# Gallstone Disease Prediction with Explainable Artificial Intelligence

Augustus Clark Raphael P. Rodriguez
*Department of Physical Sciences and Mathematics*
*University of the Philippines Manila*
Manila, Philippines
aprodriguez7@up.edu.ph

James Angelo R. Dela Cruz
*Department of Physical Sciences and Mathematics*
*University of the Philippines Manila*
Manila, Philippines
jrdelacruz@up.edu.ph

Harry William R. Acosta II
*Department of Physical Sciences and Mathematics*
*University of the Philippines Manila*
Manila, Philippines
hracosta@up.edu.ph

Jasper Anthony G. Perillo
*Department of Physical Sciences and Mathematics*
*University of the Philippines Manila*
Manila, Philippines
jgperillo@up.edu.ph

Ma. Sheila A. Magboo
*Department of Physical Sciences and Mathematics*
*University of the Philippines Manila*
Manila, Philippines
ORCID: 0000-0002-6221-7892

Vincent Peter C. Magboo
*Department of Physical Sciences and Mathematics*
*University of the Philippines Manila*
Manila, Philippines
ORCID: 0000-0001-8301-9775

***Abstract*—Gallstone Disease is a common gastrointestinal disease worldwide that can lead to further complications if left untreated. Hence, early and accurate diagnosis is essential in improving patient quality of life. In recent times, development has progressed in Gallstone diagnosis, allowing for relatively accurate diagnoses; however, high costs and inaccuracies in specific demographics continue to impede its reliability.This study proposes the use of a data-driven machine learning risk prediction model leveraging a dataset comprising a broad range of clinical and demographic variables. The study examined seven machine learning models: Random Forest, Support Vector Machines, Gradient Boosting, Logistic Regression, AdaBoost, Decision Trees, and Bagging Classifier. Additionally, the study explored four different configurations for the aforementioned models, differing in terms of means of feature selection. A baseline configuration, Correlation-based configuration, ANOVA F-score based configuration, and a mixed configuration, combining the two primary means of feature selection. The results of the cross-validated model evaluation uncovered that Logistic Regression resulted in the most consistent performance throughout the configurations. The study incorporates Explainable Artificial Intelligence , augmented upon the best performing model , to enhance transparency and improve understanding and trust in the model predictions. Similarly, the study validated the effectiveness of integrating machine learning models augmented with explainers as an automated and reliable method for handling Gallstone classification tasks and risk assessment, as well as for integrating such tools into the clinical decision-making process.**

***Index Terms*—Gallstone Disease, Explainable AI, Machine Learning, Prediction, Clinical Decision Support**

## I. Introduction

Gallstone disease (GSD) is a common gastrointestinal disease worldwide. It is the term used for the development of calculi in the gallbladder and other parts of the biliary tract. The stones frequently result in cholecystitis, pancreatitis, and biliary tract obstruction. Recent epidemiological studies found gallstones in about 6.1% of the global population, showing greater susceptibility among South Americans, females, and older people compared to Asian people [1].Multiple known risk factors contribute to the pathogenesis of GSD. They include lifestyle, sex, age, metabolic syndrome, and obesity. A study on Chinese adults has shown that metabolically healthy and unhealthy obesity were at the highest risk. [2]. A study in Taiwan also confirmed that central obesity, measured by the waist-height ratio, predicts GSD, especially among females [3].

There is an advancement in predictive modeling aimed at enhancing the early detection of GSD through non-invasive techniques. A recent study has developed a machine-learning (ML) model that utilizes bioimpedance and laboratory examination, achieving an accuracy of approximately 85.42% in gallstone prediction. The primary predictive factors were vitamin D, C-reactive protein, total body water, and lean body mass [4]. Despite these advancements, population-specific models are still necessary to enhance early detection and intervention programs. This study aims to develop and validate a risk prediction model for gallstone disease based on a moderately sized dataset with a broad range of clinical and demographic variables. To enable proactive GSD management and alleviate the health burden of the disease, this study tries to identify the primary predictors and build an accurate model.

