

ANOVA–Mutual Information Feature Selection with CatBoost for Gallstone Disease Classification

Shuvra Baral*
Kantipur Engineering College
Lalitpur, Nepal
shuvrabaral@kec.edu.np

Jianchu Lin
Huai'yn Institute of Technology
Huai'an, China
jianchulin@gmail.com

Jialin Ma
Huai'yn Institute of Technology
Huai'an, China
majl@hyit.edu.cn

* Corresponding author

Abstract—Gallstone disease is a common biliary disorder for which non-invasive, ultrasonography-independent risk assessment may support clinical decision-making. This study proposes an ANOVA–Mutual Information (ANOVA–MI) feature-selection framework combined with CatBoost, evaluated on 319 patients and 38 candidate predictors from the UCI Gallstone dataset. The framework fuses min–max normalized ANOVA F-statistics and mutual-information scores through a tunable weight α , ranks features, and selects compact subsets inside a nested repeated stratified cross-validation protocol (5×5 outer folds, 3-fold inner Optuna tuning) to prevent information leakage. At $\alpha = 0.3$, the method retains 15 features (60.5% dimensionality reduction) and achieves an F1-score of 0.800 ± 0.021 and ROC-AUC of 0.861 ± 0.032 , slightly outperforming the full 38-feature CatBoost baseline as well as ANOVA-only and MI-only selection. The corresponding sensitivity and specificity are 78.2% and 82.6%, respectively. Compared with three recent studies on the same dataset, the proposed approach attains the highest recall and F1-score while using fewer or a comparable number of features. SHAP analysis identifies C-reactive protein and vitamin D as the dominant predictors. These results demonstrate that a model-agnostic hybrid of statistical discrimination and information-theoretic relevance can substantially reduce feature dimensionality while preserving competitive predictive performance under rigorous nested evaluation.

Keywords—Gallstone Disease, Feature Selection, ANOVA, Mutual Information, CatBoost, Nested Cross-Validation

I. INTRODUCTION

Gallstone disease, also known as cholelithiasis, is a prevalent disorder of the biliary system. Although many gallstones remain asymptomatic, symptomatic disease can result in biliary colic, acute cholecystitis, choledocholithiasis, cholangitis, and pancreatitis [1]. Gallstone-related symptoms may include sudden or intensifying abdominal pain, back or shoulder pain, nausea, and vomiting, while severe abdominal pain, jaundice, fever, or chills may indicate complications requiring medical attention [2]. Consequently, reliable identification of individuals at elevated risk of gallstone disease can support clinical assessment and facilitate appropriate diagnostic investigation.

Ultrasonography remains an important diagnostic modality for gallstones because it is comparatively accessible and does not involve ionizing radiation. Nevertheless, diagnostic imaging can be affected by operator experience, patient characteristics, image quality, and resource availability. Recent research has therefore investigated machine-learning methods for both

image-based and non-imaging gallstone detection. Systematic evidence published in recent years indicates increasing interest in artificial intelligence for cholelithiasis diagnosis and risk assessment [3].

The dataset contains 319 individuals, including 161 participants diagnosed with gallstone disease and 158 without gallstone disease, represented using 38 variables which consists of demographic, bioimpedance, and laboratory measurements [4]. This relatively balanced dataset provides an opportunity to investigate whether routinely collected clinical measurements can support machine-learning-based gallstone classification without relying directly on medical images.

Esen et al. [4] introduced the dataset and demonstrated the feasibility of machine-learning-based prediction using bioimpedance and laboratory information. Subsequent studies have investigated ensemble learning, feature reduction, explainable artificial intelligence, and optimization techniques on the same dataset [5–7]. Random Forest with Sand Cat Swarm Optimization reduced the original 38 variables to 13 features while maintaining comparable predictive performance [5]. More recently, CatBoost optimized using a Sea Lion Optimization Algorithm was investigated using 19 selected variables [6]. These studies demonstrate that feature reduction can be valuable for gallstone diagnosis classification; however, they also highlight the need for systematic and reproducible feature-selection strategies.

Feature selection is particularly important in clinical machine learning because the objective is not necessarily to maximize the number of variables used by the model. Redundant, weakly informative, or highly correlated variables can increase model complexity, obscure interpretation, and potentially reduce generalization [8]. A compact feature representation may therefore provide computational and interpretability advantages while maintaining competitive predictive performance [9–12]. Nevertheless, the effectiveness of a feature-selection method depends strongly on the criterion used to define feature relevance.

Statistical filter methods such as analysis of variance (ANOVA) provide a simple and interpretable mechanism for identifying variables whose distributions differ significantly between diagnostic groups. For a continuous feature X_j , the ANOVA F-statistic quantifies the ratio between between-group and within-group variation. Thus, ANOVA can identify

features exhibiting strong univariate group discrimination. However, statistical significance alone does not necessarily imply high predictive utility. A feature may show a statistically significant difference between gallstone and non-gallstone groups while providing limited additional information because of redundancy with other variables.

Information-theoretic feature-selection methods provide a complementary perspective. Entropy quantifies uncertainty in a random variable, whereas mutual information (MI) measures the reduction in uncertainty about one variable obtained from another. For a predictor X_j and the diagnostic target Y , MI is defined as $I(X_j; Y) = H(Y) - H(Y|X_j)$, thereby providing an entropy-derived measure of feature–target dependence. Unlike ANOVA, which primarily evaluates between-group variation, MI can capture statistical dependencies that may not be adequately represented by differences in group means. Therefore, combining ANOVA-based discrimination with mutual information provides complementary evidence for ranking candidate predictors.

