ISSN 2723-6471

Multi Domain Feature Fusion and Boosting Based Learning for Robust Gallbladder Ultrasound Image Classification

Gede Angga Pradipta[1,*], Pharan Chawaphan[2], Sutikno[3], Putu Desiana Wulaning Ayu[4], Dandy Pramana Hostiadi[5], Made Liandana[6]

[1,5] *Department of Magister Information Systems, Institut Teknologi Dan Bisnis STIKOM Bali, Indonesia*

[2]*Mechatronics Engineering Department, Faculty of Technical Education, Rajamanggala University of Technology Thanyaburi, Thailand*

[3]*Department of Informatics, Faculty of Science and Mathematics, Diponegoro University, Indonesia*

[4]*Information Technology Department, Politeknik Negeri Bali, Indonesia*

[6]*Department of Informatics and Computer, Institut Teknologi Dan Bisnis STIKOM Bali, Indonesia*

(Received: February 1, 2026; Revised: April 5, 2026; Accepted: July 10, 2026; Available online: August 9, 2026)

Abstract

Accurate identification of gallbladder conditions is critically important in the medical sector, as early detection of diseases such as gallstones, cholecystitis, carcinoma, and polyps can substantially improve treatment outcomes, reduce complications, and guide timely surgical or therapeutic interventions. Existing literature on gallbladder disease classification still presents notable gaps. Most prior works rely solely on single-domain feature extraction either deep CNN-based spatial descriptors or handcrafted statistical/texture features without exploiting the complementary strengths of multi-domain feature fusion. This study addresses these gaps by proposing a hybrid framework that combines advanced preprocessing, multi-domain feature fusion, feature selection, and ensemble classification. The preprocessing pipeline applies Non-Local Means (NLM) denoising, Contrast Limited Adaptive Histogram Equalization (CLAHE), frequency-domain low-pass filtering, and Gabor filtering to enhance image quality and highlight diagnostically relevant structures. Features are extracted by fusing deep spatial descriptors from a pretrained Inception V3 network with handcrafted statistical and texture-based features, including Gray Level Dependency Matrix (GLDM) measures. Dimensionality reduction is performed using ANOVA k-best selection to retain the most discriminative attributes. The refined features are classified using LightGBM, XGBoost, Histogram Gradient Boosting, and AdaBoost, enabling a comprehensive performance comparison. Experiments conducted on the balanced UIdataGB dataset (10,692 annotated images across nine diagnostic categories) demonstrate that LightGBM, XGBoost, and Histogram Gradient Boosting achieve near-perfect performance, with accuracies exceeding 98.6% and AUC values of 0.98 across all classes, while AdaBoost shows markedly lower discriminative capability. The results suggest that gradient boosting approaches are a promising option for multi-class gallbladder disease detection, particularly when combined with multi-domain feature fusion and appropriate preprocessing and feature selection techniques.

Keywords: Gallbladder, Ultrasound Imaging, Multi-Domain Feature Fusion, Boosting Algorithms, Feature Selection

1. Introduction

The gallbladder (GB) is a small, pear-shaped organ located just beneath the liver in the upper right region of the abdomen. The main role of the gallbladder is to store bile, a liver-produced fluid that facilitates the digestion and absorption of fats in the small intestine, supports cholesterol elimination, and exhibits antimicrobial activity [1]. The bile stored in the gallbladder becomes concentrated, making it more effective in digesting fats when released into the small intestine after eating, particularly after the consumption of fatty foods. The gallbladder connects to the liver and small intestine through the bile ducts, and it plays a key role in efficient digestion, although the body can function without it. Common conditions affecting the gallbladder include gallstones, which are hardened deposits of bile components that can block the flow of bile and cause pain or infection. Cholecystitis, or inflammation of the gallbladder, is another common condition, often caused by gallstones. Other less frequent conditions include gallbladder polyps, gallbladder cancer, and biliary dyskinesia, a condition where the gallbladder does not empty bile

***Corresponding author: Gede Angga Pradipta (angga_pradipta@stikom-bali.ac.id)**
DOI: https://doi.org/10.47738/jads.v7i3.

properly. In cases of chronic gallbladder issues, surgical removal (cholecystectomy) is often recommended, and the body adjusts by having bile flow directly from the liver to the small intestine. Ultrasound imaging is widely regarded as the primary diagnostic modality for gallbladder diseases due to its non-invasive nature, cost-effectiveness, accessibility, and ability to provide real-time visualization [2], [3]. It uses sound waves to create images of the gallbladder, allowing doctors to detect gallstones, inflammation, and other abnormalities. During the ultrasound, a technician moves a transducer across the abdomen, which sends sound waves to the gallbladder and captures the returning echoes to form an image. This diagnostic tool is especially helpful for visualizing gallstones, gallbladder wall thickening (which can indicate inflammation), and other structural changes like polyps or tumours. Ultrasound screening is safe, quick, and often used as the first-line diagnostic test when gallbladder disease is suspected. However, the accurate interpretation of ultrasound images depends heavily on the expertise of radiologists and may vary between observers, underscoring the need for automated approaches in medical image analysis [4], [5].

Early identification of bladder health conditions is crucial, as medical intervention is often required to prevent potential adverse outcomes for the patient. 2-D images of the gallbladder organ sometimes exhibit variations and conditions that require precision and accuracy from medical personnel for diagnosis and poor medical outcomes and improved patient symptoms may result from errors or delays in diagnosis [6], [7]. Subjectivity and human error remain key issues in the analysis of gallbladder conditions by medical professionals. Therefore, it is essential to develop an artificial intelligence-based model to support the analysis and diagnosis of gallbladder conditions. Despite its widespread use, the interpretation of ultrasound images remains highly dependent on the expertise of radiologists and sonographers. Manual diagnosis may suffer from inter-observer variability and subjectivity, especially when the visual differences between pathological categories are subtle. These limitations have motivated the development of computer-aided diagnosis (CAD) systems that aim to assist clinicians by providing automated and objective analysis of ultrasound images. However, automated gallbladder disease classification from ultrasound images presents several significant computational challenges. Ultrasound images are inherently affected by speckle noise, low contrast, and imaging artifacts, which can obscure important anatomical structures and degrade feature quality. In addition, different gallbladder diseases often exhibit subtle variations in texture, shape, and echogenic patterns, making it difficult for machine learning models to distinguish between classes.

Another challenge lies in the high intra-class variability and inter-class similarity among ultrasound images, which increases the difficulty of reliable classification. Furthermore, many existing studies rely primarily on single-domain feature extraction approaches, either deep convolutional neural network (CNN) features or handcrafted texture descriptors. While deep CNN models are effective in learning spatial representations, they may overlook certain statistical or textural characteristics that are diagnostically relevant in ultrasound images. Conversely, handcrafted features may capture local texture patterns but lack the ability to model complex spatial hierarchies. Consequently, relying on a single feature domain may limit the discriminative capability of the classification model. To address these challenges, this study proposes a hybrid gallbladder ultrasound image classification framework that integrates advanced preprocessing techniques, multi-domain feature fusion, feature selection, and ensemble boosting classifiers. The proposed approach combines deep spatial features extracted from a pretrained Inception V3 network with handcrafted statistical and texture-based descriptors, statistical based feature and Gray Level Dependency Matrix (GLDM) features. By integrating complementary feature domains and applying ANOVA-based feature selection to retain the most informative attributes, the proposed framework aims to enhance the robustness and discriminative capability of the classification model. The selected features are then evaluated using several boosting-based classifiers, including LightGBM, XGBoost, Histogram Gradient Boosting, and AdaBoost, to identify the most effective learning strategy for gallbladder disease classification.

2. Literature Review

Early detection plays a crucial role in secondary prevention, which focuses on early diagnosis and intervention to improve patient outcomes and quality of life. By employing advanced technologies, such as artificial intelligence (AI), researchers aim to identify diseases before they reach critical stages. Several studies have investigated convolutional neural network (CNN)-based approaches for liver steatosis assessment in ultrasound imaging [8] [9]. These studies aimed to improve diagnostic accuracy by reducing reliance on subjective human interpretation. One approach applied

transfer learning with a pretrained deep CNN to extract grayscale and texture descriptors from B-mode ultrasound images, while another developed a cascaded deep neural network architecture to sequentially process liver regions and refine classification performance. Both studies utilized large ultrasound datasets and reported improvements in sensitivity and specificity compared with manual diagnosis, including stronger correlation with biopsy results and reduced false-positive rates. These findings demonstrate the potential of CNN-based models as a robust baseline for organ-level disease detection in ultrasound imaging, which is also relevant to gallbladder image analysis. In another ultrasound application, automated grading of prenatal hydronephrosis was performed using segmented kidney ultrasound images to address variability in radiologist-based grading [10]. The method incorporated a segmentation stage to isolate kidney regions, followed by a CNN classifier, achieving classification accuracy above 90%. Collectively, these studies demonstrate the potential of CNN-based models as robust baselines for organ-level disease detection and grading in ultrasound imaging, which is also relevant to gallbladder image analysis. Furthermore, a deep learning-based decision support system has been proposed for renal cancer diagnosis using CNN-based feature extraction combined with fully connected classification layers [11] . The reported results on renal ultrasound datasets, with accuracies exceeding 93%, indicate that deep models are capable of handling variations in echotexture and anatomical distortion.

