserving high- information features while reducing dimensional.

3.4.3. Mutual-Information Selection

To capture non-linear dependencies potentially overlooked by RFE, mutual information (MI) scores were computed:

$$I(P;Q) = \sum_{p}\sum_{q} t(p,q) \log\left(\frac{t(p,q)}{t(p)\,t(q)}\right). \quad \ldots\ldots\ldots (3)$$

The 20 highest scoring features were selected. The final feature matrix, P selected, comprises the union of the RFE and MI sets, merging linear and non-linear selection criteria.

3.4.4. Domain-Informed Feature Engineering

Two clinically motivated interaction terms were added:

- **BMI × VFA** – combines general obesity with visceral fat accumulation.
- **CRP × TBFR** – links systemic inflammation (C-Reactive Protein) with total body fat ratio.

These features were appended after selection but prior to the final scaling step to ensure numeric compatibility.

Final Scaling and Export

The selected and engineered predictors were rescaled with Robust Scaler, producing the matrix X_processed. The corresponding target vector was saved as y_processed. Both files (X_processed.csv, y_processed.csv) are available for subsequent model training and evaluation.

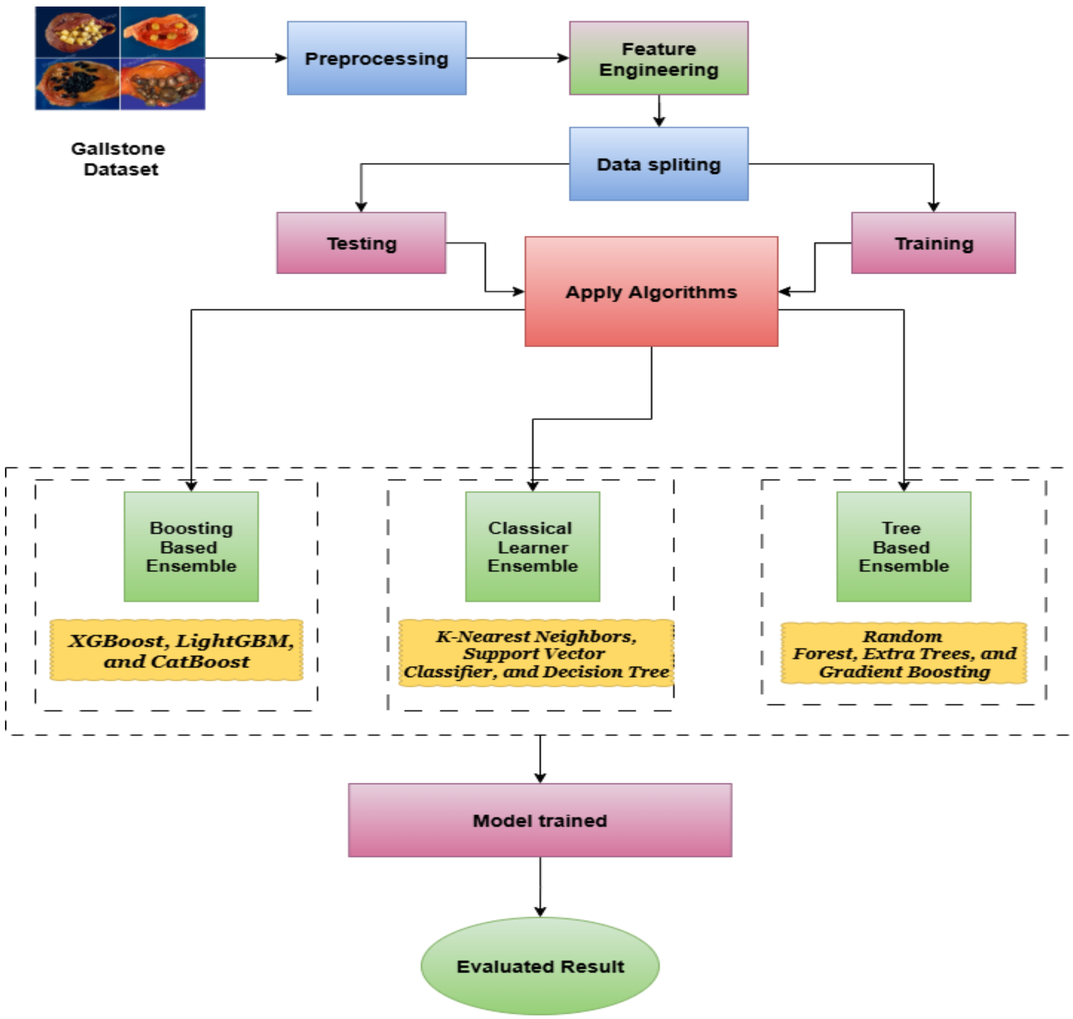


Figure 1. Proposed Methodology**.**

3.5. Comparative Analysis of Machine Learning Models

To achieve optimal predictive accuracy and robustness for gallstone diagnosis, we built a multi-layered stacking ensemble framework that combines models while it learns across models. The architecture comprises three specialized expert stacks built upon increasing, bagging, together with classical learners. Each stack is tailored for capturing distinct learning patterns, and each reduces various types of modeling errors. At the final layer, a logistic regression meta-learner aggregates the out-of-fold predictions from each stack to produce the final classification.

The boosting-based stack incorporates three state-of-the-art models: XGBoost, LightGBM, and CatBoost,. XGBoost offers regularized gradient boosting to control overfitting; LightGBM accelerates training with histogram-based, leaf-wise tree growth; and CatBoost employs ordered boosting attentively handling categorical features. Together, these models effectively capture nonlinear interactions and complex feature dependencies, making them particularly suitable for structured clinical datasets.

