



Gallstone Disease Prediction With Explainable Artificial Intelligence

Pureti Anusha¹, M. Radhika²

¹M.Tech Student, Department Of Computer Science And Engineering, Mam Women'S Engineering College, Kesanupalli, Narasaraopeta, Palnadu District, Andhra Pradesh – 522601, India.

²Associate Professor, Department Of Computer Science And Engineering, Mam Women'S Engineering College, Kesanupalli, Narasaraopeta, Palnadu District, Andhra Pradesh – 522601, India.

Abstract- Gallstone disease (GSD) is one of the most frequently encountered digestive disorders in adults, and it often stays silent until it produces pain, infection or blockage of the bile ducts. Because its early symptoms overlap with many other abdominal conditions, a timely and confident diagnosis is not always possible, and imaging is not available at every point of care. This work presents a machine learning framework that estimates gallstone risk from routinely recorded demographic, bioelectrical impedance and blood test parameters, and that explains every prediction it makes. A public dataset of 319 patients from the UCI repository is cleaned, scaled and encoded, after which four feature configurations are prepared: all attributes, attributes chosen through a correlation heatmap, attributes ranked by the ANOVA F-test, and the overlap of the last two. Seven classifiers, namely Random Forest, Support Vector Machine, Gradient Boosting, Logistic Regression, AdaBoost, Decision Tree and a Bagging ensemble, are trained under stratified five-fold cross-validation, giving 28 model-configuration pairs in total. The Support Vector Machine trained on correlation-selected features gives the strongest result, with 83.33% accuracy, 85.42% sensitivity and an area under the ROC curve of 88.80%. SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) are then attached to this model. Both methods point to Vitamin D, C-reactive protein, bone mass, aspartate aminotransferase and hepatic fat grade as the dominant drivers of risk. The outcome is a transparent, low-cost screening aid that can help physicians flag likely gallstone cases early and understand why the system reached each decision.

Keywords: Gallstone Disease, Cholelithiasis, Machine Learning, Explainable Artificial Intelligence (XAI), SHAP, LIME, Support Vector Machine, Feature Selection, ANOVA F-test, Bioimpedance, Clinical Decision Support, Healthcare Analytics.

I. INTRODUCTION

Gallstones are hardened deposits that develop inside the gallbladder when the chemical balance of bile is disturbed, typically through an excess of cholesterol, a shortage of bile salts, or sluggish emptying of the gallbladder. The condition affects roughly one in ten adults worldwide, and its share is rising with changing diets, obesity and longer life expectancy [1]. Many carriers remain without symptoms for years, but a considerable number eventually develop biliary colic, acute cholecystitis, obstructive jaundice or gallstone pancreatitis, each of which may require emergency care and surgical removal of the gallbladder. Identifying people at high risk before such complications arise is therefore valuable both for the patient and for the health system [3], [7].

In everyday practice, diagnosis is built on the patient's history, physical examination, liver function tests and abdominal ultrasonography. Ultrasound is accurate, yet it depends on equipment and a trained operator, and it is usually requested only after symptoms appear. Laboratory results and body composition readings, on the other hand, are inexpensive, are frequently collected during routine check-ups, and carry subtle signals about metabolic and inflammatory status. Combining these signals manually is difficult because their relationship with gallstone formation is non-linear and spread across many variables. Machine learning (ML) is well suited to this kind of multi-variable

pattern recognition and has already shown strong results in diabetes, cardiac and cancer screening [5].

A key obstacle to adopting ML in hospitals is the black-box nature of many models. A clinician is unlikely to act on a risk score unless the reasoning behind it can be inspected and matched with medical knowledge. Explainable Artificial Intelligence (XAI) addresses this gap by revealing how much each input contributed to a decision. Despite this, only a small number of gallstone studies built on tabular clinical data have combined systematic feature selection with formal explanation methods [3], [6].

