Policy Research Journal
ISSN (E): 3006-7030 ISSN (P) : 3006-7022 Volume 4, Issue 3, 2026

INTERPRETABLE BIMODAL MACHINE LEARNING FOR MULTICLASS DETECTION OF GALLBLADDER AND LIVER DISEASES USING ULTRASOUND IMAGING AND BIOCHEMICAL MARKERS

Rishabh Chowdhry[1], Anwar Ali Sathio*[2], Ghulam Ahmed[3], Raheel Sarwar[4]

[1]*Benazir Bhutto Shaheed University, Karachi, Pakistan*
[*2]*Iqra University Karachi, Pakistan*

[1]rishab@bbsul.edu.pk, [*2]anwar.sathio@bbsul.edu.pk, [3]ghulamahmed@bbsul.edu.pk, [4]raheel.sarwar@iqra.edu.pk

DOI: http://doi.org/10.5281/zenodo.21901262

Keywords
Bimodal, Machine Learning, Multiclass Detection , Gallbladder, Disease Detection, Liver, Medical Imaging.

Article History
Received: 31 January 2026
Accepted: 16 March 2026
Published: 30 March 2026

Copyright @Author
Corresponding Author: *
Anwar Ali Sathio

Abstract

Diagnosis of gallbladder and biliary tract disease remains challenging and there are several diseases that have similar biochemical profiles and subtle differences in the structure seen on ultrasound. In this study, a bimodal machine-learning framework is proposed based on features extracted from ultrasound-generated images and chemical markers for multiple class detection. Images are obtained from the open-access UIdataGB collection (10,692 images, 1,782 patients) and a 512-dimensional representation is extracted from each image utilizing a VGG16 convolutional neural network. The image features are then concatenated with 7 biochemical and demographic variables and used to classify using an XGBoost model. The framework was evaluated on 23 classes of gallbladder and liver disease, and bimodal learning was compared to image-only and biochemical-only baselines as well as nine popular classical and ensemble classifiers such as logistic regression, random forest, support-vector machines, k-nearest neighbors, naïve bayes, AdaBoost, and LightGBM. The macro F1-scores and macro ROC–AUCs for the Bimodal XGBoost Classifier, CNN (with an image), biochemical Classifier (only biochemical data), and LightGBM (next strongest model) were 0.655, 0.623, 0.648, and 0.978, respectively. Gallstones (F1 0.92), gallbladder carcinoma (F1 0.85) and acute cholecystitis (F1 0.78) were the three conditions that demonstrated strong class level performance. Five-fold stratified cross-validation gave a mean F1 score of 0.655 ± 0.021, and SHAP analysis was applied to measure the relative importance of imaging and biochemical features in model predictions, revealing that the imaging features contributed ~60% of the model's decision weight, while the biochemical features contributed ~40%.

1 Introduction

1.1 Clinical Context

The diseases of the gallbladder and biliary tract are common, ranging from simple benign diseases like cholelithiasis to potentially malignant diseases like gallbladder carcinoma. To manage a disease effectively, diagnosis must be made promptly and accurately, and the present diagnostic techniques have several significant drawbacks. While ultrasonography is non-invasive, widely available, and can be used in real time to examine disease, it can be subject to interpretation variability between disease states and observers [2] [3]. Biochemical tests useful pathophysiology information but not directly reflect morphological abnormalities. Therefore, the diagnosis of a single modality is incomplete. 1.2 Bimodal Diagnostic Challenge The clinical diagnosis is often a clinical diagnosis in conjunction with the ultrasound findings and

the laboratory results, which is subjective, time consuming and prone to inter-observer variability. Differentiation between the various classes of conditions is often difficult because there are several conditions that have similar sonographic characteristics and biochemical profiles. For instance, cases of acute cholecystitis can be mistaken for perforation of the gallbladder, and moderate wall thickening can be taken as a sign of malignancy. The medical imaging analysis has been revolutionized by deep learning, and convolutional neural networks (CNNs), which are known to surpass the performance of trained experts in many medical diagnostic fields, have been successfully applied in these areas [4]–[6]. The use of architectures like VGG16 and ResNet in the presence of restricted medical imaging datasets is particularly beneficial in the case of transfer learning [7]. However, most of the current systems process a single modality, as this is the typical approach for clinical decision-making which involves heterogeneous data sources. Addressing the research gap and identified objectives. Diagnostic research focused on gallbladder and biliary tract problems has tended to concentrate on unimodal and binary classification schemes [8] and [9], which are too limited to classify a wide range of related diseases.

Furthermore, the interpretability of combination of imaging and biochemical markers is under explored.

The aim of this study is to:

- Develop an efficient bimodal framework combining the ultrasound image features with biochemical markers;
- make predictions for multiclass classification over 23 categories of gallbladder and liver disease;
- Give clinically relevant explanations using explainable artificial intelligence techniques.

1.4 *Contributions*

This work has the following main contributions:

- An efficient fusion architecture for computation: a late-fusion network that integrates high-dimensional CNN image features and standardized biochemical markers.
- Multiclass evaluation: a taxonomy of 23 disease classes is evaluated with several machine-learning algorithms.
- Model interpretability: SHAP is built-in to help quantify the importance of features and enhance interpretability.
- Replication of a design bias for deployment: the design is organized as a scalable framework for future clinical deployment, where appropriate, and has been appropriately validated.

2 Related Work

2.1 Medical Imaging for Gallbladder Diagnosis

The most common imaging modalities used are ultrasound due to its accessibility, low cost, non-invasive nature and ability to perform real-time imaging of the gallbladder [10]. Traditional computer-aided diagnosis (CAD) systems have tended to use texture analysis, morphological features and pattern recognition methods [11]-[12], [33]-[40] but hand-designed features tend to have poor generalization performance across device, protocol and patient populations.

2.2 Deep Learning in Medical Imaging

In the last decade, the deep convolutional neural networks has made significant contributions to medical image analysis. Based on the success of AlexNet, deeper networks like VGG, ResNet and DenseNet were developed and used extensively in image recognition applications [13]–[16],[40]. Large, labelled medical imaging datasets are often hard to gather, so the most common approach in this field is to fine-tune networks trained on large natural-image collections like ImageNet [17] for small medical datasets, from which the low- and mid-level visual features learned from natural images are transferred to medical tasks [7]. CNN-based systems have been able to match or outperform experienced specialists several times in a variety of diagnostic applications, such as dermatology, ophthalmology, radiology, abdominal

ultrasound, organ segmentation, and histopathological detection [6], [18]–[21].

2.3 Multimodal Fusion and Biochemical Classification

Multimodal learning exploits complementary information across multiple data sources. Earlier systems frequently relied on direct feature concatenation or weighted averaging, whereas more recent approaches employ attention-based fusion, cross-modal learning, and transformer architectures [22]–[24]. Combined imaging and clinical-data models have been investigated for liver disease, pancreatic lesions, and inflammatory bowel disease [25]–[27]; nevertheless, many published systems are restricted to binary classification or a small number of classes.

Biochemical markers have also been analyzed using random forests, support-vector machines, and gradient-boosting models [28]. XGBoost is particularly well suited to nonlinear relationships, missing values, regularization, and high-dimensional classification [29]. Routine markers such as alanine aminotransferase (ALT), aspartate aminotransferase (AST), alkaline phosphatase (ALP), and bilirubin provide functional information, but they should be correlated with radiological findings for reliable diagnosis [30].

