Gümüşhane University Journal of Science and Technology

GUFBD / *GUJS* (2026) 16(2): 682-697 doi: 10.17714/gumusfenbil.1857964 Research Article

Minimizing unnecessary radiological imaging in gallstone diagnosis: a danets-based approach using bioimpedance and clinical data

Mustafa Hakan BOZKURT*[1,2]

[1]*Karadeniz Technical University, Faculty of Engineering, Department of AI and Data Engineering, 61080, Trabzon*

[2]*Maveria Information Technologies, Trabzon Teknokent, 61080, Trabzon*

• Received: 06.01.2026 • Accepted: 11.06.2026

Abstract

While ultrasonography and computed tomography (CT) remain the gold standards for gallstone diagnosis, their utility in primary screening is limited by high costs, ionizing radiation exposure, and acoustic attenuation in high-BMI patients. Bioelectrical impedance analysis (BIA) provides a cost-effective, non-invasive alternative; yet, traditional machine learning models often struggle to capture the high-dimensional, nonlinear metabolic relationships inherent in BIA data. This study evaluates a Deep Abstract Networks (DANets) framework, optimized for tabular data, to integrate BIA metrics with routine clinical parameters for gallstone classification. Unlike conventional tree-based or shallow architectures, the employed DANets model utilizes an abstraction layer to autonomously extract hierarchical features, mapping nonlinear correlations between hepatic biomarkers and body composition. The framework was validated on a clinical dataset of 319 patients. Rigorous methodological protocols, including 10-fold cross-validation and Bayesian hyperparameter optimization, were implemented to ensure robust evaluation. DANets achieved an accuracy of 88.08% (±3.91%) and an F1-score of 87.11%, demonstrating a statistically significant performance improvement over baseline gradient boosting (XGBoost: 83.07%) and multi-layer perceptron (MLP: 81.83%) models. Notably, the model yielded a precision of 93.27%. This rate suggests the model could help reduce the number of unnecessary radiological examinations by minimizing false positive alarms. In summary, the evaluated DANets approach serves as a promising proof-of-concept for a cost-effective clinical screening tool, warranting further external validation.

Keywords: Bioimpedance analysis, Clinical decision support system, Deep abstract networks, Gallstone detection

1. Introduction

The growth of AI applications has made AI more important and useful in medicine, especially when it comes to diagnosing diseases. Owing to improvements in machine learning and deep learning, these applications are becoming increasingly effective. These technologies are important for early disease detection and for helping doctors initiate treatment immediately. The mounting significance of artificial intelligence applications in medicine is exemplified by the early diagnosis of significant diseases, including gallstones. Gallbladder diseases present with significant clinical manifestations, including abdominal distension, fever, frequent episodes of nausea, jaundice (characterized by yellowing of the skin and sclera), and abdominal pain, typically localized to the right upper quadrant (Bozdag et al., 2025). The disease under investigation by physicians when analyzing the etiology of pain, typically located in the right upper abdomen, is facilitated by blood tests such as a complete blood count (CBC), liver function tests (LFT), and the measurement of enzymes including lipase and amylase. (Sebghatollahi et al., 2023). Imaging methods are also employed in the diagnostic process. Although ultrasonography (US) demonstrates high sensitivity, computed tomography (CT), magnetic resonance cholangiopancreatography (MRCP), or endoscopic retrograde cholangiopancreatography (ERCP) are utilized in cases of uncertainty (Sarker et al., 2025).

The utilization of imaging methodologies may not consistently generate definitive outcomes (AlAmodi et al., 2020) For instance, in cases of obesity or intestinal gas, the gallbladder may not be clearly visualized, and even if visible, stones that have migrated into the main bile duct (choledochal duct) may not be discernible (Feldman et al., 2020). In the context of computed tomography, patients are exposed to radiation. This is a highly risky

*Mustafa Hakan BOZKURT; mhakanbozkurt@ktu.edu.tr

condition for pregnant women and young patients. Administering contrast medium to individuals with kidney problems or a history of iodine allergy may be dangerous. Participants reported experiencing adverse effects, such as nausea and itching. MRCP is a very good way to get detailed pictures of the bile ducts and pancreatic ducts, but it is not very easy to get because it costs more than other imaging methods. It is essential to acknowledge that its application may be contraindicated in individuals with pacemakers or metallic prostheses. An endoscope is used for endoscopic retrograde cholangiopancreatography (ERCP). Its uses go beyond just diagnosing to include treatments as well Bioimpedance analysis (BIA) is a promising method for identifying individuals at high risk of gallstones because it is inexpensive and non-invasive. It has been shown to link changes in metabolic markers to the risk of gallstones (Bostan & Yılmaz, 2023). Bioimpedance analysis (BIA) is a cost-effective and non-invasive technique that assesses body composition by measuring the electrical resistance and reactance characteristics of bodily tissues. This quality makes it a good tool for screening and early diagnosis of gallstone risk (Branco et al., 2023).

Despite the success of imaging-based diagnostic tools, they present challenges such as high healthcare costs, radiation risks, and operator dependency. Conversely, while BIA and clinical laboratory tests offer cost-effective and non-invasive alternatives, current studies utilizing these tabular datasets are constrained by the representational limits of classical machine learning models. Therefore, this work is framed as an applied analysis study on a publicly available clinical gallstone dataset. Deep Abstract Networks (DANets) architecture is adopted to data without structural changes and compare it against strong standard baselines under a transparent evaluation protocol, with the goal of quantifying predictive performance and situating results alongside prior tabular-model reports where comparisons are meaningful.

First, we report a systematic empirical evaluation of an existing DANets implementation on a publicly available gallstone dataset that combines BIA signals with clinical laboratory markers, providing reproducible performance benchmarks under a fixed preprocessing and validation protocol

Second, we summarize comparative accuracy relative to selected prior studies cited in this paper, while noting differences in cohorts, features, and modeling pipelines. In another study by (Salem et al., 2023), an accuracy of 83% was achieved using 50 features consisting of blood and clinical risk factors, while (Esen et al., 2024) achieved a similar accuracy rate of 85.42% with fewer and different types of features (including bioimpedance) using the highest RF algorithm. In this study, a higher accuracy rate of 88.08% was obtained using DANets compared to those in the literature. Third, different types of Machine learning (ML) learning approaches (classical, tree-based, boosting and deep learning) are tested. ML and deep learning (DL) algorithms have demonstrated significant success in the automatic detection of gallstones. Within the scope of this study, the classification performances of the Multi-Layer Perceptron (MLP), Random Forest (RF), XGBoost, CatBoost, and DANets models were compared. CatBoost and DANets are included as representative gradient boosting and tabular deep learning implementations, alongside classical baselines, for comparative benchmarking.

2. Related Works

Existing research on the automated detection of gallstones can be broadly categorized into two domains: predictive modeling using tabular clinical data and image-based diagnosis using deep learning.

In the domain of tabular clinical data, studies have predominantly relied on traditional machine learning algorithms to evaluate metabolic and bioimpedance (BIA) markers. (Salem et al., 2023) identified key clinical features using the XGBoost algorithm, achieving 83% accuracy. Building on the NHANES database, (Deng et al., 2025) compared logistic regression, Naive Bayes, MLP, and SVM to predict gallstone formation based on physical indices. Utilizing an open-source BIA dataset, (Esen et al., 2024) achieved 85.42% accuracy with Gradient Boosting. On the same dataset, (Chakraborty and Mukherjee, 2025) applied a hybrid Adaptive LASSO and BART model, while (Sarker et al., 2025) utilized an optimized RF classifier with Sand Cat Swarm Optimization (SCSO), achieving an F1-score of 79.07%.

Parallel to these developments, significant algorithmic advancements have occurred in medical imaging for gallbladder diseases, which are heavily driven by deep learning. (Ye et al., 2026) developed a multimodal framework using EfficientNet-B3 and attention mechanisms for CT images. (Bozdag et al., 2025) utilized Content-Based Image Retrieval (CBIR) systems with CNN models like InceptionV3 to detect gallbladder diseases from ultrasound images. High-performance CNNs, such as the MSFE-GallNet-X proposed by (Nabil

et al., 2025), have achieved 99.63% accuracy on ultrasound datasets. Furthermore, advanced approaches like (Basu et al., 2024) and (Hasan et al., 2025) have introduced masked autoencoders for video streams and vision transformers for spatial correlation. The comparison of the recent studies are given in Table 1.