The present work, which builds on the approach of Rodriguez et al. [15], develops an interpretable gallstone risk predictor using a moderately sized public dataset [12]. The main contributions are: (i) a reproducible preprocessing pipeline for mixed clinical, bioimpedance and laboratory variables; (ii) a comparison of four feature selection configurations across seven classifiers; (iii) identification of the best performing model using accuracy, sensitivity, specificity, precision, F1-score and AUC; and (iv) global and patient-level explanations produced through SHAP [13] and LIME [14]. The remainder of the paper reviews related work, describes the methodology, contrasts the existing and proposed systems, presents the architecture and results, and closes with a summary and future directions.



II. SURVEY OF LITERATURE

Research on computer-assisted gallbladder diagnosis has grown quickly over the last few years. Two broad streams can be seen: studies that analyse ultrasound images with deep learning, and studies that model structured clinical and laboratory records with classical ML. The works most relevant to this project are discussed below.

Ning et al. carried out a large meta-analysis of population studies published between 2000 and 2024 and estimated the global prevalence of gallbladder stones. Their findings confirm that the disease burden is substantial and growing in many regions, which strengthens the case for affordable screening tools that can reach people before complications develop [1].

Bozdog et al. proposed a content-based image retrieval system for ultrasound scans, using engineered features and a cosine similarity measure to separate nine categories of gallbladder disease. The system achieved high retrieval quality, and the authors suggested it could support radiologists in busy clinical settings. However, the method still requires an imaging session for every patient [2].

Esen et al. were among the first to predict gallstones directly from bioimpedance and laboratory data. Among the algorithms they tested, Gradient Boosting gave the best accuracy of 85.42%, and their importance analysis highlighted Vitamin D, C-reactive protein, total body water and lean mass. The same group later released the underlying records as a public dataset, which is used in this study [7], [12].

Sarker, Tiang and Nahid optimised a Random Forest with the Sand Cat Swarm Optimization algorithm and evaluated four feature reduction settings, reporting accuracies between 78.32% and 81.25%. They applied SHAP and counterfactual (DiCE) explanations and found C-reactive protein, Vitamin D and aspartate aminotransferase to be the most influential variables, demonstrating that interpretability and good accuracy can coexist on this type of data [3].

Kanwal et al. built heterogeneous stacked ensembles that combine boosting learners, tree-based learners and classical learners, with a Logistic Regression meta-classifier that merges out-of-fold predictions. Their design reported a very high accuracy of 97% on gallstone detection, although such stacked systems are computationally heavier and harder to explain to clinicians [6].

Salem et al. examined metabolomic profiles together with clinical attributes and found that XGBoost performed best when the top fifty metabolites were joined with five clinical conditions. The study shows that biochemical markers carry

predictive information, but metabolomic testing is not routine in most hospitals [8].

Several image-based studies have also been reported. Kumar et al. designed an explainable convolutional neural network for cholelithiasis that replaces the flattening layer with global average pooling and adds post-hoc visual explanations [4]. Hong et al. compared ML algorithms on ultrasound images gathered from hospitals and public repositories, with Random Forest reaching 96.33% accuracy [10]. Ge et al. combined a CNN, a gated recurrent unit and a multilayer perceptron on ultrasound cine clips to separate normal, non-urgent and urgent gallbladder findings, reaching 91% accuracy for normal versus abnormal cases [11]. A scoping review by Wang et al. concluded that AI is becoming an essential companion to sonography in gallbladder assessment [9], and Obaid et al. compared several AI techniques for gallbladder disease diagnosis and confirmed the promise of data-driven approaches [5].

Overall, the literature shows that imaging-based methods dominate, while prediction from routine clinical and laboratory measurements has received less attention. Few studies examine how feature selection influences different classifiers on this data, and even fewer provide both global and individual explanations. These gaps motivate the present work.