## II. Relevant Literature

One research conducted discovered that by utilizing ridge regression-based machine learning models, the presence of hepatocellular carcinoma in individuals with chronic viral hepatitis indicates the potential use of such models in early intervention strategies [5]. Researchers have found that machine learning (ML) algorithms based on hemoglobin level

parameters and specific gravity can predict chronic kidney disease (CKD) with a detection rate exceeding 97%, which is clinically significant [6]. A study demonstrated that ML plays a pivotal role in cervical cancer prediction by using imbalanced data and various sampling methods [7]. Designing models to handle unbalanced data effectively results in a higher detection rate of high-risk individuals. These findings demonstrate the effectiveness and applicability of ML methods in various medical disorders, highlighting their role in the development of modern healthcare.

Several studies have utilized machine learning to enhance gallstone disease prediction by incorporating clinical and laboratory data. A model has been developed based on bioimpedance and metabolic markers from 319 subjects, identifying vitamin D deficiency, elevated C-reactive protein (CRP), total body water, and lean mass as critical predictors. Their gradient boosting classifier achieved an accuracy of 85.42%, comparable to traditional ultrasonography, but with the added advantages of being non-invasive and cost-effective. Additional findings suggest a connection between gallstone formation and systemic inflammation, obesity-related factors, and alterations in body composition, including bone mass and the total body fat ratio. These approaches highlight the value of integrating physiological, biochemical, and demographic features to improve the early detection of gallstones. The study's results demonstrate that incorporating comprehensive datasets into machine learning models for gallstone prediction is crucial. These integrative methodologies are valuable in improving diagnostic capabilities.

Integrating Explainable Artificial Intelligence (XAI) within healthcare systems meets the need for transparency and confidence in AI-driven decisions. Traditional AI models often act like a "black box" and tend to be uninterpretable, making it difficult for clinicians to determine the cause of some predictions. Healthcare professionals can validate and trust AI predictions using XAI techniques such as Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). These XAI techniques provide insights into the decision-making processes of the model [8]. This transparency is significant, as it helps identify potential biases and ensures that AI recommendations are aligned with ethical standards and medical expertise, thereby facilitating greater adoption in clinical settings.

In a systematic review, a study noted the use of XAI in neurodegenerative disease prediction, highlighting the application of SHAP to identify the most significant cognitive and imaging features related to Alzheimer's disease, which enables early and precise diagnosis [9]. Another study also used SHAP and LIME on breast cancer prediction models. They demonstrated how explainers facilitated increased transparency by identifying the most informative biomarkers that resulted in a patient's diagnosis [10]. These studies suggest the potential applicability of XAI in enhancing clinical trust and decision-making precision in machine learning models used in the healthcare industry. Although the use of XAI in gallstone disease research is in its infancy, it has potential. Few experiments have been conducted specifically using SHAP and LIME to predict and classify gallstone disease, despite the increasing number of studies utilizing machine learning for medical diagnosis. Researchers have developed machine learning models to predict the development of gallstones using various clinical and demographic variables. Using XAI on these models can render them more interpretable, allowing clinicians to determine which factors are most significant in predicting gallstone risk. More individualized patient care and targeted prevention interventions may result from this information. The application of XAI in gallstone research can lead to more transparent and efficient diagnostic tools as the field of research develops. To help clinicians understand what features (e.g., demographics, laboratory work, symptoms) contribute the most to predictions, the current study uses SHAP and LIME to explain model predictions in gallstone classification. This study fills a previously underrepresented field within medical research by maximizing the transparency of machine learning models and their applications in diagnosis.

With their strong classification performance, machine learning algorithms such as Random Forest, Support Vector Machine, Gradient Boosting, and Logistic Regression have become widely used throughout the healthcare industry. For example, a study that utilized explainable machine learning-based prediction demonstrated the accuracy of such models in predicting diabetic nephropathy.The XGBoost model achieved an AUC of 0.966, with SHAP identifying key serum metabolites contributing to predictions. [11]. Another study demonstrated how ensemble techniques, such as Gradient Boosting Decision Trees, can effectively handle tabular medical data with high accuracy [12]. Such studies affirm the use of these algorithms in gallstone disease prediction, where imbalanced datasets and numerous patient factors are typical liabilities. Through the elimination of redundant input data, feature selection techniques, therefore, play a significant role in making models more accurate. Among the most common methods used are correlation analysis and ANOVA F-value, which have been effective in improving diagnostic models in different medical studies. Through a focus on the most significant clinical factors, a study involving feature selection for heart disease demonstrated that utilizing these methods in predicting heart disease yielded more interpretable and practical models. The study showed that Mutual Information (MI) feature selection led to the best results, with Neural Networks achieving 82.3% accuracy and 0.94 recall. Logistic Regression and Random Forest also performed well and consistently across MI, ANOVA, and Chi-Square, with accuracies of 82.1% and 80.99%. In contrast, k-NN had poor results ( 51% accuracy), showing its sensitivity to irrelevant features. These findings support the importance of feature selection, especially for enhancing complex models like those the researchers were exploring [13]. Through the improvement of model transparency in medical use, XAI techniques such as SHAP and LIME are becoming increasingly vital. However, a review on XAI using LIME and SHAP noted their significance in improving clinician trust and comprehension