Handling dataset imbalance remains a critical challenge in ultrasound image analysis, as medical datasets often exhibit uneven class distributions due to the rarity of certain pathological conditions. This issue is particularly prominent in ultrasound imaging, where data acquisition depends heavily on clinical availability and expert annotation. Several studies have addressed this problem by integrating data-level and algorithm-level strategies to improve classification performance. A systematic study has demonstrated that class imbalance significantly degrades model performance, particularly in multi-class classification tasks, emphasizing the importance of resampling strategies such as oversampling and class reweighting to improve minority class representation [12]. In the context of ultrasound imaging, weakly supervised data augmentation frameworks that integrate grayscale and Doppler ultrasound images have been proposed to improve classification and localization performance. Such approaches utilize Doppler-derived spatial priors to guide augmentation, effectively reducing background noise and enhancing feature learning, particularly in scenarios with limited and heterogeneous datasets [13]. In addition, Generative Adversarial Networks (GANs) have been widely employed to synthesize realistic ultrasound images, thereby improving classification accuracy through increased data diversity. This approach highlights the potential of synthetic data generation in overcoming the limitations of small and imbalanced medical datasets [14]. Moreover, hybrid strategies that combine preprocessing, augmentation, and balanced learning have shown promising results in ultrasound-based diagnosis. Recent advances further extend this direction by integrating optimization techniques into generative frameworks. For example, optimization-driven generative models such as GAN-PSO combine Generative Adversarial Networks with Particle Swarm Optimization to automatically optimize hyperparameters and improve the quality of generated medical images. These approaches have been shown to enhance structural similarity and perceptual realism of synthetic data while also improving downstream classification performance in medical imaging tasks [15].

Image enhancement and handcrafted feature extraction have also contributed significantly to ultrasound-based diagnostic systems. Enhancement filters, including unsharp masking, Wiener filtering, and median filtering, have been examined to improve ultrasound image clarity for gastric disease detection, with unsharp masking showing the strongest ability to enhance diagnostically relevant structures [16]. Such enhancement strategies are directly relevant to gallbladder ultrasound preprocessing, where contrast improvement is essential for reliable feature extraction. Other studies have employed shape descriptors, statistical texture features, and classical classifiers such as support vector machines to detect abnormalities in ultrasound images [17] . Wavelet transforms combined with active contour segmentation have also been used to classify focal and diffuse liver diseases into multiple categories, achieving 91% accuracy [18]. In breast lesion detection, texton-based features, local configuration patterns, and statistical measures have produced accuracy, sensitivity, and specificity above 96% [19]. These findings indicate that handcrafted descriptors remain valuable, particularly when integrated with robust segmentation and carefully designed preprocessing pipelines. Recent developments further demonstrate the adaptability of deep learning across various ultrasound tasks, including recognition, segmentation, and classification. End-to-end CNN-based feature extraction has been shown to be highly transferable across different organs and imaging contexts [20]. In liver cirrhosis diagnosis, an end-to-end deep learning framework integrated preprocessing and classification into a single pipeline, improving

diagnostic automation from ultrasound images [21]. In addition, synthetic aperture ultrasound data enhancement combined with deep learning has been applied to improve image resolution and contrast, addressing inherent limitations of ultrasound acquisition devices [22]. These studies support the use of specialized deep architectures for gallbladder imaging, where low contrast, speckle noise, anatomical variability, and limited resolution can affect classification performance.

More specifically, several studies have focused on biliary and gallbladder-related disease diagnosis using machine learning and deep learning approaches. A multi-omics machine learning pipeline has been used to analyze human bile samples from benign and malignant biliary strictures, enabling classification of cancer risk profiles [23]. In gallstone analysis, CNN models have been developed to predict chemical composition using IoT-acquired medical data, supporting non-invasive diagnostic prediction [24] . For gallbladder carcinoma, tumour marker detection combined with a genetic algorithm-optimized backpropagation neural network has demonstrated high diagnostic precision and prognostic value [25]. In biliary atresia diagnosis, ensemble CNN models using gallbladder ultrasound images achieved high patient-level performance, with specificity of 93.9% and sensitivity of 93.1%, outperforming human experts [26]. Similar deep learning approaches for biliary atresia detection from gallbladder images also reported sensitivity and specificity above 90% [27] . These models highlight the potential of AI-assisted ultrasound analysis for early detection of rare but clinically critical gallbladder-related conditions. Segmentation, object detection, video-based analysis, and ensemble learning have further strengthened gallbladder ultrasound research. A fire-controlled multi-scale pulse-coupled neural network was proposed for gallbladder segmentation and showed better anatomical boundary preservation than conventional OTSU and iterative segmentation methods [28]. For gallbladder cancer detection, focused masked autoencoders have been applied to ultrasound video sequences, improving temporal feature learning compared with static image-based analysis [29]. Object detection-based approaches have also been explored by fusing Faster R-CNN and YOLOv8 outputs, achieving 92.62% classification accuracy for gallbladder cancer and outperforming either model individually [30]. In gallbladder polyp classification, an ensemble CNN combining ResNet, InceptionV3, and DenseNet achieved an AUC of 0.9082 and an accuracy of 87.61% in distinguishing true polyps from pseudo polyps smaller than 20 mm [31]. Furthermore, GBCNet with multi-scale second-order pooling was introduced to emphasize regions of interest in noisy ultrasound images, improving robustness against ultrasound artifacts [32].

Despite the significant progress in ultrasound-based disease classification, several critical limitations remain. First, most existing studies rely on single-domain feature representations, either deep spatial features extracted from convolutional neural networks (CNNs) or handcrafted statistical and texture descriptors. While deep learning models are effective in capturing high-level semantic patterns, they often overlook fine-grained statistical and textural characteristics that are diagnostically relevant in ultrasound images. Conversely, handcrafted features are limited in representing complex spatial structures. This separation leads to suboptimal feature representation and reduced discriminative capability, particularly in multi-class classification tasks involving subtle inter-class variations. Second, although preprocessing techniques such as noise reduction, contrast enhancement, and segmentation have been widely explored, they are often applied independently rather than as part of an integrated pipeline. The lack of a unified preprocessing strategy limits the ability to systematically enhance feature quality across different stages of the analysis. Third, while ensemble learning methods particularly boosting algorithms have demonstrated strong performance in general machine learning tasks, their role in multi-domain feature fusion for medical ultrasound classification remains insufficiently explored. Most prior works either rely on a single classifier or do not provide a comparative evaluation of multiple boosting strategies within a unified framework. To address these limitations, this study proposes a multi-domain feature fusion framework for gallbladder ultrasound image classification that integrates complementary information from multiple feature representations within a unified pipeline.

The main contributions of this study lie in the development of a comprehensive framework for gallbladder ultrasound image classification that integrates multiple complementary components. First, the study introduces a multi-domain feature representation by combining deep spatial features extracted from a pretrained InceptionV3 model with handcrafted statistical descriptors and GLDM-based texture features, enabling a more complete characterization of ultrasound image properties. In addition, an integrated preprocessing pipeline is proposed, incorporating Non-Local Means (NLM) denoising, CLAHE contrast enhancement, frequency-domain filtering, and Gabor-based texture

ISSN 2723-6471

enhancement to systematically improve image quality prior to feature extraction. To address the high dimensionality of the fused feature space, the framework applies ANOVA-based feature selection to retain the most discriminative features while reducing redundancy and noise. Furthermore, the study performs a systematic evaluation of multiple boosting-based classifiers, including LightGBM, XGBoost, Histogram Gradient Boosting, and AdaBoost, within a unified framework to identify the most effective classification strategy for gallbladder ultrasound data. Table 1 presents a summary of previous studies on ultrasound-based disease detection, categorized by methodological approach, including CNN-based classification, dataset handling strategies, image enhancement techniques, and gallbladder-specific AI applications. Based on the identified gap, this study proposes a multi-domain feature fusion framework that combines Inception V3-based deep spatial descriptors, statistical features, and texture-based measures following three-stage image enhancement using Non-Local Means denoising, CLAHE contrast enhancement, and frequency-domain and Gabor filtering. Dimensionality reduction is performed using ANOVA k-best feature selection to retain the most discriminative features. The refined feature set is then classified using multiple boosting algorithms LightGBM, XGBoost, Histogram Gradient Boosting, and AdaBoost allowing for a rigorous comparative analysis. This integrated framework addresses limitations in prior work by enhancing feature diversity, reducing redundancy, and systematically benchmarking state-of-the-art boosting methods for robust gallbladder ultrasound image classification.

Table 1. Thematic Summary of Previous Studies in Ultrasound-Based Disease Detection

Category	Study	Method	Task	Key Result
CNN-based ultrasound classification	Byra et al. [8]	Transfer learning CNN	Liver steatosis	Improved diagnostic accuracy
	Rhyou & Yoo [9]	Cascaded DNN	Liver steatosis	Reduced false positives
	Mahmud et al. [10]	CNN + segmentation	Hydronephrosis grading	>90% accuracy
	Tounsi et al. [11]	CNN decision support	Renal cancer	Captured echotexture patterns
Dataset handling & imbalance	Buda et al. [12]	Oversampling, class weighting	Multi-class image classification (general DL)	Resampling improves minority class representation and overall accuracy
	Keatmanee et al. [13]	Weakly supervised data augmentation with grayscale + doppler integration	Ultrasound image classification & localization	Improved feature learning and noise reduction using Doppler-guided augmentation, particularly effective for limited and heterogeneous datasets
	Frid-Adar et al. [14]	GAN-based synthetic data augmentation	Liver lesion classification (ultrasound)	improved CNN classification performance through realistic synthetic image generation
	Rais et al. [15]	GAN-PSO (GAN + Particle Swarm Optimization)	Medical image augmentation & classification	improved downstream classification via optimized generative process
Image enhancement & feature engineering	Arora & Mittal [16]	Filtering methods	Gastric ultrasound	Effective enhancement
	Glory Precious & Selvan [17]	Texture + SVM	Breast cancer	Effective handcrafted features
	Buettner et al. [18]	Wavelet + active contour	Liver disease	91% accuracy
	Acharya et al. [19]	Texton features	Breast lesion detection	>96% accuracy
Deep learning frameworks	Wang et al. [20]	Review of DL in ultrasound	Multiple tasks	CNN adaptability
	Yang et al. [21]	End-to-end DL	Cirrhosis diagnosis	Integrated pipeline

	Mao et al. [22]	Synthetic aperture + DL	Image enhancement	Improved resolution
	Urman et al. [23]	Multi-omics ML	Biliary stricture	Cancer risk prediction
	Yao et al. [24]	CNN + IoT data	Gallstone composition	Non-invasive analysis
	Chang et al. [25]	GA-NN	Gallbladder carcinoma	High precision
	Zhou et al. [26]	Ensemble CNN	Biliary atresia	Outperformed experts
	Obaid et al. [33]	CNN models	Gallbladder classification	>90% sensitivity
	Cao et al. [28]	CFC-MSPCNN	Gallbladder segmentation	Better boundary preservation
Gallbladder-specific AI	Basu et al. [29]	Masked autoencoder	Cancer detection (video)	Uses temporal information
	Dadjouy & Sajedi [30]	YOLOv8 + Faster R-CNN	Cancer detection	Improved detection
	Kim et al. [31]	CNN + metadata	Gallbladder Polyp classification	AUC 0.908
	Basu et al. [32]	GBCNet	Gallbladder Cancer detection	Robust ROI learning

3. Methodology

The proposed model presents a hybrid framework for ultrasound image classification, encompassing preprocessing, multi-domain feature extraction, feature selection, and ensemble classification using boosting algorithms. The preprocessing stage enhances image quality through speckle noise reduction, contrast enhancement, frequency domain filtering, and Gabor filtering. Combination of deep spatial features (Inception V3), statistical descriptors, and texture-based features (GLDM) is extracted to represent image content comprehensively. To reduce dimensionality and retain discriminative features, ANOVA-based feature selection is applied. For classification, several state-of-the-art boosting algorithms namely AdaBoost, Histogram Gradient Boosting, XGBoost, and LightGBM are evaluated to determine the most effective learner for this task. Performance is assessed using standard metrics derived from confusion matrices, including accuracy, precision, recall, and F1-score, providing a rigorous comparative analysis of boosting techniques in medical image classification. Figure 1 present the overall process diagram of proposed gallbladder US image classification model.

