Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: jeeemi.org; Vol. 7, No. 4, October 2025, pp: 1089-1111 e-ISSN: 2656-8632
RESEARCH ARTICLE OPEN ACCESS

Gallbladder Disease Classification from Ultrasound Images Using CNN Feature Extraction and Machine Learning Optimization

Ryan Adhitama Putra[1], Gede Angga Pradipta[2], and Putu Desiana Wulaning Ayu[2]

[1] Magister Program, Department of Magister Information Systems, Institut Teknologi dan Bisnis STIKOM Bali, Denpasar, Indonesia
[2] Department of Magister Information Systems, Institut Teknologi dan Bisnis STIKOM Bali, Denpasar, Indonesia

Corresponding author: Ryan Adhitama Putra (e-mail: 232011014@stikom-bali.ac.id), **Author(s) Email:** Gede Angga Pradipta (e-mail: angga_pradipta@stikom-bali.ac.id), Putu Desiana Wulaning Ayu (e-mail: wulaning_ayu@stikom-bali.ac.id)

Abstract **Gallbladder diseases, including gallstones, carcinoma, and adenomyomatosis, may cause severe complications if not identified correctly and in a timely manner. However, ultrasound image interpretation relies heavily on operator experience and may suffer from subjectivity and inconsistency. This study aims to develop an automated and optimized classification model for gallbladder disease using ultrasound images, aiming to improve diagnostic reliability and efficiency. A key outcome of this research is a thorough assessment of how feature selection combined with hyperparameter tuning influences the accuracy of classical machine learning models that use features extracted via CNN-based feature extraction. The proposed pipeline enhances diagnostic accuracy while remaining computationally efficient. The method involves extracting deep features from ultrasound images using a pre-trained VGG16 CNN model. The features are subsequently reduced using the SelectKBest method through Univariate Feature Selection. Multiple popular classification models, specifically SVM, Random Forest, KNN, and Logistic Regression were tested using both original settings and adjusted hyperparameters through grid search. A complete evaluation of model performance was conducted using the test set, employing key performance indicators including overall prediction correctness (accuracy), actual positive rate (recall), positive prediction accuracy (precision), F1-score, and the ROC curve's corresponding area value. Evaluation results suggest that the SVM approach, combined with selected features and hyperparameter tuning, achieved the highest performance: 99.35% accuracy, 99.32% precision, 99.35% recall, and 99.33% F1-score, with a relatively short computation time of 18.4 seconds. In conclusion, feature selection and hyperparameter tuning significantly enhance classification performance, making the proposed method a promising candidate for clinical decision support in gallbladder disease diagnosis using ultrasound imaging.**

Keywords **Gallbladder Disease Classification; Ultrasound Images; CNN Feature Extraction; Feature Selection; Hyperparameter Tuning.**

I. Introduction

Positioned under the liver, the gallbladder is a compact organ with a pear-like shape, anatomically divided into the fundus, body, and neck, which joins the cystic duct [1]. This organ primarily stores and concentrates bile synthesized by the liver during interdigestive phases. Once food enters the digestive tract, the gallbladder contracts to deliver bile into the small intestine, where it plays a crucial role in breaking down fats and supporting their absorption [2]. However, despite its small size, the gallbladder is prone to various disorders that can significantly impact human health [3][4].

Gallbladder disease is primarily influenced by lifestyle factors, such as a high-fat diet, obesity, and age-related changes [5]. Recent investigations have indicated a global upward trend in the incidence of gallbladder-related disorders. For example, data from routine health screenings in Liaoning, China, reported that 2.30% of the population exhibited gallstones, while 6.64% presented with gallbladder polyps. The prevalence of these conditions demonstrated a steady year-by-year increase, reaching its highest point in 2020 [6]. In 2019, over 52 million new instances related to gallbladder and bile duct disorders were recorded

Manuscript received July 13, 2024; Revised September 5, 2025; Accepted September 24, 2025; date of publication September 26, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i4.1030

worldwide, reflecting a 97% increase since 1990, with women and populations in low-HDI countries disproportionately affected [7]. In 2022, the number of new cases reached 122,491, accompanied by 89,055 deaths reported worldwide, with Northeastern India and Southern Chile showing the highest incidence rates. Incidence rates were notably higher in females than males, while the lowest rates were found in Uganda and South Africa [8]. Similar trends were observed in Germany, where gallstone prevalence rose from 3.8% to 10.8% over a decade, with higher rates in women and peak incidence occurring in adults aged 41–50 years [9]. In Indonesia, a recent study at Dr. Moewardi Regional Hospital in Surakarta reported that out of 86 patients reviewed between January and April 2023, 43 (50%) were diagnosed with gallstones, highlighting a significant local disease burden and a strong association with excess body mass index (BMI) [10].

Considering the potential risks associated with gallbladder disease, early detection using non-invasive imaging techniques such as ultrasonography (USG) is essential [11]. Ultrasound has been widely recognized as a safe, accessible, and cost-effective diagnostic tool, capable of identifying both symptomatic and asymptomatic cases of gallbladder diseases [12]. However, interpreting ultrasound images visually is strongly influenced by the operator's expertise and image clarity, both of which can significantly impact diagnostic accuracy [13][14]. Therefore, machine learning technology is needed to help address these limitations by supporting more accurate and consistent interpretation of ultrasound images [15][16].

In the last few years, researchers have proposed multiple techniques or approaches to improve the accuracy of classification performance in gallbladder disease diagnosis. For instance, Obaid et al. evaluated deep neural network models on a large ultrasound dataset comprising 10,692 images from 1,782 patients. They found that MobileNet achieved the highest accuracy at 98.35% in classifying nine types of gallbladder diseases [17]. Building on the same dataset, Bozdağ et al. introduced a feature-engineered content-based image retrieval (CBIR) system that outperformed six benchmark models by integrating features from multiple pre-trained CNN architectures. Achieving an Average Precision (AP) score of 0.94, the developed system demonstrates strong potential as a reliable diagnostic aid, particularly in healthcare environments lacking specialist radiological expertise [18].

In a separate study, Shuvo and Chowdhury demonstrated that an ensemble of CNN architectures, specifically VGG16, VGG19, XceptionNet, and ResNet50, significantly improved the classification accuracy for gallbladder cancer using ultrasound images. Their findings indicated that VGG19 and XceptionNet outperformed the others, achieving the highest classification accuracy of 85.44% [19]. In another study, Dadjouy and Sajedi introduced a comprehensive hierarchical model that integrates feature fusion within a dual CNN architecture incorporating uncertainty estimation, which achieved a classification accuracy of up to 92.62% for gallbladder cancer detection [20].

While deep learning approaches have yielded positive results, there are still notable limitations in existing diagnostic tools for gallbladder disease. Despite their success, end-to-end CNN architectures frequently need extensive annotated datasets and considerable processing capabilities, limiting their application in healthcare systems with constrained resources. For instance, Bozdağ et al. mentioned minimizing time-consuming processes, but did not specify the actual processing time [18], while Obaid et al. reported that their model required at least 540 seconds for image processing [17]. Shuvo and Chowdhury achieved only 85.44% accuracy using a CNN ensemble, indicating room for improvement [19]. Dadjouy and Sajedi highlighted the need for larger datasets and, in another study, limited their focus to a single disease from a single dataset, which restricts model generalizability [20].

Given these limitations, alternative strategies are essential for high diagnostic accuracy while reducing dependency on large datasets and heavy computational requirements. One promising direction is the development of combined methods leveraging CNNs for feature extraction alongside conventional machine learning classifiers. In addition, while most existing studies emphasize end-to-end deep learning frameworks, there has been relatively limited investigation into hybrid methods that integrate CNN-based feature representation followed by machine learning classification models. Among various CNN architectures, VGG16 VGG16 architecture is widely utilized for feature extraction tasks due to its simplicity, transferability, and effectiveness in capturing hierarchical features [21]. VGG16's architecture, consisting of thirteen convolutional operations followed by three dense layers, enables it to serve as a robust backbone for extracting rich image representations [22]. Previous studies have demonstrated that combining features derived from CNNs with classical machine learning classification models, including SVM, Random Forest, KNN, and Logistic Regression, can enhance classification performance and strengthen model generalizability. Waluyo et al. reported that a CNN-KNN hybrid approach outperformed a pure CNN model in detecting Mycobacterium tuberculosis in medical images [23]. Similarly, Biswas and Islam achieved high classification performance for brain tumors using a CNN-SVM hybrid model [24]. Saleh et

Manuscript received July 13, 2024; Revised September 5, 2025; Accepted September 24, 2025; date of publication September 26, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i4.1030

al. successfully implemented a CNN-Random Forest pipeline for lung nodule classification [25]. Furthermore, Kuntiyellannagari et al. introduced a hybrid CNN aimed at detecting brain tumors by applying various traditional classifiers, such as Logistic Regression, with promising results [26].To address this gap, the present study introduces a hybrid classification model that employs CNN for feature extraction with VGG16 from ultrasound images, followed by classification using four classification models, namely SVM, Random Forest, KNN, and Logistic Regression.

An essential component in achieving optimal performance in hybrid classification frameworks is the application of effective feature selection combined with hyperparameter tuning [27]. Feature selection is essential in identifying the most relevant attributes in high-dimensional data, enhancing model accuracy, reducing complexity, and minimizing overfitting, especially in medical imaging, where subtle distinctions are diagnostically crucial [28]. Various techniques are available for selecting features, such as Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), and regularization methods like LASSO, along with univariate feature selection approaches [29][30][31]. Feature selection methods like SelectKBest assist in detecting the key features within large-scale datasets by ranking them according to statistical criteria, making it a simple yet effective way to improve model performance while reducing overfitting [32]. The impact of this approach has been empirically demonstrated in several studies. For instance, Jain and Saha evaluated classifiers both with feature selection applied and without it, then reported performance gains of up to 26.5% in accuracy, 70.9% in F-measure, and 26.74% in AUC-ROC, together with a decrease in average training duration by 62 seconds [33]. Similarly, Julkaew et al. reported a classification accuracy of 92.05% after feature selection, which was higher than the 90.79% accuracy obtained using all available features [34].

Meanwhile, adjusting hyperparameters refines the model configuration, often turning average performance into high accuracy [35]. Similarly, hyperparameter tuning can be performed using Random Search, Bayesian Optimization, or Grid Search. However, this study adopts Grid Search because it systematically explores multiple hyperparameter configurations to identify the optimal setup for optimizing model performance, thereby enhancing generalization and robustness, particularly in complex classification scenarios [36][37]. Taufiq et al. reported that Grid Search achieved superior accuracy, with average improvements of 0.5% in accuracy, 0.67% in precision, 0.83% in recall, and 0.33% in F1-score. The benefits were highly algorithm-dependent, with SVM showing a substantial accuracy gain of approximately 30%, KNN and Decision Tree improving by around 4–5%, and Random Forest and Logistic Regression exhibiting gains of less than 1.1% [38]. Similarly, Sukamto et al. demonstrated that applying Grid Search increased the accuracy of the Decision Tree by 1.52% and the KNN by 1.49% [39].

To address the challenges in diagnosing gallbladder diseases from ultrasound images, this study focuses on developing an automated and optimized classification model that improves diagnostic reliability and efficiency by integrating CNN-based feature extraction with machine learning classification

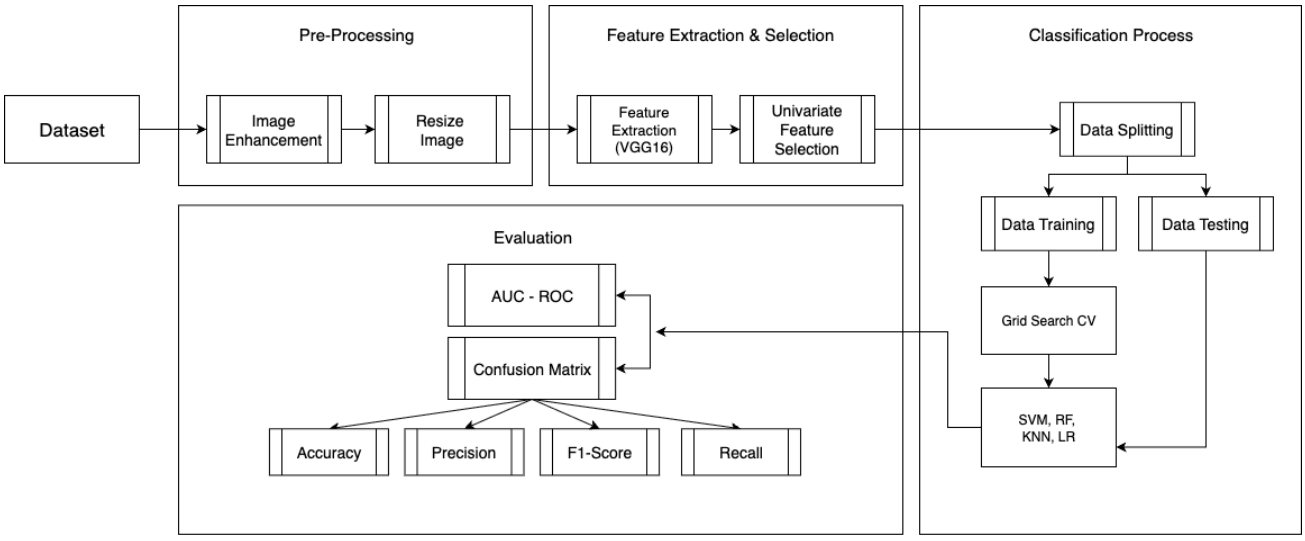

Fig. 1. The methodology flowchart of our proposed model consists of data pre-processing, CNN-based feature extraction followed by feature selection, classification using machine learning algorithms, and performance evaluation

Manuscript received July 13, 2024; Revised September 5, 2025; Accepted September 24, 2025; date of publication September 26, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i4.1030

models (Fig. 1). The proposed method leverages a pre-trained VGG16 architecture, a type of Convolutional Neural Network (CNN), for deriving deep representations of ultrasound data, and subsequently applies Univariate Feature Selection via the SelectKBest technique to reduce feature space and preserve the most informative attributes. These selected features are then classified using four classification models: SVM, Random Forest, KNN, and Logistic Regression, each evaluated in both default and grid-tuned hyperparameter configurations. This study contributes by 1) introducing a hybrid framework that efficiently integrates features extracted by CNN with machine learning classification models, 2) the integration of feature selection to improve classifier accuracy and computational efficiency, 3) the application of systematic hyperparameter tuning to optimize model performance, and 4) the achievement of high classification accuracy with efficient computation time demonstrates the method's potential as a supportive tool in clinical decision-making. The system is evaluated with multiple performance metrics, including overall prediction correctness (accuracy), true positive rate (recall), positive prediction accuracy (precision), F1-score, and the ROC curve's corresponding area value. This study introduces a replicable methodology with broad applicability, aimed at enhancing diagnostic reliability and objectivity in the classification of gallbladder diseases using ultrasound imaging.

