Received: 26 April 2026 | Revised: 17 August 2026 | Accepted: 11 November 2026 | Published online: 26 November 2026

RESEARCH ARTICLE

Medinformatics
2026, Vol. 00(00) 1–15
DOI: 10.47852/bonviewMEDIN620210151

BON VIEW PUBLISHING

Metabolic Dysfunction-Associated Steatotic Liver Disease Diagnosis Using EfficientNet Transfer Learning and Ultrasound Images

A'adel Sayyahi[1], Seyed Enayatallah Alavi[1,*] and Morteza Jaderyan[1]

[1] *Department of Computer Engineering, Shahid Chamran University of Ahvaz, Iran*

Abstract: Metabolic dysfunction-associated fatty liver disease (MASLD) is a prevalent, progressive hepatic disorder affecting approximately 30% of the population in developed countries, underscoring the need for rapid, accurate, noninvasive diagnostic tools. Ultrasound imaging remains a primary diagnostic modality; however, patient privacy is addressed directly in this work. This study presents a privacy-aware deep learning framework for detecting MASLD from ultrasound images. Given an initial dataset of only 55 images, extensive enhancement (rotation, brightness and contrast adjustment, and CLAHE) was applied to expand the training dataset to 1550 samples. To preserve patient privacy, a Mask R-CNN model was used to detect and remove foreign and potentially identifying objects from the images before processing. Subsequently, the image quality was enhanced through a homomorphic filter based on the gray wolf optimizer, which suppressed speckle noise and achieved a peak signal-to-noise ratio of 72.9 dB. The pre-trained EfficientNet network (B0-B7 and V2-S, V2-M, V2-L) was evaluated as a transfer learning backbone for 2-class classification tasks: binary discrimination between normal and fatty liver, and 4-class grading (Normal; grades I-III). EfficientNetB6 achieved 100% classification accuracy with an area under the curve (AUC) of 1.0 in the binary task, while EfficientNetB5, B6, and B7 each achieved 99.87% accuracy in the 4-class task, with EfficientNetB5 having the highest AUC of 0.9994. By simultaneously achieving diagnostic accuracy and privacy preservation, the proposed framework establishes a robust and clinically applicable approach for early diagnosis of MASLD.

Keywords: metabolic dysfunction-associated steatotic liver disease, ultrasound images, EfficientNet, Mask R-CNN, Grey Wolf optimizer

1. Introduction

The liver acts as the major metabolic organ in the human body, regulating an extensive range of anabolic and catabolic pathways that are highly regulated in order to maintain body homeostasis. In particular, some of the vital roles performed by the liver include protein synthesis and breakdown; production of bile acid from cholesterol for the digestion and absorption of lipids; carbohydrate and fat metabolism; and detoxification of drugs and xenobiotics [1]. Hepatic steatosis occurs when liver cells accumulate excessive amounts of triglycerides, which creates the medical condition known as fatty liver disease. Physicians used to diagnose this condition as nonalcoholic fatty liver disease (NAFLD), which has now changed to metabolic dysfunction-associated steatotic liver disease (MASLD) because its metabolic disturbances show strong connections to disease development [2]. MASLD includes multiple liver diseases, which range from basic fat accumulation to metabolic dysfunction-associated steatohepatitis (MASH) and advanced stages of fibrosis and cirrhosis that result in liver cancer (HCC) development [3]. The condition of MASLD has become one of the most common chronic liver illnesses that exists throughout the globe. Recent studies about disease patterns show that 30 to 38 percent of adults worldwide develop MASLD, whose prevalence increases together with the worldwide obesity and diabetes epidemics [4–6]. Patients with MASLD usually experience a slow progression of their condition, which shows no symptoms. Most patients show initial signs of simple steatosis, but a large number of them eventually develop MASH together with advanced fibrosis and cirrhosis [7].

Hepatocellular carcinoma related to MASLD can emerge without the presence of cirrhosis, which emphasizes the necessity of detecting the disease in its early stages and maintaining ongoing patient observation. Liver biopsy has long been considered the most accurate method for diagnosing MASLD and determining fibrosis. The procedure requires surgical intervention through a surgical cut and carries risks such as patient discomfort, bleeding, inaccurate results, and variability among doctors [8]. Clinicians have leaned toward noninvasive screening in recent years. They say these options are safer for patients and tend to give steadier findings than liver biopsy. Biopsy is still used as the main reference for staging cirrhosis. Still, it is an invasive step. It can also produce results that vary from case to case, so patients need repeat visits.