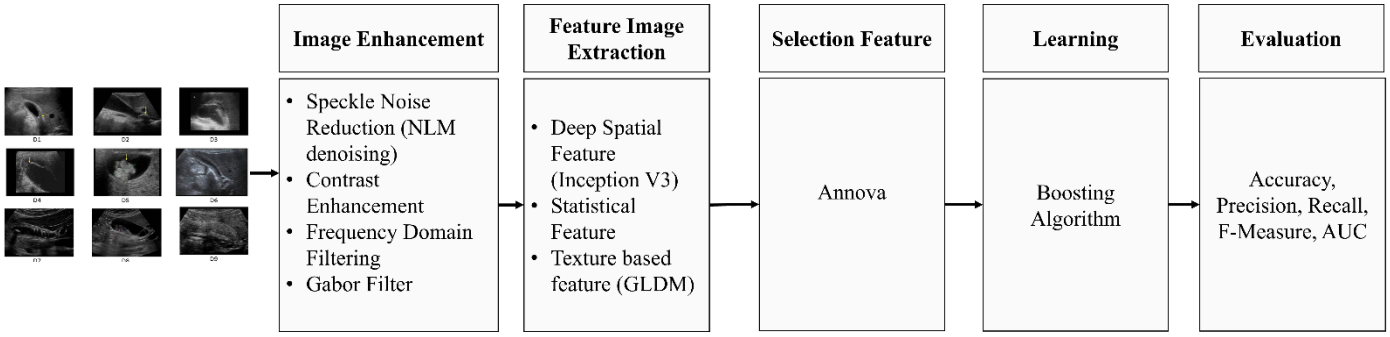

Figure 1. Process Diagram of gallblader US image classification

3.1. Non-Local Means (NLM) Denoising

Non-Local Means (NLM) is an image denoising technique that removes noise while preserving important structural details by exploiting the redundancy of similar image patches within an image. Unlike conventional denoising methods that rely only on local neighborhood statistics, NLM computes the denoised pixel intensity as a weighted average of

pixels that exhibit similar local patterns across the entire image. Let $I(p)$denote the intensity value of a pixel located at position $p$in the input grayscale image. The denoised intensity value at pixel p, denoted as $\hat{I}(p)$, is computed as a weighted average of the intensities of pixels $q$within a predefined search window Ω. The denoising process can be expressed as

$$\hat{I}(p) = \sum_{q \in \Omega} w(p,q)\, I(q) \tag{1}$$

The weight $w(p,q)$ represents the similarity between the neighborhoods centered at pixels $p$and q. These weights are non-negative and normalized such that their sum equals one, ensuring that the output intensity remains within a valid range. The similarity weight between pixels $p$and $q$is calculated based on the similarity between two square patches centered at those pixels. Let $P(p)$and $P(q)$ denote the image patches centered at pixels p and q, respectively. The weight function is defined as

$$w(p,q) = \frac{1}{Z(p)} \exp\left(-\frac{\| P(p) - P(q) \|^2}{h^2}\right) \tag{2}$$

In this formulation, $\| P(p) - P(q) \|^2$ represents the squared Euclidean distance between the two image patches, which measures their structural similarity. The parameter $h$is a filtering parameter that controls the degree of smoothing, where larger values result in stronger noise suppression but may slightly blur image details. The term $Z(p)$ is a normalization constant that ensures the weights sum to one and is computed as the sum of all exponential similarity terms within the search window. Through this formulation, pixels with similar neighborhood structures contribute more strongly to the denoised value, while dissimilar pixels receive lower weights.

3.2. Contrast Limited Adaptive Histogram Equalization (CLAHE)

CLAHE improves the local contrast of an image while preventing excessive amplification of noise in homogeneous regions. The method operates on small image tiles where histogram equalization is performed independently with a clipping mechanism to limit contrast amplification. Let $H(i)$ denote the histogram count at intensity level i. The clipped histogram $H_c(i)$is defined as

$$H_c(i) = \begin{cases} H(i), & \text{if } H(i) \le T_c \\ T_c, & \text{if } H(i) > T_c \end{cases} \tag{3}$$

T_c represents the clip limit, which restricts the maximum histogram bin height to avoid over amplification of noise. The total number of clipped histograms counts that must be redistributed is given by

$$E = \sum_{i=0}^{L-1} \max(0, H(i) - T_c) \tag{4}$$

Edenotes the excess histogram counts resulting from clipping, and L represents the number of gray levels in the image. The excess values are redistributed evenly across all histogram bins before computing the cumulative distribution function used for contrast enhancement.

3.3. Frequency Domain Transformation (2D FFT)

The two-dimensional discrete Fourier transform (2D-DFT) converts the spatial-domain representation of an image into the frequency domain, enabling the analysis and filtering of image frequency components. For a grayscale image $f(x,y)$of size $M \times N$, the 2D-DFT is defined as

$$F(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y)\, e^{-j2\pi\left(\frac{ux}{M} + \frac{vy}{N}\right)} \tag{5}$$

$f(x,y)$represents the intensity value at spatial coordinates (x, y), $F(u,v)$represents the frequency-domain coefficient at frequency coordinates (u, v), and $j$denotes the imaginary unit. The indices $M$and Ncorrespond to the image dimensions in the horizontal and vertical directions, respectively.

3.4. Gabor Filter

Gabor filters are widely used for texture analysis because they can capture spatial frequency and orientation information simultaneously. A two-dimensional Gabor filter is defined as a sinusoidal plane wave modulated by a Gaussian envelope. The mathematical formulation of the 2D Gabor filter is given by

$$G(x,y) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right)\cos\left(2\pi\frac{x'}{\lambda} + \phi\right) \quad (6)$$

The rotated coordinates $x'$and $y'$are defined as

$$x' = x\cos\theta + y\sin\theta$$
$$y' = -x\sin\theta + y\cos\theta$$

In this formulation, λ represents the wavelength of the sinusoidal component, θ denotes the orientation of the filter, σ controls the width of the Gaussian envelope, γ defines the spatial aspect ratio, and ϕ represents the phase offset. By varying the orientation and wavelength parameters, Gabor filters can effectively capture directional texture information in ultrasound images.

3.5. Deep Spatial Feature Extraction Using Inception V3 Architecture

Deep convolutional neural networks (CNNs) have significantly advanced medical image analysis, including ultrasound-based disease detection. Among them, Inception V3, introduced by Szegedy et al., is a widely used CNN architecture known for its efficient multi-scale feature extraction through parallel convolutional branches and factorized convolutions. Instead of relying on a single receptive field size, the Inception architecture processes input features through multiple convolutional paths with different kernel sizes, allowing the model to capture both fine-grained texture patterns and broader anatomical structures. Such multi-scale representations are particularly useful in ultrasound imaging, where tissue textures and lesion boundaries often appear subtle and heterogeneous. In this study, Inception V3 is employed solely as a deep feature extractor rather than an end-to-end classifier. Ultrasound images are first resized to 299 × 299 pixels to match the network input requirements. Since the original ultrasound images are grayscale, they are converted into pseudo-RGB format by replicating the grayscale channel across three channels.

The images are then passed through the pretrained Inception V3 network initialized with ImageNet weights, with all parameters frozen to ensure consistent feature extraction across experiments. Deep feature representations are extracted from the Mixed10 layer, corresponding to the final Inception block before the classification head. This layer produces feature maps of size 8 × 8 × 2048, which capture high-level semantic representations of the input image. To obtain a compact representation suitable for machine learning classifiers, Global Average Pooling (GAP) is applied across the spatial dimensions, resulting in a 2048-dimensional feature vector for each ultrasound image. No additional convolutional or fully connected layers are applied after this stage. These deep features are subsequently integrated with handcrafted statistical and GLDM texture descriptors as part of the proposed multi-domain feature fusion framework. The overall architecture of Inception V3 used for feature extraction is illustrated in figure 2, while the layer-wise configuration relevant to the feature extraction process is summarized in table 2.

Figure 2. Inception V3 architecture used for feature extraction.