An overview of this study is arranged in the following manner. Section II provides details about the utilized dataset and describes the preprocessing procedures, extraction of features using CNN, selection of relevant features, and the applied machine learning models, including their tuned parameters. Section III reports classification outcomes assessed through various evaluation metrics for both baseline and optimized models. Section IV elaborates on the experimental findings, particularly the influence of combining feature selection with hyperparameter tuning, and also addresses the study's limitations. Lastly, the fifth section summarizes the study's goals, principal results, and potential avenues for future research.

II. Methodology

This study begins by utilizing a dataset of ultrasound images related to gallbladder disease (Fig. 2). The initial stage involves data preprocessing, which includes two key steps: image enhancement, to improve visual clarity and highlight essential features, and image resizing, to ensure uniform dimensions for consistent input into subsequent processes. Once the images are preprocessed, feature extraction is performed using a CNN model. This step automatically captures deep visual patterns from the ultrasound images, serving as high-level feature representations. Subsequently, a feature selection step is performed to minimize dimensional complexity by preserving only the most informative attributes for classification, which

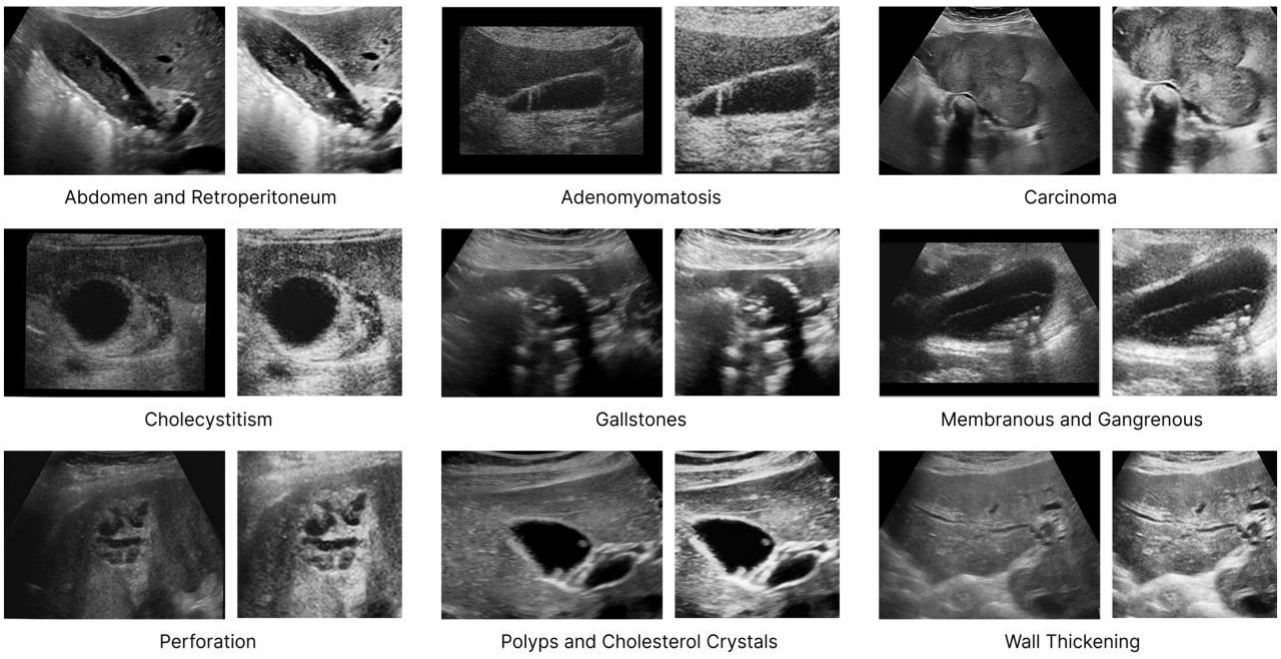

Fig. 2. Gallbladder ultrasound dataset comprising 9 diagnostic categories. The images on the left represent the original inputs, while those on the right reflect the outcomes after preprocessing procedures were applied

Manuscript received July 13, 2024; Revised September 5, 2025; Accepted September 24, 2025; date of publication September 26, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i4.1030

contributes to improved computational efficiency and may lead to better predictive accuracy.

Following the preprocessing stage, data splitting is performed by allocating 80% for training and 20% for testing, enabling the model to be assessed using independent samples to evaluate its ability to generalize effectively. In the classification stage, four classical machine learning models are employed: SVM, Random Forest, KNN, and Logistic Regression. For performance refinement, hyperparameter tuning is carried out using Grid Search, which systematically explores multiple parameter combinations to identify the most effective configuration. Finally, the evaluation phase uses two main metrics: the confusion matrix and the ROC curve's corresponding area value. For further illustration of the research workflow, see Fig. 1.

A. Dataset

This research uses an openly available dataset from Mendeley Data named the Gallbladder Diseases Dataset. The dataset is composed of 10,692 images with the following class distribution: 1,170 images of abdomen and retroperitoneum, 1,164 images of adenomyomatosis, 1,590 images of carcinoma, 1,146 images of cholecystitis, 1,326 images of gallstones, 1,224 images of membranous and gangrenous cholecystitis, 1,062 images of perforation, 1,020 images of polyps and cholesterol crystals, and 990 images of wall thickening. Over four years, imaging data were obtained from four medical institutions in Baghdad, Iraq, using ultrasound equipment including Philips Affiniti 70, Siemens Acuson X700, Canon Viamo c100, and Philips CX50. Training medical staff performed acquisition, with disease classification conducted by radiologists and quality verification by senior specialists [40].

In the original dataset, all images (600x450 pixels, 24-bit depth) were resized to 1200x900 pixels while preserving the aspect ratio, followed by noise reduction via median filtering, normalization to zero mean and unit variance, and data augmentation (rotations, flips, translations, brightness/contrast adjustments). Categorical labels were numerically encoded for model compatibility. The dataset offers a robust benchmark for multi-class classification tasks involving gallbladder ultrasound imaging through this standardized preprocessing and balanced representation across nine disease categories. Furthermore, the inclusion of clinically realistic variability in grayscale images enhances their relevance for evaluating the robustness of automated diagnostic systems.

B. Pre-Processing Data

In this study, the pre-processing stage involves of two main techniques: image enhancement and image resizing. This work uses the CLAHE method to boost contrast quality in grayscale ultrasound images of gallbladder diseases. Initially developed to enhance images with low contrast, this method is also utilized to address the issue of noise amplification that can occur when using the Histogram Equalization approach. It works by dividing the image into compact regions, typically 8-by-8 pixels, and performing histogram equalization independently within each region [41].

The CLAHE method relies on two primary configuration settings: Clipping Limit (CL) and Block Size (BS), both of which significantly influence the quality of the enhanced image [42]. Raising the Clipping Limit (CL) typically causes the image to appear more luminous, especially in cases where the original image exhibits low brightness levels. An elevated CL setting leads to a more uniform histogram distribution, thereby extending the dynamic range and improving overall image contrast [43]. The mathematical formulation of CLAHE can be expressed as shown in Eq. (1) [44].

$$\beta = \frac{p}{q}1 + (1 + \frac{\alpha}{100}S_{max}) \quad (1)$$

In this equation, β represents the CLAHE output value, p denotes the total number of pixels within a block, q corresponds to the block's dynamic range, S_{max} refers to the maximum allowable slope, and α indicates the clipping factor [44]. In this study, a clipping limit 2.0 is applied, and the tile grid size is configured as (8, 8).

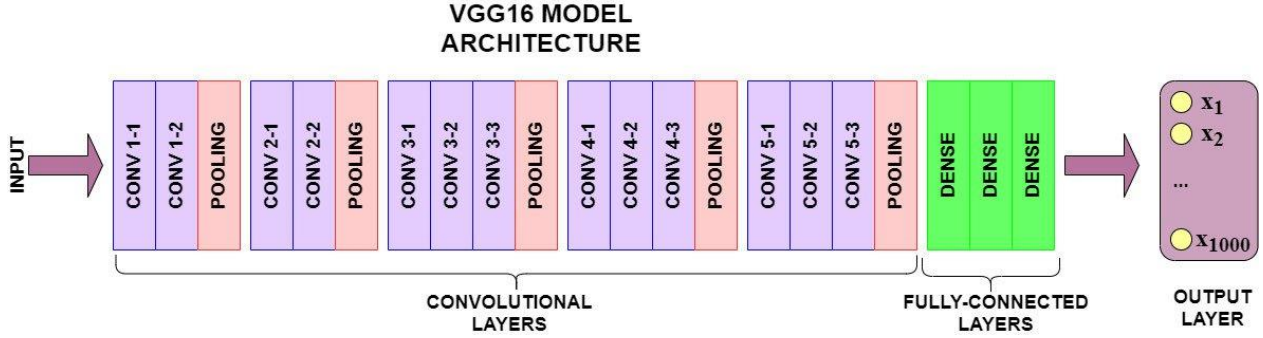

Fig. 3. Convolutional architecture of VGG16 used in feature extraction process

Manuscript received July 13, 2024; Revised September 5, 2025; Accepted September 24, 2025; date of publication September 26, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i4.1030

Setting the limit at 2.0 ensures that the contrast enhancement is sufficiently strong to reveal underlying features, while simultaneously controlling the degree of histogram clipping to avoid excessive contrast in uniform regions.

The image resizing procedure is implemented to standardize all samples in the dataset to a fixed size of 224 × 224 pixels. This step is crucial because models built upon advanced neural computation techniques, especially CNN architectures commonly used in deep learning, generally require uniform input dimensions to facilitate efficient training [45]. Fig. 2 illustrates the result of the data preprocessing procedure, including applying the CLAHE technique for contrast enhancement and the subsequent image resizing step.

C. Feature Extraction

Extracting features serves as a fundamental component in image classification, especially within medical imaging, as it involves converting raw visual inputs into meaningful numerical representations that encapsulate the image's key patterns, textures, and structural characteristics [46]. This study employs the pre-trained VGG16 convolutional model for feature extraction, leveraging its proven effectiveness in numerous visual recognition tasks [47]. VGG16 was initially developed by researchers affiliated with the Visual Geometry Group from the University of Oxford and trained using the extensive ImageNet dataset, which comprises more than one million images spanning 1,000 distinct object classes. This architecture is recognized for its clean and uniform structure, composed of 13 layers for convolution followed by three layers that are fully connected, all utilizing compact 3×3 filters throughout the network [22]. The structural design of the VGG16 network used in this research is illustrated in Fig. 3, and the convolutional operation can be mathematically expressed as shown in Eq. (2) [48].

$$C_j^l = \varphi\left(\sum_{i=1}^{M^{l-1}} C_i^{l-1} \times k_{ij}^l + b_j^{l-1}\right) \quad (2)$$

In this equation, × represents the convolution operation that characterizes the relationship between the weights of the ith and jth features across the $(l-1)$th and lth layers, b_j denotes the bias component, and φ refers to the activation function [48]. For the purposes of this research, VGG16 functions as a feature extractor and is configured to enhance its applicability to medical image interpretation. The model is initialized using weights previously learned from training on the ImageNet benchmark dataset (weights='imagenet'), enabling it to utilize generalized visual representations acquired from a large and diverse collection of images. To adapt the model for feature extraction, the final

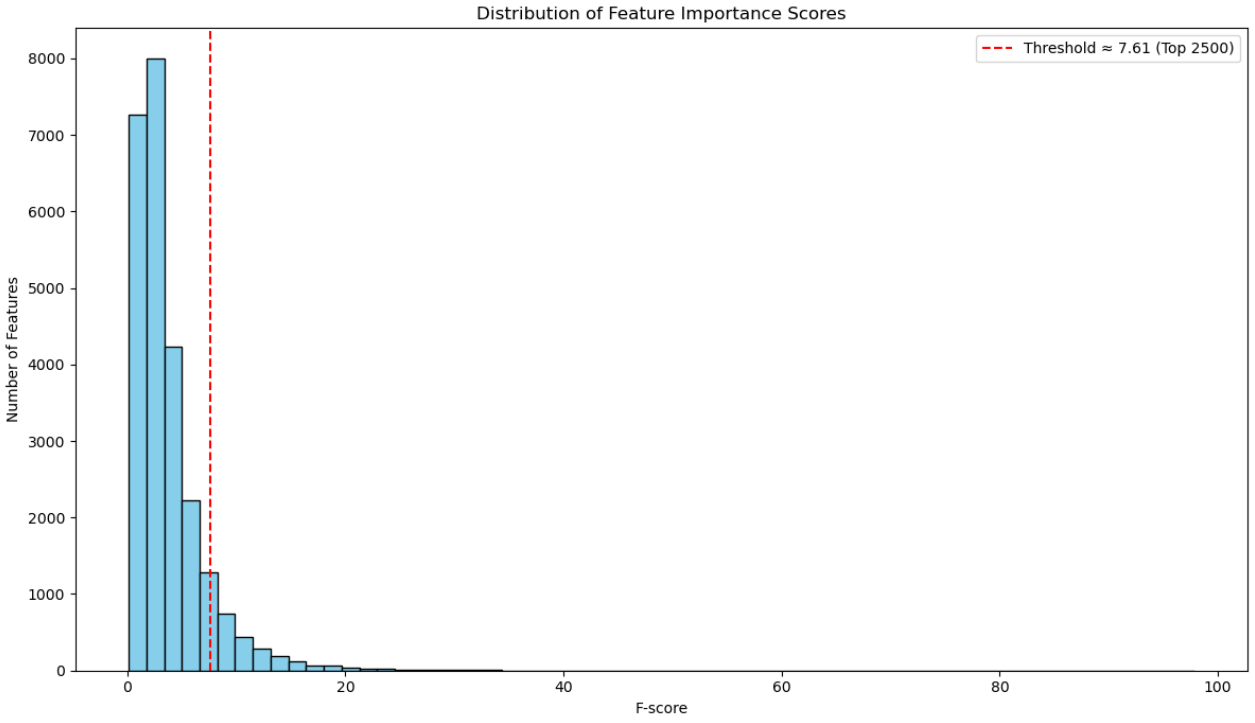

Fig. 4. Distribution of Feature Important Scores. The threshold line marks the cutoff separating the top 2,500 most discriminative features from the remaining less relevant ones.