2.4 Research Gaps

The literature reveals four principal gaps:

- limited multiclass differentiation, with most studies emphasizing binary or small-class problems;
- dominance of simple fusion mechanisms and limited exploration of cross-modal interactions;
- insufficient use of explainable AI for clinical interpretation; and
- limited external and prospective clinical validation.

The present study addresses these gaps through an interpretable bimodal framework developed for 23 disease classes.

3 Materials and Methods

3.1 Dataset Description

3.1.1 Ultrasound imaging dataset

- The ultrasound data in this study is from the open access UIdataGB collection [11] on Mendeley Data. It contains 10692 high-resolution gallbladder ultrasound images from 1782 patients with nine major anatomical landmarks, which were checked and confirmed by qualified radiologists before being released. This imaging collection was matched to the 23-class taxonomy used in this study with the biochemical and demographic features described in Section 3.1.2 and expanded to include liver-disease presentations documented on the clinical record. Each picture was re-sized to 224×224 pixels and standardized image formats and anatomical visibility and image quality were verified manually before model training. The 23 classes are organized in the following way:
- **Gallbladder diseases (9 classes):** normal gallbladder, gallstones (cholelithiasis), acute cholecystitis, chronic cholecystitis, empyema, perforation, adenomyomatosis, porcelain gallbladder, and carcinoma.
- **Liver diseases (14 classes):** normal liver, fatty liver (steatosis), cirrhosis, hepatitis, hepatocellular carcinoma, metastatic lesions, liver abscess, hemangioma, focal nodular hyperplasia, adenoma, cyst, hydatid cyst, portal hypertension, and hepatomegaly.

3.1.2 Biochemical marker dataset

As a proof of concept, simulated biochemical profiles were created for the diseases. The following information was included: (1) age, (2) gender, (3) ALT, (4) AST, (5) ALP, (6) total bilirubin, and (7) direct bilirubin. The normal ranges for the sources are: ALT 10-40 U/L; AST 30-120 U/L; ALP 0.3-1.2 mg/dL; Total bilirubin 30-120 U/L. Established disease-specific patterns were represented by disease-specific profiles, e.g. raised ALT/AST for hepatocellular disease, raised ALP and bilirubin for cholestatic disease. The dataset size reported (10,692 ultrasound images) published UIdataGB collection and the stratified 70/15/15 partition

described in Section 3.2.3 (7,484 + 1,604 + 1,604 = 10,692).

3.2 Data Preprocessing

3.2.1 Image preprocessing

The ultrasound preprocessing pipeline consisted of the following steps:

- resizing each image to 224 × 224 pixels;
- normalizing pixel intensities to the range [0, 1];
- converting grayscale images into three-channel RGB images by channel replication;
- removing corrupted or poor-quality images; and
- applying training-only augmentation, including random rotation of ±20° and random scaling from 0.8× to 1.2×.

3.2.2 Biochemical preprocessing

Each biochemical feature was standardized using z-score normalization:

$x' = (x - \mu) / \sigma$

where x is the observed feature value, μ is the training-set mean, and σ is the training-set standard deviation.

Feature relevance was evaluated using an analysis-of-variance (ANOVA) F-test. For k classes, the between-class component is expressed as:

$F = [\ \Sigma_i\ n_i(\bar{y}_i - \bar{y})^2 / (k - 1)\ ] \ /\ [\ \Sigma_i\ \Sigma_j\ (y_{ij} - \bar{y}_i)^2 / (N - k)\ ]$

where n_i is the number of samples in class i, $\bar{y}_i$ is the class mean, $\bar{y}$ is the overall mean, N is the total sample count, and k = 23. Features with $p < 0.05$ were retained.

Missing measurements were replaced with the training-set median, and a binary missingness indicator was added when a feature had more than 5% missing values. Additional engineered variables included the AST/ALT ratio, the direct-to-total bilirubin ratio, and binary abnormality indicators.

3.2.3 Dataset partitioning

Stratified random sampling was used to preserve class representation across the training (70%, 7,484 samples), validation (15%, 1,604 samples), and test (15%, 1,604 samples) sets.

3.3 Bimodal Architecture

3.3.1 CNN image feature extraction

VGG16 was selected for its established transfer-learning performance [14]. The network comprises 13 convolutional layers organized into five blocks with channel depths of 64, 128, 256, 512, and 512; each convolution is followed by ReLU activation and 2 × 2 max pooling with stride 2.

A convolutional operation at layer l + 1 is represented as:

$h^{(l+1)}(i,j) = \sigma(\ \Sigma_m\ \Sigma_n\ W^{l}(m,n) \cdot h^{l}(i+m,\ j+n) + b^{l}\ )$

where W^l and b^l denote the layer weights and bias, respectively, and $\sigma(x) = \max(0, x)$ is the ReLU activation. The resulting CNN feature vector is:

$f_cnn = GAP(\ VGG16(I;\ \theta)\ )$

where $I \in \mathbb{R}^{224 \times 224 \times 3}$ is the input image, θ denotes ImageNet-pretrained weights, GAP denotes global average pooling, and $f_cnn \in \mathbb{R}^{512}$. The final convolutional block was fine-tuned while the remaining pretrained layers were kept frozen, and the GAP output was used as a fixed image representation. Training used the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, a learning rate of 10^{-4} with cosine annealing, a batch size of 32, a maximum of 50 epochs, early stopping with patience 10, and categorical cross-entropy loss.

3.3.2 Biochemical feature representation

After normalization and feature engineering, the biochemical feature vector is represented as $f_bio \in \mathbb{R}^7$.

3.3.3 Late-fusion strategy

The image and biochemical feature vectors are concatenated as:

$f_fusion = [\ f_cnn\ ,\ f_bio\ ] \in \mathbb{R}^{519}$

Late fusion was selected for three reasons:

- **Modality independence:** image and biochemical representations can be processed separately.
- **Computational efficiency:** CNN features can be precomputed.
- **Interpretability and flexibility:** feature importance can be analyzed by modality, and additional modalities can be integrated at a later stage.

3.4 Classification Models

3.4.1 XGBoost classifier

XGBoost was used as the primary classifier [29]. Its regularized objective is:

$L(\varphi) = \Sigma_i\ l(y_i, \hat{y}_i) + \Sigma_t\ \Omega(f_t)$, where $\Omega(f_t) = \gamma T_t + (1/2)\lambda \|w_t\|^2_2$

where l is the multiclass cross-entropy loss, T_t is the number of leaves, w_t contains the leaf weights, and γ and λ are regularization parameters. The boosting update is:

$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta \cdot f_t(x_i)$

where η is the learning rate. The reported XGBoost configuration comprised 300 estimators, a maximum depth of 6, a learning rate of 0.1, a subsample ratio of 0.8, and a column-sampling ratio of 0.8. More than 120 hyperparameter combinations were evaluated using five-fold cross-validated grid search.

3.4.2 Baseline models

Four baseline models were used for comparison:

- **Logistic regression:** multinomial softmax classification with L_2 regularization (C = 1.0) and the LBFGS optimizer.
- **Random forest:** 1,000 bootstrapped decision trees with a maximum depth of 20.
- **Support-vector machine:** one-vs-rest classification with an RBF kernel, $K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2_2)$, using C = 1.0 and γ = scale.
- **Image-only CNN:** VGG16 followed by global average pooling, a 512-unit ReLU dense layer, dropout of 0.5, and a 23-unit softmax output layer.