Researchers and clinicians have worked on several noninvasive approaches. Many use blood markers, or elastic-based imaging, or other medical imaging methods [9]. For hepatic steatosis, the key imaging patterns have been checked against

***Corresponding author:** Seyed Enayatallah Alavi, Department of Computer Engineering, Shahid Chamran University of Ahvaz, Iran. Email: Se.Alavi@scu.ac.ir

© The Author(s) 2026. Published by BON VIEW PUBLISHING PTE. LTD. This is an open access article under the CC BY License (https://creativecommons.org/licenses/by/4.0/).

standard tissue findings. When these methods are used, they report accuracy above 80 percent for moderate-to-severe hepatic steatosis [10]. Together, these points support ultrasonography as the top noninvasive choice for first identification and follow-up of MASLD in both clinical and research work. Other methods are also used. These include MRI and controlled attenuation parameter with elastography. MRI proton density fat fraction, or MRI-PDFF, is used as a noninvasive benchmark for liver fat. MR elastography can give fat measures and fibrosis assessment. However, it is costly and not widely available, so it is not used as often. CAP (via FibroScan) provides a quantitative measure of steatosis, but typically requires additional hardware and is more effective for higher-grade fat. Computed tomography (CT) can detect moderate-to-severe steatosis as reduced attenuation, but routine CT use is limited by cost and radiation exposure. The combination of MRI and CAP provides objective measurement, but ultrasound serves as the practical primary method for MASLD screening because it provides affordable and easily available testing [11–13].

Deep learning (DL) is one of the most important methods in computer-aided diagnosis (CAD), which shows advanced performance in image classification, segmentation, and disease detection in several medical image modalities [14, 15]. The use of artificial intelligence in the analysis of ultrasound images has significantly increased diagnostic accuracy, objectivity, and reproducibility, and decreased dependence on the operator [16–18]. Using artificial intelligence techniques, it is possible to create algorithms that help non-expert users create high-quality medical images and improve differential diagnosis for a variety of pathologies [19]. The main motivation for this research is precisely this potential of AI methods. In recent years, deep learning methods, and in particular convolutional neural networks (CNNs), have shown promising results in automated medical image analysis, including ultrasound-based liver disease assessment. Several studies have investigated the application of CNN architectures to tasks such as hepatic steatosis grading, fatty liver diagnosis, and benign versus malignant liver lesion classification. These studies vary considerably in terms of dataset size, imaging device type, preprocessing strategies, and evaluation criteria; they range from learning texture features using standard CNNs [20] to transfer learning with pre-developed models such as VGG16 and Inception v3 [21–25], as well as automated strategies for liver lesion classification [24].

Beyond the reported performance measures, these studies share several common limitations in reproducibility that motivated the present study. First, most were based on small, single-center datasets (ranging from approximately 55 to 205 patients). Second, from the perspective of direct clinical application, the evaluation in these studies is largely limited to internal validation, without considering an independent or external test set; this makes it difficult to assess the generalizability of the reported accuracy to new patients, devices, or clinical settings. Finally, few of these works explicitly take steps to protect privacy prior to model training, while the sensitive nature of patient ultrasound data requires such consideration. The present study attempts to fill precisely these gaps: in this study, a systematic range of EfficientNet versions (from B0 to B7 as well as V2-S/M/L) was evaluated on two-class and four-class classification tasks, a Mask R-CNN-based step was included to anonymize the data before processing, and, as will be explained in the relevant section, patient-level validation and external validation are proposed as directions for future research, rather than as a final estimate of generalizability based on the results of internal cross-validation. Recent systematic reviews have similarly shown that the main obstacles to the clinical transfer of AI-based diagnostic systems for MASLD remain limitations in external validity, heterogeneity of datasets, and insufficient multicenter evaluation [15].

In this paper, we present an intelligent system for the diagnosis of nonalcoholic fatty liver disease based on transfer learning with EfficientNet architecture and data augmentation techniques on ultrasound images. Our proposed approach aims to provide a noninvasive, accurate, and efficient diagnostic tool that can support the early diagnosis and clinical management of MASLD.