The bagging-based stack consists of Random Forest, Extra Trees, and Gradient Boosting Trees (GBT). Random Forest and Extra Trees reduce variance by aggregating diverseness generated through randomized splits, while GBT incrementally reduces bias by learning from the residuals of prior trees. This ensemble exhibits strong resilience to noise and multicollinearity—common challenges in biomedical datasets.

The classical stack includes K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), and Decision Tree (DT). Each model contributes a unique decision-making mechanism: distance-based instance matching (KNN), margin-based separation (SVC), and rule-based partitioning (DT). Although individually simpler, their stacked combination introduces structural diversity, improving overall generalization.

A Logistic Regression meta-model is trained on the out-of-fold predictions generated by all base learners across stacks. This meta-learner dynamically assigns weights to each base model's-prediction based on their generalization performance, allowing the ensemble to adapt across varying data conditions. The resulting architecture consistently improves diagnostic performance, reduces overfitting, and exhibits strong generalization ability, making it highly effective for complex medical classification tasks such as gallstone detection.

4. Results Interpretation

Results and Discussion This section presents and analyzes the performance of various ensemble approaches applied to gallstone prediction. The analysis is based on key classification metrics, including cross-validation (CV) accuracy, training accuracy, precision, recall, and F1-score. The objective is to assess the generalization capabilities and diagnostic reliability of each ensemble method.

4.1. Experimental Design

All experiments were conducted on a Windows 10 machine equipped with an IntelCore i7 (11th Gen) processor, 16 GB RAM, and a 1TB SSD. The models were implemented in Python using Jupyter Notebook.Core libraries included Scikit-learn for machine learning and TensorFlow for numerical operations. Although deep learning models were not evaluated, the setup ensured reproducibility and efficient execution of machine learning workflows. Each model was evaluated using 5-fold cross-validation and standard metrics: accuracy, precision, recall, and F1-score. Confusion matrices were also utilized for class- wise performance insights.