III. RESEARCH METHODOLOGY

The research follows a step-by-step pipeline that moves from raw patient records to an explained prediction. The first step is data acquisition. The study uses the Gallstone dataset published in the UCI Machine Learning Repository [12], which contains 319 patient records described by close to forty attributes. These attributes cover demographic details such as age, gender and comorbidities; bioelectrical impedance measurements such as total body water, lean mass, bone mass, visceral fat and hepatic fat grade; and laboratory values such as glucose, lipid profile, liver enzymes, C-reactive protein and Vitamin D. The target attribute, Gallstone Status, marks each person as positive (1) or negative (0). The two classes are almost equal in size (about 158 positive and 161 negative), so oversampling methods such as SMOTE are not needed, and the records contain no missing values.

The second step is preprocessing. Attributes are first grouped by type into numerical, categorical and laboratory variables so that each group can be treated appropriately. Continuous variables are standardised with a StandardScaler so that every feature has zero mean and unit variance, which is important for distance- and margin-based learners such as the Support Vector Machine. Categorical variables, including hepatic fat grade and comorbidity indicators, are converted with one-hot encoding. Extreme values found during exploratory analysis are kept in the data, since in a clinical context an unusual laboratory reading often reflects real pathology rather than a



recording error. The third step is feature selection, which produces four configurations. The Baseline configuration keeps all attributes and acts as a reference. The Heatmap configuration uses a Pearson correlation matrix to retain the features whose absolute correlation with Gallstone Status lies above the 75th percentile, while also dropping attributes that are strongly correlated with one another to reduce multicollinearity.

The ANOVA configuration applies SelectKBest with the F-test and keeps the top 25% of features by F-score. The Mixed configuration retains only those features chosen by both the heatmap and the ANOVA methods.

The fourth step is model training. For every configuration, the data is split into 70% training and 30% testing portions with stratification so that both classes keep their original proportions. Seven classifiers are trained: Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GB), Logistic Regression (LR), AdaBoost (AB), Decision Tree (DT) and Bagging Classifier (BC). Stratified five-fold cross-validation on the training portion is used to obtain stable estimates and to limit overfitting. This yields 28 model–configuration combinations.

The fifth step is evaluation. Each combination is scored with six measures: accuracy, recall (sensitivity), specificity, precision, F1-score and the area under the receiver operating characteristic curve (AUC). Recall receives particular attention because a missed gallstone case can lead to untreated complications, while precision indicates how well the model avoids unnecessary referrals. The final step is explanation. The best performing combination is passed to two XAI techniques. SHAP, which is grounded in cooperative game theory, distributes each prediction among the input features and produces a global ranking through mean absolute Shapley values together with a beeswarm plot showing the direction of each effect [13].

LIME fits a simple surrogate model around a single patient record and lists the features that pushed that particular prediction towards or away from gallstone disease [14]. Together, these views connect the model's behaviour with known clinical reasoning.

IV. THE CURRENT SYSTEM

At present, gallstone disease is mainly confirmed through a clinical workflow. A patient who reports upper abdominal pain, bloating, nausea or fever is examined by a physician, blood tests are ordered, and abdominal ultrasound is used to visualise the stones. CT, MRCP or endoscopic procedures may follow for complicated cases. This approach is reliable once symptoms have appeared, but it is reactive. People with silent stones are usually discovered only by chance during

scans done for other reasons, and the need for imaging equipment and skilled sonographers limits access in rural and resource-poor areas. On the computational side, most existing AI systems for gallbladder disease are designed around ultrasound or CT images. Convolutional networks and retrieval systems trained on such images can be accurate, but they still need an imaging study before any prediction is made, so they do not support early, low-cost screening. The smaller group of studies that use tabular clinical data often train a single algorithm on all available attributes without testing how feature selection affects results, and some rely on complex ensembles that are difficult to deploy.

Another major limitation of the current systems is the lack of transparency. Many reported models output only a class label or a probability. Clinicians cannot see which laboratory or body composition values led to the decision, cannot check whether the reasoning agrees with medical knowledge, and therefore find it hard to trust or justify the output. Where explanations are provided, they are frequently limited to a single global importance chart that says little about an individual patient. In summary, the existing approaches are reactive, depend heavily on imaging, give little insight into the role of feature selection, and rarely explain their decisions at both the population and patient level. These shortcomings create the need for the interpretable, data-driven framework proposed in this work.