Despite these rapid advancements in imaging, a critical methodological gap remains. The most sophisticated deep learning architectures are almost exclusively tailored to computer vision tasks. In contrast, research utilizing structured, high-dimensional clinical and BIA datasets remains heavily confined to traditional tree-based algorithms and shallow neural networks. Rather than proposing a new model family, we benchmark established DANets against classical, tree-based, boosting, and MLP baselines using a unified preprocessing and validation pipeline on this dataset.

Table 1. Comparison of recent studies on gallstone detection

Study	Modality	Methodology	Key Metrics
Salem et al. (2023)	Tabular	XGBoost	Accuracy: 83.0%
Esen et al. (2024)	Tabular	Gradient Boosting	Accuracy: 85.42%
Sarker et al. (2025)	Tabular	RF + SCSO	F1-Score: 79.07%
Nabil et al. (2025)	Ultrasound	MSFE-GallNet-X (CNN)	Accuracy: 99.63%
Bozdag et al. (2025)	Ultrasound	CNN (InceptionV3)	AP: 0.94
Ye et al. (2026)	CT + Clinical	EfficientNet-B3	AUC: 0.96
This Study	Tabular (BIA + Clinical)	DANets	Accuracy: 88.08%, Precision: 93.27%

3. Material and method

The methodology of this study is based on the development, optimization and comparative evaluation of machine learning models aimed at predicting the presence of gallstones. The study utilized a specialized clinical dataset comprising 39 features, including demographic data, comorbidities, anthropometric measurements, biochemical parameters, and body composition analyses from 319 patients. The target variable was defined as a binary classification (present/absent), indicating the patients' gallstone status.

3.1. Data set definition and preprocessing

This study utilized the publicly available gallstone dataset (Esen et al., 2024), which comprises 39 features across 319 patient samples. These features encompass demographic and clinical history (e.g., age, gender, comorbidities), detailed bioimpedance and body composition metrics (e.g., body mass index, total body water, and visceral fat area), and comprehensive biochemical laboratory results (e.g., liver enzymes, lipid profiles, C-reactive protein, and Vitamin D). The primary objective is to classify gallstone status as either healthy or unhealthy. The dataset features a naturally balanced class distribution of 161 healthy and 158 unhealthy samples. Because the classes are inherently balanced, complex resampling or data balancing preprocessing strategies were not required. Although there were no missing data, a preprocessing pipeline was established for the experiments. The median value was used to impute the missing values for continuous numerical features. The most frequently occurring value (mode) was used for binary and categorical features. This approach preserves the original distribution of categories without creating synthetic or invalid fractional values. Following the imputation process, the numerical features were standardized to ensure consistency across the models.

A descriptive summary of all features, including the means and standard deviations for continuous variables and frequency counts for binary features, is presented in Table 2. This table is grouped by class labels to illustrate the clinical and biochemical distributions of the dataset. Upon examining the data, we observed that there were no missing values in any of the features or samples. The missing value rate is 0.0% for each feature. Although median and most frequent value imputation was retained as a precaution during the preprocessing stage, imputation is not necessary.

Data preparation was carried out systematically and reusably with the help of a continuously operating modular automation pipeline that underwent preprocessing stages to maintain consistency. This pipeline implements custom transformations for various data types, including numerical, categorical, Boolean, and date-time. For the categorical variables, the values that were not present were filled by the most frequently occurring value, and One-Hot Encoding was used.

Table 2. Descriptive statistics and missing data percentages for all variables in the Gallstone dataset. Continuous features are reported as Mean ± SD; binary/ordinal features as count and proportion. 0 = Healthy, 1 = Gallstone.

Feature	Mean ± SD / n (%)	Missing (%)
Gallstone Status	n = 158 (49.5%)	0.0%
Age	48.07 ± 12.11	0.0%
Gender	n = 157 (49.2%)	0.0%
Comorbidity	n = 107 (33.5%)	0.0%
Coronary Artery Disease (CAD)	n = 12 (3.8%)	0.0%
Hypothyroidism	n = 9 (2.8%)	0.0%
Hyperlipidemia	n = 8 (2.5%)	0.0%
Diabetes Mellitus (DM)	n = 43 (13.5%)	0.0%
Height	167.16 ± 10.05	0.0%
Weight	80.56 ± 15.71	0.0%
Body Mass Index (BMI)	28.88 ± 5.31	0.0%
Total Body Water (TBW)	40.59 ± 7.93	0.0%
Extracellular Water (ECW)	17.07 ± 3.16	0.0%
Intracellular Water (ICW)	23.63 ± 5.35	0.0%
Extracellular Fluid/Total Body Water (ECF/TBW)	42.21 ± 3.24	0.0%
Total Body Fat Ratio (TBFR) (%)	28.27 ± 8.44	0.0%
Lean Mass (LM) (%)	71.64 ± 8.44	0.0%
Body Protein Content (Protein) (%)	15.94 ± 2.33	0.0%
Visceral Fat Rating (VFR)	9.08 ± 4.33	0.0%
Bone Mass (BM)	2.80 ± 0.51	0.0%
Muscle Mass (MM)	54.27 ± 10.60	0.0%
Obesity (%)	35.85 ± 109.80	0.0%
Total Fat Content (TFC)	23.49 ± 9.61	0.0%
Visceral Fat Area (VFA)	12.17 ± 5.26	0.0%
Visceral Muscle Area (VMA) (Kg)	30.40 ± 4.46	0.0%
Hepatic Fat Accumulation (HFA)	1.15 ± 1.06	0.0%
Glucose	108.69 ± 44.85	0.0%
Total Cholesterol (TC)	203.50 ± 45.76	0.0%
Low Density Lipoprotein (LDL)	126.65 ± 38.54	0.0%
High Density Lipoprotein (HDL)	49.48 ± 17.72	0.0%
Triglyceride	144.50 ± 97.90	0.0%
Aspartate Aminotransferase	21.68 ± 16.70	0.0%
Alanine Aminotransferase	26.86 ± 27.88	0.0%
Alkaline Phosphatase (ALP)	73.11 ± 24.18	0.0%
Creatinine	0.80 ± 0.18	0.0%
Glomerular Filtration Rate (GFR)	100.82 ± 16.97	0.0%
C-Reactive Protein (CRP)	1.85 ± 4.99	0.0%
Hemoglobin (HGB)	14.42 ± 1.78	0.0%
Vitamin D	21.40 ± 9.98	0.0%

The transformations were arranged in the column transformer` structure of the scikit-learn library and explicitly assigned to the training data in every fold of the cross-validation procedure to avoid data leakage.

During the model creation and comparison phase, a total of five machine learning algorithms from different families were used: Deep Abstract Networks (DANets), which are deep learning networks; XGBoost and CatBoost (Shyam et al., 2020), which are gradient boosting models; Random Forest, which is an ensemble learning technique; and Multi-Layer Perceptron (MLP), which is an artificial neural network technique. The performance of these models was evaluated using stratified k-fold cross-validation and a holdout test set. While the algorithms were initially compared using 5-fold cross-validation, 10-fold cross-validation and a test set containing 30% of the data were used during the optimization and final evaluation stages to obtain a more reliable result.

3.1.1. Methodologic design

The experimental design consists of three independent evaluation protocols, each with a distinct purpose.

- Stage 1: Baseline 5-Fold Cross-Validation
- Stage 2: Holdout Evaluation
- Stage 3: Hyperparameter Optimization via 10-Fold Grid Search

The research flow is shown in Figure 1.