2. Materials and Methods

This section describes the methodology and materials used in the development of an intelligent system for the diagnosis of MASLD using ultrasound images. The proposed approach consists of two main steps: preprocessing and processing.

2.1. Dataset collection

The dataset used consists of liver ultrasound images of 55 morbidly obese patients who underwent laparoscopic surgery at the Medical University of Warsaw, Poland. The mean age of the group was 40 years, 20% of the patients were male, and their mean body mass index (BMI) was reported to be 45.26. All images were acquired in B-Mode with a resolution of 636 × 434 pixels using a GE Vivid E9 ultrasound machine (GE Healthcare, Norway) equipped with a 2.5 MHz sector probe; harmonic imaging was also enabled under the default settings for abdominal imaging, and the images were saved in DICOM format for offline analysis. To account for normal physiological changes and possible patient movements during imaging, ten images were acquired from each patient at fixed time intervals [26].

Ground-truth labeling was performed histologically: a liver biopsy was taken from each patient during surgery, and a pathologist measured the degree of fat infiltration in hepatocytes to determine the severity of the disease. Accordingly, two labeling schemes were defined. First, for the 2-class classification, patients were divided into two groups: fatty liver (fat infiltration >5%) and healthy liver (fat infiltration <5%) according to the Clinical Research Network criteria [27], with 38 patients in the fatty liver group and 17 in the healthy group (class ratio 38:17). In the second design, images were classified into four severity grades based on the percentage of hepatic fat infiltration and were used in ultrasound-based steatosis grading studies [28]: grade 0 or Normal (less than 5%), grade I or mild (5% to 33%), grade II or moderate (33% to 66%), and grade III or severe (more than 66%).

2.2. Preprocessing

Preprocessing is the first step in preparing ultrasound images for MASLD diagnosis, which aims to improve image quality and remove unwanted artifacts, in order to increase the accuracy of subsequent analyses. This step includes two main techniques: object removal using Mask R-CNN and speckle noise reduction using homomorphic filtering. These techniques have been selected to preserve only the features related to liver tissue for feature extraction and classification steps, while also overcoming common challenges in ultrasound imaging such as artifacts and noise.

2.2.1. Mask R-CNN-based anonymization of medical images for privacy protection

Ultrasound images often contain text annotations, including patient name, age, date, and other sensitive information, which can be either printed or digitally embedded in the image. The presence of these annotations can interfere with subsequent image processing; therefore, their removal is an important step in the preprocessing process to focus the analysis on areas relevant to medical diagnosis. To address this problem, Mask R-CNN, a deep learning-based model for instance-based segmentation presented by He et al. [29], was used. This model, which is effectively an extension of the Faster R-CNN framework, improves object recognition accuracy by adding pixel-wise segmentation capabilities and allows for accurate identification and removal of sensitive text information without damaging areas of diagnostic value in the image.

The architecture of this model consists of four main parts. First, a pre-trained feature extraction network based on ResNet-50 is used, which extracts features at different scales, thereby facilitating the recognition of objects of different sizes. Second, a region proposal network (RPN) suggests region candidates using anchor blocks at different scales and aspect ratios; Each of these regions is associated with a probability score and bounding box coordinates. Third, an ROI alignment module refines these proposed regions while preserving accurate spatial coordinates to provide a basis for class prediction, bounding box regression, and mask generation; this approach overcomes the limitations and spatial distortions of the traditional ROI fusion method. Fourth, a series of binary mask predictions separates the pixels belonging to the object from the background and create accurate masks for each region.

2.2.2. Speckle noise reduction

Speckle noise is a common type of image noise that is commonly observed in images obtained from ultrasonic imaging systems, such as ultrasound and radar. This type of noise appears as random point patterns with uniform dispersion throughout the image, and its origin is the constructive and destructive interference of reflected waves from discontinuous surfaces or fine textures in the imaging environment. In this study, to remove speckle noise, we used a homomorphic filter in the frequency domain, an approach that is consistent with deep learning-based noise reduction strategies presented in recent ultrasound imaging papers [30]. This filter is designed by combining two high-pass and low-pass paths, and its $\boldsymbol{d_0}$ parameter is optimized using the Gray Wolf algorithm [31]. This method removes low-frequency noise and preserves high-frequency details by combining Butterworth high-pass and Gaussian low-pass filters. To fine-tune the cutoff frequency ($\boldsymbol{d_0}$), the GWO is used to both suppress noise well and preserve edges. As a result, the image quality is improved, and the feature extraction and classification are more accurate. The GWO parameters such as population, number of iterations, $\boldsymbol{d_0}$ search range, and exploration and exploitation coefficients are listed in Table 1.