 ISSN 2723-6471

Table 2. Layer-Wise Architecture of Inception V3 Used for Feature Extraction

Stage	Layer / Module	Key Operations	Output Size
Input	Input Layer	Resize image	299 × 299 × 3
	Conv 3×3 (stride 2)	Low-level feature extraction	149 × 149 × 32
	Conv 3×3		147 × 147 × 32
	Conv 3×3		147 × 147 × 64
Stem	Max Pool 3×3 (stride 2)	Downsampling	73 × 73 × 64
	Conv 1×1	Channel expansion	73 × 73 × 80
	Conv 3×3		71 × 71 × 192
	Max Pool 3×3 (stride 2)	Downsampling	35 × 35 × 192
Inception-A	Mixed0–Mixed2	Multi-scale parallel convolutions	35 × 35 × 256
Reduction-A	Mixed3	Strided conv + pooling	17 × 17 × 768
Inception-B	Mixed4–Mixed7	Factorized conv (1×7, 7×1)	17 × 17 × 768
Reduction-B	Mixed8	Spatial reduction	8 × 8 × 1280
Inception-C	Mixed9–Mixed10	High-level semantic extraction	8 × 8 × 2048
Pooling	Global Average Pooling	Spatial aggregation	2048-D vector

3.6. Statistical-Based Feature Extraction in Grayscale Ultrasound Images

Statistical-based feature extraction is a fundamental technique in grayscale ultrasound image analysis, offering a computationally efficient and interpretable means of summarizing image characteristics. This method involves the computation of various statistical measures that describe the distribution and variability of pixel intensities within an image or a region of interest (ROI). Common first-order statistical features include the mean, which provides the average intensity level, and the standard deviation, which quantifies contrast or intensity variation. Higher-order moments such as skewness and kurtosis describe the asymmetry and peaked Ness of the intensity distribution, respectively. Entropy is another widely used feature that measures the randomness or texture complexity in an image, making it particularly useful in distinguishing tissue structures with irregular patterns. Variance, closely related to standard deviation, further captures the degree of spread in intensity values. Additional features such as the minimum and maximum intensities help define the dynamic range of the grayscale spectrum.

3.7. Texture-Based Feature Extraction Using Gray Level Dependency Matrix (GLDM)

Texture-based feature extraction is a critical component in image analysis, particularly for medical imaging applications such as ultrasound, where texture plays a vital role in characterizing tissue structures. One effective method for texture quantification is the Gray Level Dependency Matrix (GLDM). Unlike traditional second-order methods such as the Gray-Level Co-occurrence Matrix (GLCM), which measure the frequency of pixel pair relationships, GLDM assesses the spatial dependencies of gray levels by evaluating how often a pixel with a certain intensity value depends on neighbouring pixels within a predefined distance. In GLDM, for a given image, the dependency of each pixel is calculated based on the number of neighbouring pixels that have gray level values within a defined intensity difference (threshold). These dependencies are computed across various directions and distances to capture anisotropic texture patterns. The resulting matrix records the count of dependencies for each combination of gray level and dependency degree, producing a compact representation of texture regularity and contrast. GLDM quantifies how many neighbouring pixels are dependent on a reference pixel, given a specified Gray-level difference threshold. For an image I of size $M \times N$, the GLDM is a matrix $P(g, d)$ where g is a gray level, d is the number of neighboring pixels (dependency) that differ from the center pixel by less than a predefined threshold α. The dependency counts for a pixel (i, j) is calculated as Eq 3.7:

$$D(i,j) = \sum_{(u,v)\in\mathcal{N}(i,j)} \delta(|I(i,j) - I(u,v)| < \alpha) \quad (7)$$

$N(i,j)$ represents the neighborhood of the pixel, and $\delta(\cdot)$ is the indicator function that returns 1 if conditions is true, otherwise 0, α is a predefined threshold. The matrix $P(g,d)$ counts how many times a pixel with gray level g has exactly d dependent neighbours in the image.

In this study, the Gray Level Difference Method (GLDM) was used to extract texture descriptors that characterize spatial intensity variations within the ultrasound images. The GLDM computation was performed using a set of commonly adopted parameter configurations in medical image texture analysis. Specifically, the pixel pair distance parameter $d$was set to 1, which allows the method to capture fine local texture variations between neighboring pixels. To account for directional texture patterns, four standard directional orientations were considered, namely 0°, 45°, 90°, and 135°. These directions correspond to horizontal, diagonal, vertical, and anti-diagonal spatial relationships between pixel pairs, respectively. In addition, the gray-level difference threshold parameter α was set to 0, which corresponds to the conventional GLDM formulation where all gray-level differences are considered without threshold filtering. The GLDM matrix was computed for each direction, and the resulting texture descriptors were averaged across the four orientations to obtain rotation-invariant texture representations. From the resulting GLDM distributions, several statistical texture features were derived, including contrast, energy, entropy, mean difference, and homogeneity, which together characterize the spatial variation and structural heterogeneity present in ultrasound images.

3.8. Feature Selection Using ANOVA K-Best

This study applies an Analysis of Variance (ANOVA) K-best feature selection method to reduce the dimensionality of the fused feature space and retain the most informative features. The fused feature vector contains 2065 features, including 2048 deep features extracted from InceptionV3, 12 statistical features, and 5 GLDM-based texture features. The ANOVA method evaluates the significance of each feature by measuring the variance between classes relative to the variance within classes. It then computes an F-score for each feature to quantify its discriminative power.. For each feature, an F-score is computed as:

$$F = \frac{\text{variance between classes}}{\text{variance within classes}} \quad (8)$$

Features with higher F-scores are considered more discriminative for distinguishing between different classes. In this study, the SelectKBest method with ANOVA F-test was used, where the scoring function is f_classif, the number of selected features (k) was empirically determined through experimentation, and the optimal value of k was found to be in the range of 400–500 features, based on classification performance analysis.

4. Results and Discussion

4.1. Dataset description

The UIdataGB dataset[34] comprises 10,692 annotated ultrasound images, carefully verified by medical doctors and experienced radiologists. The dataset is organized into nine categories based on anatomical landmarks, with each category containing approximately 1,200 images, ensuring a balanced distribution across classes. Summary of the common GB pathologies is highlighted below and presented in figure 3.

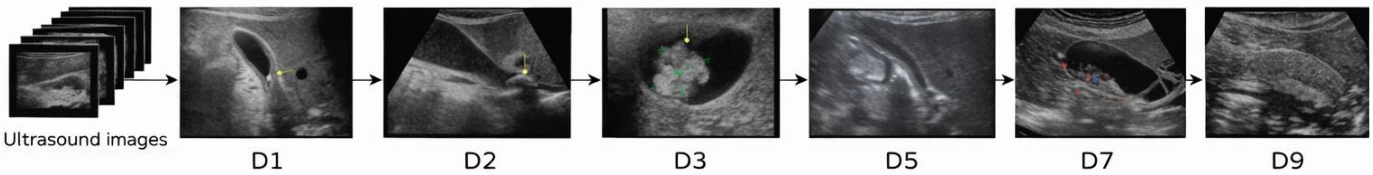

Figure 3. Gallbladder diseases (D1:D9): (D1) Gallstones, (D2) Cholecystitis, (D3) Gangrenous Cholecystitis, (D4) Perforation of GB, (D5) Polyps and Cholesterol Crystals, (D6) Gallbladder-Wall Thickening, (D7) Adenomyomas of the GB, (D8) Carcinoma, (D9) Intraabdominal and Retroperitoneum

The data was collected from 1,782 patients, including 6,246 images from female patients (average age: 63.4 years) and 4,446 images from male patients (average age: 59.6 years). The image resolution is 900×1200 pixels, and each class is stored in a separate folder labelled according to its anatomical content.

4.2. Image Enhancement and Preprocessing

Ultrasound images used in gallbladder analysis often suffer from low contrast and speckle noise, which hinder feature extraction. To address this, a three-stage image enhancement pipeline was implemented: noise reduction, contrast enhancement, and frequency domain filtering. First, speckle noise was reduced using Non-Local Means Denoising (h=10, template size=7, search window=21), with a 5×5 median filter Ultrasound images used in gallbladder analysis often suffer from low contrast and speckle noise, which hinder feature extraction. To address this, a three-stage image enhancement pipeline was implemented: noise reduction, contrast enhancement, and frequency domain filtering. First, speckle noise was reduced using Non-Local Means Denoising (h=10, template size=7, search window=21), with a 5×5 median filter used for comparison. Next, Contrast Limited Adaptive Histogram Equalization (CLAHE) improved local contrast using a clip limit of 2.0 and a tile grid size of (8,8), enhancing subtle structures without amplifying noise. In the final stage, a 2D Fast Fourier Transform (FFT) with a circular low-pass filter (radius=30 pixels) was applied to suppress high-frequency noise and highlight anatomical features. Optionally, Gabor filtering (frequency=0.6) at four orientations (0, π/4, π/2, 3π/4) was used to enhance directional structures like tissue boundaries. Figure 4 presents the results of this preprocessing pipeline on gallbladder ultrasound images.

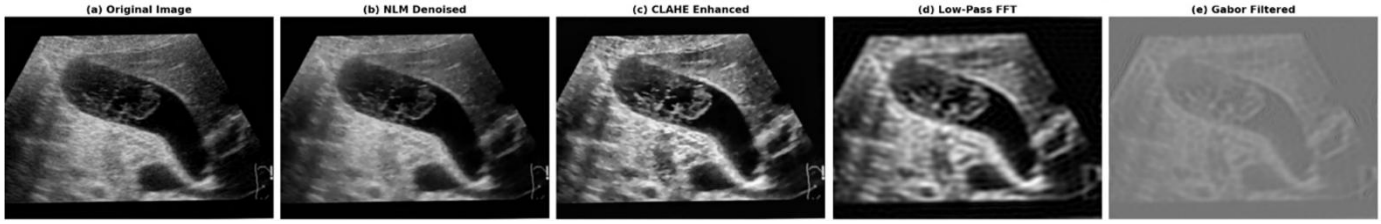

Figure 4. Gallbladder US image processing results. (a) original image, (b) NLM Denoised results, (c) CLAHE enhanced, (d) Low-pass FFT images results, (e) Gabor filtered.

To capture complementary information from ultrasound images, this study extracts three types of features: first-order statistical features, texture features based on the Gray Level Difference Method (GLDM), and deep spatial features obtained from a pretrained InceptionV3 model. Each feature category captures different characteristics of the image, ranging from global intensity distribution to spatial texture patterns and high-level semantic representations. The dimensionality of each feature subset is summarized in table 3.