V. THE SUGGESTED SYSTEM

The suggested system is an explainable gallstone risk prediction framework that works on routinely collected clinical, bioimpedance and laboratory data. Instead of waiting for imaging, it estimates the probability of gallstones from measurements that are already part of a general health check, making it suitable as an early screening and triage aid.

The system cleans and encodes the patient data, compares four feature selection strategies, trains seven well-established classifiers under identical conditions, and selects the most dependable model using a balanced set of clinical performance measures. Experimental comparison of all 28 combinations shows that the choice of features has a strong effect on some algorithms.

With all attributes, the Support Vector Machine was the weakest model, but once the correlation-based subset was used it became the strongest. Random Forest and Logistic Regression remained consistently good across configurations, whereas Decision Tree was the least accurate in every setting. Table I lists the results for the correlation heatmap configuration, which produced the best model overall.

Table I. Performance Of Classifiers On Correlation-Selected Features (%)

Model	ACC	Recall	SPE	Prec.	F1	AUC
RF	82.29	83.33	81.25	81.63	82.47	88.63
SVM	83.33	85.42	81.25	82.00	83.68	88.80
GB	77.08	83.33	70.83	74.07	78.43	87.41
LR	82.29	85.42	79.17	80.39	82.83	85.33
AB	77.08	83.33	70.83	74.07	78.43	86.55
DT	70.83	81.25	60.41	67.24	73.58	70.83
BC	79.17	77.08	81.25	80.43	78.72	88.52

The Support Vector Machine with correlation-selected features achieved 83.33% accuracy, 85.42% recall, 81.25% specificity, 82.00% precision, an F1-score of 83.68% and an AUC of 88.80%. The high recall is especially valuable because it means few true gallstone cases are missed, while the solid precision limits false alarms that could lead to unnecessary investigations. This performance is comparable with earlier work on the same type of data, which reported 85.42% [7] and 81.25% [3] accuracy. The chosen model is then explained. The SHAP global ranking places Vitamin D, C-reactive protein, bone mass and aspartate aminotransferase at the top. The beeswarm view shows that low Vitamin D, low bone mass and raised C-reactive protein push predictions towards the gallstone class, while the opposite values push them towards the healthy class.

At the patient level, LIME shows, for example, that a record predicted positive with about 81% confidence was driven mainly by reduced bone mass and moderate hepatic fat accumulation, whereas a record predicted negative was supported by low C-reactive protein, higher bone mass and absence of hyperlipidaemia. These explanations agree with known links between inflammation, metabolic health, fatty liver and gallstone formation, which increases clinical trust in the model.

VI. SYSTEM ARCHITECTURE

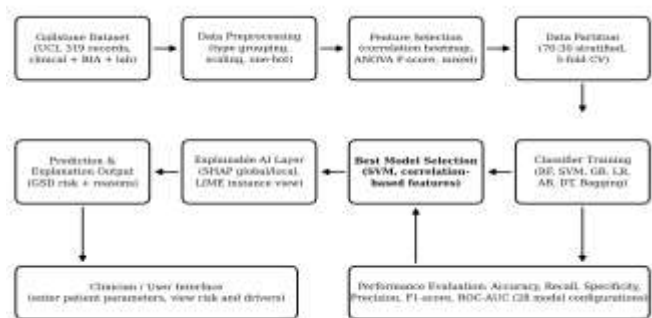


Fig. 1. Architecture of the explainable gallstone disease prediction system

The architecture of the proposed system is organised as a sequence of modules, as shown in Fig. 1. The data module loads the UCI Gallstone dataset, which holds demographic, bioimpedance and laboratory attributes for 319 individuals along with their confirmed gallstone status. The preprocessing module separates attributes by type, standardises continuous values, encodes categorical values with one-hot vectors, and verifies that no records are missing. Clinically meaningful outliers are retained so that the model learns from realistic patient variation.