These parameters were carefully selected to ensure efficient convergence toward the optimal cutoff frequency. The following metrics were used to evaluate the performance of the optimization process:

Peak Signal-to-Noise Ratio (PSNR): This metric quantifies image quality in decibels (dB), calculated as:

$$PSNR = 20 \cdot log10\left(\frac{255}{\sqrt{MSE}}\right) \tag{1}$$

Higher PSNR values indicate superior image fidelity and effective noise reduction.

Mean Squared Error (MSE): MSE measures the average squared difference between the original and denoised images, defined as:

$$MSE = \frac{1}{N}\sum_{i,j}(x(i,j) - \hat{x}(i,j))^2 \tag{2}$$

where x(i, j) is the original image, x̂(i, j) is the denoised image, and N is the total number of pixels. Lower MSE values signify better denoising performance.

Root Mean Squared Error (RMSE): RMSE provides the error magnitude, computed as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i,j}(x(i,j) - \hat{x}(i,j))^2} \tag{3}$$

Lower RMSE values reflect closer alignment between the original and denoised images.

We tested 55 images after applying the GWO-based homomorphic filter. The PSNR came out to 72.9 dB, and the MSE was 0.0037 dB. After filtering, the speckle noise in high frequencies drops a lot. At the same time, the parenchymal details stay in place. This is shown in Figure 1. The cutoff value $\boldsymbol{d_0}$ is not fixed. It runs from 21.9 to 59.9. This range sets GWO apart from histogram equalization methods. Looking at PSNR across the set, the scores form a shape close to a normal curve. The average is around 73 dB. This suggests the method stays steady across all images. There is also a clear trend with $\boldsymbol{d_0}$. When d0 gets larger, PSNR tends to rise too. For smaller $\boldsymbol{d_0}$ values, roughly 25 to 35, PSNR is usually near 71 to 72 dB. When $\boldsymbol{d_0}$ goes above 50, PSNR often goes past 73 dB. Overall, the results point to a positive link between $\boldsymbol{d_0}$ and PSNR. The method supports better image reconstruction through adaptive choice of parameters.

2.3. Processing

Processing is the most important step and the core of our proposed MASLD detection method. This step first performs

Table 1 GWO optimization parameters

Parameter	Value/Range	Description
Population Size	20	Number of wolves (candidate solutions)
Maximum Iterations	50	Number of optimization cycles
$\boldsymbol{d_0}$ Bounds	[5, 100]	Search range for cutoff frequency
Control Parameter ($\vec{\boldsymbol{a}}$)	0 to 2 (linear)	Control parameter for exploration/exploitation
Random Vectors ($\vec{\boldsymbol{r}_1}$, $\vec{\boldsymbol{r}_2}$)	(0, 1)	Random vectors for variability

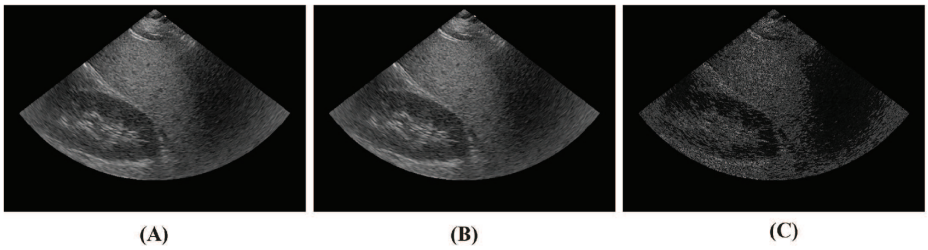

Figure 1 Visual comparison of (A) original ultrasound image, (B) denoised output, and (C) difference map highlighting noise removal.

data augmentation to increase the diversity of the dataset, followed by feature extraction, in which the pre-trained EfficientNet model acts as a support network and efficiently extracts hierarchical features. Finally, a single-task learning (STL) method is used for classification to optimize the model for the task at hand. Together, these steps improve the generalization ability of the model, reduce overfitting, and increase the classification accuracy.