Table 3. Summary of extracted feature representations and their dimensionality

Feature Category	Feature Type	Number of Features	Description
Statistical Features	First-order statistics	12	Global intensity distribution descriptors such as mean, variance, skewness, kurtosis, entropy, energy, RMS, minimum, maximum, range, standard deviation, and coefficient of variation
Texture Features	GLDM features	5	Texture descriptors derived from Gray Level Difference Method including contrast, energy, entropy, mean difference, and homogeneity
Deep Features	InceptionV3 (Global Average Pooling layer)	2048	High-level spatial features automatically extracted by the pretrained CNN capturing structural and semantic patterns
Total Features (Before Selection)	**Feature Fusion**	**2065**	**Concatenation of statistical, GLDM, and deep CNN features**

4.3. Experimental on Boosting Algorithm

The classification of gallbladder diseases using machine learning algorithms was evaluated on a comprehensive dataset consisting of 10,692 annotated ultrasound images. The dataset includes nine diagnostic categories: abdomen retroperitoneum, adenomyomatosis, carcinoma, cholecystitis, gallstones, membranous and gangrenous, perforation, polyps and cholesterol, and wall thickening. Each category contains approximately 1,200 images, resulting in a relatively balanced dataset across classes. Prior to model training, the dataset was partitioned using stratified sampling, where 80% of the images were allocated for training (8,553 images) and 20% were reserved for testing (2,139 images). The stratified split ensured that the proportional representation of each class was preserved in both subsets. It should be noted that the support values reported in the classification results correspond to the number of samples in the test set, rather than the total number of images in each class. Since approximately 20% of the images from each category were used for testing, the support values range from about 200 to 300 samples per class, which is consistent with the original class distribution. All boosting-based classifiers are implemented using fixed hyperparameter configurations, as summarized in table 4. Early stopping is enabled where applicable, and random seeds are fixed across all experiments to ensure deterministic behavior.

Table 4. Hyperparameter Settings for Boosting Classifiers

Classifier	Hyperparameters
LightGBM	num_leaves=31, max_depth=10, learning_rate=0.05, n_estimators=300, subsample=0.8, colsample_bytree=0.8, random_state=42, early_stopping_rounds=50
XGBoost	max_depth=8, learning_rate=0.05, n_estimators=300, subsample=0.8, colsample_bytree=0.8, objective=multi:softprob, eval_metric=mlogloss, seed=42, early_stopping_rounds=50
HistGradientBoosting	max_depth=10, learning_rate=0.05, max_iter=300, l2_regularization=0.01, random_state=42

The LightGBM model, configured with default parameters including a learning rate of 0.1 and 31 maximum leaves per tree, was first employed using the Gradient Boosting Decision Tree framework with multiclass classification as the objective and multi_logloss as the evaluation metric. This model demonstrated strong overall performance, achieving a high accuracy of 98.2%, and particularly excellent F1-scores for classes like membranous and gangrenous and polyps and cholesterol, both scoring a perfect 100% in recall and F1-score. However, the perforation class exhibited a relatively lower recall of 93.4%, suggesting some difficulty in detecting this more subtle and possibly visually ambiguous condition. The performance results of the LightGBM algorithm are shown in table 5.

Table 5. LightGBM Performance Report

Class/Label	Metric				
	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Support
abdomen retroperitoneum	98.4	98.2	100	99.5	234
adenomyomatosis	98.3	97.2	99.2	98.2	233
carcinoma	98.3	97.3	99.2	98.2	318
choleystitis	98.6	99.2	96.5	97.4	229
gallstones	98.6	98.3	99.2	98.2	265
membranous and gangrenous	98.2	99.3	100	100	245
perforation	98.3	97.2	93.4	95.4	213
polyps and cholesterol	98.3	100	100	100	204
wall thickening	98.5	99.2	98.2	99.4	198

In contrast, the AdaBoost classifier performed significantly worse in comparison, despite being trained on the same balanced dataset. The performance of AdaBoost was particularly weak for several clinically important classes. Notably, the adenomyomatosis class yielded a concerningly low F1-score of just 3.3%, and cholecystitis followed with an F1-

 ISSN 2723-6471

score of 17.2%, despite having moderate precision. This pattern indicates that while AdaBoost could identify some instances correctly, it failed to retrieve the majority, resulting in very low recall. Other classes such as abdomen retroperitoneum, gallstones, perforation, and wall thickening also displayed poor performance, with F1-scores ranging only between 19.2% and 26.2%, reflecting the model's general inability to handle multiclass medical image classification with sufficient sensitivity and specificity. These results clearly demonstrate the limitations of AdaBoost in dealing with complex visual patterns inherent in ultrasound imaging data. The performance results of the AdaBoost algorithm are shown in table 6.

Table 6. Adaboost Performance Report

Class/Label	Metric				
	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Support
abdomen retroperitoneum	30.2	22.3	17.3	19.2	234
adenomyomatosis	30.3	28.2	36.3	3.3	233
carcinoma	30.2	27.3	46.3	34.3	318
choleystitis	30.2	28.3	12.3	17.2	229
gallstones	30.3	24.2	16.4	19.2	265
membranous and gangrenous	34.3	35.3	50.2	41.2	245
perforation	30.2	22.7	19.2	20.2	213
polyps and cholesterol	34.3	62.1	50.2	55.3	204
wall thickening	30.3	31.4	22.2	26.2	198

Significant improvement in performance was observed when the XGBoost classifier was applied. This model yielded substantially higher predictive accuracy and generalization ability across all nine classes. F1-scores ranged impressively from 93.3% to 99.3%, including dramatic performance gains in previously challenging categories. For example, the adenomyomatosis class, which performed very poorly under AdaBoost, achieved an F1-score of 97.3%, while a more robust and dependable model for diagnostic applications. The final evaluation employed the Histogram Gradient Boosting classifier, which further solidified the findings observed with XGBoost. This model achieved a consistent overall accuracy of 98.6%, and its performance across individual classes was exceptionally uniform and stable. F1-scores across all nine categories ranged from 95.4% to 100%, with particularly outstanding results in complex and critical classes such as membranous and gangrenous and wall thickening, achieved F1-scores of 99.2%. The performance results of the XGBoost algorithm are shown in table 7.

Table 7. XGBoost Performance Report

Class/ Label	Metric				
	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Support
abdomen retroperitoneum	98.6	97.3	99.3	98.2	234
adenomyomatosis	98.6	99.2	98.2	98.2	233
carcinoma	98.6	97.4	98.4	98.2	318
choleystitis	98.6	100	95.4	97.4	229
gallstones	98.6	97.4	98.2	98.4	265
membranous and gangrenous	98.6	99.3	100	99.2	245
perforation	98.6	97.4	93.4	95.4	213
polyps and cholesterol	98.6	100	100	100	204
wall thickening	98.6	99.3	100	99.2	198

The model demonstrated high stability and generalization capabilities across diverse pathological conditions, making it a highly reliable tool for clinical decision support systems. In summary, while AdaBoost proved inadequate for the task, both XGBoost and Histogram Gradient Boosting significantly outperformed it, with Histogram Gradient Boosting offering the highest classification consistency and overall robustness. cholecystitis class improved to 96.4%, both with high precision and recall. The perforation class, which was another weak point under AdaBoost (F1-score: 22.2%), demonstrated a major recovery with an F1-score of 93.3% under XGBoost. The performance results of the Histogram Gradient Boosting algorithm are shown in table 8.

Table 8. Histogram Gradient Boosting Performance Report

Class/ Label	Metric				
	Accuracy (%)	**Precision (%)**	**Recall (%)**	**F1-score (%)**	**Support**
abdomen retroperitoneum	97.4	96.3	99.1	97.4	234
adenomyomatosis	97.4	97.2	96.3	97.3	233
carcinoma	97.4	97.2	98.5	98.4	318
choleystitis	97.4	98.3	94.2	96.4	229
gallstones	97.4	98.3	98.4	98.4	265
membranous and gangrenous	97.4	95.2	100	99.2	245
perforation	97.4	98.3	91.2	93.3	213
polyps and cholesterol	97.4	98.3	100	99.2	204
wall thickening	97.4	98.3	99.3	99.3	198

Additionally, classes like membranous and gangrenous and polyps and cholesterol, which had already performed well, continued to excel with F1-scores of 99.2% each. These results affirm that XGBoost is particularly well-suited to capturing complex patterns in medical imaging data, making it these findings emphasize the effectiveness of gradient boosting-based models in the field of ultrasound-based gallbladder disease detection and support their adoption in future computer-aided diagnosis systems. The performance of four ensemble-based classification algorithms LightGBM, XGBoost, AdaBoost, and HistGradientBoosting was also evaluated using multi-class ROC curves, as illustrated in figure 5. For each model, the ROC curve was generated for all diagnostic categories, and the corresponding Area Under the Curve (AUC) values were computed. LightGBM, XGBoost, and HistGradientBoosting demonstrated near perfect classification performance, with all diagnostic classes achieving an AUC of 1.00, indicating flawless separation between the positive and negative classes across all decision thresholds. The ROC curves for these models were tightly clustered near the top left corner of the plot, reflecting high sensitivity and specificity for all classes. In contrast, AdaBoost exhibited substantially lower discriminative capability, with AUC values ranging from 0.66 to 0.95 depending on the class. Notably, classes such as adenomyomatosis (AUC = 0.68), carcinoma (AUC = 0.66), perforation (AUC = 0.66), and wall thickening (AUC = 0.69) showed considerably reduced performance, suggesting difficulty in accurately distinguishing these conditions.