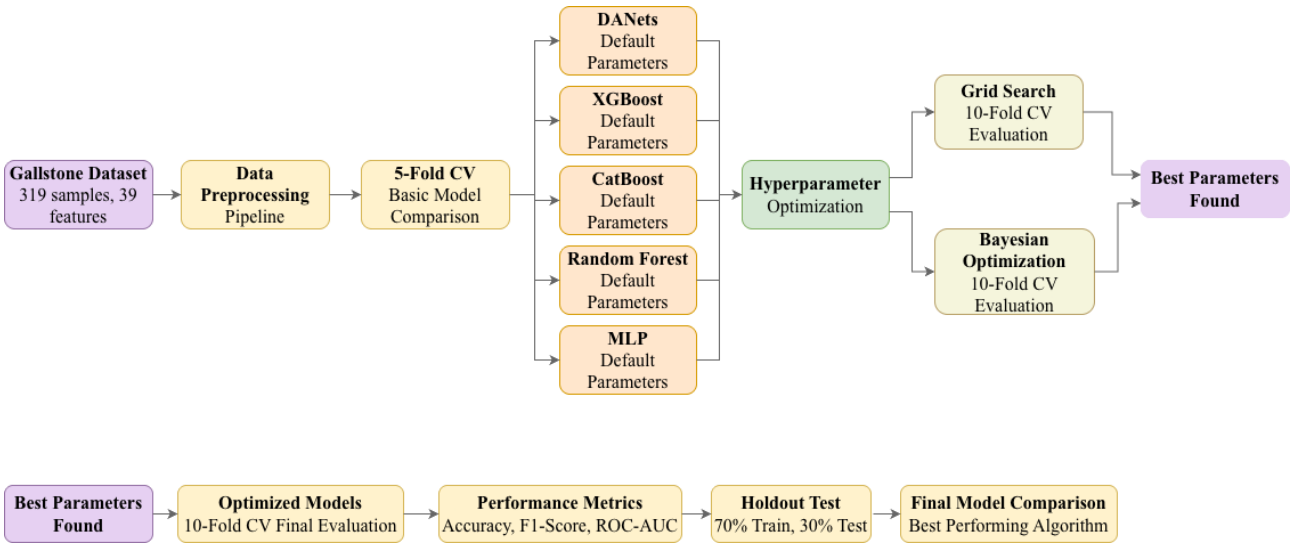

Figure 1. Research flow of the paper

In the baseline 5-fold cross-validation experiment, all five models were first evaluated using 5-fold stratified cross-validation with their default (baseline) hyperparameter configurations. This stage provides an estimate of the model stability and generalization under uniform conditions without any data-driven tuning.

In the second experiment, the full dataset was split once into a training set (70%) and a test set (30%) using stratified sampling (random_state = 42). The models were trained on the training partition using their default hyperparameters and evaluated on the held-out test partition. The preprocessing pipeline (imputation and scaling) was fitted exclusively on the training partition and applied via transform-only to the test partition, ensuring that no information from the test set influenced any parameter in the pipeline. This stage provides a single-split, independent performance estimate under baseline conditions.

In 3rd experiment, a grid search was conducted on the full dataset using 10-fold stratified cross-validation to identify the best hyperparameter configuration for each model. Within each CV fold, the preprocessing was fitted solely on the fold's training indices. The combination yielding the highest mean accuracy was then re-evaluated using a final 10-fold CV to report stable performance estimates. This stage reports the best achievable performance under optimized conditions, as estimated by cross-validation.

In this experiment, a hyperparameter grid search was conducted independently for each model using 10-fold stratified cross-validation, with mean accuracy as the optimization criterion. All combinations were exhaustively evaluated using (full grid search). For DANets, the search covered the number of ABSTLAY layers (8, 10, 12), base output dimension (64, 128, 256), number of feature groups per layer k (3, 5, 7), dropout rate (0.1, 0.2, 0.3), maximum number of training epochs (100, 150, 200), and early stopping patience (30, 40, 50), yielding 729 combinations in total; the batch size was fixed at 256. For XGBoost, the search space comprised the number of estimators (200, 400, 600), tree depth (3, 5, 7), learning rate (0.01, 0.05, 0.1), subsample ratio (0.7, 0.8, 0.9), column subsampling ratio (0.7, 0.8, 0.9), L2 regularization lambda (0, 1, 5), and L1 regularization alpha (0, 0.1, 1), totalling 2187 combinations. CatBoost was tuned over the number of iterations (400, 600, 800), tree depth (4, 6, 8), learning rate (0.01, 0.05, 0.1), and L2 leaf regularization (1, 3, 5), yielding 81 combinations. For RF, the grid covered the number of trees (100, 300, 500), maximum depth (10, 20, 30), minimum samples to split a node (2, 5, 10), minimum samples per leaf (1, 2, 4), and feature-selection criterion (sqrt, log2, None), resulting in 243 combinations. Finally, for the MLP, eight hyperparameters were tuned: hidden layer architecture ((64,), (128,), (64, 32)), activation function (ReLU, tanh, logistic), solver (Adam, SGD, L-BFGS), L2 regularization strength (1e-4, 1e-3, 1e-2), batch size (32, 64,

128), learning rate schedule (constant, adaptive, invscaling), initial learning rate (1e-4, 1e-3, 1e-2), and maximum iterations (200, 300, 500), totalling 6561 combinations. The optimal hyperparameter configuration identified for each model is presented in Table 4.

Stages 2 (holdout) and 3 (grid search) are independent: the optimized hyperparameters from Stage 3 are not applied to the holdout evaluation in Stage 2. Consequently, the holdout test results represent a baseline performance estimate, whereas the grid search results represent the upper-bound performance achievable with tuned configurations. All three stages shared the same random seed (random_state = 42) and stratified splitting to ensure reproducibility. We additionally note that extensive hyperparameter search on a single finite sample—even when preprocessing is fitted strictly within cross-validation folds—can still induce optimistic bias in the reported cross-validated performance relative to strictly prospective or externally locked testing, because many model configurations are evaluated on the same underlying observations. Accordingly, Stage 3 results should be interpreted as performance under an explicit, fully specified search protocol (with search spaces reported in the manuscript), complementary to the more conservative holdout estimate obtained without applying Stage-3-tuned hyperparameters to the holdout split.

Within the scope of the study, five different models were evaluated: MLP, Random Forest, Catboost, XGBoost, and DANets. This allowed both traditional methods, such as MLP and ensemble-based methods, to be tested. These models are methods that perform well in the literature. DANets is a modern approach based on automatic feature grouping logic using an abstract layer. Details of the methods are provided in the following sections.

3.2. Multilayer perceptron (MLP)

An MLP is a classical artificial neural network (ANN) architecture that performs nonlinear transformations through one or more hidden layers (Singh & Banerjee, 2019). The model is defined for each neuron as in Equation 1.

$$y = f(\sum_{i=1}^{n} w_i x_i + b) \tag{1}$$

Here x_i: entry features, w_i: weights, b: bias term, f: activation function like ReLU, Sigmoid, Tanh, etc. In gallstone detection, MLP can process statistical features extracted from laboratory values, patient demographic information, and ultrasonography images. The parameters were set as follows: hidden layer sizes are 100, activation is ReLU, Adam Optimizer, alpha is 0.0001, constant learning rate, and max iteration number is 200.

3.3. Random forest (RF)

RF is an ensemble method created by training multiple decision trees using random samples and feature subsets (Han et al., 2018). The prediction for each tree is defined as in Equation 2.

$$\acute{y} = \frac{1}{N}\sum_{i=1}^{N} T_i(x) \tag{2}$$

Here $T_i(x)$ the estimate for each tree, NIt is the total number of trees. This method was chosen because it is robust against overfitting and performs well on small to medium-sized clinical datasets. The parameters were set as follows: number of estimators are 100, criterion is Gini, max depth is None, minimum samples split is 2, minimum samples leaf is 1, and bootstrap is True.

3.4. Extreme gradient boosting (XGBoost)

XGBoost is constructed by sequentially training weak learners (decision trees) in a gradient-based manner (Chen & Guestrin, 2016). The objective function is specified in Equation 3.

$$Obj(\theta) = \sum_i l(y_i + \acute{y}_i^{(t)}) + \sum_k \Omega(f_k) \tag{3}$$

The term "adjustment" is defined in Equation 4.