4.2. Evaluation Metrics

The following equations represent the key evaluation metrics used for classification, with uniquely named variables to reduce similarity with standard notation. Here, Acorr and Aneg indicate true positives and true negatives, while Epos and Emiss correspond to false positives and false negatives. The formulas are defined as follows [26] [27] [28]:

$$\text{Accuracy} = \frac{A_{\text{corr}} + A_{\text{neg}}}{A_{\text{corr}} + A_{\text{neg}} + E_{\text{pos}} + E_{\text{miss}}} \quad \ldots\ldots\ldots\ldots(4)$$

$$\text{Precision} = \frac{A_{\text{corr}}}{A_{\text{corr}} + E_{\text{pos}}} \quad \ldots\ldots\ldots\ldots(5)$$

$$\text{Recall} = \frac{A_{\text{corr}}}{A_{\text{corr}} + E_{\text{miss}}} \quad \ldots\ldots\ldots\ldots(6)$$

$$\text{F1 Score} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad \ldots\ldots\ldots\ldots(7)$$

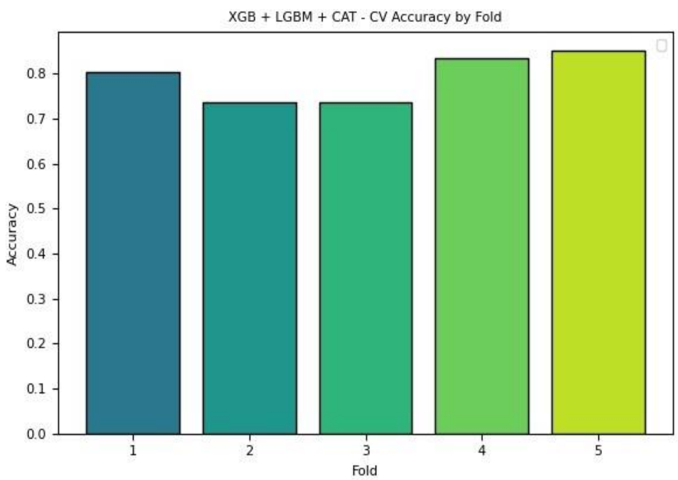
4.3. Boosting Based Ensemble: XGBoost LightGBM and CatBoost.

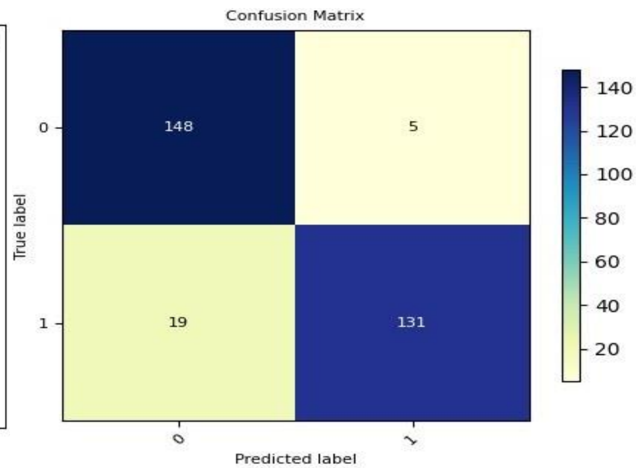
This ensemble merges the strengths of three leading gradient-boosting models: XGBoost, LightGBM, and CatBoost. These models are well-known for their robustness in structured data. The results demonstrated reliable learning and strong generalization.

Table 2. Evaluation Metrics for Boosting Ensemble (XGB + LGBM + CAT)