Despite recent progress, several methodological gaps remain in existing gallstone ML studies. First, feature selection is frequently treated as a preprocessing operation without sufficiently examining whether the selected variables provide statistically and information-theoretically complementary evidence. Second, methods based exclusively on univariate statistical significance may overlook informative dependencies that are not reflected by differences in group means, whereas information-theoretic methods alone do not explicitly account for the magnitude of between-group variation. Third, the number of retained features is often predetermined rather than systematically investigated, making it difficult to determine whether a compact subset provides a genuine performance advantage over the complete feature space. Finally, evaluation based primarily on F_1 -score or ROC-AUC can obscure clinically relevant differences in sensitivity and specificity, particularly when the objective is diagnostic classification.

To address these limitations, this study proposes a hybrid ANOVA–Mutual Information (ANOVA–MI) feature-selection framework combined with CatBoost for gallstone disease classification. The proposed framework first evaluates the statistical discriminative ability of individual features using the ANOVA F-statistic and then estimates their information-theoretic relevance using mutual information derived from entropy. The resulting ANOVA and MI scores are independently normalized using Min–Max normalization and combined through a tunable weighted fusion mechanism. Features are subsequently ranked according to the resulting hybrid relevance score, and multiple feature-subset sizes are evaluated using CatBoost. Importantly, feature selection is performed exclusively within the training data during cross-validation to prevent information leakage.

The proposed framework is evaluated against CatBoost using the complete feature set, ANOVA-only feature selection, MI-only feature selection, and the proposed integrated ANOVA–MI feature selection. In addition to accuracy, precision, sensitivity, specificity, F1-score, balanced accuracy,

Matthews correlation coefficient (MCC), ROC-AUC, and precision–recall AUC are considered to provide a comprehensive assessment of diagnostic performance. SHAP-based analysis is further employed to investigate the contribution of the selected features to the final CatBoost predictions.

The principal contributions of this study are as follows:

- 1) ANOVA–MI filter method is developed by combining normalized ANOVA F-statistics with MI score through a tunable weighted fusion mechanism.
- 2) The proposed method integrates statistical between-group discrimination with information-theoretic feature–target dependence rather than relying on a single feature-selection criterion.
- 3) Multiple feature-subset sizes are systematically evaluated to determine whether reducing the original 38-dimensional feature space can maintain or improve CatBoost classification performance.

The remainder of this paper is organized as follows. Section II presents the related work. Section III describes the dataset, preprocessing procedure, proposed ANOVA–MI feature-selection method, and CatBoost classification framework. Section IV presents the experimental results and comparative evaluation and discusses the findings. Finally, Section V concludes the study and outlines directions for future research.

II. RELATED WORK

Esen et al. introduced a structured clinical gallstone dataset comprising demographic, bioimpedance, and laboratory variables from 319 patients and demonstrated the feasibility of machine-learning-based prediction using ANOVA and model-based feature importance with Gradient Boosting [4]. The dataset subsequently became publicly available through the UCI Machine Learning Repository. A systematic review of machine-learning applications in cholelithiasis diagnosis reported increasing use of artificial intelligence for ultrasound and other imaging modalities, though tabular, non-imaging approaches such as the present study remain comparatively underexplored [3].

Recent studies using the same 319-patient dataset have produced several competing approaches, motivating close attention to methodological differentiation. Sarker et al. proposed a Random Forest classifier optimized using Sand Cat Swarm Optimization, reducing the original feature set and using SHAP and DiCE for interpretation [5]. Shrestha et al. optimized a Rotational Forest classifier using the Bald Eagle Search algorithm, reporting 78% accuracy and an AUC of 0.867 with the full feature set, and a comparable 75.78% accuracy and AUC of 0.860 using a Bald-Eagle-Search-reduced 17-feature subset - a result that, consistent with the ablation findings of the present study, suggests that metaheuristic feature reduction on this dataset tends to preserve rather than improve performance [13]. Das et al. evaluated CatBoost optimized using Sea Lion Optimization and reported a reduced feature set [6]. Indirani et al. investigated ensemble learning using clinical and biochemical variables [7]. These studies establish a strong

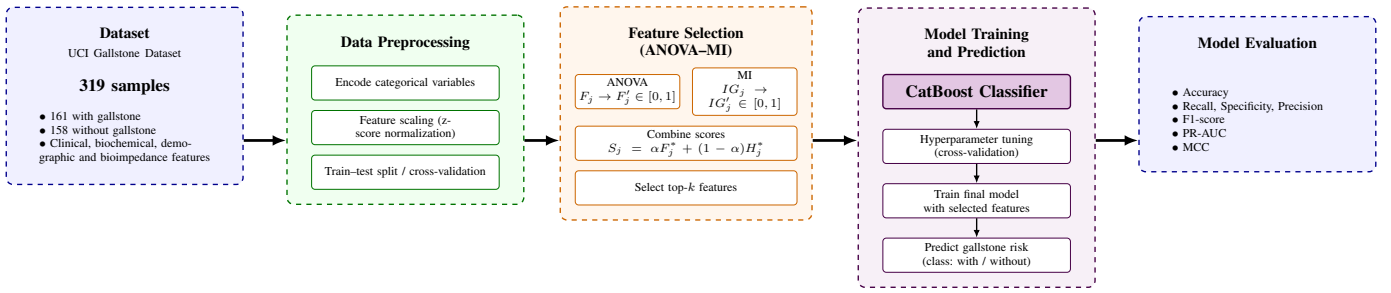


Fig. 1. Proposed ANOVA–MI feature-selection and CatBoost-based framework for interpretable gallstone disease classification.

empirical baseline but also indicate that the relatively small sample size makes methodological rigor – particularly cross-validation protocol and statistical validation of any feature-selection gain - essential.