2.3.1. Data augmentation

Data augmentation significantly increases the size of the training set, reducing the sensitivity of the model to changes in the ultrasound images. These changes tend to boost classification results. If we do not use augmentation, the model may behave like it is guessing when it sees new inputs from the initial dataset. After augmentation, the dataset size becomes 1550 images. Table 2 shows the exact settings for each augmentation method.

2.3.2. Transfer learning with EfficientNet

Transfer learning has been an effective approach for medical image classification, especially when dealing with limited datasets (such as our method, where the original dataset was augmented to 1550 data points). By leveraging pre-trained deep learning models, transfer learning allows for feature extraction from large-scale datasets that can then be fine-tuned for specific tasks. In this study, EfficientNet was used as the main network for feature extraction and MASLD classification of ultrasound images. Both versions of EfficientNet, EfficientNetV1 (B0-B7) and EfficientNetV2 (S, M, L), were used to evaluate their performance on binary and multiclass classification tasks.

EfficientNetV1, introduced by Tan & Le [32], is a family of CNN architectures that uses a hybrid scaling strategy to efficiently scale the depth, width, and resolution of the input while maintaining optimal performance. The base model, EfficientNet-B0, was designed using Neural Architecture Search (NAS) and then systematically scaled to produce larger variants (B0–B7), each of which gradually increases in size and complexity. This scaling approach enables EfficientNet models to achieve higher accuracy with fewer parameters compared to traditional architectures. In this study, different variants of EfficientNetV1 were tested to analyze their performance in binary classification (normal liver vs. fatty liver) and 4-class classification (normal, grade I, grade II, and grade III).

EfficientNetV2 is a newer design in the EfficientNet line. It aims to run faster and use less compute, but still keep good accuracy [33]. Compared with EfficientNetV1, it tends to score better. That comes from changes in both the model layout and the way training is done. One key change is the Fused-MBConv blocks. These are used to cut latency and speed up execution. Another change is how training images are handled. The input size is increased step by step, rather than staying fixed. EfficientNetV2 also works well when the model is small. In addition, it uses updated scaling choices to trade off accuracy and performance in a more controlled way. The family is split into several sizes. EfficientNetV2-S is the lightweight option. EfficientNetV2-M is in the middle for cost and capability. EfficientNetV2-L is the larger option, built for higher accuracy and more complex representations.

2.3.3. Single-Task Learning

In this work, a Single-Task Learning (STL) approach was adopted, where each classification task was trained separately. Different from multi-task learning, which involves sharing features among multiple tasks, STL is concerned with improving performance for just one objective and thus achieves greater task-specific accuracy together with a simpler model. The models were fine-tuned using ultrasound liver images, and the final

Table 2 Data augmentation techniques

Technique	Description
Brightness Adjustment	Adjusted with coefficients (0.85, 0.95, 1.05, 1.15)
Contrast Adjustment	Modified with coefficients (0.85, 0.95, 1.05, 1.15)
Gamma Correction	Varied with values (0.85, 0.95, 1.05, 1.15)
Rotation	Applied at ±5°, ±10°
Translation (Shift)	Shifted by (±5, ±5), (±10, ±10) pixels
Zoom In	Scaled by 1.1
Cropping & Resizing	Cropped (400 × 300) and resized to (434 × 636)
Flipping	Horizontal mirroring
CLAHE	Enhances contrast in low-light regions
Combinations	e.g., Rotation + Contrast + Brightness

classification layer was adjusted according to the number of target classes (Binary or Grading).

Feature extraction was done with a Global Average Pooling step. After that, a fully connected layer was used. The size of this dense layer depends on which EfficientNet version was picked. For instance, EfficientNet-B0 gives 1280 values from the GAP layer. Higher variants, such as EfficientNet-B7, provide a larger set of features. EfficientNetV2 also changes the feature size based on its own model setup. This helps the system make better choices in medical image analysis while keeping the computation cost under control.