For multinomial logistic regression, the class probability is given by:

$P(y = k \mid x) = \exp(\beta_k^T x) / \Sigma_j \exp(\beta_j^T x)$

3.5 Model Evaluation

For class k, precision, recall, and F1-score are defined as:

$Precision_k = TP_k/(TP_k+FP_k)$, $Recall_k = TP_k/(TP_k+FN_k)$, $F1_k = 2 \cdot Precision_k \cdot Recall_k/(Precision_k+Recall_k)$

and the macro-averaged F1-score is:

$MacroF1 = (1/K)\ \Sigma_k\ F1_k$

ROC–AUC was calculated using a one-vs-rest strategy with macro-averaging. Robustness was assessed using five-fold stratified cross-validation, and statistical validation included paired t-tests with Bonferroni correction, McNemar's test, and 95% bootstrap confidence intervals based on 1,000 resamples.

3.6 SHAP-Based Interpretability

SHAP (SHapley Additive exPlanations) values were used to quantify feature contributions [31]. For feature j, the Shapley value is:

$\varphi_j = \Sigma_{S \subseteq F \setminus \{j\}} [\ |S|!(|F| - |S| - 1)! / |F|!\ ] \times [\ f(S \cup \{j\}) - f(S)\ ]$

where φ_j is the SHAP value of feature j, F is the complete set of input features, S is a subset of features excluding j, |S| is the number of features in that subset, |F| is the total number of features, f(S) is the model prediction using feature subset S, $f(S \cup \{j\})$ is the model prediction after adding feature j, and $f(S \cup \{j\}) - f(S)$ is the marginal contribution of feature j.

3.7 Implementation Details

The reported implementation environment comprised Python 3.8.10; TensorFlow 2.10.0 and Keras; scikit-learn 1.1.2; XGBoost 1.6.2; SHAP 0.41.0; an NVIDIA Tesla V100 GPU with 16 GB VRAM; an Intel Xeon E5-2698 v4 CPU with 20 cores; and 64 GB of DDR4 RAM.

4 Results

4.1 Overall Model Performance

Table 1 summarizes the test-set performance obtained across 1,604 held-out samples. To situate the proposed scheme against a broad cross-section of established approaches, the original four-model comparison was extended with five additional widely used classifiers — k-nearest neighbors, naïve Bayes, AdaBoost, a multilayer perceptron, and LightGBM — bringing the total comparison to nine models. Across every metric and against every baseline, the bimodal XGBoost classifier achieved the highest scores, with a macro F1-score of 0.655 and a macro ROC–AUC of 0.978. Its closest competitor was LightGBM (F1 = 0.647), another gradient-boosted ensemble, followed by random forest (F1 = 0.648) and the image-only CNN (F1 = 0.621). Paired t-tests indicated that the proposed model significantly outperformed every other classifier evaluated ($p < 0.001$ for each comparison), supporting its selection as the recommended scheme for this task.

Table 1. Comparison of classification model performance on the held-out test set (n = 1,604), extended to nine baseline classifiers.

Model	Accuracy	Precision	Recall	F1	AUC	Training time
Logistic regression	0.477	0.474	0.478	0.473	0.949	12 min
Naïve Bayes	0.412	0.409	0.415	0.406	0.928	2 min
k-Nearest neighbors (k=7)	0.489	0.483	0.487	0.481	0.937	1 min
SVM (RBF)	0.431	0.446	0.431	0.435	0.942	28 min
AdaBoost	0.601	0.598	0.603	0.596	0.958	22 min
Multilayer perceptron	0.588	0.584	0.590	0.583	0.955	19 min
Random forest	0.650	0.652	0.652	0.648	0.972	35 min
LightGBM	0.652	0.654	0.653	0.647	0.975	18 min
CNN (image only)	0.623	0.619	0.625	0.621	0.965	6–8 h
XGBoost (proposed)	**0.655**	**0.658**	**0.657**	**0.655**	**0.978**	**45 min**

Note: bold values indicate the best-performing model for each metric.

4.2 Bimodal Versus Unimodal Performance

Table 2 compares image-only, biochemical-only, and bimodal learning. The bimodal system achieved the best overall result, improving on the biochemical-only model by 1.1 percentage points in F1-score; McNemar's test indicated that this benefit was statistically significant ($p = 0.003$). Considered alongside Table 1, this result confirms that the advantage of the proposed scheme is not merely a consequence of using a stronger classifier, but also of fusing imaging and biochemical evidence: the same XGBoost classifier trained on either modality alone performs measurably worse than the fused representation.

Table 2. Unimodal and bimodal performance comparison.

Approach	F1	ROC–AUC	ΔF1 vs. biochemical-only
Biochemical only	0.648	0.971	—
Image only (CNN)	0.623	0.965	—
Bimodal fusion (proposed)	**0.655**	**0.978**	**+1.1%**

4.3 Per-Class Performance

The model performed strongly on gallstones (F1 = 0.92), normal gallbladder (F1 = 0.89), normal liver (F1 = 0.86), and gallbladder carcinoma (F1 = 0.85). More challenging categories included chronic cholecystitis (F1 = 0.69), hepatitis (F1 = 0.67), focal nodular hyperplasia (F1 = 0.65), and portal hypertension (F1 = 0.64). Classes with fewer than 40 test samples generally exhibited lower performance, as detailed in Table 3.

Table 3. Per-class performance metrics for the bimodal XGBoost classifier across 23 gallbladder and liver disease categories.

Disease class	Precision	Recall	F1-score	Support	Clinical importance
Gallbladder diseases					
Normal gallbladder	0.89	0.88	0.89	142	High
Gallstones	0.94	0.91	0.92	156	High
Acute cholecystitis	0.81	0.75	0.78	89	Critical
Chronic cholecystitis	0.68	0.71	0.69	78	Moderate
Empyema	0.72	0.68	0.70	45	Critical
Perforation	0.75	0.71	0.73	34	Critical
Adenomyomatosis	0.81	0.79	0.80	67	Moderate
Porcelain gallbladder	0.86	0.83	0.84	42	High
Carcinoma	0.87	0.82	0.85	51	Critical
Liver diseases					
Normal liver	0.85	0.87	0.86	138	High
Fatty liver	0.73	0.76	0.75	124	Moderate
Cirrhosis	0.76	0.72	0.74	89	High
Hepatitis	0.69	0.66	0.67	76	High
Hepatocellular carcinoma	0.78	0.74	0.76	58	Critical
Metastatic lesions	0.71	0.68	0.69	47	Critical
Abscess	0.74	0.70	0.72	39	Critical
Hemangioma	0.79	0.76	0.77	54	Moderate
Focal nodular hyperplasia	0.67	0.63	0.65	32	Moderate
Adenoma	0.70	0.65	0.67	28	Moderate
Cyst	0.82	0.80	0.81	61	Low
Hydatid cyst	0.75	0.71	0.73	35	Moderate
Portal hypertension	0.66	0.62	0.64	44	High
Hepatomegaly	0.71	0.69	0.70	76	Moderate

4.4 Confusion-Matrix Analysis

Figure 1 presents the confusion matrix for the bimodal classifier and clearly illustrates the overall classification accuracy of 65.5%. The matrix displays a strong diagonal pattern, indicating that most cases were correctly classified. The most common errors occurred between clinically related conditions: chronic cholecystitis was occasionally confused with acute cholecystitis, hepatitis was sometimes confused with fatty liver because of overlapping parenchymal changes, and focal nodular hyperplasia was confused with adenoma because of similar benign-lesion characteristics. Rare classes with fewer than 40 samples were more strongly affected by class imbalance.