Most directly related to the present work, Kusuma et al. proposed a hybrid ANOVA-SHAP feature-selection framework for the same dataset, combining a statistical filter criterion with model-based SHAP importance under a tunable weighting scheme and evaluating multiple weight configurations ($\alpha = 0.3, 0.5, 0.7$) across several classifiers [14]. While methodologically adjacent, their approach is model-dependent by construction: SHAP importance requires a trained classifier, so the feature-selection stage cannot be decoupled from a specific downstream model and must be recomputed for each candidate classifier. The present study instead pairs the statistical ANOVA criterion with an information-theoretic MI criterion, both computable directly from the data without fitting a predictive model. This yields a fully model-agnostic, training-free filter that can be validated independently of the downstream classifier and is, unlike prior work on this dataset, assessed under a nested cross-validation protocol across feature-selection conditions rather than a single train-test split.

Feature selection is used to identify a subset of variables that provides useful information for a predictive task while eliminating redundant or weak variables. Recent medical studies have demonstrated that feature selection can affect predictive performance, computational cost, and interpretability [11, 15, 16]. Pacheco et al. investigated multistart tabu search for feature selection in medical applications [17]. Klonecki and Teisseyre studied feature selection under budget constraints in medical datasets [11]. Vommi and Battula proposed a hybrid filter-wrapper strategy using fuzzy feature-selection mechanisms [18]. Zhong et al. proposed an entropy-weighted feature-selection method combining copula correlation and the maximal information coefficient, motivating an integrated approach in which statistical discrimination and information content are considered jointly [9].

CatBoost is a gradient-boosting algorithm based on symmetric decision trees and ordered boosting. Its ability to model nonlinear relationships and interactions makes it suitable for structured clinical datasets, and it has been applied to several healthcare problems in recent years. Safaei et al. developed E-

CatBoost for ICU mortality prediction [19]. Hu et al. applied CatBoost to sexually transmitted infection risk prediction using NHANES data [20]. Srinivasu et al. incorporated CatBoost into an explainable breast-cancer framework [21]. Ensemble-based classifiers more broadly have shown strong performance on comparable binary medical diagnosis tasks; for instance, an AdaBoost ensemble combining SVM and decision-tree base learners achieved high precision and recall for breast-cancer malignancy classification [22]. The use of CatBoost for gallstone classification is particularly relevant because recent work has already demonstrated its feasibility [6]. Therefore, the contribution of the present work is neither the use of CatBoost itself nor the general idea of a weighted hybrid feature-selection score, which Kusuma et al. have already explored with a model-dependent variant [14]; rather, it is the specific pairing of a model-agnostic, training-free ANOVA–MI fusion criterion with a rigorously validated nested-CV evaluation of whether that criterion can simplify CatBoost-based gallstone classification without sacrificing predictive performance.

III. METHODOLOGY

The proposed framework integrates statistical and information-theoretic feature selection with CatBoost classification for gallstone disease prediction. The overall workflow consists of six major stages: (i) dataset preparation, (ii) data preprocessing, (iii) ANOVA–MI feature selection, (iv) CatBoost model training and prediction, (v) comprehensive performance evaluation, and (vi) model interpretation using SHAP. The complete workflow is illustrated in Fig. 1.

The primary objective is to determine whether a compact feature subset selected using complementary statistical and information-theoretic criteria can preserve or improve the predictive performance of CatBoost compared with using all available variables. Since the dataset contains only 319 observations, particular attention is given to preventing information leakage during feature selection and model optimization.

A. Dataset Description

The experiments are conducted using the UCI Gallstone dataset, which contains clinical, demographic, bioimpedance, and laboratory measurements collected for gallstone disease prediction [4]. The dataset comprises 319 individuals, including 161 subjects without gallstone disease and 158 subjects

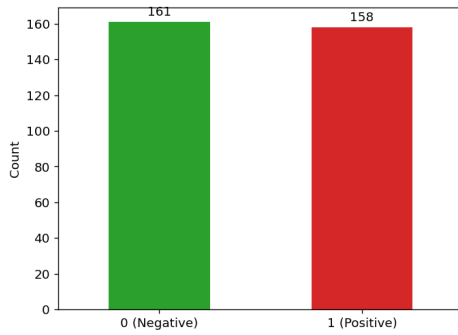


Fig. 2. Class distribution of Gallstone showing 161 negative and 158 positive cases out of 319 total samples.