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2 \tag{4}$$

Here, leaf number is the weight regularization coefficient. The parameter values were set as follows: number of estimators are 100, learning rate is 0.1, max depth is 6, subsample is 1, colsample by tree is 1, gamma is 0, and reg lambda is 1.

3.5. Categorical boosting (CatBoost)

CatBoost is a gradient boosting algorithm that performs particularly well on datasets with a high concentration of categorical variables (e.g., patient gender, age group, dietary habits, etc.). Unlike other boosting methods, it uses Target Statistics (Ordered Target Encoding) and Oblivious Tree structure (Prokhorenkova et al., 2018). In each iteration, the tree function f(x) that minimizes the loss function L is added as in Equation 5.

$$F_t(x) = F_{t-1}(x) + \eta f_t(x) \tag{5}$$

Here η learning rate. CatBoost applies Ordered Target Encoding for categorical features. In the applied ~~formula~~ Equation 6, the target value of previous observations, the smoothing parameter, represents the current category frequency.

$$encoding(x) = \frac{\sum_{j<i} y_j + aP}{n_i + a} \tag{6}$$

Oblivious Tree (Symmetric Tree) uses the same features and threshold values at every level. This structure reduces the computational cost of the model and prevents overfitting.

The parameter value assignments were implemented as follows: iterations are 1000, learning rate is 0.03, depth is 6, l2 leaf reg is 3, loss function is Logloss, border count is 254, bootstrap type is Bayesian, and eval metric is Accuracy. CatBoost provides high accuracy by processing laboratory findings (e.g., bilirubin, ALT, AST, alkaline phosphatase), demographic variables, and image-based features together. Additionally, in case the data is imbalanced (e.g., few positive cases), balancing can be performed with the class weights parameter.

3.6. Deep abstract networks (DANets)

It is a deep learning-based algorithm specifically designed for table data (Chen et al., 2022). The Abstract Layer (ABSTLAY), a core component of the algorithm, automatically groups related features to produce higher-level abstract features. The ABSTLAY layer consists of three main steps: feature selection, feature abstraction, and output merging. During the feature selection stage, a learnable mask M is used for an input vector $f \in R^m$, and this process is defined by Equation 7. The selected features are then abstracted to obtain the representation of Equations 8 and 9. When multiple groups are processed in parallel, the final output is calculated as shown in Equation 10.

$$M = entmax_\alpha(W_{mask}), f' = M \odot f \tag{7}$$

$$q = \sigma(BN(W1f')) \tag{8}$$

$$f^* = ReLU(q \odot BN(W_2 f')) \tag{9}$$

$$F_o = \sum_{k=1}^{K} p_k \, o \, s_k(f) \tag{10}$$

Here, k denotes the number of grouped feature sets. Fundamentally, DANets enables hierarchical abstraction of features in the table by stacking ABSTLAY layers. Additionally, shortcut paths were added to the model to enable direct information transfer from raw features, which increased the diversity of information in the deeper layers of the model.

The complete forward pass of DANets proceeds as follows. The raw input feature vector f ∈ $\mathbb{R}^d$ is simultaneously fed into each ABSTLAY and preserved via shortcut (residual) connection. In each ABSTLAY, the learnable soft mask M performs feature selection by assigning per-feature importance scores (Equation 7). The k selected feature groups are each projected through a fully-connected layer followed by a ReLU activation (Eqs. 8–9) to produce k abstract representations. These are concatenated and added to the shortcut-connected

raw features to form the layer output (Equation 10). By stacking eight such layers, the network builds progressively higher-order abstractions of the clinical and bioimpedance input features.

DANets was selected for this study over alternative tabular deep learning methods such as TabNet (Arik & Pfister, 2021) and NODE (Popov et al., 2020) because its parallel multi-group feature abstraction mechanism is particularly well-suited to the structure of the Gallstone dataset. The 39 features contain strongly correlated physiological subgroups; BIA body-composition indices (TBW, ECW, ICW, muscle mass, fat mass) and liver/metabolic function markers (ALT, AST, ALP, cholesterol, triglycerides); whose inter-feature interactions the ABSTLAY mechanism can discover and exploit explicitly, without requiring hard feature splits as in tree-based models or treating all features independently as in standard MLPs.

In this study, we adopted the original DANet architecture introduced by (Chen et al., 2022) without any structural modifications. Accordingly, the emphasis is on systematic hyperparameter search and comparison against XGBoost, CatBoost, RF, and MLP on the same public dataset, using a transparent three-stage protocol (baseline CV, holdout test, and 10-fold grid-search CV) to support reproducible performance reporting.

In medical applications such as gallstone data classification, DANets can automatically learn the complex relationships between biochemical and clinical indicators. For example, correlations between liver function tests (ALT, AST) and bile duct indicators (bilirubin, ALP) can be grouped using mask weights to create high-level representations of stone presence or type. This helps the model to achieve a deeper semantic representation compared to classical tree-based methods, thereby improving diagnostic accuracy. The baseline configuration used in the 5-fold cross-validation and holdout evaluations is: num_layers=8, base_outdim=64, k=5 (feature groups per ABSTLAY), drop_rate=0.1, max_epochs=120, patience=30 (early stopping). The optimal configuration identified via grid search (Stage 3) is: layer=12, base_outdim=256, k=3, drop_rate=0.1, as reported in Table 4.

To ensure full reproducibility, the complete training configuration is as follows. The optimizer used is QHAdam (quasi-hyperbolic momentum Adam) with a learning rate of lr=0.008, weight_decay=1e-5, and quasi-hyperbolic parameters ν_1=0.8 and ν_2=1.0. The learning rate is decayed via a StepLR scheduler (γ=0.95, step_size=20). The batch size is 256 and the virtual batch size is 256. The loss function is Binary Cross-Entropy. Training is performed with CUDA acceleration on an NVIDIA RTX 4090 GPU; all other models are evaluated on CPU. A fixed random seed of 42 is applied to all data splitting, model initialisation, and cross-validation fold generation to guarantee deterministic results. The grid search explored the following hyperparameter space: layer ∈ {8, 10, 12}, base_outdim ∈ {64, 128, 256}, k ∈ {3, 5, 7}, drop_rate ∈ {0.1, 0.2, 0.3}, max_epochs ∈ {100, 150, 200}, patience ∈ {30, 40, 50}, and batch_size=256 fixed.

A comprehensive hyperparameter optimization process was applied to maximize the predictive power of the models. In this process, the Grid Search method, which scans a wide search space, was first used, followed by the Bayesian Optimization technique (using the Hyperopt library), which provides more efficient search. As an optimization goal, performance criteria consisting of the average Accuracy, F1-Score, and ROC-AUC metrics were used to provide a comprehensive evaluation of the classification performance. This approach targets a balanced and robust model by jointly optimizing classification accuracy, sensitivity to class imbalance, and discriminative ability.

To ensure the validity and reliability of the study's results, the final performance of the models was evaluated using the Accuracy, F1-Score, and Area Under the Curve (ROC-AUC) metrics. The mean and standard deviation values of the cross-validation results were reported for each metric. To ensure the reproducibility of the study, a fixed initial value (random seed=42) was used in all experiments, stratified sampling was adopted to preserve class distributions during data set splitting, and preprocessing steps were applied independently for each cross-validation fold. This rigorous experimental protocol strengthens the objectivity and scientific validity of the findings.

4. Results

In this section it is presented analyses of the performance evaluations of models developed on the gallbladder disease dataset. First, the dataset and experimental setup are detailed, followed by a comparison of model performances using various metrics.

4.1. Experimental setup

Three different types of experiments were conducted to provide a comprehensive evaluation and fair comparison:

- 5-Fold Cross-Validation is used to evaluate the initial stability of the models.
- Parameter Selection (grid search) with 10-fold Cross-Validation. It has been applied for hyperparameter optimization and final performance evaluation.
- Holdout Test. The dataset has been divided into 70% training and 30% testing using the stratified sampling method.