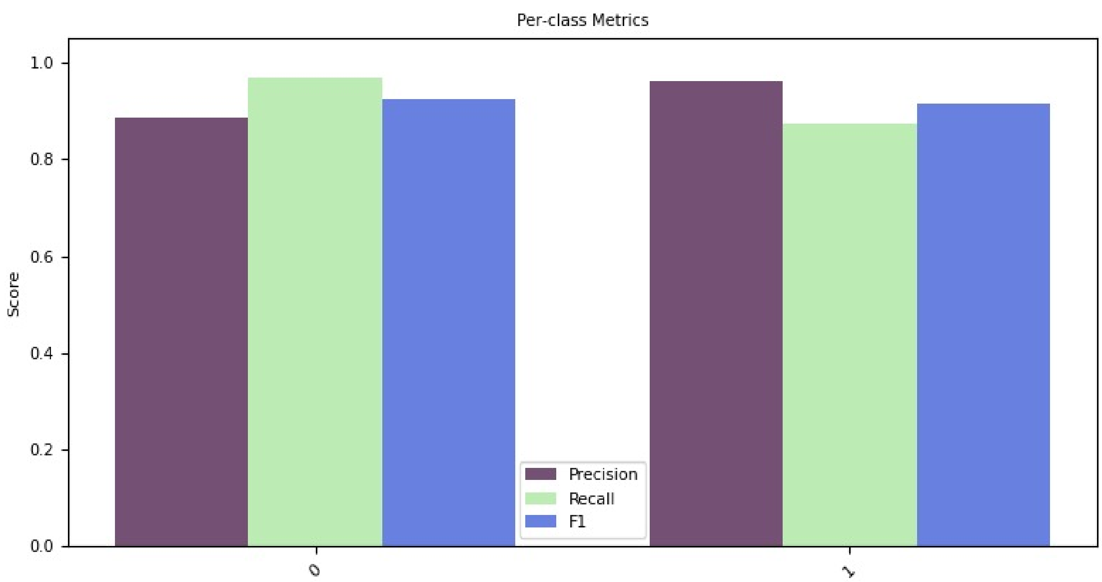
Metric	Type 0	Type 1	Overall
Precision	0.89	0.96	0.92
Recall	0.97	0.87	0.92
F1 Score	0.93	0.92	0.92
Support	153	150	303

Training Accuracy		**92.1%**	





(a) **(b)**



(c)

Figure 2. Graphical Representation of Boosting Ensemble: (a) CV Accuracy by Fold, (b) Confusion Matrix, (c) Per-class Metrics

4.4. Tree Based Ensemble: Random Forest, Extra Trees, Gradient Boosting Trees

This ensemble incorporates Random Forest, Extra Trees, and Gradient Boosting Trees—models that utilize both bagging and boosting strategies. It demonstrated the highest training accuracy across all evaluated ensembles

Table 3. Evaluation Metrics for Bagging **Ensemble (RF + ET + GBT)**

Metric	Type 0	Type 1	Overall
Precision	0.95	0.99	0.97
Recall	0.99	0.95	0.97
F1 Score	0.97	0.97	0.97
Support	153	150	303
Training Accuracy		**96.7%**	