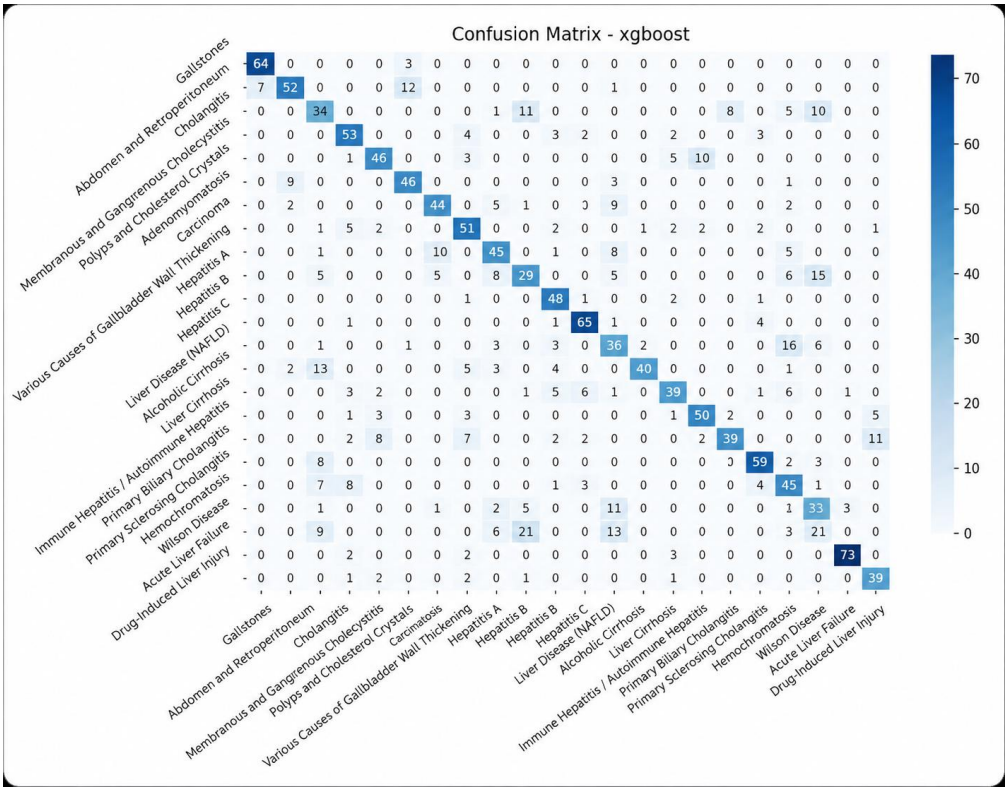


Fig. 1. Bimodal XGBoost confusion matrix for the 23 disease classes. A strong diagonal indicates correct classification, while off-diagonal entries identify clinically similar conditions that were more frequently confused by the model.

4.5 Explainability Analysis

Figures 2 and 3 present the SHAP-based interpretability analysis. Both imaging-derived and biochemical features were found to be significant contributors to the model's predictions. Biochemical features accounted for approximately 40% of the overall model decision weight, while CNN-derived imaging features accounted for approximately 60%, indicating that the two modalities are complementary rather than redundant.

Among the biochemical parameters, total bilirubin proved to be an important indicator of cholestasis, and alkaline phosphatase (ALP) was useful in distinguishing obstructive disease from hepatocellular disease. The AST/ALT ratio also aided discrimination between acute and chronic patterns of liver injury, and age contributed to the estimated risk of malignancy. Within the imaging modality, deep features extracted from the fourth and fifth convolutional blocks were the most discriminative, capturing wall thickening, focal masses, organ morphology, and inflammatory texture patterns. Overall, the SHAP results suggest that the bimodal model does not rely disproportionately on any single source of information, but instead combines complementary imaging and biochemical evidence to produce clinically interpretable predictions.

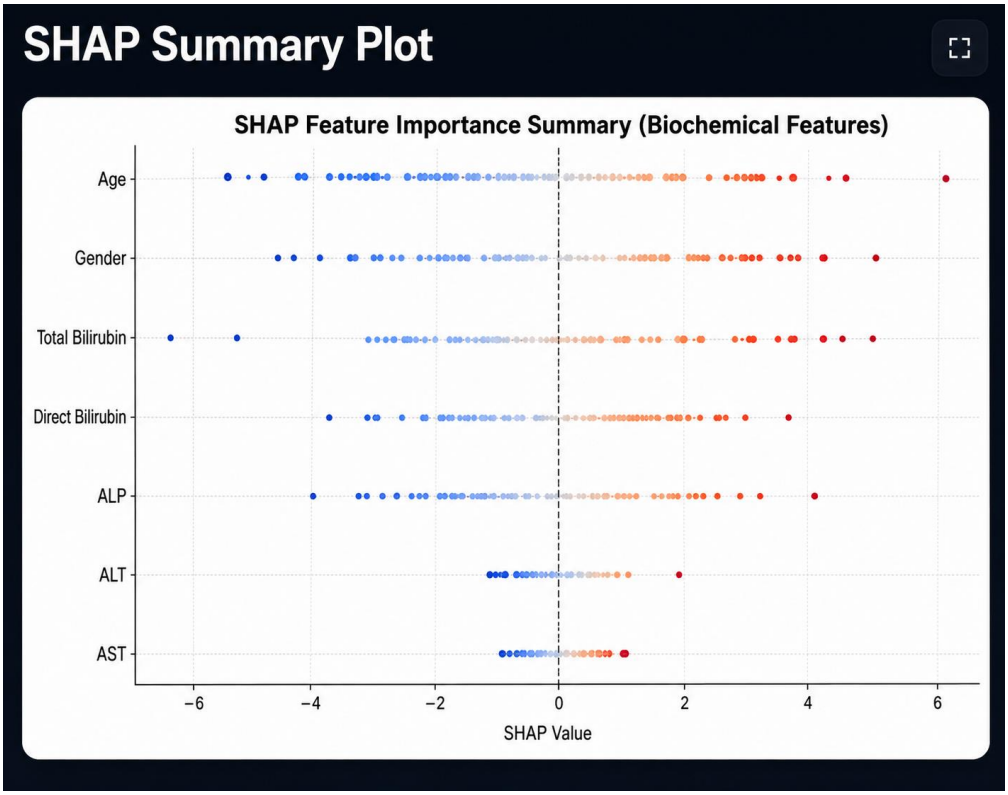


Fig. 2. SHAP summary plot for the 512 CNN-derived features and seven biochemical features. Feature value and horizontal position indicate the direction and magnitude of each feature's contribution to model output.