These three experimental stages are independent of one another: Stage 1 (5-fold CV) and Stage 3 (holdout) use default hyperparameters and are not influenced by the hyperparameter search conducted in Stage 2 (grid search). The holdout test set is defined solely by the 70/30 stratified split and is never used as a training or validation partition in Stage 2. In all stages, preprocessing transformers (StandardScaler, OneHotEncoder, SimpleImputer) are fitted exclusively on the training indices of each fold or split, and only transform operations are applied to the corresponding test or validation indices.

All experiments were conducted using the same random seed value (random seed = 42). This was done to ensure reproducibility. The training of the DANets model is performed with CUDA acceleration on an NVIDIA RTX 4090 GPU, while other models are run on the CPU.

4.2. Cross-validation results

Five-fold cross-validation was applied to measure the stability and generalization ability of the models. Table 3 shows the mean and standard deviation values of the performance metrics across all folds.

Table 3. 5-Fold Cross-validation results

Model	Accuracy	F1-Score	Precision	Sensitivity	Specifity	ROC-AUC
DANets	79.64%±3.34%	77.69%±3.50%	85.30%±5.41%	71.53%±3.91%	83.33%	83.38%±4.13%
XGBoost	79.33%±3.54%	78.52%±4.00%	80.89%±4.00%	76.65%±6.54%	83.33%	85.61%±3.80%
CatBoost	79.32%±3.59%	77.82%±4.85%	82.50%±2.35%	74.11%±8.08%	83.33%	86.21%±3.17%
RF	79.31%±2.34%	78.15%±3.31%	81.58%±1.57%	75.34%±6.60%	79.17%	84.74%±3.74%
MLP	76.13%±9.45%	75.64%±10.35%	75.78%±9.27%	75.89%±12.67%	72.92%	80.40%±8.40%

4.3. Grid search optimization results

For comprehensive hyperparameter optimization, 10-fold cross-validation was performed on all models using Grid Search. Table 4 summarizes the best performance achieved at the end of this optimization process and the best hyperparameter sets that provided this performance.

Table 4. Optimized model performance (10-fold cross-validation with grid search)

Model	Accuracy	F1-Score	Precision	Sensitivity	Specificity	ROC-AUC	Best Parameters
DANets	88.08%± 3.91%	87.11%± 4.53%	93.27%± 5.00%	82.21%± 7.50%	93.79%	87.87%± 6.97%	layer=12, base_outdim=256, k=3, drop_rate=0.1, max_epochs=100, patience=40
XGBoost	83.07%± 5.44%	82.64%± 5.69%	84.80%± 9.05%	81.79%± 9.35%	84.47%	87.72%± 5.86%	n_estimators=600, max_depth=7, learning_rate=0.1, subsample=0.7
MLP	81.83%± 6.34%	80.55%± 6.96%	86.01%± 9.10%	76.75%± 10.33%	85.09%	88.84%± 7.23%	hidden_layer_sizes=(64,), activation='logistic', solver='adam', alpha=0.001
CatBoost	80.57%± 6.35%	79.96%± 6.97%	82.34%±9.21%	79.29%± 11.77%	81.99%	85.95%± 5.01%	iterations=400, depth=8, learning_rate=0.1, l2_leaf_reg=3
RF	79.62%± 5.99%	79.09%± 6.77%	80.39%±7.60%	78.58%± 9.67%	80.75%	86.55%± 5.50%	n_estimators=500, max_depth=20, min_samples_split=2, min_samples_leaf=2.

4.4. Deeper analysis

The ranking of models based on accuracy metrics according to 10-fold cross-validation results using grid search optimization is as follows:

1. **DANets**: 88.08%±3.91% (The Best)
2. **XGBoost**: 83.07%±5.44%
3. **MLP**: 81.83%±6.34%
4. **CatBoost**: 80.57%±6.35%
5. **RF**: 79.62%±5.99%

The results demonstrate that the DANets model has both the highest accuracy and the lowest standard deviation, indicating both the model's better performance and high stability. In particular, the precision value of 93.27% is of great importance for medical diagnostic applications where minimizing false positives is critical.

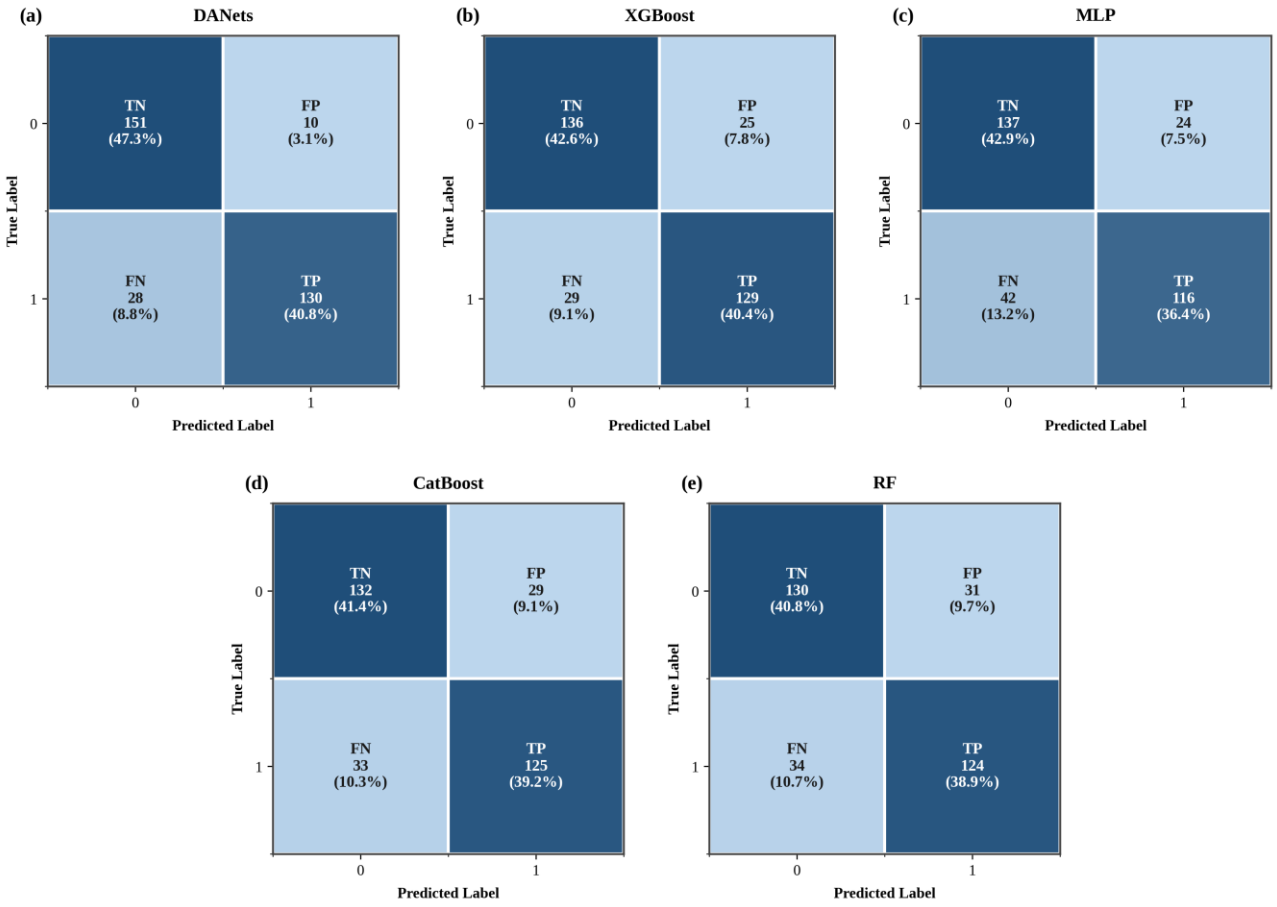

Figure 2. Confusion matrixes of the models **(a)** DANets, **(b)** XGBoost, **(c)** MLP, **(d)** CatBoost, and **(e)** RF

The ROC (Receiver Operating Characteristic) curves and AUC (Area Under the Curve) values of the models also support these findings. DANets achieved an AUC value of 87.87% ± 6.97%. Among all models, MLP achieved the highest ROC-AUC (88.84% ± 7.23%), while DANets led in Accuracy, F1-Score, and Precision. Figure 2 and 3 shows a comparison of performance metrics for all models based on 10-fold cross-validation results, and Figure ~~7~~ 4 shows the ROC curves for the same experimental setup.