with gallstone disease, as illustrated in Fig. 2. It contains 38 predictor variables representing demographic, bioimpedance, and laboratory measurements. The dataset is approximately class-balanced, with gallstone-negative and gallstone-positive cases accounting for 50.5% and 49.5% of the observations, respectively. The dataset contains no missing values, with all 319 observations available for the recorded variables. The predictor set consists of 7 binary, 2 ordinal, and 29 continuous variables. The prediction task is formulated as a binary classification problem. Let $Y \in \{0, 1\}$, where $Y = 1$ represents the presence of gallstone disease and $Y = 0$ represents its absence. Let $\mathbf{X} = [X_1, X_2, \dots, X_p]$ denote the $p = 38$ candidate predictor variables. The objective is to learn a classification function $\hat{Y} = f(\mathbf{X})$ that accurately distinguishes between the two diagnostic classes using a reduced, information-rich subset of $\mathbf{X}$, such that the number of retained predictors is minimized without compromising – and, where possible, improving – the predictive performance obtained from the complete 38-variable feature set.

ANOVA and mutual information (MI) provide complementary perspectives on feature–target association and neither criterion is uniformly superior. For a continuous predictor X_j , one-way ANOVA quantifies the extent to which the mean of the predictor differs between the gallstone-positive and gallstone-negative groups through the ratio of between-group to within-group variance. Consequently, features exhibiting substantial and consistent differences in group means can obtain large ANOVA F -statistics. However, ANOVA is primarily sensitive to differences in group means and may assign relatively low relevance to features whose predictive information is expressed through nonlinear or non-monotonic dependencies, differences in distributional shape, or variance rather than through a mean shift. In contrast, MI provides an information-theoretic measure of statistical dependence. For predictor X_j and target Y , it is defined as

$$MI_j = I(X_j; Y) = H(Y) - H(Y | X_j) \quad (1)$$

where $H(Y)$ denotes the entropy of the target and $H(Y | X_j)$ denotes the conditional entropy of the target given X_j . MI therefore measures the reduction in uncertainty about the diagnostic class provided by a predictor and can capture de-

pendencies that are not necessarily reflected by differences in group means. Hence, a feature may exhibit strong information-theoretic dependence with the target while showing relatively weak mean separation, whereas another feature may exhibit strong mean separation but comparatively limited additional information captured by MI.

The two criteria also have different practical considerations. The ANOVA F -statistic provides an analytically defined measure of group discrimination without requiring discretization, but its reliability can be affected by outliers, highly skewed distributions, unequal variances, and near-constant predictors. MI, on the other hand, can capture a broader range of statistical dependencies but may be sensitive to the estimation procedure. In the present study, continuous variables are discretized before MI estimation as described in Section III-A; consequently, the resulting MI values may depend on the adopted discretization strategy, particularly given the relatively small sample size ($n = 319$). These complementary characteristics motivate the fusion of ANOVA and MI rather than relying on either criterion independently.

The proposed fusion combines the normalized ANOVA and MI scores into a single model-independent relevance score. By integrating evidence from both between-group mean discrimination and information-theoretic dependence, the hybrid ranking is intended to reduce reliance on the limitations of any single univariate criterion. Importantly, neither ANOVA nor MI requires a trained classifier to derive the feature ranking. Therefore, the feature-selection procedure can be independently applied within each training fold of the nested cross-validation framework, with the resulting feature subset subsequently evaluated by the downstream CatBoost classifier. This separation prevents feature-selection information from being transferred from validation data into the training process and enables a model-independent assessment of whether the selected compact feature subsets can preserve or improve classification performance relative to the complete 38-feature representation.

B. Data Preprocessing

Prior to feature selection and model training, the 38 predictor variables are separated into categorical and continuous groups based on cardinality: variables with seven binary and 2 ordinal are treated as categorical and passed to the downstream models in their native encoding, while the remaining 29 variables are treated as continuous. Continuous variables are standardized using z-score normalization, $x' = (x - \mu)/\sigma$, where μ and σ are estimated exclusively from the training partition of each cross-validation fold and subsequently applied to the corresponding validation and test partitions. Fitting the scaler on training data only, rather than on the full dataset prior to splitting, prevents information from the held-out folds from influencing the standardization parameters – a common and avoidable source of optimistic bias in small-sample clinical prediction studies.

C. Proposed ANOVA-MI Feature-Selection Framework

For each candidate predictor X_j , $j = 1, \dots, p$, two relevance scores are computed independently.

1) *ANOVA F-statistic*: The one-way ANOVA F-statistic compares the diagnostic groups $k \in \{0, 1\}$:

$$F_j = \frac{\sum_{k=0}^1 n_k (\bar{x}_{jk} - \bar{x}_j)^2 / (K - 1)}{\sum_{k=0}^1 \sum_{i=1}^{n_k} (x_{jik} - \bar{x}_{jk})^2 / (N - K)}, \quad (2)$$

where $\bar{x}_{jk}$ is the mean of feature j within group k , $\bar{x}_j$ is the overall mean, n_k is the group size, N is the total sample size, and $K = 2$ is the number of diagnostic groups.