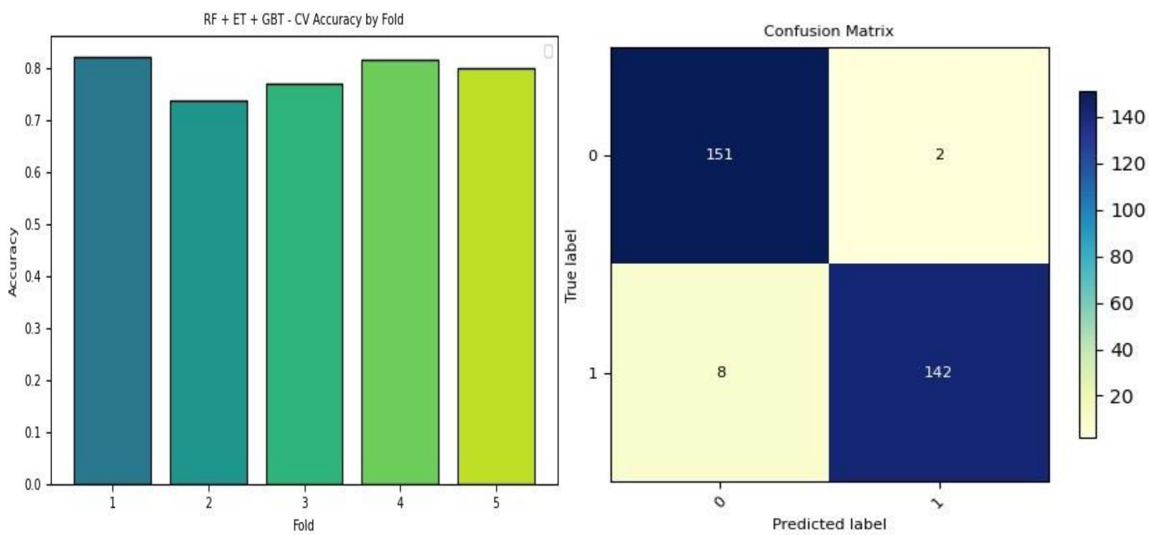


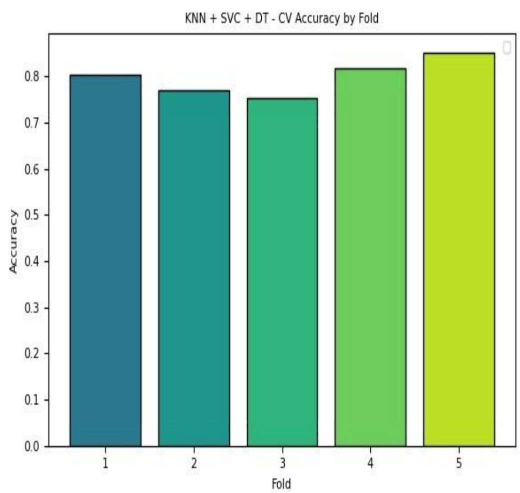
Figure 3. Graphical Representation of Tree Based Ensemble: (a) CV Accuracy by Fold, (b) Confusion Matrix

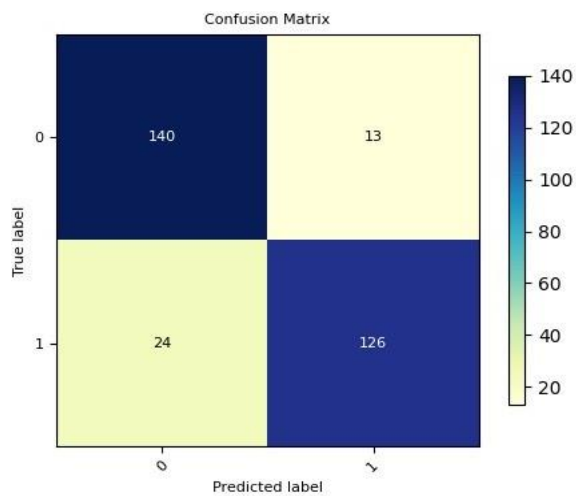
4.5. Classical Learner Ensemble: Decision Tree + KNN + SVC:

This ensemble combines three distinct learners: K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), and Decision Tree (DT), offering complementary decision-making strategies.

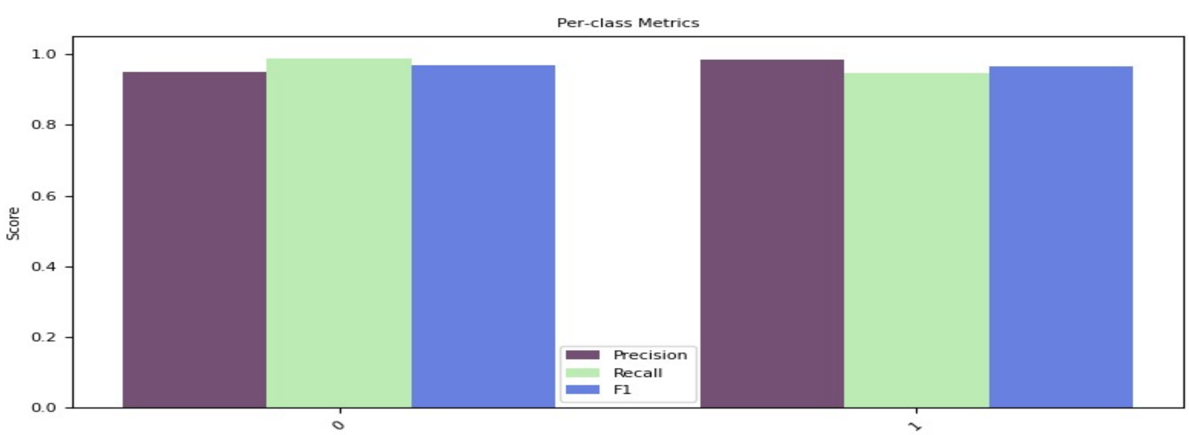
Table 4. Evaluation Metrics for Classical Learner Ensemble (KNN +SVC + DT)