ISSN 2723-6471

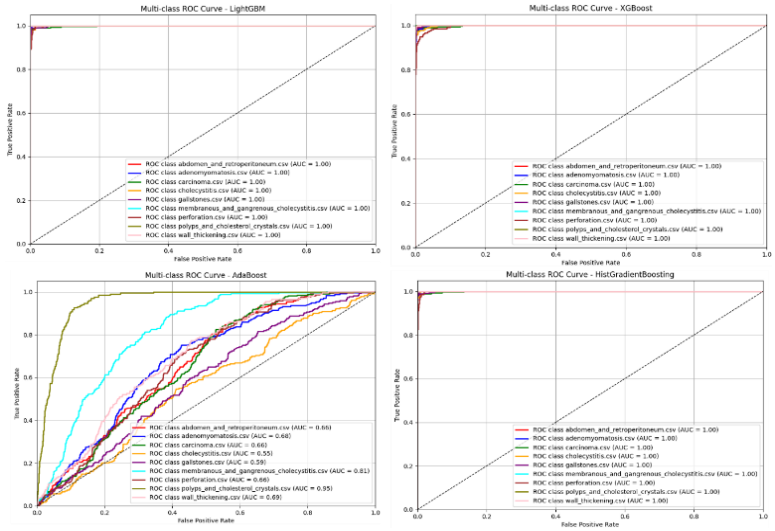

Figure 5. Multi-class Receiver Operating Characteristic (ROC) curves with Area Under the Curve (AUC) values for LightGBM, XGBoost, AdaBoost, and HistGradientBoosting classifiers across all diagnostic categories.

Meanwhile, gallstones (AUC = 0.95) and polyps with cholesterol crystals (AUC = 0.95) retained relatively strong classification accuracy. The stark performance disparity between AdaBoost and the gradient boosting variants (LightGBM, XGBoost, HistGradientBoosting) indicates that the latter approaches are more effective at capturing the complex, non-linear patterns present in the dataset. This is likely due to their ability to optimize leaf wise growth strategies (LightGBM) or histogram-based binning (HistGradientBoosting), which enhance feature discrimination in high dimensional spaces. To further validate the robustness of the proposed classification framework and address concerns regarding the near-perfect AUC values, an additional experiment was conducted using 5-fold stratified cross-validation on the training dataset as shown in table 9. This experiment specifically focused on the LightGBM classifier, which demonstrated the highest performance among the evaluated models in the initial experiments. In this procedure, the training dataset was partitioned into five folds while preserving the class distribution using stratified sampling. During each iteration, four folds were used for model training, and the remaining fold was used for validation. This process was repeated five times so that each fold served as the validation set once. For each fold, the same preprocessing pipeline and feature extraction procedures were applied, including Non-Local Means denoising, CLAHE contrast enhancement, frequency-domain filtering, and Gabor filtering, followed by the extraction of deep features from the pretrained InceptionV3 network (2048 features) together with handcrafted statistical and GLDM texture features.

The fused feature representation was then refined using ANOVA k-best feature selection before being classified using the LightGBM algorithm. The evaluation metrics, including accuracy, precision, recall, F1-score, and AUC, were computed for each fold. The results demonstrate consistently high performance across all folds with only minor variations, indicating that the proposed framework generalizes well across different data partitions and does not rely on a specific train-test split.

Table 9. 5-Fold Cross-Validation Performance of LightGBM algorithm

Fold	Accuracy	Precision	Recall	F1-Score	AUC
Fold 1	0.989	0.989	0.988	0.988	0.999
Fold 2	0.991	0.971	0.991	0.991	1
Fold 3	0.992	0.991	0.989	0.989	0.999
Fold 4	0.982	0.982	0.991	0.991	1
Fold 5	0.985	0.989	0.989	0.989	0.999
Average	0.990 ± 0.001	0.990 ± 0.001	0.989 ± 0.001	0.989 ± 0.001	0.999 ± 0.001

ISSN 2723-6471

Furthermore, figure 6 illustrates the relationship between the number of features selected using the ANOVA k-best method and the classification accuracy of the LightGBM model. The results show a steep increase in accuracy as the number of selected features (k) increases from 100 to approximately 500, with accuracy rising from 0.80 to around 0.94. Beyond k ≈ 500, accuracy plateaus fluctuated slightly between 0.94 and 0.96 up to the maximum tested k value of 3000. This suggests that a relatively small subset of features (≈ 500) is sufficient to achieve near optimal performance, while adding more features provides minimal additional accuracy gains. Beyond this point, the accuracy gradually decreases as more features are included. This behavior suggests that additional features introduce redundancy and potential noise, which negatively affects model generalization. Therefore, selecting an optimal number of features is critical to achieving a balance between model complexity and predictive performance.

This study conducts an ablation analysis to assess the individual impact of each component in the proposed gallbladder ultrasound image classification framework. The analysis aims to quantify the contribution of the preprocessing pipeline, multi-domain feature extraction, feature selection, and boosting algorithm to the overall system performance.

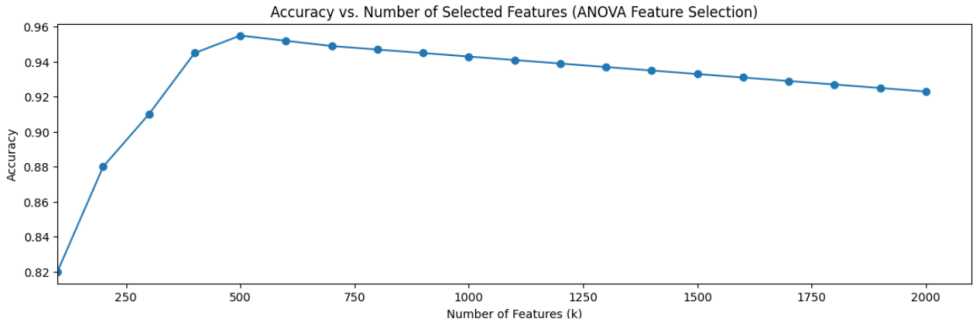

Figure 6. Plot between number features and accuracy in Light GBM

Each variant was evaluated independently on the UIdataGB dataset using identical training testing splits and evaluation metrics. The results in table 10 clearly demonstrate that each module contributes meaningfully to the overall classification performance. Removing preprocessing steps led to a noticeable decline in accuracy, especially when CLAHE or NLM denoising were omitted, confirming their importance in enhancing local contrast and suppressing ultrasound speckle noise. The multi-domain feature fusion (deep spatial + statistical + texture) provided the largest improvement. Using CNN features alone reduced accuracy by 4.4%, and handcrafted features alone by 5.8%. This confirms that integrating complementary feature domains yields richer representations for discriminating subtle tissue variations. The inclusion of ANOVA k-best feature selection improved accuracy by approximately 3%, reducing redundancy and enhancing generalization. Finally, among the ensemble classifiers, Histogram Gradient Boosting achieved the most consistent and highest performance, while AdaBoost failed to model complex non-linear patterns effectively. These findings validate that every component of the proposed architecture preprocessing, multi-domain fusion, feature selection, and gradient-boosting ensemble collectively drives the superior performance observed in the full model.

Table 10. Ablation Study Result on The UIDataGB Dataset

Model	Metric			
	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
HistGradientBoosting Full Model (All Modules)	98.6	98.4	98.2	98.3
HistGradientBoosting Without NLM Denoising	96.4	96.1	95.9	96.0
HistGradientBoosting Without CLAHE Enhancement	95.7	95.3	94.8	95.0
HistGradientBoosting Without Gabor Filtering	97.1	96.9	96.5	96.7
HistGradientBoosting Deep CNN Features	94.2	93.9	93.5	93.7
HistGradientBoosting Statistical + Texture	92.8	92.4	92.1	92.2
HistGradientBoosting Without ANOVA	95.6	95.3	95.0	95.1

AdaBoost	30.2	25.1	20.4	22.1
XGBoost	97.4	97.0	96.7	96.8
LightGBM	98.2	98.0	97.8	97.9
HistGradientBoosting	98.6	98.4	98.2	98.3

4.4. Performance Comparison on Several CNN Network Model and Proposed Model

This study evaluates the effectiveness of the proposed framework through a comparative experiment with several modern convolutional neural network architectures, including Inception V4, EfficientNet, and ConvNeXt, all evaluated under the same experimental protocol to ensure a fair comparison on the UIdataGB ultrasound dataset. The proposed approach processes ultrasound images using a preprocessing pipeline comprising Non-Local Means denoising, CLAHE contrast enhancement, frequency-domain filtering, and Gabor filtering. It then extracts deep spatial features from a pretrained InceptionV3 network and integrates them with handcrafted statistical descriptors and GLDM texture features to form a multi-domain feature representation. The resulting feature vector was then refined using ANOVA k-best feature selection and classified using ensemble boosting algorithms. For comparison, the alternative CNN architectures (Inception V4, EfficientNet, and ConvNeXt) were evaluated using a conventional end-to-end learning strategy, where ultrasound images were directly fed into the networks and the final classification layers were adapted to match the diagnostic categories of the dataset. All models were evaluated using the same dataset, stratified 80/20 train–test split, input resolution, and evaluation metrics, including accuracy, precision, recall, F1-score, and AUC, ensuring a consistent and fair experimental setting. The results indicate that all three CNN backbones achieved strong discriminative performance on the dataset, with classification accuracies exceeding 94%. Inception V4 demonstrated stable performance but required higher computational resources. EfficientNet achieved a competitive balance between accuracy and computational efficiency, reaching approximately 96% accuracy while maintaining relatively lower model complexity. ConvNeXt obtained the highest performance among the evaluated single-backbone CNN models, achieving approximately 97% accuracy, with particularly strong results in classes such as carcinoma and wall thickening, where subtle tissue patterns are important for diagnosis. When compared with these end-to-end CNN architectures, the proposed multi-domain feature fusion framework achieved slightly higher overall performance, reaching approximately 98.6% accuracy with consistently high AUC values across diagnostic categories. This improvement may be attributed to the integration of complementary information sources, where deep spatial representations extracted by CNNs are combined with statistical and texture-based descriptors, and further refined using feature selection and ensemble boosting techniques. While modern CNN architecture provides strong baseline performance, the results suggest that incorporating domain-specific handcrafted descriptors together with ensemble learning can contribute to improved discrimination in gallbladder ultrasound image classification. The quantitative results of this comparative analysis are summarized in table 11.