in complex models [14]. On the other hand, another study on XAI successfully utilized both tools in diabetes prediction to identify key features that impact model decisions [15]. Combining SHAP and LIME realizes the need for interpretable outcomes in gallstone disease, making predictions accurate and comprehensible to medical professionals.

## III. Materials and Methods

The study follows a structured pipeline composed of a defined set of stages outlined in Figure 1. The researchers downloaded the dataset from the archive provided by the UC Irvine Machine Learning Repository. Exploratory Data Analysis was employed to understand the dataset and prepare it for further analysis and evaluation. Categorical features were encoded using one-hot encoding, while numerical features, such as bioimpedance and laboratory measurements, were standardized to ensure uniform scales during training. Pre-processing steps included verifying missing values and segregating features.

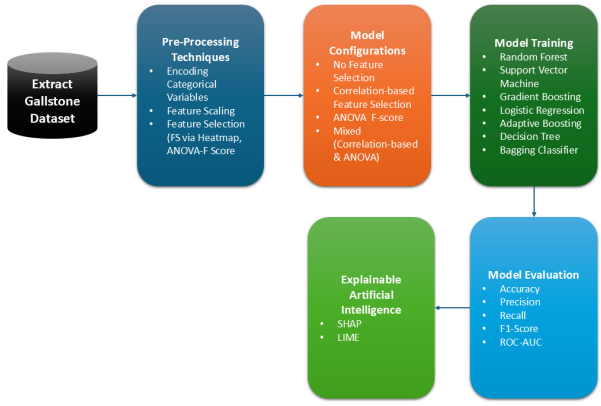


Fig. 1. Machine Learning Pipeline for Gallstone Disease Prediction Study

### A. Dataset Specification

The dataset used in this study was sourced from a publicly available repository containing clinical and laboratory data for gallstone disease classification. It comprises 319 patient records without missing values. Each record includes numerical and categorical features derived from clinical examinations and laboratory tests. The target variable, labeled "Gallstone Status," indicates gallstone presence (1) or absence (0). Exploratory data analysis confirmed balanced class distribution.

### B. Pre-processing Steps

The dataset underwent thorough pre-processing prior to modeling. Initially, features were separated by data type into numerical, categorical, and laboratory-related categories, facilitating tailored pre-processing approaches. Numeric features were standardized using StandardScaler, transforming values to a mean of zero and unit variance. Categorical variables underwent one-hot encoding (`OneHotEncoder`) to transform discrete features into a suitable numerical format. Due to balanced class distribution, class-imbalance handling methods like SMOTE were unnecessary. Potential outliers identified during exploratory analysis were retained, given their clinical relevance.

### C. Feature Selection and Class Imbalance Handling

To enhance model interpretability and predictive performance, multiple feature selection strategies were carefully explored. Initially, a correlation analysis was conducted using a correlation heatmap to identify features strongly correlated with the target variable, "Gallstone Status." A threshold corresponding to the 75th percentile of absolute correlations was used to retain the most relevant features. Additionally, the ANOVA F-score method, implemented through the `SelectKBest` technique, was applied. This approach statistically assessed the discriminative power of each feature between the two target classes, selecting the top 25% of features based on their F-scores. Finally, a combined "Mixed Selection" approach was performed, integrating both correlation and ANOVA methods. This robust approach selected features that appeared consistently across both selection methods, ensuring the chosen features were statistically significant and highly correlated with gallstone disease prediction.