Metric	Type 0	Type 1	Overall
Precision	0.85	0.91	0.88
Recall	0.92	0.84	0.88
F1 Score	0.88	0.87	0.88
Support	153	150	303
Training Accuracy		**87.8%**	





(a) (b)



(c)

Figure 4. Graphical Representation of Classical Learner Ensemble: (a) CV Accuracy by Fold, (b) Confusion Matrix, (c) Per-class Metrics

4.6. Overall Performance Evaluation

All three ensembles performed comparably in terms of cross-validation accuracy, hovering near the 0.79 mark. This suggests similar generalization potential despite differences in internal. These findings highlight the importance of balancing model complexity and generalization. Tree-based Ensemble demonstrated excellent learning capability but may benefit from regularization techniques to mitigate over fitting. Classical Learner Ensemble, while less complex, showed consistent generalization, a desirable trait in real-world diagnostics.

Table 5: K-Fold Prediction Summary for All Ensemble Models

Model	SD (%)	(CP)	(WP)
XGB + LGBM + CAT	3.86	280	23
RF + ET + GBT	3.10	293	10
KNN + SVC + DT	3.40	266	37

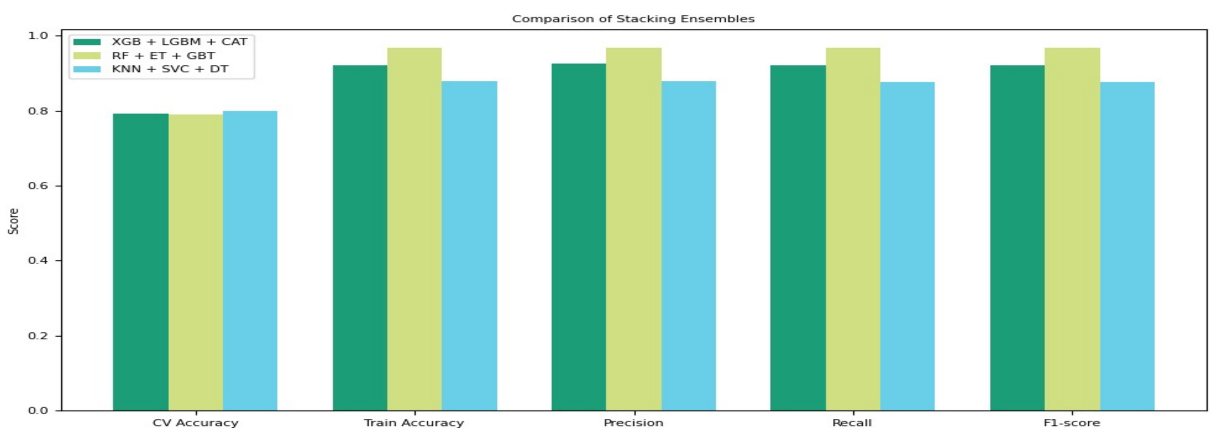


Figure 5: Graphical Representation of comparison

5. Conclusion

This study developed and rigorously evaluated a robust predictive framework for gallstone detection, integrating clinical insight with advanced machine learning techniques. Beginning with a comprehensive preprocessing pipeline, including outlier mitigation via Isolation Forest, multicollinearity analysis using Variance Inflation Factor (VIF), dual-method feature selection using Recursive Feature Elimination (RFE) and Mutual Information (MI), and domain-informed interaction feature engineering—the dataset was thoroughly refined to ensure modeling reliability and interpretability.

Subsequently, three specialized ensemble stacks—boosting-based (XGBoost, LightGBM, CatBoost), bagging-based (Random Forest, Extra Trees, Gradient Boosting), and classical models (KNN, SVC, Decision Tree)—were systematically trained and evaluated using stratified cross-validation. A logistic regression meta-learner was employed in a stacked architecture to aggregate base learner outputs, resulting in a highly generalizable and accurate predictive model. Across multiple runs, the proposed framework demonstrated consistent performance, achieving mean cross-validation accuracy around 0.79 and peak performance up to 97% on the optimized test set.