The feature selection module generates four views of the data: the complete attribute set, a correlation-heatmap subset, an ANOVA F-score subset and their intersection. Each view is passed to the data partition module, which performs a stratified 70:30 split and five-fold cross-validation. The training module fits seven classifiers on each view, and the evaluation module records accuracy, recall, specificity, precision, F1-score and AUC for all 28 combinations. The model selection module compares these scores and keeps the Support Vector Machine trained on correlation-selected features as the final predictor.

The explainability module wraps the final predictor with SHAP and LIME. SHAP supplies a global ranking of features and a beeswarm view of how high and low values affect risk, while LIME gives a short list of the factors behind each individual prediction. The output module combines the predicted class, the risk probability and these explanations into a single report. Through a simple user interface, a clinician can enter a patient's parameters and immediately see both the estimated gallstone risk and the reasons behind it, allowing the result to be checked against clinical judgement before any further tests or treatment are ordered.

VII. SUMMARY

This work presented an explainable machine learning framework for predicting gallstone disease from routine clinical, bioimpedance and laboratory measurements. Using a public dataset of 319 patients, four feature selection configurations and seven classifiers were evaluated under stratified cross-validation. The Support Vector Machine trained on correlation-selected features delivered the best balance of results, with 83.33% accuracy, 85.42% sensitivity and an AUC of 88.80%. The study also showed that feature selection can change the ranking of algorithms considerably, which highlights the need to test several configurations rather than a single fixed pipeline.

By attaching SHAP and LIME to the selected model, the system explains its reasoning at both the population and individual level. Vitamin D, C-reactive protein, bone mass, aspartate aminotransferase and hepatic fat grade emerged as the most important indicators, which is consistent with existing medical understanding of inflammatory and



metabolic risk. These transparent outputs make the tool easier for physicians to trust and can support earlier referral for ultrasound and timely treatment, particularly where imaging is not readily available.

The main limitations are the small dataset, which comes from a single centre, and the absence of external validation. Future work will gather larger multi-centre data, test the model on independent cohorts, explore hyperparameter tuning and stacked ensembles, and add further inputs such as imaging findings, genetic markers and longitudinal records. Deploying the framework as a web-based decision-support service is another planned step towards practical clinical use.

REFERENCES

1. Q. Ning, F. Liu, Y. Fang, X. Zhu, J. Liu and Z. Li, "Estimating global prevalence of gallbladder stones in general population from 2000 to 2024: systematic review and meta-analysis," *Annals of Medicine*, vol. 57, no. 1, 2025, doi: 10.1080/07853890.2025.2570795.
2. A. Bozdog, 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.
3. 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.
4. 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.
5. A. M. Obaid, A. Turki, H. Bellaaj and M. Ksantini, "Diagnosis of gallbladder disease using artificial intelligence: a comparative study," *Int. J. Computational Intelligence Systems*, vol. 17, p. 46, 2024.
6. 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.
7. 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, e37258, 2024.
8. N. M. Salem, K. M. Jack, H. Gu et al., "Machine and deep learning identified metabolites and clinical features associated with gallstone disease," *Computer Methods and Programs in Biomedicine Update*, vol. 3, 2023.
9. 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*, 2024.
10. C. Hong, I. Zafar, M. M. Ayaz et al., "Analysis of machine learning algorithms for real-time gallbladder stone identification from ultrasound images in clinical decision support systems," *Int. J. Computational Intelligence Systems*, vol. 18, p. 73, 2025.
11. 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.
12. J. I. Esen, H. Arslan, S. Aktürk, M. Gülşen, N. Kültekin and O. Özdemir, "Gallstone [Dataset]," *UCI Machine Learning Repository*, 2024.
13. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
14. M. T. Ribeiro, S. Singh and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.
15. A. C. R. P. Rodriguez, J. A. R. Dela Cruz, H. W. R. Acosta II, J. A. G. Perillo, M. S. A. Magboo and V. P. C. Magboo, "Gallstone disease prediction with explainable artificial intelligence," in *Proc. Int. Conf. Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA)*, 2026, doi: 10.1109/ACDSA67686.2026.11468066.