Manuscript received July 13, 2024; Revised September 5, 2025; Accepted September 24, 2025; date of publication September 26, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i4.1030

dense layers are removed (include_top=False), retaining only the convolutional part of the network. The input size is set to (224, 224, 3), which matches the resized image dimensions and the RGB color channels. The pooling parameter is set to none, meaning no additional global pooling layer is added after the final convolutional block, allowing direct access to the high-dimensional feature map from the last max pooling layer.

The last convolutional layer is chosen specifically because it encodes the most abstract and semantically rich features, capturing complex patterns and morphological details that are highly relevant for gallbladder disease classification while discarding low-level noise present in earlier layers. With this configuration, the number of features extracted from each image is approximately 25,088. These features are obtained by forwarding the input image through the VGG16 architecture up to the last convolutional and max-pooling layers, producing a 3D tensor of size 7 × 7 × 512 flattened into a one-dimensional vector. The model functions as a fixed feature encoder by eliminating the final classification layers and keeping just the convolutional feature extraction components, producing compact yet informative descriptors that capture the textural and morphological characteristics relevant to the classification process.

D. Feature Selection

After feature extraction, a feature selection process is conducted to isolate features with the highest discriminatory value and reduce the overall dimensionality of the dataset. This stage plays a vital role in discarding irrelevant or redundant attributes, thereby enhancing the model's efficiency while minimizing the potential for overfitting [49]. Univariate feature selection was selected due to its simplicity, computational efficiency, and proven ability to improve model performance, particularly in high-dimensional datasets. Several studies have shown that univariate approaches can yield classification results comparable to, or better than, more complex multivariate methods [50][51]. In this study, the SelectKBest method combined with the ANOVA F-test (f_classif) was employed to evaluate the degree of statistical association between individual features and the output classes. The ANOVA F-test effectively identifies features that display substantial variance across different class labels, making it particularly suitable for classification problems [50].

From the initial 25,088 features extracted using the VGG16 architecture, the top 2,500 features with the highest F-scores were selected. This dimensionality reduction significantly lowers computational complexity while preserving the most informative characteristics of the images. As a result, the selected feature subset improves model efficiency and enhances generalization performance by reducing overfitting and focusing the learning process on the most relevant data. Feature selection is guided by the ANOVA F-test, which measures the ability of each feature to distinguish between classes by analyzing the ratio of inter-class variance (Mean Square Between, MSB) to intra-class variance (Mean Square Within, MSW). The F-score is calculated by the following Eq. (3) [52].

$$F = \frac{Between\ Group\ Variance\ (MSB)}{Within\ Group\ Variance\ (MSW)} \quad (3)$$

A higher F-score indicates that a feature has a significant difference in mean values across classes and small variation within each class, making it more useful for distinguishing between categories. Conversely, features with low F-scores exhibit similar distributions across all classes and are considered uninformative. Therefore, selecting features with the highest F-scores ensures that only the most class-discriminative features are retained, contributing to better classification outcomes. The distribution of F-score values used in the feature selection process is illustrated in Fig. 4, where the threshold line clearly marks the cutoff point separating the top 2,500 most informative features from the remaining less relevant ones.

E. SVM

SVM has been extensively used due to its strong capability in managing various classification challenges [53]. Through the use of kernel functions, SVM is capable of managing classification tasks within high-dimensional feature spaces [53]. It can effectively handle data with both linear and non-linear distributions [54]. In linear scenarios, SVM aims to establish a decision boundary that maximizes the separation margin between classes. If the data distribution is non-linear, kernel methods are employed to transform the input space, allowing the model to identify a suitable separating surface [55]. The kernel function is essential for determining the similarity between input samples [56]. In this study, three kernel types were utilized: linear, Radial Basis Function (RBF), and polynomial, with their respective formulations presented in Eq. (4), Eq. (5), and Eq. (6) [56] [57].

Linear. $$K(x_i, x_j) = x_i^T x_j \quad (4)$$

RBF $$K(x_i, x_j) = (x_i^T x_j + 1)^d \quad (5)$$

Polynomial $$K(x_i, x_j) = \exp(-\sigma \parallel x_i - x_j \parallel^2) \quad (6)$$

F. Random Forest

Random Forest is an ensemble-based machine learning method that builds multiple decision trees, where each tree is trained using a different bootstrap-resampled subset of the original data. At every node split, rather than assessing all available features, the algorithm selects a random subset, reducing

Manuscript received July 13, 2024; Revised September 5, 2025; Accepted September 24, 2025; date of publication September 26, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i4.1030

correlation between trees and improving model diversity. This randomness contributes to greater resilience against overfitting and supports stronger generalization performance. Each individual tree produces a separate prediction, and the final outcome is determined by combining these predictions using majority voting in classification problems or averaging in the case of regression [58][59]. The Random Forest algorithm's formula is used in Eq. (7) [60].

$$Gini\ (S) = 1 - \sum pi2\ k\ i = 1 \quad (7)$$

where pi denotes the probability that S belongs to the class i, and k refers to the total number of classes or categories in the dataset. Thus, pirepresents the proportion of data instances belonging to a particular class. The algorithm proceeds through the following phases [60].

1) Randomly select samples from the dataset.
2) Construct a decision tree for each sample and generate predictions from each tree.
3) Tally the prediction frequencies for each class.
4) Select the class with the highest frequency as the final output.

G. KNN

The k-Nearest Neighbors (KNN) algorithm is known for its flexibility, primarily because it does not depend on strict assumptions about the underlying distribution of the data. Its classification capability stems from measuring the distance between data points and assigning a class based on the most frequently occurring label among the nearest samples. As a non-parametric, instance-based learner, KNN consistently performs well across diverse datasets due to its straightforward yet effective decision-making strategy [61].

k-Nearest Neighbors (KNN) uses distance metrics to determine the closeness between data points. KNN identifies neighbors based on the chosen distance metric, significantly affecting similarity assessment and classification results [62]. The analysis incorporated Euclidean, Manhattan, and Minkowski distance functions, computed as shown in Eq. (8), Eq. (9), and Eq. (10), respectively [63].

Euclidean $$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad (8)$$

Manhattan $$\sum_{i=1}^{k} | x_i - y_i | \quad (9)$$

Minkowski $$\left(\sum_{i=1}^{k}(x_i - y_i)^q\right)^{1/q} \quad (10)$$

H. Logistic Regression

Logistic Regression is a commonly utilized method in classification problems, designed to model the association between multiple independent variables and a categorical dependent variable, either nominal or ordinal in type [64]. It performs well in both binary and multiclass classification scenarios by estimating class probabilities based on the input attributes. This technique is often utilized in predictive analytics to assess the chance that a given sample falls into a particular category [65]. The underlying mechanism involves using a logistic (sigmoid) function that maps the prediction result to a probability score ranging from 0 to 1, thus enabling clear and interpretable classification outcomes [66]. The Logistic Regression algorithm's formula is used in Eq. (11).

Table 1. Hyperparameter Tuning Configuration

Model	Parameter	Value
SVM	kernel	Linear, RBF, Polynomial
Random Forest	max_depth	1 - 10
	n_estimators	1 - 300
	max_features	Sqrt, Log2
	min_sample_split	1 - 10
	min_sample_leaf	1 - 10
KNN	n_neighbors	1 - 5
	distance	Minkowski, Euclidean, Manhattan
Logistic Regression	solver	Liblinear, Lbfgs, Saga
	max_iter	1 - 200
	penalty	None, L1, L2

Manuscript received July 13, 2024; Revised September 5, 2025; Accepted September 24, 2025; date of publication September 26, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i4.1030

$$log\left(\frac{Pbj}{1-Pbj}\right) + \beta_1 X_{1j} + \beta_2 X_{2j} + \cdots + \beta_n X_{nj} \quad (11)$$

In this equation, β denotes the coefficient (slope) associated with the n independent variable, while X_{nj} refers to the value of the n independent variable in the j record. The variable n indicates the total number of independent variables, and j represents the number of records in the dataset [67].

I. Grid Search Hyperparameter Tuning

Grid Search is performed using a predefined set of parameter values to achieve optimal accuracy and AUC performance [68]. It systematically explores all specified hyperparameter combinations to determine the most suitable value for each parameter, Eq. (12) [69].

$$\text{Parameter} = \arg max_{\theta \, \epsilon \, G} \quad (12)$$

The expression $\theta \, \epsilon \, G$ indicates that every possible configuration of tuning parameters (θ) within the grid set (G) is considered. The function f(θ) serves as a performance function used to measure how well the model performs under a specific parameter configuration. The Grid Search process involves several key steps [70], which include:

1) Automatically constructs combinations of parameter values based on the range defined for each hyperparameter. For instance, if three parameters are each assigned three values, grid search examines all 3 × 3 × 3 = 27 possible setups.
2) Assesses and compares all candidate configurations of hyperparameters.
3) Selects the optimal set of parameters that yields the highest performance according to the chosen metric.

In the implementation phase, hyperparameter tuning for the classification models, including SVM, Random Forest, KNN, and Logistic Regression, was conducted using the Grid Search technique with cross-validation. This method systematically evaluates all possible combinations of predefined hyperparameter values to determine the parameter setting that yields the best predictive outcome on the training dataset. The complete hyperparameter configurations used for each algorithm are detailed in Table 1.

J. Confusion Matrix

Within the field of machine learning, the effectiveness of a classification model is typically evaluated using a confusion matrix, which provides a structured summary of the model's prediction outcomes in comparison to the true class labels [71]. This matrix facilitates the examination of how well predicted categories align with their corresponding ground truth values [72]. The confusion matrix consists of four key components that summarize classification outcomes: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) [73].

True Positives are cases where the model correctly predicts the presence of a disease, while True Negatives occur when it accurately identifies the absence of a specific disease. False Positives arise when a disease is incorrectly predicted, and False Negatives occur when the model fails to detect an existing disease. These outcomes form the basis for the performance metrics in this study.

In multiclass classification, a one-vs-all confusion matrix is built for each class to obtain TP, TN, FP, and FN, with metrics computed per class and averaged using the macro-average to give all classes equal weight regardless of frequency. The evaluation metrics considered in this study are based on confusion matrix parameters used to assess each performance indicator. Each evaluation metric is defined by the formulas listed in Eq. (13), Eq. (14), Eq. (15), Eq. (16) [74].

$$Accuracy = \frac{TP+TN}{(TP+FN)+(FP+TN)} \quad (13)$$

$$Precision = \frac{TP}{TP+FP} \quad (14)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (15)$$

$$F1\ Score = \frac{2*Precision*Sensitivity}{Precision+Sensitivity} \quad (16)$$

K. Area Under The ROC Curve

This value is obtained by computing the area beneath the ROC curve, where the AUC indicates the model's capability to effectively distinguish between positive and negative class labels [75]. The classification strategy suggests that, on average, positive instances should receive higher scores than negative ones when sampled randomly. Consequently, a greater AUC value reflects a model's stronger capability to separate positive and negative classes [76] accurately .

In multiclass classification, ROC-AUC is computed using the one-vs-rest approach, generating an ROC curve and AUC for each class, then applying macro-averaging to ensure equal class contribution. AUC ranges from 0 to 1, with higher values indicating stronger model discrimination. Its computational formula is shown in Eq. (17) [77].

$$AUC = \frac{\left(\frac{TP}{TP+FN}\right)x\left(\frac{TN}{TN+FP}\right)}{2} \quad (17)$$

AUC measures a model's ability to distinguish between classes and serves as a key metric for comparing algorithms. Higher AUC indicates better prediction, with 0.90 to 1.00 being deemed outstanding, 0.80 to 0.90 indicating strong performance, 0.70 to 0.80 suggesting

Manuscript received July 13, 2024; Revised September 5, 2025; Accepted September 24, 2025; date of publication September 26, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i4.1030

moderate accuracy, 0.60 to 0.70 being considered weak, and 0.50 to 0.60 being ineffective [77].