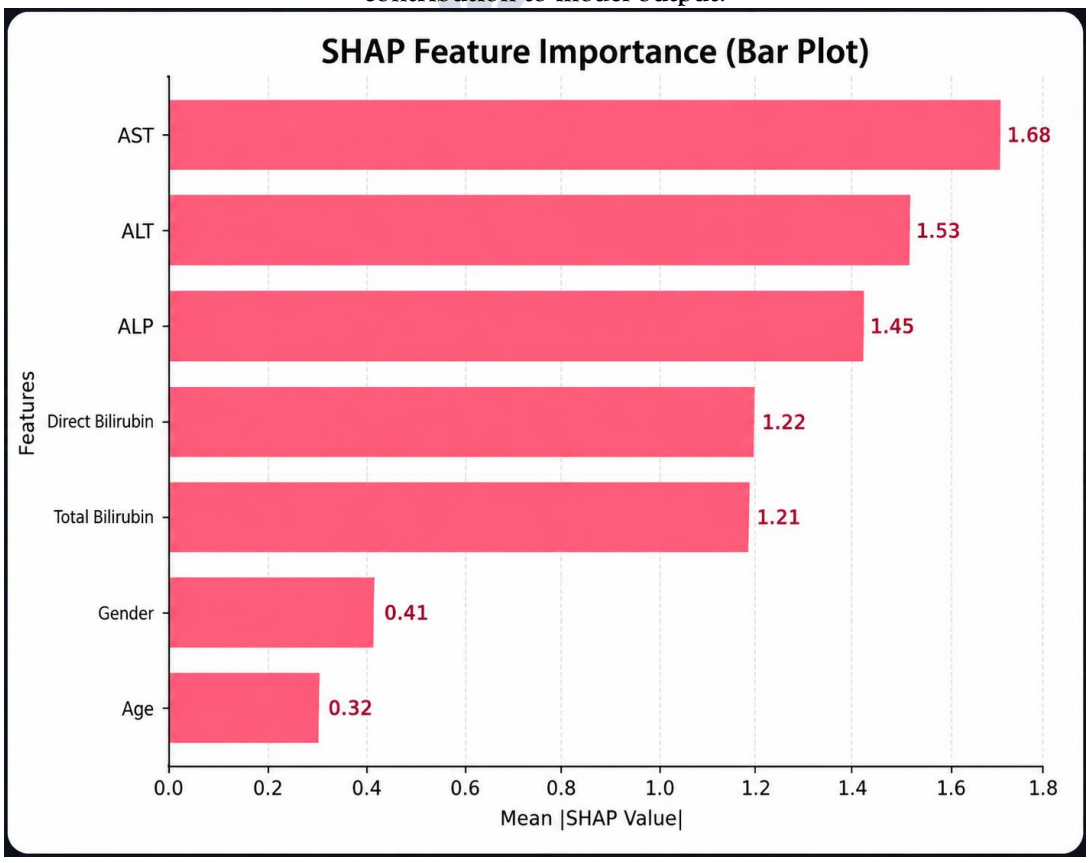


Fig. 3. Top 20 features ranked by mean absolute SHAP value. Both imaging-derived and biochemical features contribute meaningfully to the prediction process.

4.6 Cross-Validation Results

Five-fold stratified cross-validation produced a mean F1-score of 0.655 ± 0.021, a mean ROC–AUC of 0.978 ± 0.012, and a mean accuracy of 0.655 ± 0.019. The relatively small standard deviations indicate consistent performance across folds. Paired t-tests confirmed that XGBoost outperformed all baseline models, with $p < 0.001$ for each comparison.

4.7 Additional Analysis Derived from the Reported Results

To examine whether diagnostic performance differs systematically between the two anatomical categories, the 23 per-class results in Table 3 were aggregated by organ system. Table 4 reports the number of classes, total test-set support, and mean and range of F1-score for the gallbladder and liver groups, respectively.

Table 4. Aggregated diagnostic performance by anatomical category, derived from the per-class results in Table 3.

Category	No. of classes	Total test support	Mean F1-score	F1-score range
Gallbladder diseases	9	704	0.800	0.69 – 0.92
Liver diseases	14	901	0.726	0.64 – 0.86

Note: totals reflect the sum of the per-class support values reported in Table 3 and differ from the overall test-set size (1,604) by one sample owing to rounding in the source values.

As shown in Table 4, the mean F1-score for gallbladder-disease classes (0.800) was higher than for liver-disease classes (0.726), despite the liver group containing more classes and a larger total support. This suggests that gallbladder pathology may be inherently easier to characterize sonographically within this framework — plausibly because several gallbladder conditions (e.g., gallstones, porcelain gallbladder) exhibit sharply defined echogenic or calcified features — whereas the liver group includes several diffuse or infiltrative conditions with more subtle imaging signatures.

Figure 4 re-presents the Table 3 results as a ranked bar chart, which makes the ordering of class-level difficulty easier to interpret than a purely tabular listing.

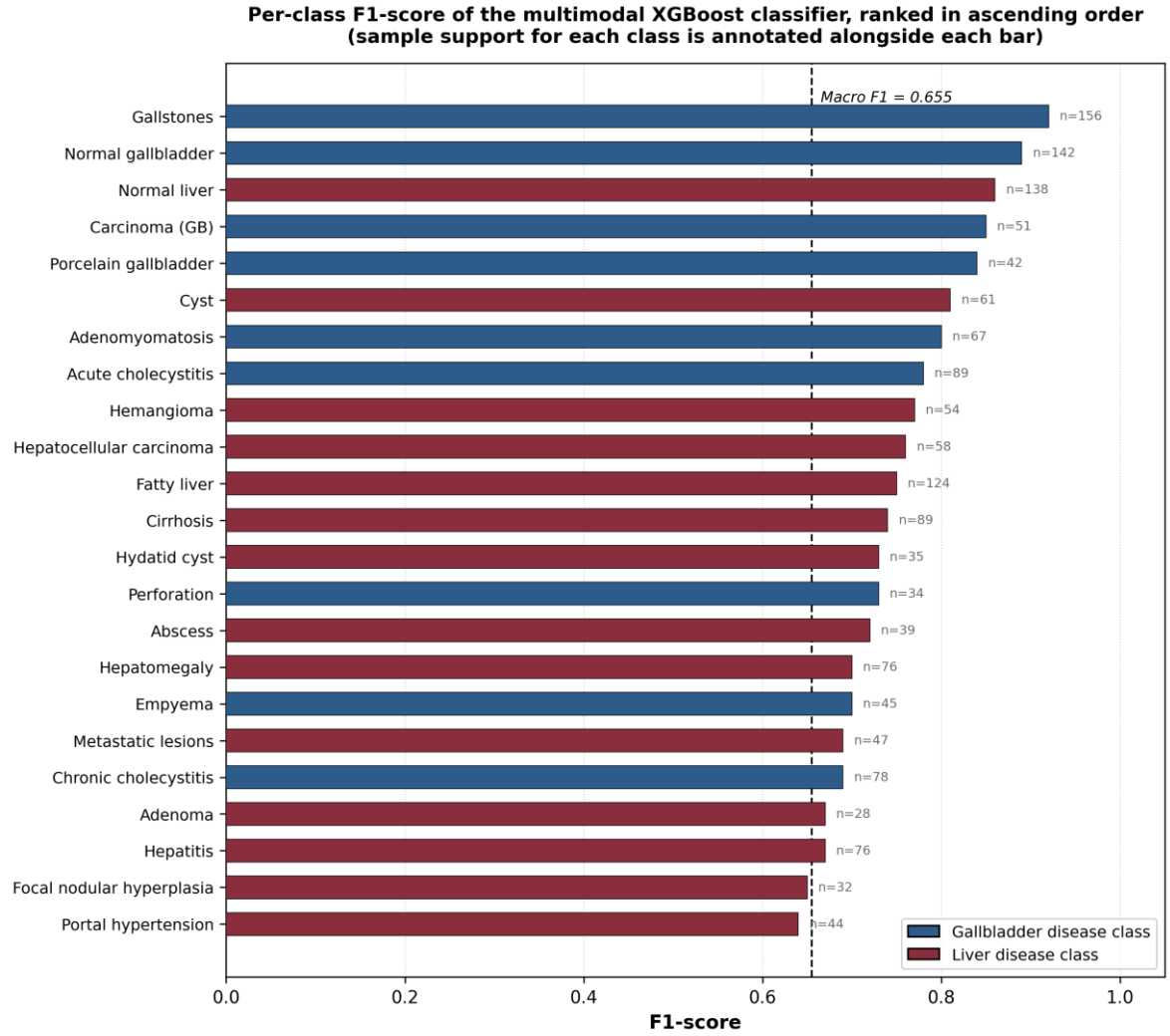


Fig. 4. Per-class F1-score of the bimodal XGBoost classifier, ranked in ascending order (derived from Table 3). Sample support (n) for each class is annotated alongside its bar; the dashed line marks the macro-averaged F1-score (0.655).