The classification errors across the tested models were evaluated using confusion matrices. Figure 2 illustrates the distribution of true and false predictions for DANets, XGBoost, MLP, CatBoost, and RF architecture in order.

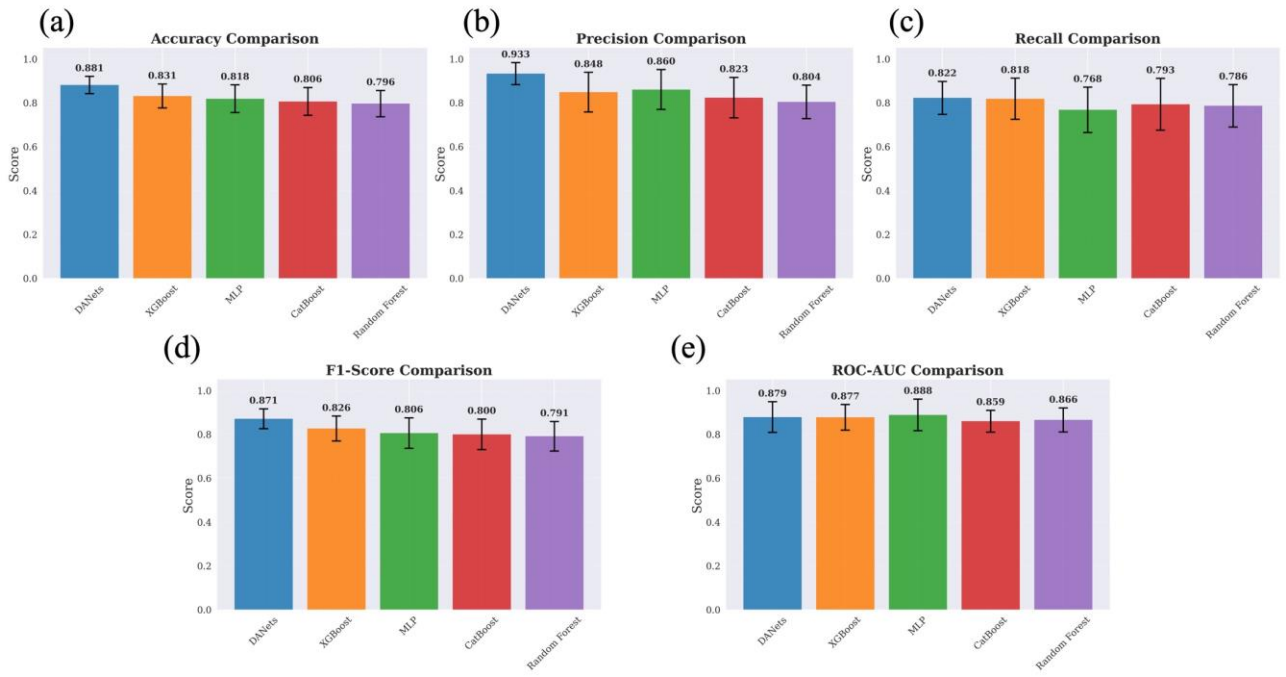

Figure 3. (a) Accuracy comparison of the models, **(b)** Precision comparison of the models, **(c)** Recall comparison of the models, **(d)** F1-Score comparison of the models, **(e)** ROC-AUC comparison of the models

Figure 3 shows the comparison of the models in terms of accuracy, F1-Score, precision, recall, and ROC-AUC metrics, respectively, based on the results of 10-fold cross-validation. The graph shows that the DANets model achieves the highest Accuracy, F1-Score, and Precision among all models. In terms of ROC-AUC, MLP achieves a marginally higher value (88.84% vs. 87.87% for DANets); thus, DANets and MLP are comparable on this metric. In particular, the high values observed for the precision and accuracy metrics indicate that the model provides reliable results by minimizing false positives. The low standard deviation values support the consistency and stability of DANets' results. Other models generally fall within similar performance ranges; however, it has been observed that RF and MLP models, in particular, exhibit lower performance in sensitivity and F1-Score metrics.

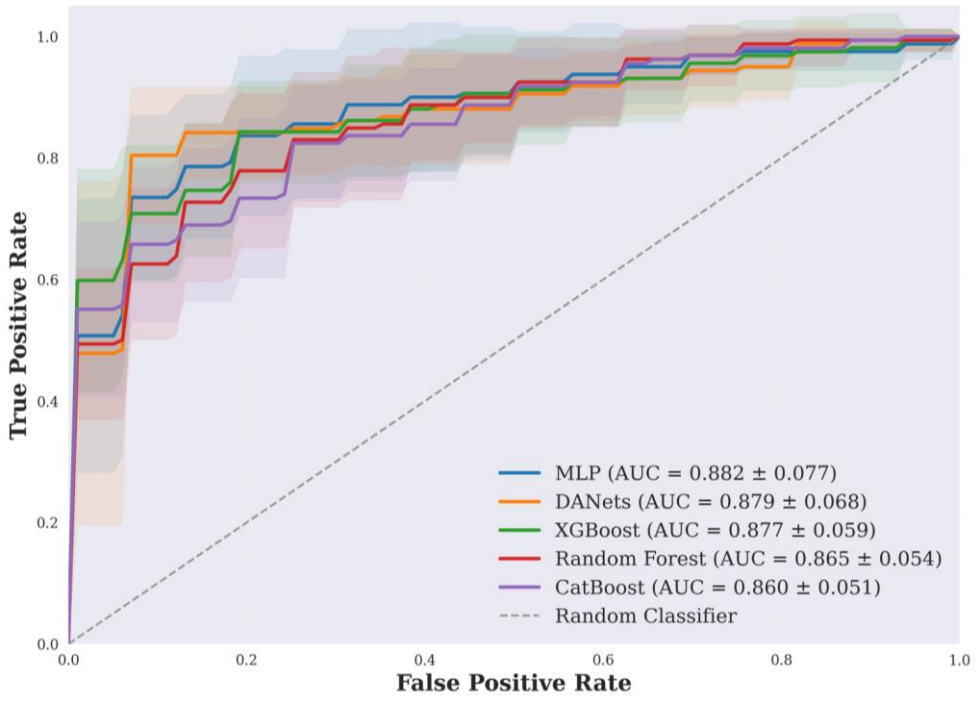

Figure 4. ROC curves of the models

Figure 4 presents the ROC curves and corresponding AUC values for five different models. The DANets model achieved an AUC of 0.879, which is comparable to those of XGBoost (0.877) and MLP (0.888). While MLP achieved the highest ROC-AUC in the 10-fold grid search evaluation, DANets demonstrated superior performance in terms of Accuracy (88.08%), F1-Score (87.11%), and Precision (93.27%), which are more clinically relevant metrics for minimising false positives in gallstone screening. The XGBoost and CatBoost models also exhibited competitive performance with high AUC values. The ROC curves visualize the sensitivity and specificity balances of the models at different threshold values, providing important clues as to which model can make more reliable decisions in clinical applications. The shapes and areas of the curves support the overall classification success of the models.

4.5. Model performance validation on holdout test

Table 5 shows the performance of all models on the holdout test set (30% of the data). The DANets model outperformed all other models by a significant margin, with an accuracy rate of 85.42%.

Table 5 Holdout test results

Model	Accuracy	F1-Score	Precision	Sensitivity	ROC-AUC
DANets	85.42	85.11	86.96	83.33	88.02
XGBoost	84.38	84.54	83.67	85.42	91.67
CatBoost	82.29	82.47	81.63	83.33	90.58
RF	81.25	81.63	80.00	83.33	89.45
MLP	80.21	80.00	80.85	79.17	87.11

4.6. Statistical Significance Analysis

To verify that the observed performance differences are not attributable to sampling variability, three non-parametric statistical tests were applied to the per-fold results of the 10-fold cross-validation (n=10 paired observations per model).
The Friedman test, a non-parametric repeated-measures test, revealed that the five models differ significantly in both Accuracy (χ^2=17.74, p=0.001) and F1-Score (χ^2=10.57, p=0.032), indicating that at least one model performs systematically differently from the others across folds.