2) *Mutual Information*: For each predictor X_j , mutual information (MI) with the target variable Y is estimated using Eq. 1, where $H(\cdot)$ denotes Shannon entropy. To estimate the joint probability distribution $p(X_j, Y)$, continuous predictors are discretized into $q = 5$ equal-frequency bins using quantile-based binning before MI calculation. The empirical MI is then computed as

$$MI_j = I(X_j; Y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}. \quad (3)$$

Categorical predictors are retained in their native encoded representation and are not subjected to additional discretization. Equal-frequency quantile binning is selected because it allocates approximately equal numbers of observations across bins, reducing the occurrence of sparsely populated bins and limiting the influence of highly skewed continuous distributions. This property is particularly useful for the relatively small dataset ($n = 319$), where sparse contingency cells can produce unstable empirical probability estimates. The bin boundaries are estimated exclusively from the training partition of each cross-validation fold and subsequently applied to the corresponding validation partition, thereby preventing information leakage during MI estimation.

3) *Score Fusion and Selection*: F_j and MI_j lie on incomparable scales, each is independently rescaled to $[0, 1]$ via min-max normalization,

$$F_j^* = \frac{F_j - \min_l F_l}{\max_l F_l - \min_l F_l}, \quad MI_j^* = \frac{MI_j - \min_l MI_l}{\max_l MI_l - \min_l MI_l}, \quad (4)$$

and combined through a tunable weighted sum,

$$S_j = \alpha F_j^* + (1 - \alpha) MI_j^*, \quad \alpha \in [0, 1]. \quad (5)$$

Features are ranked by S_j in descending order, and the top- k features are retained for classification. Both the score computation (Eqs. 2–3) and the min-max normalization (Eq. 4) are performed exclusively within the training partition of every outer cross-validation fold, so that no information from the held-out validation or test partitions influences which features are selected.

D. CatBoost Classifier

CatBoost, a gradient-boosting algorithm built on symmetric decision trees and ordered boosting [23], is used as the

downstream classifier for all feature-selection conditions. Its ordered-boosting scheme mitigates the target leakage that conventional gradient boosting can introduce during training, and its native handling of categorical variables removes the need for one-hot or target encoding of the seven categorical predictors retained in this study.

Table I show hyperparameters of CatBoost are tuned using Optuna's Tree-structured Parzen Estimator (TPE) sampler [24] over 40 trials per outer fold, with the objective set to the mean F1-score across the inner cross-validation of three folds. The best-performing configuration from the inner loop is then used to refit CatBoost on the full outer-training partition, using only the features selected for that fold, before evaluation on the corresponding held-out outer test partition.

TABLE I
HYPERPARAMETER TUNING OF CATBOOST USING OPTUNA

Hyperparameter	Search Range
iterations	[100, 1000]
depth	[4, 10]
learning_rate	[0.01, 0.3]
l2_leaf_reg	[1, 10]
bagging_temperature	[0, 1]
border_count	[32, 255]

E. Nested Cross-Validation Protocol

To obtain an unbiased estimate of generalization performance while simultaneously tuning hyperparameters and selecting features, a nested cross-validation scheme is used. The outer loop consists of 5-fold stratified cross-validation, repeated 5 times with independent fold assignments (25 outer train/test partitions in total), preserving the class balance in every partition. Within each outer training partition, an inner 3-fold stratified cross-validation loop is used exclusively for Optuna's hyperparameter search (Section III-D); the inner loop never has access to the corresponding outer test partition. Feature selection (Section III-C) and continuous-variable standardization (Section III-B) are likewise refit within each outer training partition only. This nesting ensures that every stage of the pipeline that involves a data-dependent decision – scaling, feature ranking, and hyperparameter selection – is isolated from the data used for final performance evaluation, preventing optimistic bias from information leakage. Performance metrics are reported as the mean and standard deviation across all 25 outer test partitions.

IV. RESULTS AND DISCUSSION

Table II presents the top-15 features according to the normalized ANOVA F -score, information-gain score, and the proposed hybrid criterion. CRP obtained the highest ANOVA-MI score (0.8798), followed by Vitamin D (0.6649) and AST (0.2466). The ranking demonstrates that the proposed criterion does not simply reproduce the ordering obtained from either individual criterion. For example, Vitamin D achieved

TABLE II
ANOVA F -SCORE, MUTUAL INFORMATION, AND ANOVA-MI FEATURE SCORE ON TOP 15 FEATURES WHERE $\alpha = 0.3$.

Feature	F_{score}	MI_{score}	$ANOVA - MI_{\text{score}}$
C-Reactive Protein (CRP)	0.599239	0.227974	0.879772
Vitamin D	1.000000	0.119382	0.664857
Aspartate Aminotransferase (AST)	0.127866	0.068650	0.246647
Lean Mass (LM) (%)	0.372131	0.044024	0.243925
Bone Mass (BM)	0.340946	0.045617	0.239486
Total Body Fat Ratio (TBFR) (%)	0.371160	0.039757	0.230464
Extracellular Water (ECW)	0.227545	0.048356	0.213917
Hemoglobin (HGB)	0.279195	0.037022	0.194435
Hepatic Fat Accumulation (HFA)	0.056152	0.055179	0.183557
Total Fat Content (TFC)	0.206220	0.037174	0.173011
Hyperlipidemia	0.186077	0.025895	0.132156
High Density Lipoprotein (HDL)	0.178385	0.025877	0.129794
Extracellular Fluid/Total Body Water (ECF/TBW)	0.205389	0.021861	0.125501
Visceral Fat Area (VFA)	0.138769	0.023179	0.109581
Gender	0.166681	0.017060	0.099071

the highest normalized F -score, whereas several features, including ECW and HFA, exhibited comparatively stronger information-gain contributions relative to their ANOVA scores. HFA is particularly notable, with an F -score of only 0.0562 but an MI-score of 0.0552, indicating that its relevance may not be fully captured by mean-based class separation. Similarly, ECW achieved an MI-score of 0.0484 despite a moderate F -score of 0.2275. These differences indicate complementary behavior between the statistical and information-theoretic criteria. Consequently, the hybrid ranking provides a broader relevance assessment and retains features that may be under-ranked when either criterion is applied independently.