Table 11. Performance Comparison of CNN Backbones and Proposed Model on The UIDataGB Dataset

Model	Metric				
	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
Inception V3	94.5	94.1	93.8	94.0	0.97
Inception V4	95.2	95.0	94.8	95.0	0.98
EfficientNet-B3	95.8	95.5	95.4	95.5	0.98
ConvNeXt	97.0	96.8	96.6	96.7	0.99
Proposed Multi Feature Domain and Boosting	98.6	98.4	98.2	98.3	0.99

4.5. Comparative Analysis Between Previous Studies and the Proposed Model

This section presents a comparative analysis between the proposed framework and existing studies on gallbladder disease diagnosis. To ensure methodological fairness and avoid misleading interpretations, the comparison is explicitly divided into two sub-tables, each serving a different purpose. Table 12(a) reports a strictly comparable performance

evaluation, while table 12(b) summarizes contextual reference studies that are related in topic but differ in task definition, data modality, or evaluation protocol.

4.6. Strictly Comparable Evaluation

Direct and fair comparison is only possible with the study by Ahmed Mahdi Obaid et al. [33] , as this is the only prior work that employs the same ultrasound image dataset and the same nine-class gallbladder disease classification task addressed in this research. In their study, several CNN architectures including VGG16, InceptionV3, ResNet152, and MobileNet were evaluated, with MobileNet achieving the best reported accuracy of 98.35%. Using the identical dataset and classification scope, the proposed multi-domain feature fusion framework achieves an accuracy of 98.60%, thereby slightly outperforming the best CNN based result reported by Obaid et al. In addition, while the previous study primarily reported accuracy, the proposed method provides a more comprehensive diagnostic assessment, reporting precision, recall, and ROC-AUC, with the latter reaching a value of 99.94, indicating excellent separability among the nine disease classes.

Table 12 (a). Strictly Comparable Performance Comparison on the Same Dataset and Task

Author	Method	Task	Result			
			Accuracy	Precision	Recall	AUC/ROC
Ahmed Mahdi Obaid et al. [33]	CNN-based models (VGG16, InceptionV3, ResNet152, MobileNet)	9-class gallbladder disease classification from ultrasound images	98.35 (MobileNet)	-	-	-
Proposed Model	Multi-Domain Feature Fusion + ANOVA k-best + Boosting	9-class gallbladder disease classification from ultrasound images	98.60	98.40	98.20	99.94

4.7. Comparative Analysis within the Broader Landscape of Gallbladder Ultrasound Research

In addition to the strictly comparable evaluation, several previous studies are included as contextual references to position the proposed model within the broader landscape of gallbladder ultrasound research. These studies address related diagnostic problems but differ in at least one fundamental aspect, such as data modality (image, video, or non-image data), classification scope (binary or three-class vs. nine-class), or evaluation protocol. As such, the results reported in these studies are not directly comparable to those of the proposed model and are not used to infer superiority. Dadjouy and Sajedi [30] combined YOLOv8 and Faster R-CNN for region-of-interest detection followed by gallbladder cancer classification, achieving an accuracy of 92.62%. Chang et al. [25] reported an accuracy of 94.94% using a GA-optimized backpropagation neural network based on clinical tumor markers rather than ultrasound images. Basu et al. [32] introduced GBCNet with ROI extraction and curriculum learning, reporting performance values such as 0.711–0.734 under a 10-fold cross-validation protocol; these values correspond to specific backbone configurations and differ in scale and definition from percentage-based accuracy. In a subsequent study, Basu et al. [29] proposed FocusMAE, a video-based gallbladder cancer detection framework that achieved 96.4% accuracy by leveraging spatiotemporal information from ultrasound videos. Kim et al. [31] employed an ensemble of CNNs combined with demographic metadata for gallbladder polyp classification, achieving 87.61% accuracy and an AUC of 0.908. It is important to emphasize that the inclusion of contextual reference studies in table 12(b) does not imply a state-of-the-art claim across all prior works, as the experimental settings, datasets, and objectives differ substantially. Instead, these studies are presented to illustrate the diversity of existing approaches and to highlight the increased complexity addressed by the proposed method. Within the strictly comparable setting summarized in table 12(a), the proposed framework demonstrates a measurable performance improvement over the best existing CNN-based approach on the same dataset. At the same time, it avoids reliance on video data, manual ROI annotation, or non-imaging clinical features, making it a scalable and practical solution for real world multi class gallbladder disease diagnosis from ultrasound images.

ISSN 2723-6471

Table 12(b). Contextual Reference Comparison with Related Studies

Author	Method	Data Modality & Task	Reported Performance
Dadjouy & Sajedi [30]	YOLOv8 + Faster R-CNN	Ultrasound images; ROI detection + binary cancer classification	Accuracy = 92.62%
Chang et al. [25]	GA-Optimized BPNN	Clinical tumor markers (non-image); cancer prediction	Accuracy = 94.94%
Basu et al. [32]	GBCNet + ROI + Curriculum Learning	Ultrasound images; binary/3-class cancer detection	CV accuracy = 0.711–0.734
Basu et al. [29]	FocusMAE	Ultrasound videos; cancer detection	Accuracy = 96.40%
Kim et al. [31]	Ensemble CNN + Demographic Fusion	Ultrasound images + metadata; polyp classification	Accuracy = 87.61%, AUC = 0.908
Proposed Model	Multi-Domain Feature Fusion + ANOVA k-best + Boosting	9-class gallbladder disease classification from ultrasound images	Accuracy = 98.60%, Precision= 98.40%, Recall=98.20%, AUC=99.94

5. Conclusion

This study presented a hybrid framework for gallbladder disease classification from ultrasound images by integrating multi-domain feature representations and ensemble machine learning techniques. The proposed approach combines deep spatial features extracted from a pretrained InceptionV3 network with handcrafted statistical descriptors and GLDM-based texture features, followed by ANOVA-based feature selection and classification using ensemble boosting algorithms. The experimental results on the UIdataGB ultrasound dataset demonstrate that the integration of complementary feature domains can provide effective representations for distinguishing different gallbladder conditions. To better understand the contribution of each component in the proposed pipeline, an ablation study was conducted by systematically removing individual modules from the framework. The results indicate that each stage of the pipeline contributes to the overall performance of the model. In particular, removing preprocessing components such as Non-Local Means denoising, CLAHE contrast enhancement, or Gabor filtering leads to a noticeable reduction in classification performance, highlighting the importance of image enhancement for improving ultrasound feature quality. Similarly, excluding the ANOVA-based feature selection step results in decreased accuracy, indicating that selecting the most discriminative features plays a key role in reducing feature redundancy and improving classifier effectiveness. The ablation results also show that using deep CNN features alone or handcrafted statistical and texture features alone leads to lower performance compared to their combined representation, suggesting that these feature domains provide complementary information for ultrasound image characterization. In comparative experiments, several modern CNN architectures, including Inception V4, EfficientNet, and ConvNeXt, were evaluated using an end-to-end training strategy under the same dataset and evaluation protocol. While these models achieved strong classification performance, the proposed hybrid framework produced slightly higher overall accuracy and consistently high AUC values across diagnostic categories. These findings suggest that integrating deep spatial representations with statistical and texture-based descriptors, together with ensemble learning techniques, can provide improved discriminative capability for gallbladder ultrasound image classification.

It is important to note that the experiments in this study were conducted on a single dataset (UIdataGB). Therefore, although the results indicate promising performance, further evaluation on additional datasets and multi-center ultrasound collections will be necessary to more comprehensively assess the generalization capability of the proposed framework. Future work will also explore domain adaptation strategies and lightweight model architectures to facilitate broader clinical applicability. Overall, the results of this study suggest that multi-domain feature fusion combined with ensemble learning represents a promising direction for improving automated gallbladder ultrasound image classification.

6. Declarations

6.1. Author Contributions

Conceptualization: G.A.P. and P.C.; Methodology: G.A.P., P.D.W.A., and S.; Software: P.D.W.A., S., and M.L.; Validation: M.L. and D.P.H.; Formal Analysis: G.A.P. and P.D.W.A.; Data Curation: S., G.A.P., and D.P.H.; Writing Original Draft Preparation: G.A.P., P.D.W.A., and P.C.; Writing Review and Editing: G.A.P. and P.D.W.A.; Visualization: D.P.H.; Supervision: G.A.P.; Funding Acquisition: G.A.P. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

Gallblader Diseases Dataset (https://data.mendeley.com/datasets/r6h24d2d3y/2) (Mendeley Data).

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. S. Cerçi, F. M. Ozbek, C. Cerçi, B. Baykal, H. E. Eroğlu, Z. Baykal, M. Yildiz, S. Sağlam, and A. Yeşildağ, "Gallbladder function and dynamics of bile flow in asymptomatic gallstone disease," *World Journal of Gastroenterology*, vol. 15, no. 22, pp. 2763–2767, 2009, doi: 10.3748/wjg.15.2763.

[2] C. F. Dietrich and B. Braden, "Sonographic assessments of gastrointestinal and biliary functions," Best Pract. Res. Clin. Gastroenterol., vol. 23, no. 3, pp. 353–367, 2009, doi: https://doi.org/10.1016/j.bpg.2009.03.003.

[3] S. Rashid, C. J. Das, A. Chauhan, G. Aggarwal, R. C. Joshi, R. Burget, and M. K. Dutta, "Self-attention-guided residual deep neural network with multi-scale dilated feature extraction for automated gallbladder disease diagnosis in ultrasound imaging," *Computer Methods and Programs in Biomedicine*, vol. 271, no.8, pp. 1-12, 2025, doi: 10.1016/j.cmpb.2025.109020

[4] G. Doherty, M. Manktelow, B. Skelly, P. Gillespie, A. J. Bjourson, and S. Watterson, "The Need for Standardizing Diagnosis, Treatment and Clinical Care of Cholecystitis and Biliary Colic in Gallbladder Disease," *MDPI*. vol.58, no.3, pp.1-18, 2022, doi: 10.3390/medicina58030388.

[5] L. Quinn, K. Tryposkiadis, J. Deeks, H. C. de Vet, S. Mallett, L. B. Mokkink, Y. Takwoingi, S. Taylor-Phillips, and A. Sitch, "Interobserver variability studies in diagnostic imaging: a methodological systematic review," *British Institute of Radiology*.vol.96 ,no.1148, pp.1-8,2023 doi: 10.1259/bjr.20220972.

[6] M. H. Yu, Y. J. Kim, H. S. Park, and S. I. Jung, Benign gallbladder diseases: Imaging techniques and tips for differentiating with malignant gallbladder diseases. *World journal of gastroenterology*, vol.26, no. (22), pp. 2967–2986, 2020, doi: 10.3748/wjg.v26.i22.2967.