Pairwise post-hoc comparisons were then conducted using the Wilcoxon signed-rank test (one-tailed, α=0.05) between DANets and each competing model. As shown in Table 6, DANets achieves statistically significantly higher Accuracy and F1-Score than all four alternatives. All pairwise p-values are below 0.05 for both metrics.

Bootstrap resampling (10,000 iterations, random_state=42) was used to construct 95% confidence intervals for each model's mean Accuracy and F1-Score. The lower bound of DANets' Accuracy CI (0.856) exceeds the upper bound of CatBoost (0.846) and RF (0.831), confirming that the advantage is robust. For ROC-AUC, no statistically significant pairwise difference was detected (all p > 0.09), as MLP achieved a comparable mean ROC-AUC (0.888 vs. DANets 0.879), and this metric alone should not be used to draw conclusions of superiority.

Table 6 Holdout test results

Comparison	Accuracy p	F1 p	ROC-AUC p	Significant?
DANets vs XGBoost	0.0078	0.0322	0.338	✓ (Acc, F1)
DANets vs MLP	0.0156	0.0186	0.557	✓ (Acc, F1)
DANets vs CatBoost	0.0020	0.0049	0.097	✓ (Acc, F1)
DANets vs RF	0.0020	0.0029	0.125	✓ (Acc, F1)

4.7. Literature Comparison

To establish a comparative baseline, the performance of the DANets model was evaluated against recent literature using the same clinical and bioimpedance dataset. Table 7 presents a direct comparison of these studies based on a shared cohort of 319 patients. Previous benchmarks on this dataset employed various

machine learning and hybrid techniques. A baseline accuracy of 85.42% was achieved using Gradient Boosting (Esen et al., 2024). Subsequent studies applying metaheuristic optimization and hybrid frameworks reported F1 scores of 77.44% and 85.29%, respectively (Sarker et al., 2025; Chakraborty and Mukherjee, 2025).

Table 7. Literature Comparison of the models

Study	Methodology	Accuracy	F1-Score	Precision
Esen et al. (2024)	Gradient Boosting	85.42%	86.00%	91.00%
Sarker et al. (2025)	RF-SCSO (with CV)	78.32%	77.44%	80.36%
Chakraborty & Mukherjee (2025)	Adaptive LASSO + BART	84.60%	85.29%	82.56%
This Study	**DANets**	**88.08%**	**87.11%**	**93.27%**

The results indicate that the present study achieves higher scores across the evaluated metrics compared to these existing approaches. The application of the DANets architecture yields an overall accuracy of 88.08% and an F1 score of 87.11%. In addition, the model achieves a precision of 93.27%. This specific metric is critical for the proposed clinical screening tool, as it suggests an improved capability to minimize false-positive predictions compared with the evaluated traditional algorithms. This comparative performance implies that utilizing deep abstraction layers can be advantageous for processing complex tabular medical data.

5. Discussion

In this study, the DANets (Deep Abstract Networks) framework was systematically evaluated through a domain-specific application integrating bioimpedance analysis (BIA) and clinical parameters for gallstone classification. The findings reveal that the DANets (Deep Abstract Networks) architecture, customized for tabular data, demonstrates superior performance in gallstone detection compared to traditional machine learning algorithms (XGBoost, CatBoost, RF, MLP). The 88.08% accuracy rate achieved through 10-fold cross-validation and, in particular, the high precision value of 93.27% suggests potential utility as a component of future clinical decision support systems, pending external and prospective validation.

The most striking finding of our study is that gradient boosting methods (XGBoost, CatBoost), generally considered the "gold standard" for tabular data (Patil et al., 2020), are outperformed by the DANets architecture in this dataset. It is well known in literature that deep learning models (e.g., MLP) generally lag tree-based methods for tabular data (Kucyi et al., 2018). This pattern may partly reflect how ABSTLAY groups related features in this dataset; generalisation to other cohorts and feature sets remains to be validated. This layer has enhanced the model's learning capacity by automatically grouping related features such as liver function tests (ALT, AST) and gallstones indicators and performing high-level abstractions.

This demonstrates that DANets is significantly more successful than the standard MLP architecture (81.83%) in modeling the complex and nonlinear relationships between biochemical and clinical indicators.

Comparing this study with the recent literature reveals architectural and clinical advantages. An accuracy of 83.0% was previously achieved using XGBoost on clinical and metabolomic data (Salem et al., 2023). That study used a different patient cohort and feature set; therefore, direct numerical comparisons require caution, although the methodological contrast is highly instructive. Gradient boosting models handle tabular data effectively; however, they rely on hard decision tree splits that may struggle to capture highly complex nonlinear metabolic interactions. This application of the DANets framework replaces these shallow trees with an abstraction mechanism (ABSTLAY) to group correlated features, such as liver enzymes and bioimpedance metrics, which supports an accuracy of 88.08%.

The strength of the deep abstraction framework is particularly evident when the results are benchmarked against studies using the exact same open-source BIA dataset to ensure a fair evaluation. A baseline accuracy of 85.42% was reported on this shared benchmark using a gradient boosting technique (Esen et al., 2024). More recently, F1 scores of 85.29% and 79.07% were reached by employing complex hybrid pipelines and metaheuristic optimization algorithms (Chakraborty & Mukherjee, 2025; Sarker et al., 2025). The DANets approach provides a more unified learning architecture that achieves a superior 87.11% F1 score without requiring multi-stage feature selection protocols.

These findings should also be considered in the context of recent deep learning applications in medical imaging. While CNN models applied to ultrasound or CT images report accuracies exceeding 95%, these tools are primarily used for definitive diagnosis (Nabil et al., 2025; Ye et al., 2026). This study does not aim to replace such imaging modalities; instead, a high precision of 93.27% could be a sign that it could serve as a reliable, non-invasive tool for early triage. This offers a potential path to bridge the performance gap between routine laboratory tests and advanced radiological scans, representing a cost-effective proof of concept for clinical screening.

These comparisons suggest that, on the same public BIA-inclusive dataset, the evaluated deep learning pipeline can exceed several prior tabular-model accuracy reports, while acknowledging differences in modeling choices across studies.

The most critical outcome of the study is the Precision value of 93.27% from a clinical perspective. A low False Positive rate in medical diagnosis is crucial in preventing patients from undergoing unnecessary further tests and radiation exposure (e.g., CT scans). Combining an inexpensive, rapid, and non-invasive method such as bioimpedance analysis with a high-precision artificial intelligence model offers a cost-effective solution, particularly for the early screening of asymptomatic individuals or those in risk groups (obesity, diabetes, etc.).

Although the results are promising, certain limitations should be addressed. This study was based on a single-center dataset of 319 patients and currently lacks external or prospective validation. Consequently, at this stage, the findings represent proof-of-concept applications rather than deployment-ready clinical solutions. The reliance on one publicly available tabular dataset further limits geographic and clinical-pathway diversity, statistical power for rare subgroups, and the generalizability of point estimates; quantitative results should therefore be treated as dataset-specific pending multi-site replication.

Deep learning models traditionally require exceptionally large and diverse datasets to achieve optimal performance and prevent networks from memorizing training samples. To rigorously mitigate this inherent limitation and prevent overfitting of our dataset, a 10-fold cross-validation methodology was employed. The narrow standard deviation observed in the accuracy (± 3.91 %) of the DANets model across these folds provides strong empirical evidence that the framework maintains a robust generalization capability, even when operating with a restricted sample size. A secondary limitation of deploying deep architectures in medical contexts is their reduced interpretability. While conventional tree-based algorithms like RF or XGBoost offer highly transparent node splits and direct feature importance metrics, the hierarchical abstraction layers within DANets inherently function as a "black box." This opacity can make it difficult for medical professionals to extract the clinical rationale behind a specific prediction.