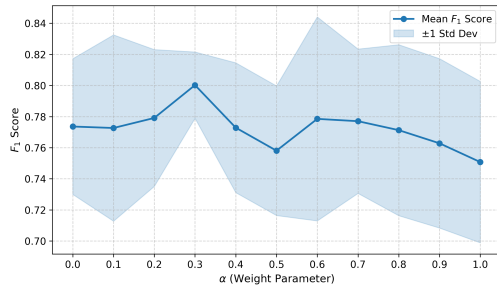


Fig. 3. Performance to the fusion weight α , showing the mean F_1 -score and ± 1 standard deviation across the evaluated configurations with 5-fold CV.

Figure 3 present the sensitivity of the proposed hybrid feature-selection approach to the fusion weight α . The results demonstrate a non-monotonic relationship between α and the resulting F_1 -score with 5-fold CV. The highest performance was obtained at $\alpha = 0.3$, achieving an F_1 -score of 0.8002 ± 0.0213 . Under the adopted formulation, this corresponds to a 30% contribution from the ANOVA criterion and a 70% contribution from the information-theoretic criterion. The MI-only configuration ($\alpha = 0$) achieved 0.7736 ± 0.0436 , whereas the ANOVA-only configuration ($\alpha = 1$) achieved 0.7508 ± 0.0519 . Thus, MI provided stronger standalone feature-selection performance than ANOVA; however, the integrated configuration at $\alpha = 0.3$ further improved the F_1 -score by 0.0266 and 0.0495 over MI-only and ANOVA-only configurations, respectively. Notably, $\alpha = 0.3$ also produced the lowest

observed standard deviation (0.0213), indicating lower performance variability under the evaluated resampling procedure. These findings suggest that the information-theoretic criterion provides the dominant relevance signal for this dataset, while a moderate ANOVA contribution supplies complementary class-discrimination information. The empirical optimum is therefore set to $\alpha = 0.3$ for subsequent feature-subset selection and CatBoost classification.

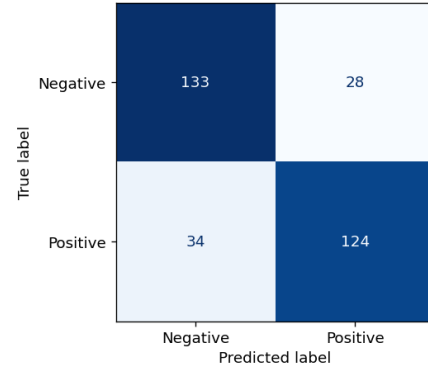


Fig. 4. Aggregated out-of-fold confusion matrix of the proposed ANOVA-MI feature-selection strategy with Nested CV of tuned CatBoost.

Figure 4 presents the aggregated out-of-fold confusion matrix obtained using the proposed ANOVA-MI feature-selection strategy followed by the tuned CatBoost classifier with 5-fold outer CV and 3-fold inner CV. Out of 319 observations, the model correctly classified 133 negative and 124 positive instances, while producing 28 false positives and 34 false negatives. This corresponds to a sensitivity of 78.48%, specificity of 82.61%, and F_1 -score of 80.00%. The relatively close sensitivity and specificity indicate balanced discrimination between the two classes, with a modestly higher ability to correctly identify negative instances. The relatively high precision and sensitivity resulted in an F_1 -score of 0.800, consistent with the mean F_1 -score obtained for the proposed method in the ablation analysis. The confusion matrix demonstrates that the proposed 15-feature ANOVA-MI representation maintains balanced classification performance while substantially reducing the original feature space. Table III shows the ablation results

TABLE III
ABLATION STUDY COMPARING CATBOOST MODELS USING THE FULL FEATURE SET AND DIFFERENT FEATURE-SELECTION STRATEGIES.

Method	Features	Accuracy	Precision	Recall	F_1 -Score	ROC-AUC
Full	38	0.803 ± 0.055	0.828 ± 0.030	0.761 ± 0.117	0.789 ± 0.069	0.860 ± 0.048
ANOVA	15	0.781 ± 0.039	0.808 ± 0.036	0.735 ± 0.093	0.766 ± 0.053	0.843 ± 0.048
MI	15	0.796 ± 0.037	0.822 ± 0.048	0.753 ± 0.039	0.786 ± 0.036	0.849 ± 0.026
ANOVA-MI (Proposed)	15	0.805 ± 0.024	0.818 ± 0.041	0.785 ± 0.038	0.800 ± 0.021	0.861 ± 0.032

by comparing full feature set ANOVA-based selection, MI-based selection, and the proposed ANOVA-MI selection at fusion weight $\alpha = 0.3$, and 5-fold CV. The proposed ANOVA-MI method uses only 15 features and corresponding to a 60.5% reduction in the feature space. Despite this substantial dimensionality reduction, its Accuracy and ROC-AUC were slightly

higher than those obtained using all 38 features, while its F1-score increased from 0.789 ± 0.069 to 0.800 ± 0.021 . Compared with ANOVA-only selection, the proposed approach improved Accuracy, Recall, F1-score, and ROC-AUC by 2.4, 5.0, 3.4, and 1.8 percentage points, respectively. Compared with MI-only selection, the corresponding improvements were 0.9, 3.2, 1.4, and 1.2 percentage points. The proposed method also exhibited the lowest observed standard deviation in F1-score, suggesting lower performance variability across the evaluated partitions. Although Precision decreased slightly compared with the full feature set, the improvement in Recall resulted in a higher overall F1-score. These findings demonstrate that combining ANOVA and MI can identify a compact subset of informative features while preserving, and in several metrics improving, the predictive performance of the full feature representation.