III. Result

A. Experimental Setup

This study utilized a dataset comprising 10,692 ultrasound images of the gallbladder, distributed across nine distinct diagnostic categories. To maintain evaluation integrity and minimize bias, the dataset was split using scikit-learn's train_test_split function with stratification to preserve class proportions. A total of 8,554 images (80%) were allocated for training and 2,138 images (20%) for testing. A fixed random seed of 42 was applied to ensure reproducibility across runs. Furthermore, 10-fold cross-validation was employed during the model tuning phase to enhance generalizability and prevent overfitting.

All experiments were performed locally on a MacBook featuring an Apple M3 processor (8-core CPU, 10-core integrated GPU, 8 GB unified memory) running macOS Sequoia. The development setup utilized Python 3.12.7 within Spyder IDE version 5.5.5, incorporating major libraries such as Pandas 2.2.0, TensorFlow 2.12.0, scikit-learn 1.5.1, Matplotlib 3.9.2, NumPy 1.23.5, and Seaborn 0.13.2. Computations were carried out exclusively in CPU mode without GPU acceleration. All scripts were executed within the Spyder interactive environment under stable runtime conditions, with no other processes running concurrently during training.

B. Classification Model Without Feature Selection and Default Tuning

This study evaluated four classification algorithms: SVM, Random Forest, KNN, and Logistic Regression, using their default hyperparameters without feature selection. To assess model robustness and generalizability, 10-fold cross-validation was performed on the training data. The results indicated that SVM achieved the highest average accuracy of 98.32%, closely followed by Logistic Regression at 98.50%, then Random Forest at 95.87%, and KNN at 77.10%. Comparable patterns were observed across precision, recall, and F1-score metrics, confirming the consistent superior performance of SVM and Logistic Regression.

On the test set, SVM achieved the highest accuracy (98.60%), followed closely by LR (98.55%) and RF (95.98%), while KNN lagged at 77.61%. These results highlight the strong performance of SVM and LR even without tuning or feature selection. However, SVM required the longest testing time (824.21 s) compared to LR (80.11 s), RF (42.18 s), and KNN (11.61 s), as shown in Table 2. Although all models provided usable predictions, factors such as testing efficiency and interpretability should also be considered when selecting the most suitable deployment model.

C. Classification Model with Feature Selection and Hyperparameter Tuning

Hyperparameter optimization using Grid Search CV was performed to enhance predictive performance for each classifier. After 10-fold cross-validation, tuned models achieved accuracies of 99.15% (SVM), 98.35% (Logistic Regression), 98.10% (Random Forest), and 93.10% (KNN), showing notable improvements over default settings. The optimal parameters for each model are as follows:

1) Support Vector Machine (SVM) achieved its best performance using a linear kernel, which is well-

Table 2. Testing result for machine learning models without feature selection and default tuning

Model	Accuracy	Precision	Recall	F1-Score	Time (s)
SVM	0.9860	0.9861	0.9860	0.9860	824.21
Random Forest	0.9598	0.9621	0.9598	0.9599	42.18
KNN	0.7761	0.7984	0.7761	0.7756	11.61
Logistic Regression	0.9855	0.9857	0.9855	0.9855	80.11

Table 3. Testing result for machine learning models with feature selection and hyperparameter tuning

Model	Accuracy	Precision	Recall	F1-Score	Time (s)
SVM + UFS + GS	0.9935	0.9932	0.9935	0.9933	18.4
Random Forest + UFS + GS	0.9818	0.9839	0.9809	0.9822	36.39
KNN + UFS + GS	0.9360	0.9415	0.9348	0.9371	1.24
Logistic Regression + UFS + GS	0.9864	0.9878	0.9856	0.9866	20.47

Manuscript received July 13, 2024; Revised September 5, 2025; Accepted September 24, 2025; date of publication September 26, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i4.1030

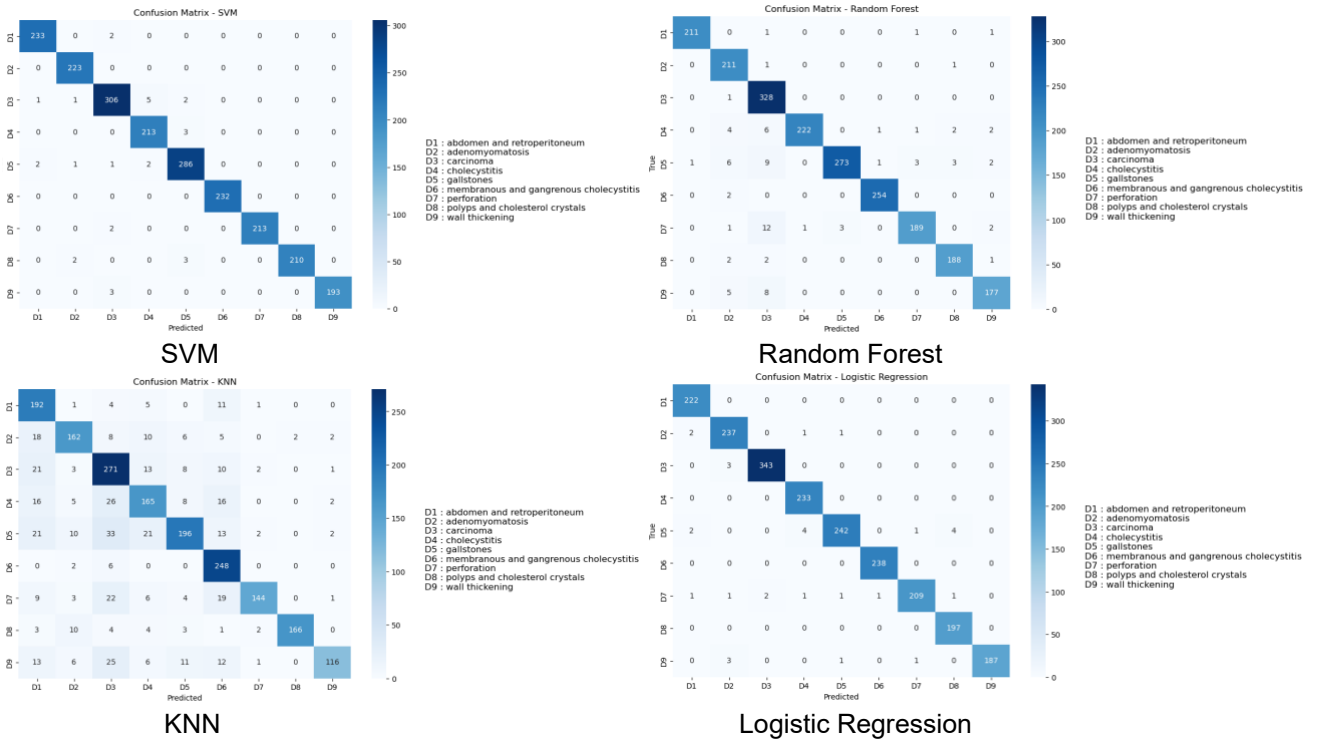

Fig. 5. Confusion Matrix for machine learning models without feature selection and default tuning

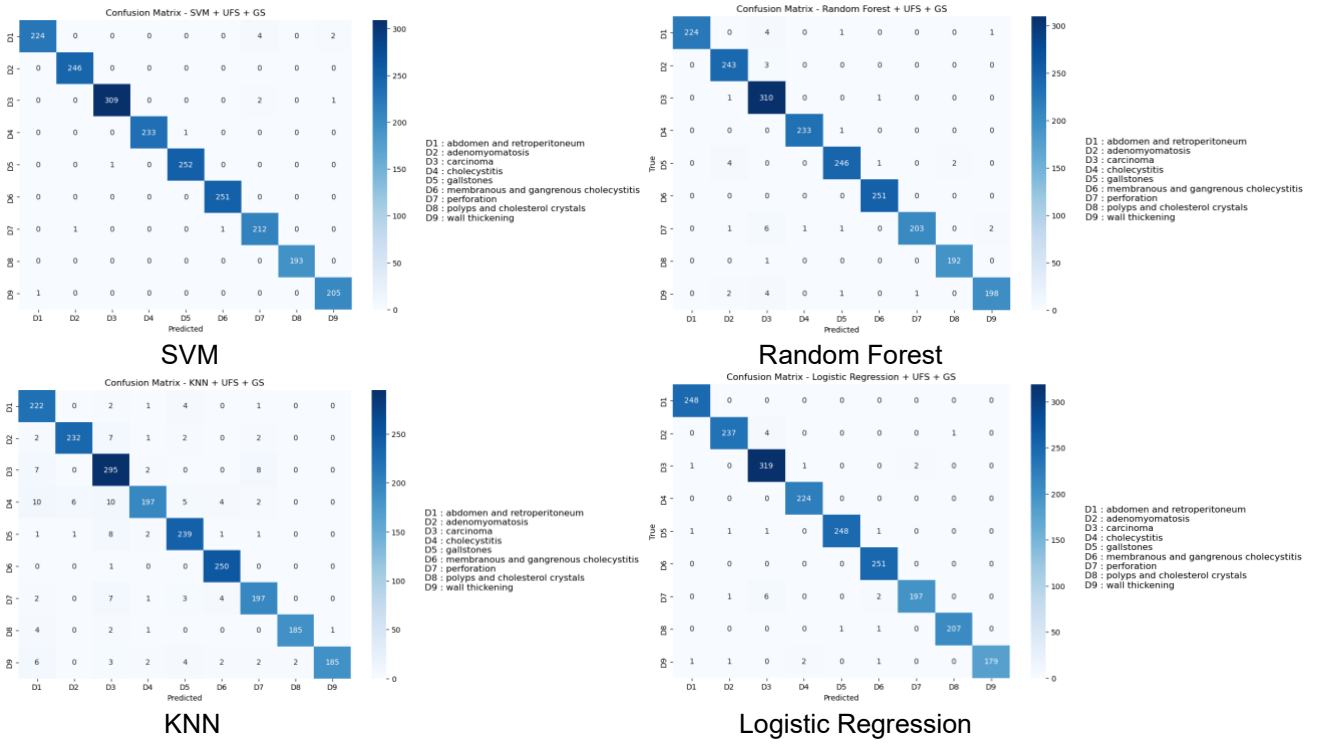

Fig. 6. Confusion Matrix for machine learning models with feature selection and hyperparameter tuning

suited for linearly separable feature spaces produced by the CNN-based extraction process.

2) k-Nearest Neighbors (KNN) achieved its best performance using three nearest neighbors (n_neighbors = 3) along with the Euclidean distance

Manuscript received July 13, 2024; Revised September 5, 2025; Accepted September 24, 2025; date of publication September 26, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i4.1030

metric, striking an effective balance between model simplicity and localized accuracy.

3) The optimal performance of the Random Forest (RF) classifier was achieved using the following configuration: 300 decision trees (n_estimators = 300), a tree depth capped at 10 (max_depth = 10), feature subset selection based on the square root strategy (max_features = 'sqrt'), at least two instances per terminal node (min_samples_leaf = 2), a split criterion requiring no fewer than four examples (min_samples_split = 4), and the use of 'balanced' class weights to compensate for imbalanced class proportions in the dataset.
4) For Logistic Regression (LR), the best result was obtained using the liblinear solver, no regularization (penalty = None), and a maximum iteration limit of 200 (max_iter = 200), which helped the model converge efficiently without explicit regularization.

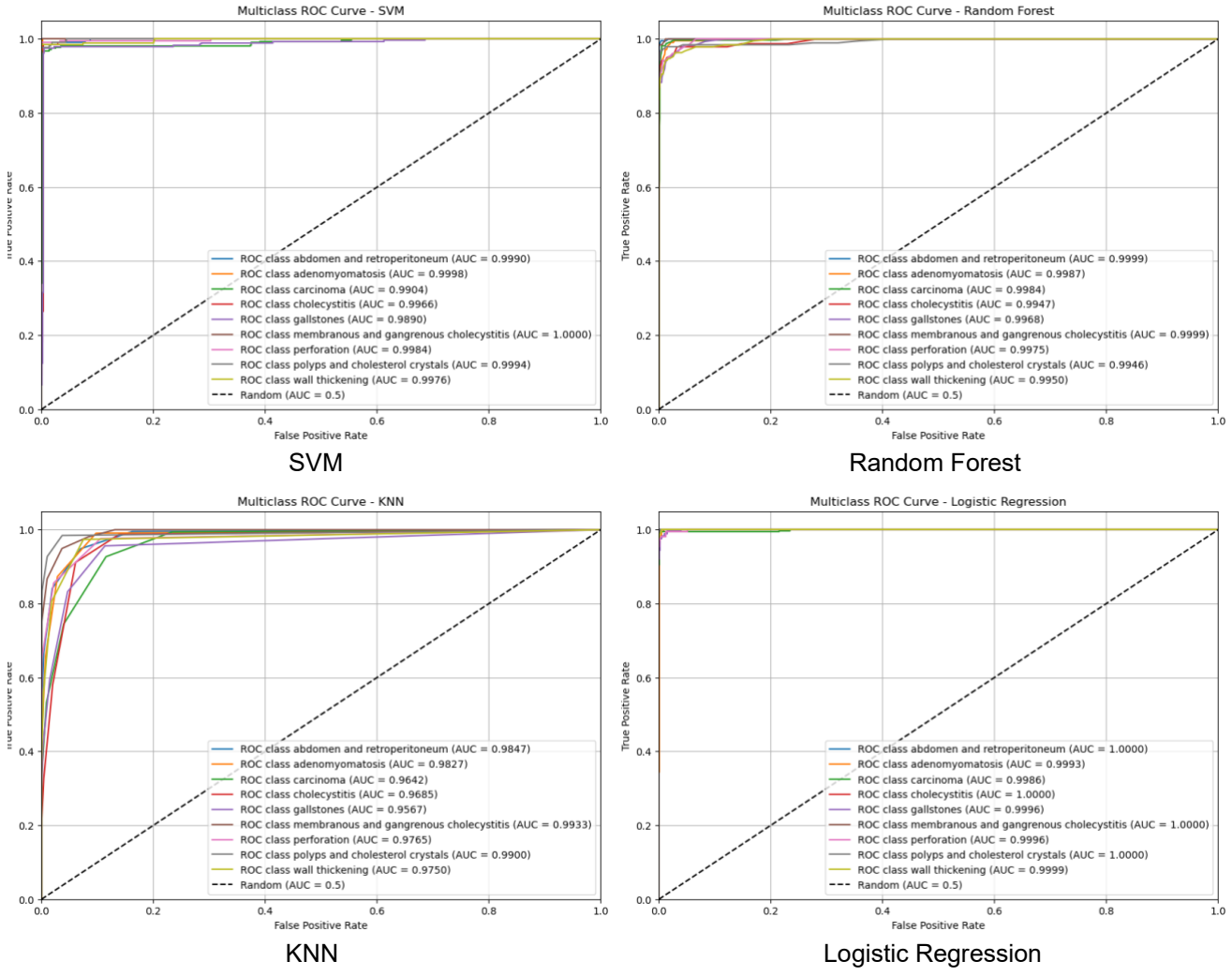

Fig. 7. AUC-ROC for machine learning modesls without feature selection and default tuning