These structural and data-driven constraints outline the trajectory for future investigations. Subsequent research must prioritize validating this proof-of-concept framework across multicenter cohorts with diverse demographic profiles. Such external validation is critical for definitively establishing the clinical reliability and generalizability of the model. Furthermore, to overcome the interpretability barrier and build clinician trust, future iterations of this work plan to integrate specialized Explainable AI (XAI) modules, such as SHAP or LIME. These tools will be crucial for translating the complex high-level abstractions of DANets into transparent and clinically actionable insights. Finally, future studies should explore multimodal learning architectures that are capable of simultaneously processing numerical bioimpedance data, biochemical laboratory results, and raw ultrasound imagery. Developing an end-to-end pipeline that fuses these distinct data modalities could theoretically push the diagnostic success rate beyond the 95% threshold, representing a step toward future clinically deployable decision-support systems, following comprehensive external validation.

6. Conclusions

In this study, we evaluate an existing DANets-based pipeline that integrates bioimpedance analysis (BIA) data with clinical parameters for early gallstone classification, following the published architecture without structural modification. The DANets model demonstrated a ~~clear~~ performance advantage over commonly used methods in the literature, such as XGBoost (83.07%) and classical MLP (81.83%), achieving 88.08% accuracy and an F1 score of 87.11% as a result of comprehensive 10-fold cross-validation tests. In particular, the high precision value of 93.27% demonstrates the model's ability to minimize false positive alarms and suggests that

the model may contribute to reducing unnecessary radiological imaging in screening settings, subject to validation in real-world clinical workflows.

In our experiments, DANets showed stronger discrimination than the tested tree-based models on several reported metrics, though relative performance may differ under other preprocessing choices or cohorts. This approach has the potential to be used as a cost-effective and non-invasive screening tool in primary healthcare settings prior to expensive and specialized ultrasonography examinations The study results indicate that AI-supported analysis of bioimpedance data serves as a proof of concept for future integration into clinical screening protocols, pending prospective multicenter validation. Future research should focus on developing multimodal structures that combine this model with ultrasound images and testing its generalizability across different populations. Ultimately, this study offers an accessible, fast, and efficient artificial intelligence solution for the management of gallbladder diseases by combining high diagnostic accuracy with low-cost data collection processes.

Declaration of ethical code

This study utilized a publicly available, retrospective, and fully anonymized dataset obtained from the UCI Machine Learning Repository (originally published by Esen et al., 2024). Because this research involves the secondary analysis of completely de-identified open-source data, institutional ethics committee approval and informed consent were officially waived.

Conflicts of interest

The authors declare that there is no conflict of interest

References

AlAmodi, M. O., AlSulaiman, A., & AlSulaiman, A. (2020). Diagnostic accuracy of ultrasonography in choledocholithiasis. *The Pan African Medical Journal*, *35*, 136.

Basu, S., Gupta, M., Madan, C., Gupta, P., & Arora, C. (2024). Focusmae: Gallbladder cancer detection from ultrasound videos with focused masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11715-11725).

Bostan, M. S., & Yılmaz, S. (2023). Prospective evolution of body compositions based on bioelectrical impedance analysis and water intake on patients with gallstone. *Turkish Journal of Trauma & Emergency Surgery*, *29*(8), 850–856.

Bozdag, A., Yildirim, M., Karaduman, M., Mutlu, H. B., Karaduman, G., & Aksoy, A. (2025). Detection of gallbladder disease types using a feature engineering-based developed CBIR system. *Diagnostics*, *15*(5), 552. https://doi.org/10.3390/diagnostics15050552

Branco, M. G., Mateus, C., Capelas, M. L., Pimenta, N., Santos, T., Mäkitie, A., ... Ravasco, P. (2023). Bioelectrical impedance analysis (BIA) for the assessment of body composition in oncology: A scoping review. *Nutrients*, *15*(22), 4792. https://doi.org/10.3390/nu15224792

Chakraborty, C., & Mukherjee, N. (2025). *Bayesian hybrid machine learning of gallstone risk*. arXiv. https://doi.org/10.48550/arXiv.2506.14561

Chen, J., Liao, K., Wan, Y., Chen, D. Z., & Wu, J. (2022). DANets: Deep abstract networks for tabular data classification and regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(4), 3930–3938. https://doi.org/10.1609/aaai.v36i4.20309

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Deng, L., Wang, S., Wan, D., Zhang, Q., Shen, W., Liu, X., & Zhang, Y. (2025). Relative fat mass and physical indices as predictors of gallstone formation: Insights from machine learning and logistic regression. *International Journal of General Medicine*, *18*, 509–527. https://doi.org/10.2147/IJGM.S507013

Esen, İ., Arslan, H., Esen, S. A., Gülşen, M., Kültekin, N., & Özdemir, O. (2024). Early prediction of gallstone disease with a machine learning-based method from bioimpedance and laboratory data. *Medicine*, *103*(8), e37258. https://doi:10.1097/MD.0000000000037258

Feldman, M., Friedman, L. S., & Brandt, L. J. (2020). *Sleisenger and Fordtran's gastrointestinal and liver disease: Pathophysiology, diagnosis, management* (11. Baskı). Elsevier.

Han, T., Jiang, D., Zhao, Q., Wang, L., & Yin, K. (2018). Comparison of random forest, artificial neural networks and support vector machine for intelligent diagnosis of rotating machinery. *Transactions of the Institute of Measurement and Control*, *40*(8), 2681–2693. https://doi.org/10.1177/0142331217708242

Hasan, M. Z., Rony, M. A. H., Chowa, S. S., Bhuiyan, M. R. I., & Moustafa, A. A. (2025). GBCHV an advanced deep learning anatomy aware model for accurate classification of gallbladder cancer utilizing ultrasound images. *Scientific Reports*, *15*(1), 7120.

Kucyi, A., Sadaghiani, S., Tambini, A., Keilholz, S., & Cohen, J. R. (2018). Spontaneous cognitive processes and the behavioral validation of time-varying brain connectivity. *Network Neuroscience*, *2*(4), 397–417. https://doi.org/10.1162/netn_a_00037

Nabil, H. R. (2025). MSFE-GallNet-X: a multi-scale feature extraction-based CNN for gallstone classification. *BMC Medical Imaging*, 25(1), 1-14.

Obaid, A. M., Turki, A., Bellaaj, H., & Ksantini, M. (2024). Diagnosis of gallbladder disease using artificial intelligence: A comparative study. *International Journal of Computational Intelligence Systems*, *17*(1), 46. https://doi.org/10.1007/s44196-024-00431-w

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, *31*.

Shyam, R., Ayachit, S. S., Patil, V., & Singh, A. (2020). Competitive analysis of the top gradient boosting machine learning algorithms. In Proceedings of the 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) (pp. 191–196). IEEE. https://doi: 10.1109/ICACCCN51052.2020.9362840

Salem, N. M., Jack, K. M., Gu, H., Kumar, A., Garcia, M., Yang, P., & Dinu, V. (2023). Machine and deep learning identified metabolites and clinical features associated with gallstone disease. *Computer Methods and Programs in Biomedicine Update*, *3*, 100106. https://doi.org/10.1016/j.cmpbup.2023.100106

Sarker, P., Tiang, J. J., & Nahid, A. A. (2025). Gallstone classification using random forest optimized by sand cat swarm optimization algorithm with SHAP and DiCE-based interpretability. *Sensors*, *25*(17), 5489. https://doi.org/10.3390/s25175489

Sebghatollahi, V., Parsa, M., Minakari, M., & Azadbakht, S. (2023). A clinician's guide to gallstones and common bile duct (CBD): A study protocol for a systematic review and evidence-based recommendations. *Health Science Reports*, *6*(9), e1555. https://doi.org/10.1002/hsr2.1555Digital Object Identifier (DOI)

Singh, J., & Banerjee, R. (2019). A study on single and multi-layer perceptron neural network. *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, 35–40.

Ye, Y., Zeng, T., Huang, J., Chipusu, K., Zhang, J., & Yu, W. (2026). Multimodal clinical integration in deep learning-enhanced CT imaging for automated detection of gallbladder stones. *Journal of Radiation Research and Applied Sciences*, 19(1), 102241.