### D. Model Configuration and Machine Learning Models

The dataset was divided into training and testing subsets using a 70-30 stratified split, ensuring balanced representation of both target classes. Numerical features were standardized using standard scaling to transform them to zero mean and unit variance, optimizing model performance. Feature selection was performed using the ANOVA F-value method with `SelectKBest`, applying a threshold at the 75th percentile to retain the most discriminative features. Four configurations were tested to evaluate how feature selection impacts model performance:

1) **Baseline Configuration:** This configuration used the raw, unprocessed dataset with no feature selection, providing a baseline for comparison.
2) **Heatmap-based Configuration:** Feature selection was based on a correlation heatmap, which identified features most strongly correlated with the target variable, "Gallstone Status." This approach ensured that only features with high correlation to the target were retained.
3) **ANOVA F-Score Configuration:** Here, feature selection was performed using the ANOVA F-value method with `SelectKBest`, selecting the top 25% of features based on their F-scores, emphasizing the most statistically significant predictors.
4) **Mixed Configuration:** This configuration combined features selected by both the correlation heatmap and ANOVA F-score methods, ensuring that only the most relevant and highly correlated features were used.

Seven machine learning algorithms were evaluated: Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GB), Logistic Regression (LR), AdaBoost (AB), Decision Tree (DT), and Bagging Classifier (BC). Models were trained

and validated using 5-fold stratified cross-validation, which ensured balanced class distribution in each fold, reducing the risk of overfitting. Each model's performance was rigorously assessed using several evaluation metrics, including accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC). This comprehensive evaluation across the four configurations enabled a detailed comparison of model effectiveness, providing valuable insights into the most robust machine learning model for predicting gallstone disease.

### *E. Explainable AI*

To enhance model interpretability and clinical trust, this study utilized Shapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). These explainability methods were applied to the Logistic Regression model from the Mixed configuration, which provided the best performance. SHAP was used to compute both global feature importance and local explanations for individual predictions. SHAP global feature importance was visualized using a bar plot to identify the features with the highest contributions to the model's predictions. Additionally, LIME was applied to explain individual predictions through beeswarm plots, which provide a visual representation of the distribution of feature impacts on the model's outputs. These methods enabled a transparent understanding of how specific clinical and laboratory parameters influenced the model's predictions, ensuring that clinicians could interpret and trust the results in the context of gallstone disease prediction.

## RESULTS AND DISCUSSION

The study evaluated the performance of seven machine learning models—Gradient Boosting (GB), Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), Bagging Classifier (BC), and AdaBoost (AB)—in predicting gallstone disease using clinical and laboratory data. Model performance was assessed across several metrics, including accuracy, precision, recall, F1-score, and AUC. The results from the four different configurations—Baseline, Heatmap-based, ANOVA F-score, and Mixed—are summarized in Tables I-IV.

In the **Baseline Configuration** as seen in Table I , Random Forest (RF) performed the best, achieving the highest accuracy (73.10%), precision (74.84%), and F1-score (72.96%). These results suggest that RF strikes a solid balance between identifying positive and negative cases, making it a strong candidate for gallstone prediction. Logistic Regression (LR) followed closely, with a high accuracy (66.84%), precision (75.75%), and AUC (78.09%), indicating it was able to correctly classify cases with good overall performance. Gradient Boosting (GB) also showed strong results, achieving a recall (62.73%), meaning it was effective at identifying positive cases. However, Support Vector Machine (SVM) exhibited a weaker performance, with a low recall (35.62%), highlighting its tendency to misclassify a significant portion of true positive cases despite its high precision (94.94%). The Decision Tree (DT) and Bagging Classifier (BC) models displayed moderate performance, with DT achieving the highest recall (70.91%), indicating it was effective at detecting positive cases but still lacked precision. AdaBoost (AB) had the weakest overall performance, failing to meet the predictive standards of the other models.

Table II details metrics from the **Heatmap-based Configuration** feature selection was performed based on the correlation heatmap, Logistic Regression (LR) continued to show strong performance, achieving an accuracy of 66.34%, a precision of 66.28%, and an AUC of 65.74%. However, its recall (61.82%) was slightly lower compared to the Baseline Configuration, suggesting that some feature selections based on the heatmap reduced the model's ability to detect positive cases. Random Forest (RF) showed a moderate drop in accuracy (58.77%), but its recall (57.27%) remained relatively robust, indicating that RF can still effectively identify gallstone cases with fewer features. Gradient Boosting (GB) and SVM showed similar performance as in the baseline configuration, but with SVM's recall (35.62%) remaining low. Decision Tree (DT) and Bagging Classifier (BC) continued to demonstrate moderate performance, with DT again being the better performer in recall but still lacking in precision.