In the subsequent evaluation, all four models were re-tested using Univariate Feature Selection (UFS) and Grid Search (GS) tuning, yielding notable improvements across accuracy, precision, recall, and F1-score. SVM + UFS + GS achieved the highest accuracy (99.35%), with a precision of 99.32%, a recall of 99.35%, and a reduced testing time of 18.4 s. Logistic Regression + UFS + GS followed with 98.64% accuracy (20.47 s), while Random Forest + UFS + GS attained 98.18% accuracy with balanced metrics and a 36.39 s testing time.

Notably, KNN + UFS + GS showed the most substantial relative improvement, increasing its accuracy from 77.61% to 93.60%, with a rapid testing time of just 1.24 seconds. These findings demonstrate that combining feature selection and hyperparameter

Manuscript received July 13, 2024; Revised September 5, 2025; Accepted September 24, 2025; date of publication September 26, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i4.1030

tuning not only boosts model accuracy but also significantly reduces testing time, making the models more efficient and viable for real-time or resource-constrained deployment scenarios, as summarized in Table 3.

D. Confusion Matrix – without Feature Selection and Default Tuning

The confusion matrix evaluation was performed on a test set of 2,139 ultrasound images using four machine learning models: SVM, Random Forest, KNN, and Logistic Regression. The SVM model demonstrated strong classification performance, correctly predicting 233 images of abdomen and retroperitoneum (2 misclassifications), 223 of adenomyomatosis (0 errors), 306 of carcinoma (9 errors), 213 of cholecystitis (3 errors), 286 of gallstones (6 errors), 232 of membranous and gangrenous cholecystitis (0 errors), 213 of perforation (2 errors), 210 of polyps and cholesterol crystals (5 errors), and 193 of wall thickening (3 errors).

The Random Forest model achieved high accuracy across most classes, with only 3 errors in abdomen and retroperitoneum, 2 in adenomyomatosis, 1 in carcinoma, 16 in cholecystitis, 25 in gallstones, 2 in membranous and gangrenous cholecystitis, 19 in perforation, 5 in polyps and cholesterol crystals, and 13 in wall thickening. However, its performance declined slightly in more complex cases such as gallstones and perforation.

In contrast, KNN recorded a higher number of errors across nearly all classes, including 22 in abdomen and retroperitoneum, 51 in adenomyomatosis, 58 in carcinoma, 73 in cholecystitis, 102 in gallstones, 8 in membranous and gangrenous cholecystitis, 64 in perforation, 27 in polyps and cholesterol crystals, and 74 in wall thickening. This indicates that KNN struggled to generalize in the multiclass setting, especially for classes with overlapping features or limited samples, due to its sensitivity to noise, outliers, and the complexity of the high-dimensional feature space extracted by VGG16.

The Logistic Regression model showed excellent classification accuracy across nearly all categories: 222 images of abdomen and retroperitoneum were correctly classified, with no misclassifications. 237 images of adenomyomatosis were correctly classified, with 4 misclassifications. 343 images of carcinoma were correctly classified, with 3 misclassifications. 233 images of cholecystitis were correctly classified, with no misclassifications. 242 images of gallstones were correctly classified, with 11 misclassifications. 238 images of membranous and gangrenous cholecystitis were correctly classified, with no misclassifications. 209 images of perforation were correctly classified, with 8 misclassifications. 197 images of polyps and cholesterol crystals were correctly classified, with no misclassifications. 187 images of wall thickening were correctly classified, with 5 misclassifications. A detailed visual representation of the confusion matrices for all models can be seen in Fig. 5.

E. Confusion Matrix – with Feature Selection and Hyperparameter Tuning

After applying feature selection combined with hyperparameter tuning, the classification performance of all models improved noticeably, as reflected in their respective confusion matrices. The SVM model achieved excellent results, correctly predicting 224 images of abdomen and retroperitoneum (6 misclassifications), 246 of adenomyomatosis (0 errors), 309 of carcinoma (3 errors), 233 of cholecystitis (1 error), 252 of gallstones (1 error), 251 of membranous and gangrenous cholecystitis (0 errors), 212 of perforation (2 errors), 193 of polyps and cholesterol crystals (0 errors), and 205 of wall thickening (1 error).

The Random Forest (RF) model also showed strong results, with 224 correct predictions for abdomen and retroperitoneum (6 errors), 243 of adenomyomatosis (3 errors), 310 of carcinoma (2 errors), 233 of cholecystitis (1 error), 246 of gallstones (7 errors), 251 of membranous and gangrenous cholecystitis (0 errors), 203 of perforation (11 errors), 192 of polyps and cholesterol crystals (1 error), and 198 of wall thickening (8 errors). While improved, the k-Nearest Neighbors (KNN) model still exhibited relatively higher error rates compared to SVM and RF. It correctly predicted 222 images of abdomen and retroperitoneum (8 errors), 232 of adenomyomatosis (14 errors), 295 of carcinoma (17 errors), 197 of cholecystitis (37 errors), 239 of gallstones (14 errors), 250 of membranous and gangrenous cholecystitis (1 error), 197 of perforation (17 errors), 185 of polyps and cholesterol crystals (8 errors), and 185 of wall thickening (21 errors).

Lastly, the Logistic Regression (LR) model demonstrated robust performance, with 248 correct predictions for abdomen and retroperitoneum (0 errors), 237 of adenomyomatosis (5 errors), 319 of carcinoma (4 errors), 224 of cholecystitis (0 errors), 248 of gallstones (4 errors), 251 of membranous and gangrenous cholecystitis (0 errors), 197 of perforation (9 errors), 207 of polyps and cholesterol crystals (2 errors), and 179 of wall thickening (5 errors). These results show that SVM and LR maintained consistently high accuracy, while RF followed closely behind. Although KNN improved, it still produced more misclassifications in several classes. A detailed visual representation of the confusion matrices for all models can be seen in Fig. 6.

F. AUC - ROC without Feature Selection and Default Tuning

The ROC (Receiver Operating Characteristic) analysis was conducted to evaluate each model's classification

Manuscript received July 13, 2024; Revised September 5, 2025; Accepted September 24, 2025; date of publication September 26, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i4.1030

performance in distinguishing between the nine diagnostic categories of gallbladder disease.

All four models demonstrated strong discriminatory capabilities, as reflected in their respective AUC scores. The Support Vector Machine (SVM) model achieved high AUC values, including 0.9990 for abdomen and retroperitoneum, 0.9998 for adenomyomatosis, 0.9904 for carcinoma, 0.9966 for cholecystitis, 0.9890 for gallstones, 1.0000 for membranous and gangrenous cholecystitis, 0.9984 for perforation, 0.9994 for polyps and cholesterol crystals, and 0.9976 for wall thickening, resulting in a mean AUC of 0.9967. The Random Forest (RF) model also delivered excellent results, with AUC scores of 0.9999 for abdomen and retroperitoneum, 0.9987 for adenomyomatosis, 0.9984 for carcinoma, 0.9947 for cholecystitis, 0.9968 for gallstones, 0.9999 for membranous and gangrenous cholecystitis, 0.9975 for perforation, 0.9946 for polyps and cholesterol crystals, and 0.9950 for wall thickening resulting in a mean AUC of 0.9973.

The k-Nearest Neighbors (KNN) model, while slightly behind in performance, still demonstrated satisfactory class separability, with AUC scores of 0.9847, 0.9827, 0.9642, 0.9685, 0.9567, 0.9933, 0.9765, 0.9900, and 0.9750 across the respective classes, yielding a mean AUC of 0.9768. Lastly, the Logistic Regression (LR) model achieved the most consistent and superior AUC values, including perfect scores (1.0000) for cholecystitis, abdomen and retroperitoneum, polyps and cholesterol crystals, and membranous and gangrenous cholecystitis, while maintaining scores above 0.9980 for all other classes, resulting in the highest mean AUC of 0.9989. These findings affirm that all models effectively distinguish between the diagnostic classes, with LR and RF exhibiting the most robust and consistent performance.

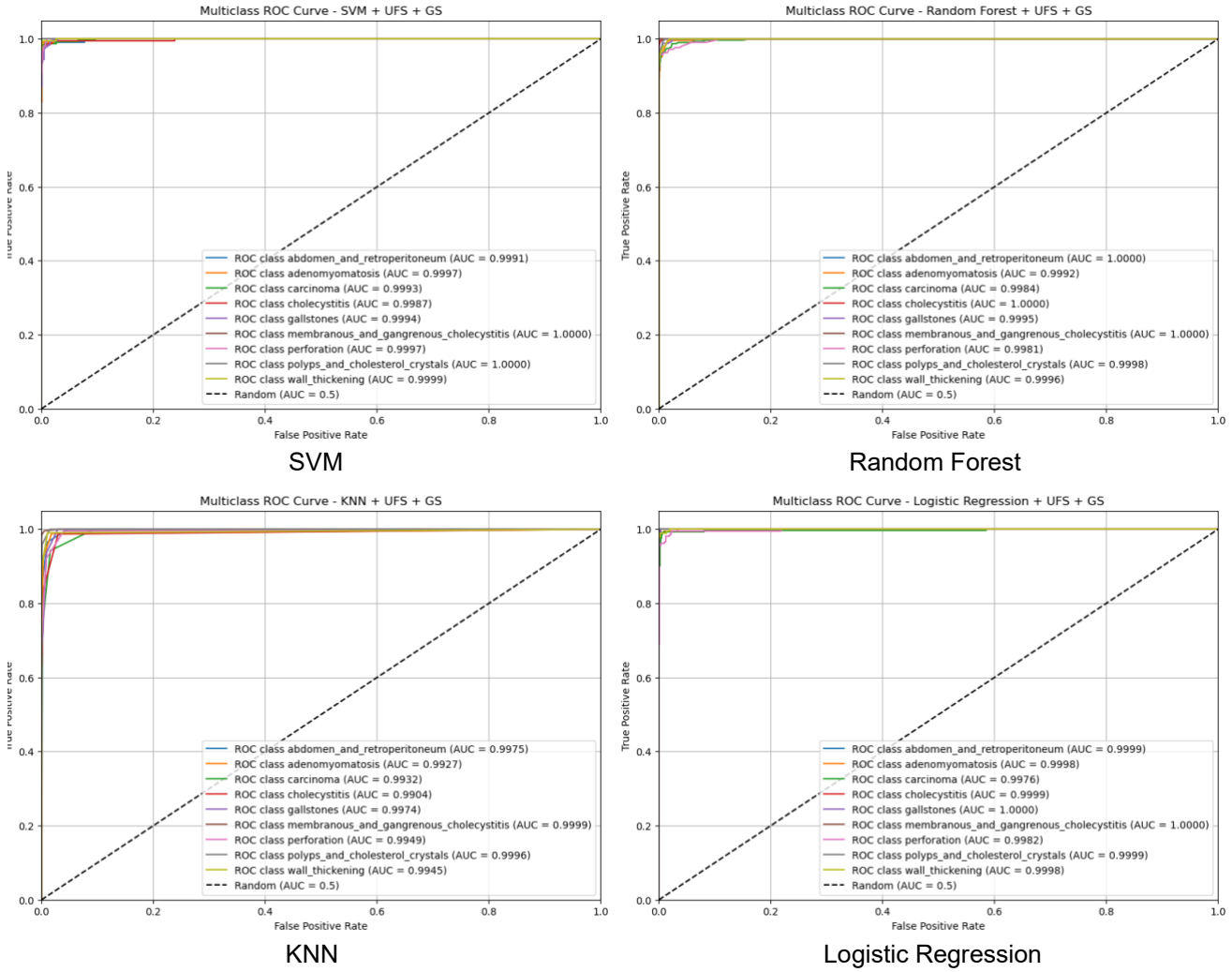

Fig. 8. AUC-ROC for machine learning models with feature selection and hyperparameter tuning

Manuscript received July 13, 2024; Revised September 5, 2025; Accepted September 24, 2025; date of publication September 26, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i4.1030

A detailed visual representation of the ROC curves for all models is provided in Fig. 7.