TABLE IV
COMPARISON OF EXISTING STUDIES AND PROPOSED FRAMEWORKS FOR GALLSTONE DISEASE.

Author	Model	Features	CV	Recall	F1-Score
Das et al. [6]	SLOA-CatBoost	19	5	0.777	0.799
Sarker et al. [5]	RF-SCSO	13	10	0.750	0.774
Shrestha et al. [13]	RoF-BFSA	17	10	0.747	0.754
Proposed	ANOVA-MI-CatBoost	15	Nested	0.785	0.800

Table IV compares the proposed ANOVA–MI–CatBoost framework with selected recent gallstone classification approaches based on the number of selected features, recall, and F1-score. The proposed method achieved the highest recall (0.785) and F1-score (0.800) among the methods included in the comparison. Compared with the SLOA-CatBoost [6], the proposed method achieved a slightly higher recall (0.785 vs. 0.777) and F1-score (0.800 vs. 0.799) while using 15 rather than 19 features. The difference in F1-score is small, indicating comparable predictive performance between the two approaches rather than a substantial performance advantage. Compared with RF-SCSO [5], the proposed method improved recall and F1-score by 3.5 and 2.6 percentage points, respectively. Similarly, compared with RoF-BFSA [13], improvements of 3.8 percentage points in recall and 4.6 percentage points in F1-score are obtained, while using fewer features (15 vs. 17). Overall, the comparison indicates that the proposed ANOVA–MI feature-selection strategy combined with CatBoost provides competitive predictive performance with a compact feature representation. In particular, its ability to achieve the highest recall and F1-score in the considered comparison while retaining only 15 of the original 38 features supports the effectiveness of combining statistical and information-theoretic feature relevance. Figure 5 illustrates the SHAP summary plot for the 15-feature ANOVA–MI–CatBoost model, identifying C-Reactive Protein (CRP) and Vitamin D as the largest SHAP distributions, indicating the strongest contributions to model predictions. Higher CRP levels correlate with a positive gallstone prediction, whereas higher Vitamin D levels drive predictions toward the negative class. ECF/TBW, VFA, Bone Mass, AST, HGB, and HFA also show

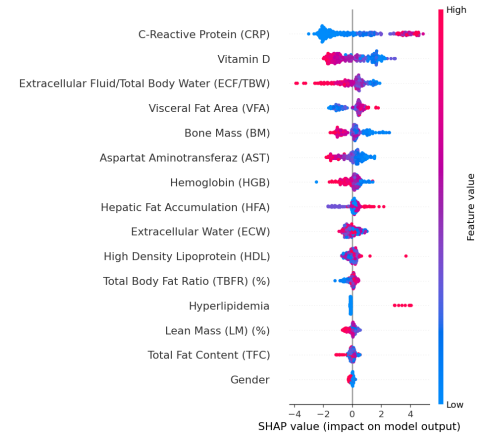


Fig. 5. SHAP plots for ANOVA-MI-CatBoost with nested cross-validation

notable contributions with varying directions, demonstrating that their effects differ across observations. In contrast, Gender has SHAP values concentrated near zero, indicating limited influence on the final predictions. Divergences between the ANOVA–MI feature ranking and SHAP importance highlight the distinction between model-agnostic univariate relevance and the nonlinear interaction effects captured by CatBoost.

V. CONCLUSION

This study proposed an ANOVA–MI feature-selection framework with CatBoost for gallstone disease prediction using 319 patients and 38 predictors. The model-independent criterion combines normalized ANOVA F -statistics and mutual information, enabling feature selection to be refit within nested repeated cross-validation without information leakage. With the optimal fusion weight $\alpha = 0.3$, feature subset size $k = 15$, and optimized MI discretization, the method reduced the 38-feature space by 60.5% while achieving an F1-score of 0.800 ± 0.021 and ROC-AUC of 0.861 ± 0.032 . The results also demonstrate that feature-selection parameters require careful optimization. Specifically, optimizing parameters such as the fusion weight (α), feature subset size (k), and the quantile bin count for MI discretization exclusively inside the inner training loop prevents data leakage and ensures unbiased performance estimates. SHAP analysis further complements the ANOVA–MI ranking by providing classifier-specific explanations, identifying CRP and vitamin D as the dominant contributors and revealing the direction and magnitude of their effects. Clinically, a compact 15-feature risk score based on routinely available laboratory and bioimpedance measurements could support triage or prioritization for ultrasonography, provided the residual false-negative rate ($\approx 21.5\%$) is carefully considered. While SHAP is used to explain the model's predictions, incorporating counterfactual explanation methods such as DiCE in future work would allow the identification of minimal, clinically modifiable changes associated with a reduced predicted gallstone risk.