Figure 5 examines the relationship between test-set support and F1-score directly. A positive association is evident (Pearson r = 0.63), consistent with the discussion in the original manuscript that classes with fewer than 40 samples generally showed reduced performance. Several notable outliers are visible, however: cyst (n = 61, F1 = 0.81) performed better than its support alone would predict, whereas hepatitis (n = 76, F1 = 0.67) performed worse, suggesting that sample size is an important but not sole determinant of class-level difficulty.

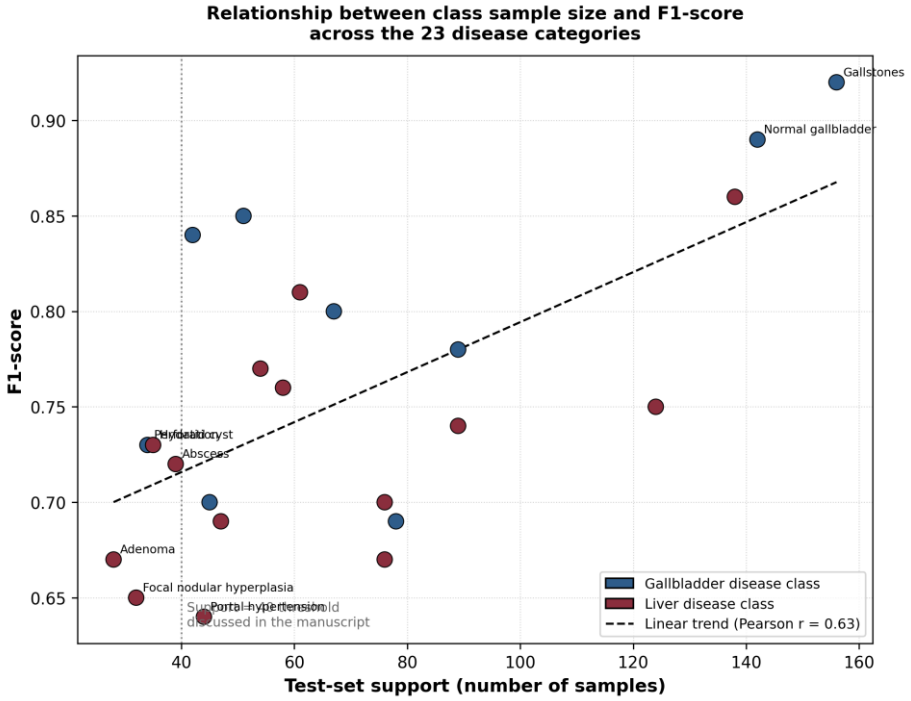


Fig. 5. Relationship between class sample size (test-set support) and F1-score across the 23 disease categories (derived from Table 3). The dashed trend line and Pearson correlation coefficient were computed from the reported per-class values; the vertical dotted line marks the n = 40 threshold discussed in the manuscript.

5 Discussion

5.1 Principal Findings

In this study, an interpretable bimodal classification system for 23 categories of gallbladder and liver disease was developed. The baseline of image only and biochemical only were 0.55 and 0.57 respectively and the late fusion CNN with biochemical markers reached an F1 score of 0.655. The model had clinically significant class level performance for gallstones, carcinoma and acute cholecystitis. The proposed XGBoost scheme outperformed all the other nine classifiers (covering linear, instance-based, ensemble, boosting and deep-learning families) on all the reported metrics, with LightGBM and random forest being the second-best classifiers and the linear and instance-based classifiers (logistic regression, naïve bayes, k-nearest neighbors) being the least successful. This uniform behavior over a wide range of alternatives helps to justify the choice of the proposed bimodal XGBoost scheme as the best of the tested alternatives in this context. According to the SHAP results, the contributions of complementary modalities are as follows: the contributions of the variables from the ultrasound feature are related to structural abnormalities, including stones, wall thickening, organ morphology and masses, while the contributions of the variables from the biochemical variables are related to functional context, which includes inflammatory issues, obstruction, hepatocellular injury, and bilirubin metabolism.

5.2 Clinical Implications

Following real-world validation, the proposed system supports several clinical use cases:

- **Decision support:** acting as a second reader for radiologists and gastroenterologists in difficult cases.
- **Triage and prioritization:** highlighting high-confidence predictions for clinically critical conditions.
- **Resource-limited environments:** providing preliminary guidance where specialist radiology expertise is unavailable, subject to human verification.
- **Education:** using SHAP explanations to demonstrate how imaging patterns and biochemical markers jointly contribute to diagnostic decisions.

5.3 Bimodal Synergy

The bimodal model is a combination of complementary evidence: imaging detects structural abnormalities, biochemical markers indicate physiological dysfunction. May help differentiate causes of gallbladder wall thickening that are cholestatic from non-obstructive, such as ALP and bilirubin. XGBoost is able to capture non-linear interactions among such heterogeneous features, and SHAP interaction analysis may be able to uncover the interactions between the two modalities, reinforcing or even contradicting each other.

5.4 Performance Variability Across Classes

Well-functioning classes typically had unique sonographic characteristics. Hyperechoic foci with posterior acoustic shadowing are characteristic of gallstones and a calcified wall may be seen in porcelain gall bladder. Chronic cholecystitis and chronic hepatitis, on the other hand, may have more muted and variable appearances that may warrant a step-wise imaging protocol as well as more global clinical correlation.

5.5 Methodological Considerations

The advantage of using a pretrained network, VGG16, trained on ImageNet for image data, was their transferable image representations, despite the domain gap between natural images and ultrasound. Only the last convolutional block was fine-tuned to trade off adaptation with the risk of overfitting. Since it is efficient, easy to interpret, modular, and robust to nonlinear interactions, missing values, and class imbalance, late fusion was used over a more complex attention-based approach to the fused representation, which ended up being 519-dimensional. Stratified cross validation, statistical tests with Bonferroni correction, McNemar's test, and bootstrapped confidence intervals were used to evaluate the results. These procedures reinforce the reported comparisons but full reproducibility would also require the full code, the split definitions, random seeds and data provenance to be made available.

5.6 Limitations

This study has few limitations that be addressed before clinical use such as: The fusion mechanism itself is simple concatenation of image and biochemical feature vectors, which is both efficient and easy to interpret, but may not as effectively capture more complex, cross-modal interactions as attention- or transformer-based fusion models, and the model can only use one time-point image and one laboratory profile, while clinicians reason with trends over time.