Table III details the metrics from the **ANOVA Configuration** improved results across models compared to the Heatmap-based configuration. Logistic Regression (LR) again performed well, achieving an accuracy of 69.51%, precision of 71.37%, and recall of 64.55%, marking it as a reliable model with balanced performance. Gradient Boosting (GB) and Random Forest (RF) also performed well with recall scores of 57.27% and 61.82%, respectively. However, SVM continued to struggle with low recall (53.64%), indicating that it misclassifies many true positive cases, despite its high precision (69.99%). Decision Tree (DT) and Bagging Classifier (BC) again showed moderate results, with DT achieving higher recall than BC, but both models still underperformed compared to LR and GB.

Table IV presents the results of the **Mixed Configuration** where features were selected using both correlation heatmap and ANOVA F-score methods, Logistic Regression (LR) once again proved to be the best-performing model, achieving accuracy (69.51%), precision (71.37%), and recall (64.55%), along with AUC (72.02%). These results demonstrate the robustness of LR across various configurations and highlight its ability to balance precision and recall effectively. Random Forest (RF) maintained a similar performance to the other configurations with accuracy (64.60%) and recall (61.82%), but its precision (67.39%) was consistently high. Gradient Boosting (GB) demonstrated solid performance in recall (57.27%), but SVM continued to struggle with low recall (35.62%), although it maintained high precision (69.99%). Decision Tree (DT) and Bagging Classifier (BC) showed consistent but relatively weaker performance, with DT performing better in recall but still showing lower discriminative power overall.

Overall, the results showed that Logistic Regression (LR) and Gradient Boosting (GB) performed better than the other models in every configuration. As evidenced by its high recall,

GB was best at detecting positive cases, although LR was superior in accuracy, precision, and AUC. Random Forest (RF) also did well, especially in recall, but its overall performance was constrained by its lower precision when compared to LR. AdaBoost performed the worst, and SVM had trouble with low recall. For gallstone prediction and additional assessment, models that prioritized recall with balanced criteria, such as precision and F1-score, were preferred.

The study evaluated four model configurations—Baseline, Heatmap-based, ANOVA F-score, and Mixed—to explore the effect of feature selection on model performance in predicting gallstone disease. All configurations used 5-fold cross-validation, ensuring robust and reliable performance estimation across the dataset. Among the models tested, Logistic Regression (LR) emerged as the best performer overall, with the highest AUC values across all configurations. This model demonstrated excellent accuracy and precision, making it an ideal candidate for further interpretability analysis.

| ML Model | Accuracy | Precision | Recall | F1 Score | AUC |
|---|---|---|---|---|---|
| RF | 73.10 | 74.84 | 72.73 | 72.96 | 80.17 |
| SVM | 47.07 | 53.51 | 37.27 | 35.62 | 49.18 |
| GB | 70.83 | 75.32 | 62.73 | 67.91 | 78.91 |
| LR | 72.18 | 75.75 | 66.36 | 70.21 | 78.09 |
| AB | 70.87 | 74.53 | 63.64 | 68.13 | 75.99 |
| DT | 65.92 | 64.29 | 70.91 | 67.09 | 66.05 |
| BC | 66.81 | 70.05 | 58.18 | 62.74 | 74.91 |

TABLE I
PERFORMANCE METRICS ON THE **BASELINE** CONFIGURATION

| ML Model | Accuracy | Precision | Recall | F1 Score | AUC |
|---|---|---|---|---|---|
| RF | 58.77 | 59.64 | 57.27 | 58.01 | 62.81 |
| SVM | 65.01 | 69.87 | 54.55 | 60.34 | 66.63 |
| GB | 59.65 | 60.98 | 57.27 | 58.72 | 67.66 |
| LR | 66.34 | 66.28 | 64.55 | 65.30 | 69.08 |
| AB | 64.11 | 65.56 | 61.82 | 63.23 | 65.74 |
| DT | 57.81 | 58.25 | 56.36 | 56.80 | 57.85 |
| BC | 60.56 | 63.60 | 54.55 | 57.69 | 64.03 |

TABLE II
PERFORMANCE METRICS ON THE **HEATMAP-BASED** CONFIGURATION.