G. AUC - ROC with Feature Selection and Hyperparameter Tuning

The ROC (Receiver Operating Characteristic) analysis was conducted to evaluate the classification performance of all models after applying feature selection and hyperparameter tuning. The Support Vector Machine (SVM) model achieved excellent results across all diagnostic categories, with AUC values of 0.9991 for abdomen and retroperitoneum, 0.9997 for adenomyomatosis, 0.9993 for carcinoma, 0.9987 for cholecystitis, 0.9994 for gallstones, 1.0000 for membranous and gangrenous cholecystitis, 0.9997 for perforation, 1.0000 for polyps and cholesterol crystals, and 0.9999 for wall thickening. These results yielded a mean AUC of 0.9995, indicating outstanding class separability. The Random Forest model also delivered robust performance with near-perfect scores: 1.0000 for abdomen and retroperitoneum, 0.9992 for adenomyomatosis, 0.9984 for carcinoma, 1.0000 for cholecystitis, 0.9995 for gallstones, 1.0000 for membranous and gangrenous cholecystitis, 0.9981 for perforation, 0.9998 for polyps and cholesterol crystals, and 0.9996 for wall thickening, resulting in a mean AUC of 0.9994.

Despite slightly lower values, the k-Nearest Neighbors (KNN) model still maintained high discriminative ability, with AUC scores of 0.9975, 0.9927, 0.9932, 0.9904, 0.9974, 0.9999, 0.9949, 0.9996, and 0.9945 across the nine categories, respectively. This corresponds to a mean AUC of 0.9956, reflecting reliable performance after optimization. The Logistic Regression (LR) model yielded the most consistent near-perfect results among all methods, with AUC values of 0.9999, 0.9998, 0.9976, 0.9999, 1.0000, 1.0000, 0.9982, 0.9999, and 0.9998 for the respective classes. These values translate to the highest overall mean AUC of 0.9994. In summary, the AUC-ROC analysis clearly demonstrated the significant boost in classification capability for all models following the application of feature selection and hyperparameter tuning. All classifiers achieved near-optimal class separability, with Logistic Regression and SVM emerging as the top-performing models, supported by consistent AUC values across all diagnostic categories. A detailed graphical representation of the ROC curves is shown in Fig. 8.

IV. Discussion

This study demonstrates the substantial impact of feature selection combined with hyperparameter tuning on the performance of machine learning models for gallbladder disease classification using ultrasound images. Four classification algorithms, SVM, Random Forest, KNN, and Logistic Regression, were evaluated under two scenarios: without optimization, and with Univariate Feature Selection (UFS) and Grid Search (GS) tuning.

Without optimization, SVM and LR already achieved high classification accuracies of 98.60% and 98.55%, respectively, though SVM incurred a notably longer testing time (824.21 seconds). KNN showed the weakest performance with 77.61% accuracy, highlighting its limited suitability under default conditions. However, after applying UFS and GS, all models exhibited substantial improvements. SVM + UFS + GS emerged as the best-performing model with a 99.35% accuracy, AUC of 0.9995, and significantly reduced testing time (18.4 seconds). KNN displayed the greatest relative improvement, increasing its accuracy by 16% and achieving the fastest execution time (1.24 seconds). This notable reduction in testing time suggests potential for improved clinical workflow efficiency, particularly in settings requiring rapid diagnostic decisions. However, it is essential to critically assess whether this efficiency gain affects diagnostic accuracy or reliability. While the optimized models maintain high accuracy, careful consideration of any trade-offs between speed and precision is necessary to ensure patient safety and diagnostic effectiveness in real-world clinical practice.

Feature selection played a crucial role in these performance gains by reducing the dimensionality of the feature space, which removed irrelevant or noisy features that could otherwise degrade classification accuracy and increase computational burden. This process allowed the models, especially SVM and Logistic Regression, to focus on the most informative features extracted from ultrasound images, improving both accuracy and speed. Nevertheless, there remains an inherent risk that the feature selection process might exclude some informative features, potentially leading to less comprehensive models that could miss subtle but clinically important patterns. This limitation highlights the need for future research to explore advanced or hybrid feature selection methods that balance dimensionality reduction with retention of critical diagnostic information, thereby further optimizing model robustness and clinical applicability.

The superior performance of SVM and LR can be attributed to their inherent ability to handle high-dimensional feature spaces and maintain strong generalization under limited noise interference, characteristics well-suited for ultrasound image data that often contains fine-grained texture patterns. SVM, in particular, excels at maximizing the decision margin, effectively separating subtle grayscale variations between gallbladder disease categories. Logistic Regression, while simpler, benefits from its linear

Manuscript received July 13, 2024; Revised September 5, 2025; Accepted September 24, 2025; date of publication September 26, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i4.1030

decision boundaries and probabilistic output, which align well with the dataset's relatively linearly separable feature space after CNN-based extraction. In contrast, KNN is more sensitive to irrelevant features and noise, leading to degraded performance in high-dimensional spaces, while Random Forest may underperform when critical discriminatory information is distributed across many correlated features. These algorithmic characteristics explain why SVM and LR maintained superior accuracy and AUC compared to KNN and RF in this study.

Analysis of the confusion matrix revealed that misclassifications predominantly occurred in complex or overlapping diagnostic categories, such as gallstones, perforation, and adenomyomatosis. These classes present subtle textural and morphological similarities in ultrasound images, increasing the difficulty of accurate discrimination. Notably, models like KNN showed higher error rates in these categories, likely due to sensitivity to noise and reliance on local distance measures in a high-dimensional feature space. Conversely, SVM and Logistic Regression consistently minimized misclassifications across all classes, reflecting their robustness in capturing nuanced differences between closely related categories. Random Forest also performed strongly but was slightly less effective in complex cases such as gallstones and perforation.

Further, AUC-ROC analysis confirmed the superior discriminatory capability of SVM and Logistic Regression, with both achieving near-perfect class separation across nine diagnostic categories, demonstrating excellent suitability for multi-class gallbladder disease classification. However, this systematic pattern of errors also highlights inherent model limitations when faced with visually ambiguous cases, suggesting potential areas for improvement, such as incorporating advanced feature selection, ensemble methods, or domain-specific data augmentation to enhance model reliability and generalizability in clinical applications.

Although AUC values close to 1 indicate excellent class separability and robust discrimination capabilities, it is crucial to interpret these results in real-world clinical practice. High AUC suggests effective differentiation between disease categories across various decision thresholds, potentially reducing misdiagnosis. Nonetheless, clinical application requires careful management of false positives and false negatives to prevent unnecessary interventions, patient anxiety, increased healthcare costs, or delayed treatment, impacting patient outcomes. Therefore, in addition to AUC, metrics such as precision, recall, and confusion matrix analyses are vital to comprehensively understand model performance on specific disease categories. The consistently high precision and recall for SVM and Logistic Regression indicate these classifiers maintain a low rate of clinically significant errors while separating classes effectively. Furthermore, successful clinical integration of these models demands rigorous validation on diverse external datasets and consideration of ultrasound image variability, operator expertise, and patient heterogeneity. Real-world deployment also requires interpretability and confidence estimation to support clinicians in making informed decisions based on model outputs. Future research should focus on prospective clinical trials and developing decision-support systems that combine these performance metrics with clinical risk factors to optimize patient safety and diagnostic accuracy.

Despite these promising metrics, it is essential to contextualize how such high accuracy and discrimination translate into clinical decision-making. The practical benefits of accurate classification must be weighed against the risks of misclassification errors, including false positives and false negatives, which can respectively lead to unnecessary interventions or delayed treatment. Such errors may impact patient outcomes, healthcare costs, and clinician confidence in AI-assisted diagnosis. Therefore, integrating these models into clinical workflows requires not only robust performance metrics but also a thorough understanding of their decision boundaries, error patterns, and interpretability to support informed clinical judgments. Future work should prioritize prospective clinical studies that evaluate these models in real-world settings and develop decision-support systems that combine algorithmic outputs with clinical risk factors to optimize patient safety and diagnostic efficacy.

This study uses a dataset of 10,692 ultrasound images, enabling a valid and direct comparison with prior works that employed the same dataset. As summarized in Table 4, Obaid et al. employed a MobileNet-based approach, achieved an accuracy of 98.35%, with recall and F1-score values of 98.30% and 98.34%, respectively. However, their AUC was limited to 0.9340, and the model required 540 seconds of testing time, an efficiency bottleneck for real-time deployment. Bozdağ et al. utilized a content-based image retrieval (CBIR) system that reported a precision of 0.9440, but did not provide accuracy, recall, time processing, or AUC metrics, which limits comprehensive performance evaluation and comparison. Shuvo and Chowdhury, using a different gallbladder cancer dataset, implemented VGG19 and XceptionNet architectures and achieved an accuracy of 85.44%, precision of 85.82%, recall of 85.19%, and F1-score of 85.25%. However, they did not report AUC or inference time. Similarly, Dadjouy and Sajedi, also using a different gallbladder cancer dataset, proposed the GBCRet method, attaining an accuracy of 92.62%

Manuscript received July 13, 2024; Revised September 5, 2025; Accepted September 24, 2025; date of publication September 26, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i4.1030

and a recall of 85.71%, without reporting precision, F1-score, AUC, or execution time. In contrast, the proposed method in this study combines Support Vector Machine (SVM) classification with Univariate Feature Selection (UFS) and Grid Search (GS) hyperparameter tuning, delivering superior performance across all evaluation metrics.

Specifically, it achieved an accuracy of 99.35%, precision of 99.32%, recall of 99.35%, F1-score of 99.33%, and an AUC near perfect at 0.9995. Moreover, the testing time was drastically reduced to 18.4 seconds, significantly improving inference efficiency. The critical differences in methodology namely, the use of CNN-based feature extraction followed by rigorous feature selection and hyperparameter optimization, likely contribute to these performance gains. Additionally, the disparity in evaluation metrics reported by previous studies highlights the necessity for standardized benchmarks to facilitate clearer, more meaningful comparisons. Therefore, while the quantitative results underscore the superior accuracy and efficiency of the current approach, this critical analysis reveals genuine contributions in terms of methodological robustness and practical applicability, advancing the state-of-the-art in gallbladder disease classification from ultrasound images.

While the outcomes appear promising, this research is not without its limitations. First, it uses a single dataset, which may not capture the diversity and variability in real-world clinical settings. This dataset's image quality and labeling are relatively consistent, unlike real-world ultrasound data that often suffers from noise, motion artifacts, and inconsistent annotations. Additionally, while feature selection helped improve performance, there is a risk that some valuable information may have been excluded during dimensionality reduction. This limitation highlights the necessity for future research to explore advanced or hybrid feature selection methods that balance dimensionality reduction with retention of critical diagnostic information, thereby further optimizing model robustness and clinical applicability. Moreover, no external validation dataset was used, limiting the results' generalizability. Furthermore, the reliance on a single dataset restricts the generalizability of the findings, as it may not fully represent the variability encountered in diverse clinical environments, including differences in ultrasound equipment, patient populations, and operator techniques. This limitation underscores the need for future studies to validate the proposed models on multiple, heterogeneous external datasets to ensure robustness and broad applicability in real-world settings.

Additionally, future research should consider validating the models on multi-center datasets to capture clinical heterogeneity better and improve external validity. Integrating multimodal data sources, such as combining ultrasound imaging with clinical records, laboratory tests, or patient history, could enhance diagnostic accuracy and provide a more holistic approach to gallbladder disease classification. These strategies would help advance the translation of current findings into practical, reliable clinical tools.

This research offers a validated and efficient pipeline for ultrasound-based diagnosis of gallbladder diseases, combining CNN-based feature extraction with optimized machine learning models. Its high performance and low testing latency demonstrate strong potential for integration into real-time diagnostic tools, especially in resource-constrained or underserved healthcare environments. These findings contribute meaningfully to AI-driven medical imaging, where model performance, efficiency, and reproducibility are critical for clinical adoption. Exploration of ensemble learning, hybrid architectures, attention-based models, transfer learning techniques, and lightweight deep learning frameworks may also yield further improvements in diagnostic performance, scalability, and real-world applicability.

Table 4. Comparison with previous research

Research	Method	Accuracy	Precision	Recall	F1 Score	AUC	Time (s)
Obaid et al [17]	MobileNet	0.9835	-	0.9830	0.9834	0.9340	540
Bozdag et al [18]	CBIR-based system	-	0.9440	-	-	-	-
Shuvo and Chowdhury [19]	VGG19 and XceptionNet	0.8544	0.8582	0.8519	0.8525	-	-
Dadjouy and Sajedi [20]	GBCRet	0.9262	-	0.8571	-	-	-
This Study	**SVM, UFS, GS**	**0.9935**	**0.9932**	**0.9935**	**0.9933**	**0.9995**	**18.4**

V. Conclusion

Manuscript received July 13, 2024; Revised September 5, 2025; Accepted September 24, 2025; date of publication September 26, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i4.1030

This study is centered on developing an automated and optimized classification model for gallbladder disease using ultrasound images by integrating CNN-based feature extraction with Univariate Feature Selection (UFS) and Grid Search (GS) hyperparameter tuning. Four models, SVM, Random Forest, KNN, and Logistic Regression were evaluated under two scenarios: default configuration and optimized setting. The goal was to build a fast, accurate, and efficient classification pipeline suitable for potential clinical deployment.