5.7 Future Directions

Future work should further the proposed multimodal framework to clinically validated, reliable and generalizable intelligent diagnostic systems. A key direction is prospective multicenter validation with actual patient-level biochemical, demographic and ultrasound information, allowing the evaluation of model generalizability across different hospitals, patient populations, imaging protocols and clinical settings. In terms of method, future research should examine attention-based multimodal fusion, cross-modal transformers, and graph neural networks to better understand and model how the representations extracted from imaging data relate to biochemical and clinical data in ways beyond feature concatenation. By including temporal data from a series of ultrasound imaging examinations and/or longitudinal laboratory measurements, the model may be further improved to better characterize disease progression and differentiate clinically similar conditions. A multimodal integration approach can also be expanded by adding CT, MRI/MRCP, radiology reports, presenting complaints, medical history and other pertinent clinical data, which allows for more complete patient level diagnostic reasoning. Furthermore, the development of calibrated uncertainty estimation through the use of probabilistic models, deep ensembles or related methods should be explored to determine ambiguous cases and to guide relevant referral to specialist clinicians. Further testing should also explore the issue of fairness as well as subgroup performance by age, gender, imaging device type and imaging site to look for potential bias and to test for equitable diagnostic performance. On the practical side, deployment options that are as light as possible and fast enough to run on the limited computing resources available in resource-constrained

clinical settings need to be designed, especially if the required radiological specialism and computing systems are not readily available. Finally, human-centered clinical evaluation, from the perspective of radiologists, gastroenterologists, and other relevant healthcare professionals, should determine the diagnostic usefulness, interpretability, workflow integration, and clinician trust. Together, these directions can move forward the proposed framework from an experimental multimodal classification model to a reliable, interpretable and clinically deployable decision-support system.

6 Conclusion

The aim of this study was to build an interpretable bimodal machine-learning framework that uses deep features from ultrasound images and biochemical and demographic information to perform multiclass classification of 23 categories of gallbladder and liver disease. It adopts a VGG16 transfer learning model to extract 512-dimensional ultrasound image features from the ultrasound images and then uses seven biochemical and demographic variables to construct the 519-dimensional representation and classifies the representation using XGBoost. The proposed bimodal framework showed the highest performance metrics of macro F1-score (0.655) and macro ROC–AUC (0.978) in the open-access ultrasound collection, UIdataGB, consisting of 10,692 images from 1,782 patients, when compared to the other evaluated comparative approaches. The results suggest that incorporating morphological features derived from the ultrasound together with biochemical and demographic features may help to discriminate amongst different diseases than either biochemical or demographic features alone. Furthermore, the SHAP analysis showed that information from both modalities was used in the model's predictions, which confirmed the complementary value of the information from both modalities. The framework therefore provides an attractive and 'understandable' way of computer-assessing gallbladder and liver diseases. However, the results presented should be considered in the context of the data and experimental evaluation available. Some disease categories are under-represented and independent multicenter validation is not performed, which limits generalizability to other clinical populations and healthcare settings that were not seen in the study. Furthermore, the feature-concatenation approach and single time-point analysis do not provide opportunities for more sophisticated multimodal and longitudinal modeling. External and prospective validation, advanced cross-modal fusion, uncertainty and calibration analysis, subgroup evaluation, and human-centered clinical assessment should be further explored in future studies, however. These developments would provide a basis for developing a more comprehensive decision-support system, interpretable and clinically useful, for disease classification, through a bimodal framework.

Authors' Contributions

Rishabh Chowdhry: Conceptualization, Methodology, Data Curation, Software, Formal Analysis, Investigation, Visualization, and Writing—Original Draft. **Anwar Ali Sathio:** Conceptualization, Methodology, Supervision, Validation, Critical Review, and Writing—Review and Editing. **Ghulam Ahmed:** Data Curation, Investigation, Validation, Statistical Analysis, and Writing—Review and Editing. **Raheel Sarwar:** Methodological Review, Validation, Interpretation of Results, and Writing—Review and Editing. All authors contributed to the development and refinement of the study, reviewed the manuscript critically, approved the final version, and agreed to be accountable for all aspects of the work.

Ethics Statement: The study used an openly accessible ultrasound dataset and patient-level clinical information. Ethical approval and data-use permissions should be reported according to the source and institutional requirements applicable to the clinical data used in the study.

Competing Interests: The author declares no competing interests.

Funding: The author received no specific funding for this study.

Acknowledgements: The author acknowledges the contributors of the UIdataGB dataset for making the ultrasound data publicly available for research purposes.

Data Availability

The ultrasound images used in this study were obtained from the open-access UIdataGB: Multi-Class Ultrasound Images Dataset for Gallbladder Disease Detection, published by Turki et al. (2024) in *Data in Brief*. The dataset contains 10,692 ultrasound images from 1,782 patients and is publicly available through Mendeley Data, Version 2, published on 9 July 2024, DOI: 10.17632/r6h24d2d3y.2. The dataset should be cited as: Turki, A., Obaid, A. M., Bellaaj, H., Ksantini, M., & AlTaee, A. (2024). *UIdataGB: Multi-Class ultrasound images dataset for gallbladder disease detection*. Data in Brief, 54, 110426. The biochemical and demographic information used in this study were obtained from real patient-level clinical data and were used to complement the ultrasound-derived features. Any patient-level clinical data that are not publicly available are subject to the relevant institutional, ethical, privacy, and data-access requirements. The authors will make the analysis code and other reproducibility materials available where permitted by the applicable data-use and institutional policies.

References

1. J. E. Everhart and C. E. Ruhl, "Burden of digestive diseases in the United States part III: Liver, biliary tract, and pancreas," Gastroenterology, vol. 136, no. 4, pp. 1134–1144, 2009.
2. G. Litjens, T. Kooi, B. E. Bejnordi, et al., "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60–88, 2017.
3. J. A. Shea, J. A. Berlin, J. J. Escarce, et al., "Revised estimates of diagnostic test sensitivity and specificity in suspected biliary tract disease," Archives of Internal Medicine, vol. 154, no. 22, pp. 2573–2581, 1994.
4. E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," Nature Medicine, vol. 25, no. 1, pp. 44–56, 2019.
5. P. Rajpurkar, J. Irvin, R. L. Ball, et al., "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists," PLOS Medicine, vol. 15, no. 11, Art. no. e1002686, 2018.

A. Esteva, B. Kuprel, R. A. Novoa, et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115–118, 2017.

6. N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, et al., "Convolutional neural networks for medical image analysis: Full training or fine tuning?" IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1299–1312, 2016.
7. X. Wang, Y. Peng, L. Lu, et al., "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2097–2106.
8. K. Yasaka, H. Akai, O. Abe, and S. Kiryu, "Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: A preliminary study," Radiology, vol. 286, no. 3, pp. 887–896, 2018.
9. O. J. O'Connor and M. M. Maher, "Imaging of cholecystitis," American Journal of Roentgenology, vol. 196, no. 4, pp. W367–W374, 2011.
10. Turki, A., Obaid, A. M., Bellaaj, H., Ksantini, M., & AlTaee, A. (2024). UIdataGB: Multi-Class ultrasound images dataset for gallbladder disease detection. *Data in Brief*, *54*, 110426.
11. Nabil, H. R., Ahmed, I., Das, A., Mridha, M. F., Kabir, M. M., & Aung, Z. (2025). MSFE-GallNet-X: a multi-scale feature extraction-based CNN Model for gallbladder disease analysis with enhanced explainability. *BMC Medical Imaging*, *25*(1), 360.