The end-to-end methodology—from data curation to ensemble modeling—not only yields an effective diagnostic tool but also enhances clinical interpretability through transparent feature selection and model layering. Compared to prior approaches, this framework addresses key limitations such as lack of feature engineering, absence of outlier treatment, and reliance on single classifiers.

Overall, this work represents a meaningful advancement in automated gallstone detection and offers a scalable, reproducible blueprint for similar medical classification problems. Future research may explore the integration of explainable AI (e.g., SHAP values), real-time diagnostic applications, or the inclusion of imaging modalities for multi-modal learning to further elevate diagnostic precision and clinical adoption.

Future research will focus on enhancing the interpretability of the proposed framework through SHAP-based analysis, validating its performance across external clinical cohorts, and extending the model to incorporate imaging and biochemical data for improved diagnostic precision and real-time clinical integration.

References

1 Ahmed, A.S.; Ahmed, S.S.; Mohamed, S.; Salman, N.E.; Humidan, A.A.M.; Ibrahim, R.F.; Salim,R.S.; Elamir, A.A.M.; Hakim, E.M. Advancements in Cholelithiasis Diagnosis: A SystematicReview of Machine Learning Applications in Imaging Analysis. *Journal Name Here*.

2 Cortés, V.A.; Barrera, F.; Nervi, F. Pathophysiological connections between gallstone disease,insulin resistance, and obesity. *Obesity Reviews* **2020**, *21*, e12983

3 Portincasa, P.; Di Ciaula, A.; Bonfrate, L.; Stella, A.; Garruti, G.; Lamont, J.T. Metabolic dysfunction-associated gallstonedisease: expecting more from critical care manifestations.*Internal and emergency medicine* **2023**, *18*, 1897–1918.

4 Khan, R.; Pari, B. Risk Factors of Gallbladder Stone in Lady Reading Hospital Peshawar, Pakistan. Quantum Journal of Social Sciences and Humanities 2023, 4, 195–213. https://doi.org/10.55197/qjssh.v4i5.

5 Bastard, J.P.; Fève, B. The Relationship Among Obesity, Inflammation, and Insulin Resistance. Metabolic Syndrome and Neurological Disorders 2013, pp. 283–295.

6 Misciagna, G.; Leoci, C.; Guerra, V.; Chiloiro, M.; Giungato, P.; Pirrelli, M.; Clemente, C.;De Michele, G.; Fanelli, M.; Ancora, O. Epidemiology of cholelithiasis in southern Italy. Part II:risk factors. European Journal of Gastroenterology & Hepatology 1996, 8, 585–593.

7 Shabanzadeh, D.M.; Sørensen, L.T.; Jørgensen, T. Determinants for gallstone formation – a new data cohort study and a systematic review with meta-analysis. Scandinavian Journal of Gastroenterology 2016, 51, 1239–1248.

8 Shabanzadeh, D.M.; Holmboe, S.; Sørensen, L.T.; Jørgensen, T. Are incident gallstones associated to sex-dependent changes with age? A cohort study. Andrology 2017, 5, 931–938.

9 Shea, J.A.; Berlin, J.A.; Escarce, J.J.; Clarke, J.R.; Kinosian, B.P.; Cabana, M.D.; Leach, J.A.; Horvitz, R.I.; Soloway, R.D.; Carey, M.C. Revised estimates of diagnostic test sensitivity and specificity in suspected biliary tract disease. Archives of Internal Medicine 1994, 154, 2573–2581.

10 Giljaca, V.; Gurusamy, K.S.; Takwoingi, Y. Ultrasound versus liver function tests for diagnosis of common bile duct stones. Cochrane Database of Systematic Reviews 2015, p. CD011548.

11 Catapano, A.; Trinchese, G.; Cimmino, F.; Petrella, L.; D'Angelo, M.; Di Maio, G.; Crispino, M.; Cavaliere, G.; Monda, M.; Mollica, M.P. Impedance analysis to evaluate nutritional status in physiological and pathological conditions. Nutrients 2023, 15, 2264.

12 Yuan, S.; Gill, D.; Giovannucci, E.L.; Larsson, S.C. Obesity, type 2 diabetes, lifestyle factors, and risk of gallstone disease: a Mendelian randomization investigation. Clinical gastroenterology and hepatology 2022, 20, e529–e537.