[7] S. Jayanthi, I. Kaur, E. L. Lydia, P. R. Babu, J. C. Yoon, C. Yoon, and G. P. Joshi, "Gallbladder disease diagnosis from ultrasound using squeeze-and-excitation capsule network with convolutional bidirectional long short-term memory," *Sci. Rep*., vol.16, no.3147, pp.1-18, 2026, doi: 10.1038/s41598-025-32978-9.

[8] M. Byra, G. Styczynski, C. Szmigielski, P. Kalinowski, Ł. Michałowski, R. Paluszkiewicz, B. Ziarkiewicz-Wróblewska, K. Zieniewicz, P. Sobieraj, and A. Nowicki, "Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, no. 12, pp. 1895–1903, Dec. 2018, doi: 10.1007/s11548-018-1843-2.

[9] S. Y. Rhyou and J. C. Yoo, "Cascaded deep learning neural network for automated liver steatosis diagnosis using ultrasound images," *Sensors*, vol. 21, no. 16, pp.1-16, Aug. 2021, doi: 10.3390/s21165304.

[10] S. Mahmud, T. O. Abbas, M. E. Chowdhury, A. Mushtak, S. Kabir, S. Muthiyal, A. Koko, A. B. Altyeb, A. Alqahtani, A. Khandakar, and S. M. Islam, "Automated grading of prenatal hydronephrosis severity from segmented kidney ultrasounds using deep learning," *Expert Systems with Applications*, vol. 255, no., pp.1-18 , Dec. 2024, doi: 10.1016/j.eswa.2024.124594.

[11] M. Tounsi, D. Y. Abdulhussain, A. T. Azar, A. Al-Khayyat, and I. K. Ibraheem, "Deep Learning Model-based Decision Support System for Kidney Cancer on Renal Images," *Engineering, Technology and Applied Science Research*, vol. 14, no. 5, pp. 17177–17187, Oct. 2024, doi: 10.48084/etasr.8335.

[12] D. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, no. 1, pp. 249–259, 2018, https://doi.org/10.1016/j.neunet.2018.07.011.

[13] C. Keatmanee, D. Songsaeng, S. Klabwong, Y. Nakaguro, A. Kunapinun, M. Ekpanyapong, and M. N. Dailey,"Enhancing weakly supervised data augmentation networks for thyroid nodule assessment using traditional and Doppler ultrasound images," *Computers in Biology and Medicine*, vol. 196, no. 1, pp. 1-13, 2025, doi: 10.1016/j.compbiomed.2025.110553.

[14] N. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *Neurocomputing*, vol. 321, no. 1, pp. 321–331, 2018, doi: 10.1016/j.neucom.2018.09.013.

[15] K. Rais, M. Amroune, M. Y. Haouam, and A. Benmachiche, "GAN-PSO: A novel approach for effective data augmentation through medical image generation," Digital Signal Processing, vol. 180, pp. 1-15, no. 106246, 2026, https://doi.org/10.1016/j.dsp.2026.106246.

[16] H. Arora and N. N. Mittal, "Image Enhancement Techniques for Gastric Diseases Detection using Ultrasound Images," in *Proc. 3rd Int. Conf. Electronics Communication and Aerospace Technology (ICECA)*, vol. 24, no.1, pp. 251–256, 2019, doi: 10.1109/ICECA.2019.8822148.

[17] J. Glory Precious and Shirley Selvan, "Detection of Abnormalities in Ultrasound Images Using Texture and Shape Features," in *IEEE International Conference on Current Trends toward Converging Technologies, IEEE*, vol. 2018, no.11, pp. 1–6. doi: 10.1109/ICCTCT.2018.8551174.

[18] R. Buettner, M. Bilo, N. Bay, and T. Zubac, "A Systematic Literature Review of Medical Image Analysis Using Deep Learning," *IEEE Symposium on Industrial Electronics & Applications (ISIEA)*, vol. 2020, no.7, pp.1-4, doi: 10.1109/ISIEA49364.2020.9188131.

[19] U. R. Acharya, K. M. Meiburger, J. E. Koh, E. J. Ciaccio, N. Arunkumar, M. H. See, N. A. Taib, A. Vijayananthan, K. Rahmat, F. Fadzli, and S. S. Leong, "A Novel Algorithm for Breast Lesion Detection Using Textons and Local Configuration Pattern Features with Ultrasound Imagery," *IEEE Access*, vol. 7, no. 2, pp. 22829–22842, 2019, doi: 10.1109/ACCESS.2019.2898121.

[20] Y. Wang, X. Ge, H. Ma, S. Qi, G. Zhang, and Y. Yao, "Deep Learning in Medical Ultrasound Image Analysis: A Review," *IEEE Access*, vol. 9, no. 4, pp. 54310–54324, 2021, doi: 10.1109/ACCESS.2021.3071301.

[21] H. Yang, X. Sun, Y. Sun, L. Cui, and B. Li, "Ultrasound Image-Based Diagnosis of Cirrhosis with an End-to-End Deep Learning model," in *Proceedings - 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020, Institute of Electrical and Electronics Engineers Inc.*, vol. 2020, no. 12, pp. 1193–1196. doi: 10.1109/BIBM49941.2020.9313579.

[22] Y. Mao, W. Lu, Y. Li, and Y. Wang, "Enhancing Ultrasound Imaging with Synthetic Aperture Data: A Deep Learning Approach for Superior Resolution and Contrast," in *Proceedings - International Symposium on Biomedical Imaging, IEEE Computer Society*, vol.2024, no. 5, pp.1-5, doi: 10.1109/ISBI56570.2024.10635178.

[23] J. M. Urman, J. M. Herranz, I. Uriarte, M. Rullán, D. Oyón, B. González, I. Fernandez-Urién, J. Carrascosa, F. Bolado, L. Zabalza, and M. Arechederra, "Pilot multi-omic analysis of human bile from benign and malignant biliary strictures: A machine-learning approach," *Cancers* , vol. 12, no. 6, pp. 1–30, Jun. 2020, doi: 10.3390/cancers12061644.

[24] C. Yao, S. Wu, Z. Liu, and P. Li, "A deep learning model for predicting chemical composition of gallstones with big data in medical Internet of Things," *Future Generation Computer Systems*, vol. 94, no.5, pp. 140–147, doi: 10.1016/j.future.2018.11.011.

[25] Y. Chang, Q. Wu, L. Chi, H. Huo, and Q. Li, "Adoption of combined detection technology of tumor markers via deep learning algorithm in diagnosis and prognosis of gallbladder carcinoma," *Journal of Supercomputing*, vol. 78, no. 3, pp. 3955–3975, Feb. 2022, doi: 10.1007/s11227-021-03843-z.

[26] W. Zhou, Y. Yang, C. Yu, J. Liu, X. Duan, Z. Weng, D. Chen, Q. Liang, Q. Fang, J. Zhou, and H. Ju, "Ensembled deep learning model outperforms human experts in diagnosing biliary atresia from sonographic gallbladder images," *Nature Communications*, vol. 12, no. 1, pp.1-14, Dec. 2021, doi: 10.1038/s41467-021-21466-z.

[27] A. M. Obaid, A. Turki, H. Bellaaj, and M. Ksontini, "Detection of Biliary Artesia using Sonographic Gallbladder Images with the help of Deep Learning approaches," in *2022 8th International Conference on Control, Decision and Information Technologies, CoDIT 2022, Institute of Electrical and Electronics Engineers Inc.*, 2022, vol. 1, no. 5, pp. 705–711, doi: 10.1109/CoDIT55151.2022.9804084.

[28] Y. Cao, J. Lian, B. Shi, and X. Yang, "A gallbladder segmentation method of a CFC-MSPCNN in ultrasound image," in *Proceedings of 2023 IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence, ICIBA 2023, Institute of Electrical and Electronics Engineers Inc.*, vol.2023, no.5, pp. 312–319, 2023, doi: 10.1109/ICIBA56860.2023.10165027.

[29] S. Basu, M. Gupta, C. Madan, P. Gupta, and C. Arora, "FocusMAE: Gallbladder Cancer Detection from Ultrasound Videos with Focused Masked Autoencoders," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE*, vol.2024, no.7, pp. 11715–11725, 2024. doi: 10.1109/CVPR52733.2024.01113.

[30] S. Dadjouy and H. Sajedi, "Gallbladder Cancer Detection in Ultrasound Images based on YOLO and Faster R-CNN," in *2024 10th International Conference on Artificial Intelligence and Robotics, QICAR 2024, Institute of Electrical and Electronics Engineers Inc*. vol.2024, no.2, pp. 227–231, 2024. doi: 10.1109/QICAR61538.2024.10496645.

[31] T. Kim, Y. H. Choi, J. H. Choi, S. H. Lee, S. Lee, and I. S. Lee, "Gallbladder polyp classification in ultrasound images using an ensemble convolutional neural network model," *J. Clin. Med*., vol. 10, no. 16, pp.1-11, Aug. 2021, doi: 10.3390/jcm10163585.

[32] S. Basu, M. Gupta, P. Rana, P. Gupta, and C. Arora, "Surpassing the Human Accuracy: Detecting Gallbladder Cancer from USG Images with Curriculum Learning," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society,* vol.2022, no.6, pp. 20854–20864,2022. doi: 10.1109/CVPR52688.2022.02022.

[33] A. M. Obaid, A. Turki, H. Bellaaj, M. Ksantini, A. AlTaee, and A. Alaerjan, "Detection of Gallbladder Disease Types Using Deep Learning: An Informative Medical Method," *Diagnostics*, vol. 13, no. 10, pp. 1-12, May 2023, doi: 10.3390/diagnostics13101744.

[34] A. Turki, A. M. Obaid, H. Bellaaj, M. Ksantini, and A. AlTaee, "UIdataGB: Multi-Class ultrasound images dataset for gallbladder disease detection," *Data Brief,* vol. 54, no.6, pp.1-9, 2024, doi: 10.1016/j.dib.2024.110426.