The key result indicated that the SVM model, combined with feature selection and optimized hyperparameters, yielded superior overall results, achieving 99.35% accuracy, an AUC score of 0.9995, and a markedly shorter testing duration of 18.4 seconds. Even without optimization, both SVM and LR showed high accuracy (98.60% and 98.55% respectively), though SVM's high computation time (824.21s) was a limitation. KNN initially performed poorly with 77.61% accuracy but saw the greatest relative improvement after optimization, reaching 93.60% with the fastest testing time (1.24s). Confusion matrix and AUC-ROC evaluations further confirmed that SVM and LR consistently outperformed other models across all nine disease categories.

A comparative analysis with prior research demonstrated that this study outperformed existing approaches in both accuracy and computational efficiency. For example, while Obaid et al. achieved 98.35% accuracy with an AUC of 0.9340 and a 540-second testing time, our approach achieved higher scores in all metrics. The findings indicate combining feature selection and tuning techniques with CNN-extracted features can lead to more accurate, consistent, and faster diagnostic systems. These results emphasize the strength of this pipeline for multiclass disease classification tasks in medical imaging.

Nonetheless, some limitations must be acknowledged. This study relied on a single, labeled dataset, which may limit its generalizability to real-world clinical environments where image quality and variability are less controlled. The feature selection process, while beneficial, could potentially exclude valuable diagnostic information due to dimensionality reduction. Additionally, the lack of external validation datasets remains a significant constraint, limiting the robustness of the findings across diverse clinical settings. Future work should prioritize validation on larger, more heterogeneous, multi-center datasets to better capture clinical variability and improve external validity. Furthermore, integrating multimodal data sources, conducting prospective clinical trials, and exploring advanced approaches such as ensemble learning, hybrid CNN-machine learning models, attention-based mechanisms, transfer learning, and lightweight deep learning frameworks could further enhance diagnostic robustness, scalability, and real-world clinical deployment, particularly in resource-constrained environments.

References

[1] A. M. Lucchese, M. C. Machry, A. N. Kalil, A. F. Oliveira, and F. J. F. Coimbra, "Exploring Anatomy, Macroscopic and Microscopic Structures, and Physiological Functions of the Gallbladder," in *Gallbladder Cancer*, Cham: Springer Nature Switzerland, 2024, pp. 1–13. doi: 10.1007/978-3-031-76746-3_1.

[2] J. L. Turumin, V. A. Shanturov, and H. E. Turumina, "The role of the gallbladder in humans," *Rev Gastroenterol Mex*, vol. 78, no. 3, pp. 177–187, Jul. 2013, doi: 10.1016/j.rgmx.2013.02.003.

[3] 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," *Medicina (B Aires)*, vol. 58, no. 3, p. 388, Mar. 2022, doi: 10.3390/medicina58030388.

[4] R. Lam, A. Zakko, J. C. Petrov, P. Kumar, A. J. Duffy, and T. Muniraj, "Gallbladder Disorders: A Comprehensive Review," *Disease-a-Month*, vol. 67, no. 7, p. 101130, Jul. 2021, doi: 10.1016/j.disamonth.2021.101130.

[5] N. M. Parra-Landazury, J. Cordova-Gallardo, and N. Méndez-Sánchez, "Obesity and Gallstones," *Visc Med*, vol. 37, no. 5, pp. 394–402, 2021, doi: 10.1159/000515545.

[6] X. Zhang, L. Guan, H. Tian, and Y. Li, "Prevalence and Risk Factors of Gallbladder Stones and Polyps in Liaoning, China," *Front Med (Lausanne)*, vol. 9, Apr. 2022, doi: 10.3389/fmed.2022.865458.

[7] Z.-Z. Li, L.-J. Guan, R. Ouyang, Z.-X. Chen, G.-Q. Ouyang, and H.-X. Jiang, "Global, regional, and national burden of gallbladder and biliary diseases from 1990 to 2019," *World J Gastrointest Surg*, vol. 15, no. 11, pp. 2564–2578, Nov. 2023, doi: 10.4240/wjgs.v15.i11.2564.

[8] M. Piñeros *et al.*, "Global variations in gallbladder cancer incidence: What do recorded data and national estimates tell us?," *Int J Cancer*, vol. 156, no. 7, pp. 1358–1368, Apr. 2025, doi: 10.1002/ijc.35232.

[9] W. Kratzer, M. Klysik, A. Binzberger, and J. Schmidberger, "Gallbladder stone incidence and prevalence in Germany: a population-based study," *Z Gastroenterol*, vol. 59, no. 08, pp. 859–864, Aug. 2021, doi: 10.1055/a-1401-2170.

Manuscript received July 13, 2024; Revised September 5, 2025; Accepted September 24, 2025; date of publication September 26, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i4.1030

[10] K. Z. P. Wibowo, I. B. B. S. Adnyana, S. Indrakila, and N. Agustriani, "Interrelation between body mass index and the occurrence of gallstones," *International Journal of Surgery Science*, vol. 8, no. 2, pp. 46–51, Jan. 2024, doi: 10.33545/surgery.2024.v8.i2a.1083.

[11] J.-Y. Choi and D. R. Chang, "Imaging Diagnosis of Diseases of the Gallbladder: US, CT, and MRI," in *Diseases of the Gallbladder*, Singapore: Springer Singapore, 2020, pp. 45–60. doi: 10.1007/978-981-15-6010-1_4.

[12] L. Mencarini, A. Vestito, R. M. Zagari, and M. Montagnani, "New Developments in the Ultrasonography Diagnosis of Gallbladder Diseases," *Gastroenterol Insights*, vol. 15, no. 1, pp. 42–68, Jan. 2024, doi: 10.3390/gastroent15010004.

[13] T. Najeebi, L. Aljailani, H. Falamarzi, and S. Alghanem, "Impact of emergency department ultrasound in diagnosing patients with right upper quadrant pain in a tertiary hospital in the Kingdom of Bahrain: A cross-sectional study," *Med Sci*, vol. 27, no. 136, pp. 1–7, Jun. 2023, doi: 10.54905/disssi/v27i136/e251ms2949.

[14] R. Farina and A. Sparano, "Errors in Sonography," in *Errors in Radiology*, Milano: Springer Milan, 2012, pp. 79–85. doi: 10.1007/978-88-470-2339-0_8.

[15] R. S. Sirisati, C. S. Kumar, P. Venuthurumilli, J. Ranjith, and K. S. Rao, "Cancer Sight: Illuminating the Hidden-Advancing Breast Cancer Detection with Machine Learning-Based Image Processing Techniques," in *2023 International Conference on Sustainable Communication Networks and Application (ICSCNA)*, IEEE, Nov. 2023, pp. 1618–1625. doi: 10.1109/ICSCNA58489.2023.10370462.

[16] X. Wang, H. Zhang, Z. Bai, X. Xie, and Y. Feng, "Current status of artificial intelligence analysis for the diagnosis of gallbladder diseases using ultrasonography: a scoping review," *Transl Gastroenterol Hepatol*, vol. 10, pp. 12–12, Jan. 2025, doi: 10.21037/tgh-24-61.

[17] 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, p. 1744, May 2023, doi: 10.3390/diagnostics13101744.

[18] A. Bozdag, M. Yildirim, M. Karaduman, H. B. Mutlu, G. Karaduman, and A. Aksoy, "Detection of Gallbladder Disease Types Using a Feature Engineering-Based Developed CBIR System," *Diagnostics*, vol. 15, no. 5, p. 552, Feb. 2025, doi: 10.3390/diagnostics15050552.

[19] S. B. Shuvo and M. Z. Chowdhury, "Classification of Gallbladder Cancer Using Average Ensemble Learning," in *2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)*, IEEE, May 2024, pp. 1450–1455. doi: 10.1109/ICEEICT62016.2024.10534480.

[20] S. Dadjouy and H. Sajedi, "Gallbladder cancer detection via ultrasound image analysis: An end-to-end hierarchical feature-fused model," *IET Image Process*, vol. 19, no. 1, Jan. 2025, doi: 10.1049/ipr2.13292.

[21] L. Li, "Deep Learning-based EEG Signal Identity Recognition Using VGGNet," in *2024 4th International Conference on Neural Networks, Information and Communication (NNICE)*, IEEE, Jan. 2024, pp. 1092–1095. doi: 10.1109/NNICE61279.2024.10498553.

[22] M. M. Bala, K. J. L. Bai, S. Kattubadi, G. Pulipati, and A. Balguri, "Deep Learning Transformations in Content-Based Image Retrieval: Exploring the Visual Geometry Group (VGG16) Model," in *2024 IEEE 4th International Conference on ICT in Business Industry & Government (ICTBIG)*, IEEE, Dec. 2024, pp. 1–4. doi: 10.1109/ICTBIG64922.2024.10911292.

[23] W. N. Waluyo, R. Rizal Isnanto, and Adian Fatchur Rochim, "Comparison of Mycobacterium Tuberculosis Image Detection Accuracy Using CNN and Combination CNN-KNN," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 1, pp. 80–87, Feb. 2023, doi: 10.29207/resti.v7i1.4626.

[24] A. Biswas and Md. S. Islam, "A Hybrid Deep CNN-SVM Approach for Brain Tumor Classification," *Journal of Information Systems Engineering and Business Intelligence*, vol. 9, no. 1, pp. 1–15, Apr. 2023, doi: 10.20473/jisebi.9.1.1-15.

[25] A. Yahya Saleh, C. Ka Chin, and R. Ameera Rosdi, "Transfer Learning for Lung Nodules Classification with CNN and Random Forest," *Pertanika J Sci Technol*, vol. 32, no. 1, pp. 463–479, Nov. 2023, doi: 10.47836/pjst.32.1.25.

[26] B. Kuntiyellannagari, B. Dwarakanath, and P. V. Reddy, "Hybrid model for brain tumor detection using convolution neural networks," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 33, no. 3, p. 1775, Mar. 2024, doi: 10.11591/ijeecs.v33.i3.pp1775-1781.

[27] M. S. Tahosin, M. A. Sheakh, T. Islam, R. J. Lima, and M. Begum, "Optimizing brain tumor classification through feature selection and hyperparameter tuning in machine learning models," *Inform Med Unlocked*, vol. 43, p. 101414, 2023, doi: 10.1016/j.imu.2023.101414.

[28] F. Mohammadi and A. J. Irani, "A Review of Feature Selection Methods for Disease Risk Prediction and healthcare : Review of Feature

Manuscript received July 13, 2024; Revised September 5, 2025; Accepted September 24, 2025; date of publication September 26, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i4.1030

Selection Methods for Disease Prediction," in *2024 11th International Symposium on Telecommunications (IST)*, IEEE, Oct. 2024, pp. 713–719. doi: 10.1109/IST64061.2024.10843597.

[29] Q. Bani Baker and M. F. Alajlouni, "Comparative Analysis of Feature Selection Techniques with Metaheuristic Grasshopper Optimization Algorithm," 2024, pp. 159–169. doi: 10.1007/978-3-031-56728-5_14.

[30] F. N. Osman, M. A. Abdul Aziz, and M. N. Taib, "Comparative Evaluation of Feature Selection Algorithms for Predictive Modeling of Academic Performance Outcomes," in *2024 IEEE 13th International Conference on Engineering Education (ICEED)*, IEEE, Nov. 2024, pp. 1–6. doi: 10.1109/ICEED62316.2024.10923823.

[31] F. Rahmat *et al.*, "Supervised feature selection using principal component analysis," *Knowl Inf Syst*, vol. 66, no. 3, pp. 1955–1995, Mar. 2024, doi: 10.1007/s10115-023-01993-5.

[32] S. Abdumalikov, J. Kim, and Y. Yoon, "Performance Analysis and Improvement of Machine Learning with Various Feature Selection Methods for EEG-Based Emotion Classification," *Applied Sciences*, vol. 14, no. 22, p. 10511, Nov. 2024, doi: 10.3390/app142210511.

[33] S. Jain and A. Saha, "Rank-based univariate feature selection methods on machine learning classifiers for code smell detection," *Evol Intell*, vol. 15, no. 1, pp. 609–638, Mar. 2022, doi: 10.1007/s12065-020-00536-z.

[34] S. Julkaew, T. Wongsirichot, K. Damkliang, and P. Sangthawan, "Improving accuracy of vascular access quality classification in hemodialysis patients using deep learning with K highest score feature selection," *Journal of International Medical Research*, vol. 52, no. 4, Apr. 2024, doi: 10.1177/03000605241232519.

[35] L. W. Rizkallah, "Optimizing SVM hyperparameters for satellite imagery classification using metaheuristic and statistical techniques," *Int J Data Sci Anal*, Apr. 2025, doi: 10.1007/s41060-025-00762-7.

[36] H. Dabool, H. Alashwal, H. Alnuaimi, A. Alhouqani, S. Alkaabi, and A. Al Ahbabi, "Comparative Analysis of Hyperparameter Tuning Methods in Classification Models For Ensemble Learning," in *2024 7th International Conference on Algorithms, Computing and Artificial Intelligence (ACAI)*, IEEE, Dec. 2024, pp. 1–5. doi: 10.1109/ACAI63924.2024.10899492.

[37] D. M. Belete and M. D. Huchaiah, "Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results," *International Journal of Computers and Applications*, vol. 44, no. 9, pp. 875–886, Sep. 2022, doi: 10.1080/1206212X.2021.1974663.

[38] A. Taufiq, S. Yulianti, A. Rahmatulloh, I. Darmawan, and R. Rizal, "Comparison of Hyperparameter Tuning Techniques on KNN Algorithm to find the Best K Value using Grid Search and Random Search Methods," in *2024 7th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, IEEE, Dec. 2024, pp. 180–186. doi: 10.1109/ISRITI64779.2024.10963519.