2.4. Evaluation metrics

In this study, we ran several classification checks to see how well our models did. The results tell us how often the model gets the right label, and how it performs on each class. The evaluation metrics used are as follows:

Accuracy: Accuracy measures the proportion of correctly classified instances out of all samples.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Where:

TP: True Positive (correctly classified positive samples)
TN: True Negative (correctly classified negative samples)
FP: False Positive (incorrectly classified negative samples as positive)
FN: False Negative (incorrectly classified positive samples as negative)

Precision: Precision quantifies the proportion of correctly predicted positive samples among all positive predictions.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

Recall (Sensitivity): Recall, also known as true positive rate (TPR), measures the proportion of actual positive cases that were correctly identified.

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

F1-Score: The F1-score is the harmonic mean of precision and recall, offering a balanced measure when the dataset is imbalanced.

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{7}$$

Area Under the ROC Curve (AUC-ROC): The AUC-ROC score evaluates the model's ability to distinguish between classes across different threshold values. It is computed as:

$$AUC = \int_{-\infty}^{+\infty} TP(FP)\, d\, FP \tag{8}$$

Class-Indifference Score (CIS): CIS measures the model's consistency in classifying all classes equally.

$$CIS = 1 - \frac{\sigma_c}{\mu_c} \tag{9}$$

Where

σ_c: is the standard deviation of class-wise accuracies
μ_c: is the mean class-wise accuracy.
A CIS value close to 1 indicates uniform classification performance across all classes.

Error Balance Index (EBI): EBI evaluates whether classification errors are evenly distributed across all classes.

$$EBI = \frac{min\,(Error\ Rate_c)}{Max\,(Error\ Rate_c)} \tag{10}$$

Where:

$Error\ Rate_c$: is the classification error rate for class c accuracies

A value close to 1 suggests balanced errors. Note that, as computed here, EBI is a per-class error-rate ratio rather than a strict min/max normalization; consequently, its value is not constrained to the [0, 1] interval and can exceed 1, with the degree of deviation from 1 in either direction reflecting the degree of imbalance between classes.

Prediction Confidence Index (PCI): PCI represents the model's confidence in its predictions, calculated as the average confidence score of correctly classified samples.

$$PCI = \frac{\sum_{i=1}^{N} P_i \times I_i}{\sum_{i=1}^{N} I_i} \tag{11}$$

Where:

P_i : is the confidence score for sample i.
I_i : is 1 if the sample is classified correctly, otherwise 0.
Higher PCI values indicate greater prediction confidence.

Class Discordance Index (CDI): CDI measures the disagreement between predicted and actual classes based on class distances.

$$CDI = \frac{1}{N} \sum_{i=1}^{N} d(Y_i, \hat{Y}_i) \tag{12}$$

Where:

$d(Y_i, \hat{Y}_i)$: is the class distance function. Lower CDI values indicate better classification.

Cost-Balance Index (CBI): CBI assesses the model's ability to minimize classification costs relative to errors.

$$CBI = \frac{1}{C} \sum_{c=1}^{C} \frac{1}{1 + Cost_c} \tag{13}$$

Where:

$Cost_c$: represents the classification cost of class c accuracies.
Higher CBI values imply better cost-effectiveness.

Adaptability Index (AI): AI evaluates how well the model generalizes to new data by comparing test and validation performance.

$$AI = 1 - |Accuracy_{test} - Accuracy_{val}| \tag{14}$$

A higher AI indicates better generalization.

3. Results

The proposed model was implemented in Python using the PyTorch library. All experiments were performed on a system with Ubuntu 24.04 (LTS) operating system and 64-bit x86_64 architecture. The hardware configuration used included a 12-core processor, 16 GB of RAM, and an NVIDIA GeForce RTX 4070 Ti SUPER graphics card with 16 GB of dedicated memory, which enabled parallel processing and accelerated deep network calculations.

The definitions and mathematical relationships of the evaluation criteria used in this study are presented in detail in Section 2.4. The performance of the proposed models was evaluated in two classification scenarios, a binary task (2 classes: Normal and Fatty) and a multiclass task (4 classes: Normal and obesity grades I-III). To ensure robust evaluation and avoid overfitting, a 10-fold cross-validation method was used; the dataset was divided into ten subsets, and in each iteration, one of the subsets was considered as test data and the other nine subsets as training data, and this process was repeated ten times. Finally, by averaging the results of these ten runs, a reliable estimate of the model performance was obtained. The models were evaluated using a comprehensive set of criteria including standard classification indices (accuracy, precision, sensitivity, F1 score and area under the ROC curve) as well as supplementary indices (CIS, EBI, PCI, CDI, CBI and AI) in both the mentioned scenarios, the results of which are summarized in Table 3.