| ML Model | Accuracy | Precision | Recall | F1 Score | AUC |
|---|---|---|---|---|---|
| RF | 59.64 | 59.80 | 60.00 | 59.39 | 66.73 |
| SVM | 64.54 | 69.09 | 53.64 | 59.55 | 69.43 |
| GB | 66.85 | 69.72 | 63.64 | 65.71 | 70.56 |
| LR | 69.51 | 71.37 | 64.55 | 67.50 | 72.02 |
| AB | 61.87 | 60.75 | 62.73 | 61.60 | 65.85 |
| DT | 61.84 | 62.35 | 57.27 | 59.19 | 61.86 |
| BC | 64.11 | 65.89 | 58.18 | 61.37 | 69.39 |

TABLE III
PERFORMANCE METRICS ON THE **ANOVA F-SCORE** CONFIGURATION.

In terms of feature selection, the Mixed configuration, which combined ANOVA F-score and Heatmap-based selection, produced the most balanced and comprehensive performance across the models. This configuration allowed for a more varied spread of results, demonstrating the power of integrating multiple selection techniques to retain both highly correlated and statistically significant features. Figure 2 revealed that key features like Total Body Fat Ratio (TBFR), Extracellular Water (ECW), and Bone Mass (BM) were the most influential in determining the predictions, with TBFR contributing the most to the model's decision-making process.

| ML Model | Accuracy | Precision | Recall | F1 Score | AUC |
|---|---|---|---|---|---|
| RF | 64.60 | 67.39 | 61.82 | 63.54 | 67.08 |
| SVM | 64.54 | 69.09 | 53.64 | 59.55 | 69.43 |
| GB | 66.85 | 69.72 | 63.64 | 65.71 | 70.52 |
| LR | 69.51 | 71.37 | 64.55 | 67.50 | 72.02 |
| AB | 61.87 | 60.75 | 62.73 | 61.60 | 65.85 |
| DT | 65.00 | 66.52 | 60.91 | 63.05 | 65.02 |
| BC | 65.01 | 69.11 | 55.45 | 60.91 | 68.43 |

TABLE IV
PERFORMANCE METRICS ON THE **MIXED** CONFIGURATION.

Figure 3 confirms these findings, illustrating how changes in feature values impacted the predictions. For instance, Extracellular Water (ECW) and Total Body Water (TBW) showed positive correlations with gallstone risk, meaning higher values increased the likelihood of a positive diagnosis. Conversely, features like Lean Mass (LM) showed an inverse relationship, where higher values correlated with a reduced risk of gallstones.

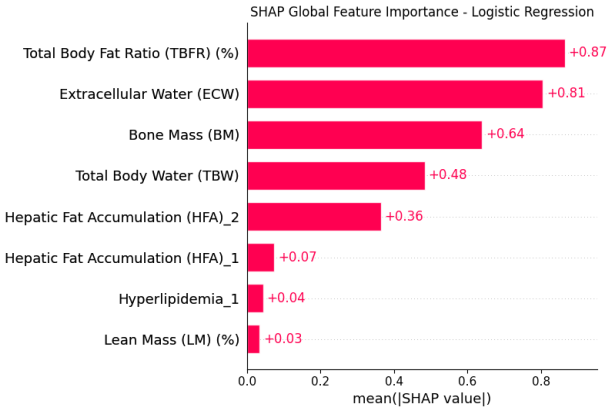


Fig. 2. SHAP Global Feature Importance Bar Plot for LR

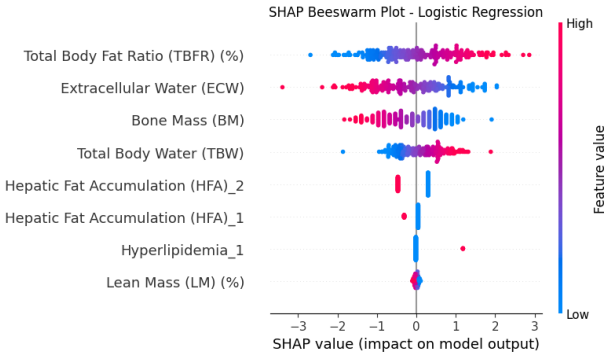


Fig. 3. Beeswarm Plot for LR

Figure 4 shows a LIME instance on a positive sample observation from the Logistic Regression (LR) model's test set. The model predicted the sample as positive for gallstone

disease (GSD) with a 44% probability. The Extracellular Water (ECW) feature and Bone Mass (BM) were the most influential in determining the positive prediction, where lower values of ECW and BM contributed to the higher probability of a positive prediction. The model identified that the reduced ECF/TBW ratio and Bone Mass values were indicative of an increased risk for gallstones, reinforcing the role of these features in identifying patients at risk. This individual positive instance however, is indicative of an uncertainty in the logistic regression model wherein the LIME explainer is reflecting a borderline instance.