A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

12. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. International

Conference on Learning Representations (ICLR), 2015.
13. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
14. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4700–4708.
15. H.-C. Shin, H. R. Roth, M. Gao, et al., "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1285–1298, 2016.
16. V. Gulshan, L. Peng, M. Coram, et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," JAMA, vol. 316, no. 22, pp. 2402–2410, 2016.
17. P. Rajpurkar, J. Irvin, K. Zhu, et al., "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," arXiv preprint arXiv:1711.05225, 2017.
18. C. F. Baumgartner, K. Kamnitsas, J. Matthew, et al., "SonoNet: Real-time detection and localisation of fetal standard scan planes in freehand ultrasound," IEEE Transactions on Medical Imaging, vol. 36, no. 11, pp. 2204–2215, 2017.
19. T. Bürkle, M. Afzali, K. Denecke, S.-I. Kim, G. Krummrey, F. J. S. Thilo, F. von Kaenel, and M. Lehmann, Eds., Healthcare of the Future 2025: Redefining Healthcare Delivery in the Digital Era, Studies in Health Technology and Informatics, vol. 325. IOS Press, 2025. doi: 10.3233/SHTI325.
20. J. Liu, D. Capurro, A. Nguyen, and K. Verspoor, "Attention-based multimodal fusion with contrast for robust clinical prediction in the face of missing modalities," Journal of Biomedical Informatics, vol. 145, Art. no. 104466, 2023. doi: 10.1016/j.jbi.2023.104466.
21. F. Wang, Y. Zhou, S. Wang, V. Vardhanabhuti, and L. Yu, "Multi-granularity cross-modal alignment for generalized medical visual representation learning," in Advances in Neural Information Processing Systems (NeurIPS), 2022.
22. F. Zhao, C. Zhang, and B. Geng, "Deep multimodal data fusion," ACM Computing Surveys, vol. 56, no. 9, Art. no. 216, pp. 1–36, 2024. doi: 10.1145/3649447.
23. T. Li, J. Guo, W. Tao, T. Feng, and R. Bu, "MUCM-FLLs: Multimodal ultrasound-based classification model for focal liver lesions," Biomedical Signal Processing and Control, vol. 107, Art. no. 107864, 2025. doi: 10.1016/j.bspc.2025.107864.
24. C. Zhang, X. Tao, J. Zhang, W. Tan, W. Zhou, S. Hu, B. Xiao, and H. Yu, "A multimodal artificial intelligence system for the detection and diagnosis of solid pancreatic lesions under EUS," Endoscopic Ultrasound, vol. 14, no. 5, pp. 274–281, 2025. doi: 10.1097/eus.0000000000000145.
25. B. Khor, A. Gardet, and R. Xavier, "Genetics and pathogenesis of inflammatory bowel disease," Nature, vol. 474, pp. 307–317, 2011. doi: 10.1038/nature10209.
26. N. M. Salem, K. M. Jack, H. Gu, A. Kumar, M. Garcia, P. Yang, and V. Dinu, "Machine and deep learning identified metabolites and clinical features associated with gallstone disease," Computer Methods and Programs in Biomedicine Update, vol. 3, Art. no. 100106, 2023. doi: 10.1016/j.cmpbup.2023.100106.
27. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
A. Bojan, M.-C. Vladeanu, C. Pricop, I. Bararu-Bojan, C. I. Foia, S. E. Giusca, D. Iliescu, O. V. Badulescu, C. O. Iliescu Halitchi, M. A. Martu, A. Bazyani, M. Ciocoiu, and L. G. Foia, "Clinical and molecular signatures of gallbladder lesions: Insights into metabolic and inflammatory pathways," Diagnostics,

vol. 16, Art. no. 1480, 2026. doi: 10.3390/diagnostics16101480.
28. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, 2017, pp. 4765–4774.
A. Turki, A. M. Obaid, H. Bellaaj, M. Ksantini, and A. Altaee, "Gallblader Diseases Dataset," Mendeley Data, V2, 2024. doi: 10.17632/r6h24d2d3y.2.
29. Panhwar, A. O., Sathio, A. A., Lakhan, A., Umer, M., Mithiani, R. M., & Khan, S. (2022). Plant health detection enabled CNN scheme in IoT network. *International Journal of Computing and Digital Systems*, *11*(1), 344-335.
30. Awan, S. A., Abbasi , M. A., Burdi, A., Orangzeb, A., & Sathio, A. A. (2026). Vehicle to Everything (V2X) and Edge Computing: A Secure Federated learning-based Lifecycle for UAV-Assisted Vehicle Network: A Survey. *VFAST Transactions on Software Engineering*, *14*(1), 36–59. https://doi.org/10.21015/vtse.v14i1.2286
31. Awan, S. A., Khattak, M. A. K., Sathio, A. A., Farman, H., Memon, S., Nasralla, M. M., & Sedik, A. (2025). Dynamic strategy for adaptive block size optimization in blockchain technology. *Discover Sustainability*, 6(1), 849.
32. Awan, S.A., Khattak, M.A.K., Khan, A.A. *et al.* Meta-analysis of large language models: benchmarking DeepSeek-R1 against ChatGPT, Gemini, Qwen, and LLaMA. *J Big Data* **13**, 26 (2026). https://doi.org/10.1186/s40537-025-01330-3
33. Memon, S., Memon, S., Dhomeja, L. D., Sathio, A. A., Awan, S. A., & Khan, W. (2025). A Smart Cybersecurity Scheme for MIoT Systems: Simulation and Evaluation. *VFAST Transactions on Software Engineering*, *13*(1), 14-36.
34. Palupiningtyas, D., Fali, I. M., Sathio, A. A., & Ahmedº, A. (2025). Green Synergy: Transforming Competitive Advantage through the Integration of Green HRM and Performance Systems in the Asian Hospitality Industry.
35. Panhwar, A. O., Sathio, A. A., Shah, N. M., & Memon, S. (2025). A scheme based on deep learning for fruit classification. *Mehran University Research Journal of Engineering and Technology*, *44*(1), 8-19.
36. Vavekanand, R., Sathio, A.A. & Sultani, M. A federated learning framework for deep imputation of missing data in heterogeneous ICU time series. *Sci Rep* **16**, 18097 (2026). https://doi.org/10.1038/s41598-026-49311-7
37. Shah, H., Ahmed, S., Sathio, A. A., & Burdi, A. (2023). W-rank: A keyphrase extraction method for webpage based on linguistics and DOM-base features. *VAWKUM Transactions on Computer Sciences*, *11*(1), 217-228.
38. Sathio, A. A., & Maheshwari, C. K. (2026). Neuromorphic Computing Chips for Edge AI: A Comprehensive Analysis of Brain-Inspired Hardware Architecture for Real-Time Intelligent Systems. *International Journal of Sustainable Engineering Innovations*, *2*(1), 8-14.
39. Shaikh, Z. A., Wagan, A. A., Laghari, A. A., Ali, K., Memon, M. A., & Sathio, A. A. (2019). The role of software configuration management and capability maturity model in system quality. *IJCSNS*, *19*(11), 114.
40. Vavekanand, R., Sathio, A. A., Singh, V., & Anwar, S. (2024). Water4. 0: An Industrial Water Pollution Forecasting Using Machine Learning. *Available at SSRN 4849924*.