3.1. Overall performance results

The quantitative performance of the models, averaged over 10 cross-validation layers, is presented in Table 3. In the 2-class scenario, EfficientNetB6 emerged as the best model, achieving perfect classification with accuracy, precision, recall, and F1 score of 1, AUC of 1, and CDI of 0. This indicates perfect separation between the normal and fatty classes. Other models, including EfficientNetB5 and EfficientNetB7, also showed excellent performance, with their accuracies exceeding 0.9987. In the 4-class scenario, EfficientNetB5 and EfficientNetB7 both achieved the highest accuracies of 0.9987, while EfficientNetB5 slightly outperformed, recording an AUC of 0.9994, indicating its ability to discriminate across multiple classes. Notably, some models exhibited high EBI/CBI values indicating class-error imbalance rather than a metric-computation anomaly.

To examine the entire EfficientNet family (from B0 to B7, plus V2S, V2M, V2L) in both scenarios, a set of visualizations was prepared for each model. According to the criteria in Table 3, we examine three models in particular: EfficientNetB6, which was the best with a perfect accuracy of 1 in the 2-class case, EfficientNetB5, which outperformed in the 4-class case with an AUC of 0.9994, and EfficientNetV2L, which is the most complex model with the largest number of parameters. Since it did not take first place in either scenario, we put its details in the Supplementary Materials section. For each of the top two models, five visualizations were prepared and are shown in a figure with subfigures (A) to (E): Figure 2 for EfficientNetB6 (2-class), Figure 3 for EfficientNetB5 (4-class), and **Supplementary Figures S1** (2-class) and **S2** (4-class) for EfficientNetV2L. These visualizations include: the confusion matrix, ROC/AUC curves, and training/-validation error and accuracy curves. The confusion matrix shows the distribution of correct and incorrect classifications; for example, in Figure 2(A), for EfficientNetB6, the classification was perfect. The ROC/AUC curves are also used to measure the ability to discriminate between classes; Figure 2(B) is for EfficientNetB6 with AUC = 1, and Figure 3(B) is for EfficientNetB5 with AUC = 0.9994. The training/validation error and accuracy curves also follow the model trend over 100 epochs and 10 layers; the training was done with the Adam optimizer and a fixed learning rate of 0.0001 for all models, with early stopping based on validation error, and the batch size was 64 for B0–B4 and V2S, 32 for B5–B6 and V2M, and 16 for B7 and V2L.

The reduction in batch size was applied only for larger architectures and to accommodate the 16 GB limit of available GPU memory, not to tune the hyperparameters; the implications of this will be discussed in Section 4. For example, Figure 2(D) for EfficientNetB6 shows that the error converges to zero and the accuracy stabilizes at a value of 1. The scatterplot of the PCI shows the relationship between prediction confidence and accuracy (for example, Figure 2(D) for EfficientNetB6, where most of the examples have confidence above 0.95); and the distribution of accuracy per class (boxplot) shows the distribution of accuracy across classes, such as Figure 3(E) for EfficientNetB5 in the 4-class scenario, where the top-ranked class shows the lowest median accuracy (around 0.994) and the highest variance, which is consistent with its CIS of 0, indicating balanced performance despite minor class-specific challenges.

3.2. Comparison with prior methods

Table 4 summarizes six previous benchmark studies [20–25] along with the proposed method. The best previous methods achieve accuracies of 98.64% [22] and 98.50% [24] in binary recognition. Our framework, while achieving 100% accuracy in the binary task (EfficientNet-B6) and 99.87% in the 4-class task (EfficientNet-B5), has a more fundamental distinction: unlike previous studies that have mainly limited themselves to binary recognition or separate assessments [23], the present study performs a 4-level clinical classification (normal, grade I to III) in an integrated and multi-class manner.