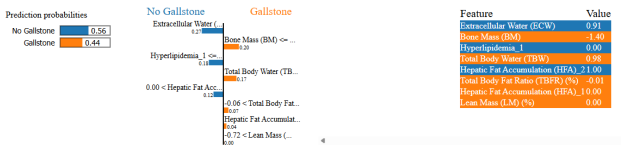


Fig. 4. LIME Output with LR on a positive sample

In contrast, Figure 5 shows a LIME instance on a negative sample observation with 96% confidence predicting the absence of gallstones. In this case, the model highlighted higher levels of Extracellular Water (ECW), Bone Mass (BM), and Total Body Water (TBW), which were considered protective features against the disease. These features, with higher values, contributed significantly to the model's decision for a negative diagnosis. The Total Body Fat Ratio (TBFR) also played a minor role in reducing the likelihood of GSD, as indicated by its influence on the model's output.


Fig. 5. Lime Output with LR on a negative sample

Gallstone disease (GSD) was found to be significantly predicted by parameters including Extracellular Water (ECW), Bone Mass (BM), and Total Body Water (TBW), according to interpretability assessments conducted on the Logistic Regression (LR) model utilizing SHAP and LIME. In particular, lower BM and ECW values were associated with a higher risk of GSD, demonstrating the clinical use of these characteristics in determining GSD risk. The model's decision-making process was better understood because of the insights from SHAP and LIME, which increased confidence in the model's predictions. The results show that Logistic Regression, enhanced with SHAP and LIME, provides an efficient and comprehensible method for predicting GSD risk by elucidating the influence of these factors both locally (for individual instances) and globally (over the entire dataset).

## Conclusions and Recommendations

This study evaluated the predictive accuracy of a selection of machine learning algorithms for GSD using various physiological and clinical characteristics. Specifically, the study aimed to address the scarcity of research on interpretable gallstone prediction models and enhance transparency in model decision-making by incorporating explainable AI techniques, such as SHAP and LIME. The study tackled a determined set of configurations in order to confidently determine how to tackle the evaluated dataset. The four configurations were designed to explore the impact of different feature selection methods on the performance of the seven models trained and evaluated, with the goal of identifying the most effective selection of features that net the best metrics and consequently yield the best interpretability when augmented with the aforementioned explainers. Through this study's methodology, the importance of a varied but consistent dataset in addition to robust feature selection has in reinforcing the value of machine learning with explainable artificial intelligence has in the clinical decision making process. With that being said, the model training and evaluation resulted in the Logistic Regression performing consistently throughout the configurations only outperformed by Random Forest in the Baseline configuration. Additionally, the configuration that utilized the mixed feature selection resulted in the most consistent spread of cross-validated model metrics combined with moderate performance. Furthermore, the augmentation of SHAP and LIME explainers provide transparent insights into feature contributions, reinforcing clinical interpretability. These findings are highly indicative of the effectiveness of data-driven tools in facilitating early diagnosis and personalized risk assessment and management.

Future studies are encouraged to expand the dataset size and diversity through relevant means, such as collecting data from multiple centers, to improve model generalizability and robustness further. Researchers may employ different feature engineering approaches by incorporating advanced data, such as imaging data, genetic markers, or longitudinal clinical measurements, which can reveal deeper patterns. In addition, various pre-processing measures should also be taken into consideration to improve model performance in conjunction with the diversified dataset. In terms of modeling, future studies should consider a different pool of models accompanied by additional ensemble and stacking techniques to further improve performacne and reduce model uncertainty, particularly for borderline instances where the model encounters uncertainty in predictions. Additionally, researchers may consider integrating more advanced explainable artificial intelligence techniques to further refine and add interpretability, particularly for higher-dimensional datasets. Lastly, the implementation of neural networks such as deep learning models, are encouraged to further understand the non-linear relationships in the data.