REFERENCES

- [1] NIDDK, *Gallstones*, <https://www.niddk.nih.gov/health-information/digestive-diseases/gallstones>, [Accessed 11-09-2026], 2017.
- [2] Mayo Clinic, *Gallstones*, <https://www.mayoclinic.org/diseases-conditions/gallstones/symptoms-causes/syc-20354214>, [Accessed 11-09-2026], 2025.
- [3] A. S. Ahmed et al., "Advancements in cholelithiasis diagnosis: A systematic review of machine learning applications in imaging analysis," *Cureus*, vol. 16, no. 8, 2024.
- [4] İ. Esen, H. Arslan, S. A. 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.
- [5] 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.
- [6] P. Das, M. A. U. Khan, P. Sarker, and A.-A. Nahid, "Gallstone disease classification using sloa-optimized catboost classifier with explainable ai," *Plos one*, vol. 21, no. 6, e0342945, 2026.
- [7] M Indirani, S Sudheer, R Mahaveerakannan, and P Ruba, "Gallstone disease prediction using clinical and biochemical features through ensemble learning techniques," *International Journal of Computational Intelligence Systems*, vol. 19, no. 1, p. 19, 2025.
- [8] B. P. Ale and A. Khadka, "Credit card fraud detection using heterogeneous ensemble techniques," *Journal of Science and Technology*, vol. 3, no. 1, pp. 31–39, 2023.
- [9] Q. Zhong, J. Shang, Q. Ren, F. Li, C.-N. Jiao, and J.-X. Liu, "Fscme: A feature selection method combining copula correlation and maximal information coefficient by entropy weights," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 9, pp. 5638–5648, 2024.
- [10] S.-H. Kim, S. Y. Park, H. Seo, and J. Woo, "Feature selection integrating shapley values and mutual information in reinforcement learning: An application in the prediction of post-operative outcomes in patients with end-stage renal disease," *Computer Methods and Programs in Biomedicine*, vol. 257, p. 108416, 2024.
- [11] T. Klonecki and P. Teisseyre, "Feature selection under budget constraint in medical applications: Analysis of penalized empirical risk minimization methods: T. klonecki and p. teisseyre," *Applied Intelligence*, vol. 53, no. 24, pp. 29943–29973, 2023.
- [12] R. K. Khadka, A. Kayastha, S. Baniya, S. Bista, and A. Khadka, "Calorie burn prediction using XGBoost with feature selection and SHAP analysis," in *Proceedings of the International Conference on Innovation in Computing, Science, Engineering and Technology (ICICSET 2025)*, vol. 2, 2025.
- [13] K. Shrestha, P. Sarker, J.-J. Tiang, and A.-A. Nahid, "Metaheuristic-based gallstone classification using rotational forest explained with shap," *Frontiers in Digital Health*, vol. 8, p. 1727559, 2026.
- [14] E. J. Kusuma, R. Nurmandhani, Y. M. Manglapy, I. Pantiawati, and E. Widianawati, "Hybrid anova-shap approach for model-agnostic feature selection in gallstone disease prediction," *INSTEK: Jurnal Informatika dan Teknologi Komputer*, vol. 11, no. 1, pp. 119–132, 2026.
- [15] H. Wang, M. Zhang, L. Mai, X. Li, A. Bellou, and L. Wu, "An effective multi-step feature selection framework for clinical outcome prediction using electronic medical records," *BMC medical informatics and decision making*, vol. 25, no. 1, p. 84, 2025.
- [16] J. Labory, E. Njomgue-Fotso, and S. Bottini, "Benchmarking feature selection and feature extraction methods to improve the performances of machine-learning algorithms for patient classification using metabolomics biomedical data," *Computational and Structural Biotechnology Journal*, vol. 23, pp. 1274–1287, 2024.
- [17] J. Pacheco, O. Saiz, S. Casado, and S. Ubillos, "A multistart tabu search-based method for feature selection in medical applications," *Scientific Reports*, vol. 13, no. 1, p. 17140, 2023.
- [18] A. M. Vommi and T. K. Battula, "A hybrid filter-wrapper feature selection using fuzzy knn based on bonferroni mean for medical datasets classification: A covid-19 case study," *Expert Systems with Applications*, vol. 218, p. 119612, 2023.
- [19] N. Safaei et al., "E-catboost: An efficient machine learning framework for predicting icu mortality using the eicu collaborative research database," *Plos one*, vol. 17, no. 5, e0262895, 2022.
- [20] M. Hu, H. Peng, X. Zhang, L. Wang, and J. Ren, "Building gender-specific sexually transmitted infection risk prediction models using catboost algorithm and nhanes data," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, p. 24, 2024.
- [21] P. N. Srinivasu et al., "Xai-driven catboost multi-layer perceptron neural network for analyzing breast cancer," *Scientific Reports*, vol. 14, no. 1, p. 28674, 2024.
- [22] S. Gurung and A. Khadka, "Performance enhancement of breast cancer detection using adaboost ensemble based on svm and decision tree," *Journal of Science and Technology*, vol. 4, no. 2, pp. 1–6, 2024.
- [23] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems*, vol. 31, 2018, pp. 6638–6648.
- [24] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2623–2631.