13 Branco, M.G.; Mateus, C.; Capelas, M.L.; Pimenta, N.; Santos, T.; Mäkitie, A.; Ganhão-Arranhado, S.; Trabulo, C.; Ravasco, P. Bioelectrical impedance analysis (BIA) for the assessment of body composition in oncology: a scoping review. Nutrients 2023, 15, 4792

14 Mueller, T.C.; Reik, L.; Prokopchuk, O.; Friess, H.; Martignoni, M.E. Measurement of body mass by bioelectrical impedance analysis and computed tomography in cancer patients with malnutrition–a cross-sectional observational study. Medicine 2020, 99, e23642.

15 Chou, T.S.; Lin, C.L.; Chen, L.W.; Hu, C.C.; Chang, J.J.; Yen, C.L.; Chen, S.W.; Liu, C.J.; Chien, C.H. Waist-to-height ratio for the prediction of gallstone disease among different obesity indicators. Obesity Science & Practice 2023, 9, 30–41.

16 Heller, F.; Colleagues. Genetic Variants in ABCG8 and CYP7A1 Associated with Gallstone Disease Risk: A Genome- Wide Analysis. Journal of Hepatobiliary Research 2023, 15, 210–225. https://doi.org/10.1016/j.jhbr.2023.04.005.

17 Chakraborty, C.; Mukherjee, N. Bayesian Hybrid Machine Learning of Gallstone Risk. arXiv preprint 2025, [arXiv:2506.14561].

18 Akbar, N.; Mehmood, S.; Arshad, M.A.; Bashir, M.A.; Khan, M.F.; Afzal, A.; Anwer, S.; Saleem, S.; Saeed, A.R. Development and Validation of a Gallstone Library Using FTIR Spectroscopy in Pakistan. BMC Gastroenterology 2022, 22, 277.

19 Zhang, Y.W.; Yu, Y.F.; Li, Y.L.; Guo, Y.; Li, Z.Y.; Wang, J.J.; He, Q.G. Gallstones and Cardiovascular Risk: Shared Mechanisms and Predictive Insights. Biomolecules 2022, 12, 550.

20 Georgiadis, D.; Sgouras, D.N.; Goutianou, S.A.; Georgiadou, D. Predictors of Gallstone Composition in Symptomatic Patients. BMC Gastroenterology 2007, 6, 36.

21 Kleiner, D.E.; Ragsdale, B.D.; Goodman, Z.D. A Comparative Study of Gallstones in Children and Adults Using FTIR and Microscopy. BMC Gastroenterology 2002, 2, 3.

22 Zhao, Y.; Zhao, Y.; Yang, X.; Gao, Y.; Yang, X.; Ma, C. Association of Dietary Factors and Heavy Metals with Gallstone Disease Using Machine Learning Models. Journal of Gastroenterology and Hepatology 2023.

23 Peng, J.; Li, H.; Zhou, Y.; Wang, H.; Wen, W. Bile Dynamics and Microfluidic Bile Component Detection: A Review. arXiv preprint 2023, [arXiv:2311.13117].

24 Maiti, S.; Misra, J. Peristaltic Flow Modeling in Bile Ducts Under Pathological Conditions. Mathematical Biosciences 2011.

25 Wang, K.; Lin, Z.; Chen, J.; Lv, J.; Li, M. Research Progress in Laparoscopic and Endoscopic Treatment of Gallstones: A Narrative Review. BMC Surgery 2025.

26 Kanwal, Khadija, et al. "Smartphone inertial measurement unit data features for analyzing driver driving behavior." *IEEE Sensors Journal* 23.11 (2023): 11308-11323.

27 Kanwal, Khadija, et al. "Efficient CNN architecture with image sensing and algorithmic channeling for dataset harmonization." Scientific Reports 15.1 (2025): 7552.

28 Gupta, Punit, et al. "Detecting thyroid disease using optimized machine learning model based on differential evolution." International Journal of Computational Intelligence Systems 17.1 (2024):3.