## References

[1] X. Wang et al., "Global Epidemiology of Gallstones in the 21st Century: A Systematic Review and Meta-Analysis," Clinical Gastroenterology and Hepatology, vol. 22, no. 8, Feb. 2024, doi: https://doi.org/10.1016/j.cgh.2024.01.051.

[2] J. Zhang et al., "Association between metabolically healthy overweight/obesity and gallstones in Chinese adults," Association between metabolically healthy overweight/obesity and gallstones in Chinese

adults, vol. 20, no. 1, Mar. 2023, doi: https://doi.org/10.1186/s12986-023-00741-4.

[3] GlobalRPH, “Obesity Classification And The Risk Of Gallstones,” GlobalRPH, Nov. 14, 2023. https://globalrph.com/2023/11/obesity-classification-risk-of-gallstones

[4] İrfan Esen, H. Arslan, Selin Aktürk Esen, Mervenur Gülşen, Nimet Kültekin, and Oğuzhan Özdemir, “Early prediction of gallstone disease with a machine learning-based method from bioimpedance and laboratory data,” Medicine, vol. 103, no. 8, pp. e37258–e37258, Feb. 2024, doi: https://doi.org/10.1097/md.0000000000037258.

[5] G. L.-H. Wong et al., “Novel machine learning models outperform risk scores in predicting hepatocellular carcinoma in patients with chronic viral hepatitis,” JHEP Reports, vol. 4, no. 3, p. 100441, Mar. 2022, doi: https://doi.org/10.1016/j.jhepr.2022.100441.

[6] Md. A. Islam, Md. Z. H. Majumder, and Md. A. Hussein, “Chronic kidney disease prediction based on machine learning algorithms,” Journal of Pathology Informatics, p. 100189, Jan. 2023, doi: https://doi.org/10.1016/j.jpi.2023.100189.

[7] Mădălina Maria Muraru, Zsuzsa Simó, and László Barna Iantovics, “Cervical Cancer Prediction Based on Imbalanced Data Using Machine Learning Algorithms with a Variety of Sampling Methods,” Applied Sciences, vol. 14, no. 22, pp. 10085–10085, Nov. 2024, doi: https://doi.org/10.3390/app142210085.

[8] Razan Alkhanbouli, Hour, F. Alhosani, and M. Can, “The role of explainable artificial intelligence in disease prediction: a systematic literature review and future research directions,” BMC Medical Informatics and Decision Making, vol. 25, no. 1, Mar. 2025, doi: https://doi.org/10.1186/s12911-025-02944-6.

[9] V. Vimbi, N. Shaffi, and M. Mahmud, “Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer’s disease detection,” Brain informatics, vol. 11, no. 1, Apr. 2024, doi: https://doi.org/10.1186/s40708-024-00222-1.

[10] S. Rao, S. Mehta, S. Kulkarni, H. Dalvi, N. Katre and M. Narvekar, ”A Study of LIME and SHAP Model Explainers for Autonomous Disease Predictions,” 2022 IEEE Bombay Section Signature Conference (IBSSC), Mumbai, India, 2022, pp. 1-6, doi: 10.1109/IBSSC56953.2022.10037324.

[11] J.-M. Yin, Y. Li, J.-T. Xue, G.-W. Zong, Z.-Z. Fang, and L. Zou, “Explainable machine learning-based prediction model for diabetic nephropathy,” arXiv.org, 2023. https://arxiv.org/abs/2309.16730

[12] Y. A. Yarkın and A. Kalayci, “Gradient Boosting Decision Trees on Medical Diagnosis over Tabular Data,” arXiv.org, 2024. https://arxiv.org/abs/2410.03705

[13] B. Ahmad, J. Chen, and H. Chen, “Feature selection strategies for optimized heart disease diagnosis using ML and DL models,” arXiv.org, 2025. https://arxiv.org/abs/2503.16577

[14] A. S. Shaikh, R. M. Samant, K. S. Patil, N. R. Patil, and A. R. Mirkale, “Review on Explainable AI by using LIME and SHAP Models for Healthcare Domain,” International Journal of Computer Applications, vol. 185, no. 45, pp. 18–23, 2023, Accessed: May 30, 2025. [Online]. Available: https://www.ijcaonline.org/archives/volume185/number45/32992-2023923263

[15] M. Panda and M. S. Ranjan, “Explainable artificial intelligence for Healthcare applications using Random Forest Classifier with LIME and SHAP,” arXiv.org, 2023. https://arxiv.org/abs/2311.05665