[39] Sukamto, Hadiyanto, and Kurnianingsih, "KNN Optimization Using Grid Search Algorithm for Preeclampsia Imbalance Class," *E3S Web of Conferences*, vol. 448, p. 02057, Nov. 2023, doi: 10.1051/e3sconf/202344802057.

[40] 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, p. 110426, Jun. 2024, doi: 10.1016/j.dib.2024.110426.

[41] S. M. Pizer *et al.*, "Adaptive histogram equalization and its variations," *Comput Vis Graph Image Process*, vol. 39, no. 3, pp. 355–368, Sep. 1987, doi: 10.1016/S0734-189X(87)80186-X.

[42] K. Zuiderveld, "Contrast Limited Adaptive Histogram Equalization," in *Graphics Gems*, Elsevier, 1994, pp. 474–485. doi: 10.1016/B978-0-12-336156-1.50061-6.

[43] A. M. Reza, "Realization of the Contrast Limited Adaptive Histogram Equalization (CLAHE) for Real-Time Image Enhancement," *The Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology*, vol. 38, no. 1, pp. 35–44, Aug. 2004, doi: 10.1023/B:VLSI.0000028532.53893.82.

[44] I. K. Seneng, P. D. W. Ayu, and R. R. Huizen, "Comparative Analysis of Augmentation and Filtering Methods in VGG19 and DenseNet121 for Breast Cancer Classification," *Jurnal Teknik Informatika (Jutif)*, vol. 6, no. 3, pp. 1131–1146, Jun. 2025, doi: 10.52436/1.jutif.2025.6.3.4397.

[45] N. Nafi'iyah and J. Maknun, "CNN Architecture for Classifying Types of Mango Based on Leaf Images," *Telematika*, vol. 14, no. 2, pp. 112–121, Aug. 2021, doi: 10.35671/telematika.v14i2.1262.

[46] P. Dehbozorgi, O. Ryabchykov, and T. W. Bocklitz, "A comparative study of statistical, radiomics, and deep learning feature extraction techniques for medical image classification in optical and radiological modalities," *Comput Biol Med*, vol. 187, p. 109768, Mar. 2025, doi: 10.1016/j.compbiomed.2025.109768.

[47] S. K. Dash *et al.*, "Ocular Disease Detection Using Fundus Images: A Hybrid Approach of Grad-CAM and Multiscale Retinex Preprocessing

Manuscript received July 13, 2024; Revised September 5, 2025; Accepted September 24, 2025; date of publication September 26, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i4.1030

With VGG16 Deep Features and Fine KNN Classification," *Applied Computational Intelligence and Soft Computing*, vol. 2025, no. 1, Jan. 2025, doi: 10.1155/acis/6653543.

[48] T.-H. Nguyen, T.-N. Nguyen, and B.-V. Ngo, "A VGG-19 Model with Transfer Learning and Image Segmentation for Classification of Tomato Leaf Disease," *AgriEngineering*, vol. 4, no. 4, pp. 871–887, Oct. 2022, doi: 10.3390/agriengineering4040056.

[49] X. Cheng, "A Comprehensive Study of Feature Selection Techniques in Machine Learning Models," *Insights in Computer, Signals and Systems*, vol. 1, no. 1, pp. 65–78, Nov. 2024, doi: 10.70088/xpf2b276.

[50] M. Islam and R. Islam, "Exploring the Impact of Univariate Feature Selection Method on Machine Learning Algorithms for Heart Disease Prediction," in *2023 International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM)*, IEEE, Jun. 2023, pp. 1–5. doi: 10.1109/NCIM59001.2023.10212832.

[51] B. B. Gupta, A. Gaurav, R. W. Attar, V. Arya, A. Alhomoud, and K. T. Chui, "Sustainable IoT Security in Entrepreneurship: Leveraging Univariate Feature Selection and Deep CNN Model for Innovation and Knowledge," *Sustainability*, vol. 16, no. 14, p. 6219, Jul. 2024, doi: 10.3390/su16146219.

[52] B. Mohammad Hasani Zade, N. Mansouri, and M. M. Javidi, "An improved beluga whale optimization using ring topology for solving multi-objective task scheduling in cloud," *Comput Ind Eng*, vol. 200, p. 110836, Feb. 2025, doi: 10.1016/j.cie.2024.110836.

[53] S. Adige, R. Kurban, A. Durmuş, and E. Karaköse, "Classification of apple images using support vector machines and deep residual networks," *Neural Comput Appl*, vol. 35, no. 16, pp. 12073–12087, Jun. 2023, doi: 10.1007/s00521-023-08340-3.

[54] X. Wu, T. Oli, J. H. Qian, V. Taylor, M. C. Hersam, and V. K. Sangwan, "An Autotuning-based Optimization Framework for Mixed-kernel SVM Classifications in Smart Pixel Datasets and Heterojunction Transistors," Sep. 2024.

[55] L. Zhu and P. Spachos, "Support vector machine and YOLO for a mobile food grading system," *Internet of Things*, vol. 13, p. 100359, Mar. 2021, doi: 10.1016/j.iot.2021.100359.

[56] M. Hussain, S. K. Wajid, A. Elzaart, and M. Berbar, "A Comparison of SVM Kernel Functions for Breast Cancer Detection," in *2011 Eighth International Conference Computer Graphics, Imaging and Visualization*, IEEE, Aug. 2011, pp. 145–150. doi: 10.1109/CGIV.2011.31.

[57] S. M. Al-azzawi, M. A. Deif, H. Attar, A. Amer, and A. A. A. Solyman, "Hyperparameter Optimization of Regression Model for Electrical Load Forecasting During the COVID-19 Pandemic Lockdown Period," *International Journal of Intelligent Engineering and Systems*, vol. 16, no. 4, pp. 239–253, Aug. 2023, doi: 10.22266/ijies2023.0831.20.

[58] L. Breiman, "Random Forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[59] A. Parmar, R. Katariya, and V. Patel, "A Review on Random Forest: An Ensemble Classifier," 2019, pp. 758–763. doi: 10.1007/978-3-030-03146-6_86.

[60] N. Z. Al Habesyah, R. Herteno, F. Indriani, I. Budiman, and D. Kartini, "Sentiment Analysis of TikTok Shop Closure in Indonesia on Twitter Using Supervised Machine Learning," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 6, no. 2, pp. 148–156, Apr. 2024, doi: 10.35882/jeeemi.v6i2.381.

[61] W. Li, Y. Chen, and Y. Song, "Boosted K-nearest neighbor classifiers based on fuzzy granules," *Knowl Based Syst*, vol. 195, p. 105606, May 2020, doi: 10.1016/j.knosys.2020.105606.

[62] V. Kalra, I. Kashyap, and H. Kaur, "Effect of Distance Measures on K-Nearest Neighbour Classifier," in *2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA)*, IEEE, Sep. 2022, pp. 1–7. doi: 10.1109/ICCSEA54677.2022.9936314.

[63] M. Mailagaha Kumbure and P. Luukka, "A generalized fuzzy k-nearest neighbor regression model based on Minkowski distance," *Granular Computing*, vol. 7, no. 3, pp. 657–671, Jul. 2022, doi: 10.1007/s41066-021-00288-w.

[64] T. Ciu and R. S. Oetama, "Logistic Regression Prediction Model for Cardiovascular Disease," *IJNMT (International Journal of New Media Technology)*, vol. 7, no. 1, pp. 33–38, Jul. 2020, doi: 10.31937/ijnmt.v7i1.1340.

[65] R. T. Yunardi, R. Apsari, and M. Yasin, "Comparison of Machine Learning Algorithm For Urine Glucose Level Classification Using Side-Polished Fiber Sensor," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 2, no. 2, pp. 33–39, Jul. 2020, doi: 10.35882/jeeemi.v2i2.1.

[66] A. Balboa, A. Cuesta, J. González-Villa, G. Ortiz, and D. Alvear, "Logistic regression vs machine learning to predict evacuation decisions in fire alarm situations," *Saf Sci*, vol. 174, p. 106485, Jun. 2024, doi: 10.1016/j.ssci.2024.106485.

[67] Putri Nabella, Rudy Herteno, Setyo Wahyu Saputro, Mohammad Reza Faisal, and Friska Abadi, "Impact of a Synthetic Data Vault for

Manuscript received July 13, 2024; Revised September 5, 2025; Accepted September 24, 2025; date of publication September 26, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i4.1030

Imbalanced Class in Cross-Project Defect Prediction," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 6, no. 2, pp. 219–230, Apr. 2024, doi: 10.35882/jeeemi.v6i2.409.

[68] B. H. Shekar and G. Dagnew, "Grid Search-Based Hyperparameter Tuning and Classification of Microarray Cancer Data," in *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*, IEEE, Feb. 2019, pp. 1–8. doi: 10.1109/ICACCP.2019.8882943.

[69] T. N. Nuklianggraita, A. Adiwijaya, and A. Aditsania, "On the Feature Selection of Microarray Data for Cancer Detection based on Random Forest Classifier," *JURNAL INFOTEL*, vol. 12, no. 3, pp. 89–96, Aug. 2020, doi: 10.20895/infotel.v12i3.485.

[70] C. Tantithamthavorn, S. McIntosh, A. E. Hassan, and K. Matsumoto, "The Impact of Automated Parameter Optimization on Defect Prediction Models," *IEEE Transactions on Software Engineering*, vol. 45, no. 7, pp. 683–711, Jul. 2019, doi: 10.1109/TSE.2018.2794977.

[71] M. Jena and S. Dehuri, "An Integrated Novel Framework for Coping Missing Values Imputation and Classification," *IEEE Access*, vol. 10, pp. 69373–69387, 2022, doi: 10.1109/ACCESS.2022.3187412.

[72] Siti Napi'ah, Triando Hamonangan Saragih, Dodon Turianto Nugrahadi, Dwi Kartini, and Friska Abadi, "Implementation of Monarch Butterfly Optimization for Feature Selection in Coronary Artery Disease Classification Using Gradient Boosting Decision Tree," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 5, no. 4, Oct. 2023, doi: 10.35882/jeeemi.v5i4.331.

[73] D. Chicco, V. Starovoitov, and G. Jurman, "The Benefits of the Matthews Correlation Coefficient (MCC) Over the Diagnostic Odds Ratio (DOR) in Binary Classification Assessment," *IEEE Access*, vol. 9, pp. 47112–47124, 2021, doi: 10.1109/ACCESS.2021.3068614.

[74] Luh Ayu Martini, G. A. Pradipta, and R. R. Huizen, "Analysis of the Impact of Data Oversampling on the Support Vector Machine Method for Stroke Disease Classification," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 7, no. 2, pp. 404–421, Mar. 2025, doi: 10.35882/jeeemi.v7i2.698.

[75] W. Islam *et al.*, "A Neoteric Feature Extraction Technique to Predict the Survival of Gastric Cancer Patients," *Diagnostics*, vol. 14, no. 9, p. 954, May 2024, doi: 10.3390/diagnostics14090954.

[76] Shalehah, Muhammad Itqan Mazdadi, Andi Farmadi, Dwi Kartini, and Muliadi, "Implementation of Particle Swarm Optimization Feature Selection on Naïve Bayes for Thoracic Surgery Classification," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 5, no. 3, pp. 150–158, Jul. 2023, doi: 10.35882/jeemi.v5i3.305.

[77] Jin Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Trans Knowl Data Eng*, vol. 17, no. 3, pp. 299–310, Mar. 2005, doi: 10.1109/TKDE.2005.50.

Author Biography

Ryan Adhitama Putra began his undergraduate studies in Information Systems at the Institut Teknologi dan Bisnis STIKOM Bali in 2019 and graduated in the 2022/2023 odd semester. During this time, he developed expertise in enterprise systems. Currently, he is pursuing a master's degree (S2) in Information Systems at the same institution, starting in 2023. His research focuses on analyzing features and selection from medical images, particularly in classification of gallbladder disease. By leveraging machine learning algorithms, he aims to enhance the accuracy and reliability of predictive models in healthcare analytics. His final project seeks to explore classification strategies to improve the performance of machine learning algorithms, addressing key challenges in medical datasets. He can be contacted at email 232011014@stikom-bali.ac.id

Gede Angga Pradipta holds a Doctor of Computer Science from the Department of Computer Science and Electronics, Faculty of Natural Sciences, Universitas Gadjah Mada (UGM), Yogyakarta, Indonesia, in 2021. He also received a bachelor's degree in computer informatics from Universitas Atma Jaya (UAJY), Yogyakarta, Indonesia, in 2012 and a master's degree in information technology from Universitas Gadjah Mada (UGM), Yogyakarta, Indonesia, in 2014. His research interests include Machine Learning, Pattern Recognition, and Image Processing. He is currently lecturing at the Department of Magister Information Systems, Institut Teknologi dan Bisnis STIKOM Bali, Indonesia. He can be contacted at email angga_pradipta@stikom-bali.ac.id

Manuscript received July 13, 2024; Revised September 5, 2025; Accepted September 24, 2025; date of publication September 26, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i4.1030

Putu Desiana Wulaning Ayu received the Dr. (Doctor) in Computer Science from the Department of Computer Science and Electronics, Faculty of Mathematics and Natural Science, Universitas Gadjah Mada, with the dissertation "Segmentation and feature extraction model on 2-D ultrasonograph images for amniotic fluid classification". Her research interests are medical image processing, machine learning, deep learning, and computer vision. She is lecturing in Magister Information Systems, Institut Teknologi dan Bisnis STIKOM Bali, Indonesia. She is a member of the Indonesian Computer, Electronics, and Instrumentation Support Society. She can be contacted at email: wulaning_ayu@stikom-bali.ac.id

Manuscript received July 13, 2024; Revised September 5, 2025; Accepted September 24, 2025; date of publication September 26, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i4.1030