The second distinguishing feature is the preprocessing pipeline. None of the methods in Table 1 have an explicit privacy-preserving step, while we have used Mask R-CNN-based anonymization to remove identity contexts and optimized denoising with GWO. From an architectural perspective, the evaluation of eleven EfficientNet variants in this study is much broader than the single architectures of previous works, allowing the identification of a performance plateau (in models B5 to B7). Finally, direct comparisons of accuracies should be made with caution, as the size of the datasets and validation strategies (such as the lack of patient-level segmentation) in all these studies, including the present work, have similar limitations, which are discussed in detail in Section 4.

3.3. Ablation and architecture comparison

To see how the performance of the different EfficientNet variants differs, we evaluated all 11 architectures in a common test framework. As shown in Table 3, all models were trained on the same dataset, with the same data augmentation strategy and cross-validation procedure. The only major difference was the backbone architecture; the batch size was also adjusted based on the memory requirements of each model. So this comparison is an architecture-level evaluation, under roughly the same training conditions.

Table 3 Performance metrics for 2-class and 4-class scenarios

Scenario	Model	Batch Size	Accu-racy	Preci-sion	Recall	F1 Score	AUC	CIS	EBI	PCI	CDI	CBI	AI
2-Class (Normal, Fatty)	B0	64	0.9981	0.9981	0.9981	0.9981	0.9980	1	0.50	0.9996	0.0039	0.50	0.9984
	B1	64	0.9987	0.9987	0.9987	0.9987	0.9979	1	0	0.9993	0.0013	0	0.9955
	B2	64	0.9974	0.9974	0.9974	0.9974	0.9970	1	1	0.9993	0.0051	1	0.9966
	B3	64	0.9981	0.9981	0.9981	0.9981	0.9980	1	0.50	0.9993	0.0039	0.50	0.9979
	B4	64	0.9968	0.9968	0.9968	0.9968	0.9954	1	4	0.9989	0.0064	4	0.9959
	B5	32	0.9994	0.9994	0.9994	0.9994	0.9990	1	0	0.9998	0.0006	0	0.9915
	B6	32	1	1	1	1	1	1	0	0.9999	0	0	0.9966
	B7	16	0.9987	0.9987	0.9987	0.9987	0.9979	1	0	0.9998	0.0013	0	0.9979
	V2S	64	0.9987	0.9987	0.9987	0.9987	0.9985	1	1	0.9998	0.0026	1	0.9976
	V2M	32	0.9987	0.9987	0.9987	0.9987	0.9985	1	1	0.9993	0.0026	1	1
	V2L	16	0.9968	0.9968	0.9968	0.9968	0.9971	1	0.25	0.9987	0.0064	0.25	0.9993
4-Class (Normal, Grade I , Grade II , Grade III)	B0	64	0.9923	0.9923	0.9923	0.9923	0.9942	0.9986	1.33	0.9910	0.0309	1.33	0.9753
	B1	64	0.9981	0.9981	0.9981	0.9981	0.9987	0.9993	1	0.9972	0.0058	1	0.9917
	B2	64	0.9961	0.9962	0.9961	0.9961	0.9973	0.9988	4	0.9989	0.0116	4	0.9928
	B3	64	0.9974	0.9974	0.9974	0.9974	0.9979	0.9992	2	0.9984	0.0077	2	0.9952
	B4	64	0.9961	0.9961	0.9961	0.9961	0.9973	0.9990	1	0.9983	0.0154	1	0.9943
	B5	32	0.9987	0.9987	0.9987	0.9987	0.9994	0.9994	0	0.9989	0.0013	0	0.9848
	B6	32	0.9987	0.9987	0.9987	0.9987	0.9986	0.9995	1	0.9983	0.0026	1	0.9920
	B7	16	0.9987	0.9987	0.9987	0.9987	0.9990	0.9995	0	0.9988	0.0013	0	0.9973
	V2S	64	0.9968	0.9968	0.9968	0.9968	0.9978	0.9993	1	0.9989	0.0129	1	0.9963
	V2M	32	0.9974	0.9974	0.9974	0.9974	0.9982	0.9995	2	0.9972	0.0077	2	1
	V2L	16	0.9961	0.9962	0.9961	0.9961	0.9968	0.9989	2	0.9976	0.0077	2	0.9991

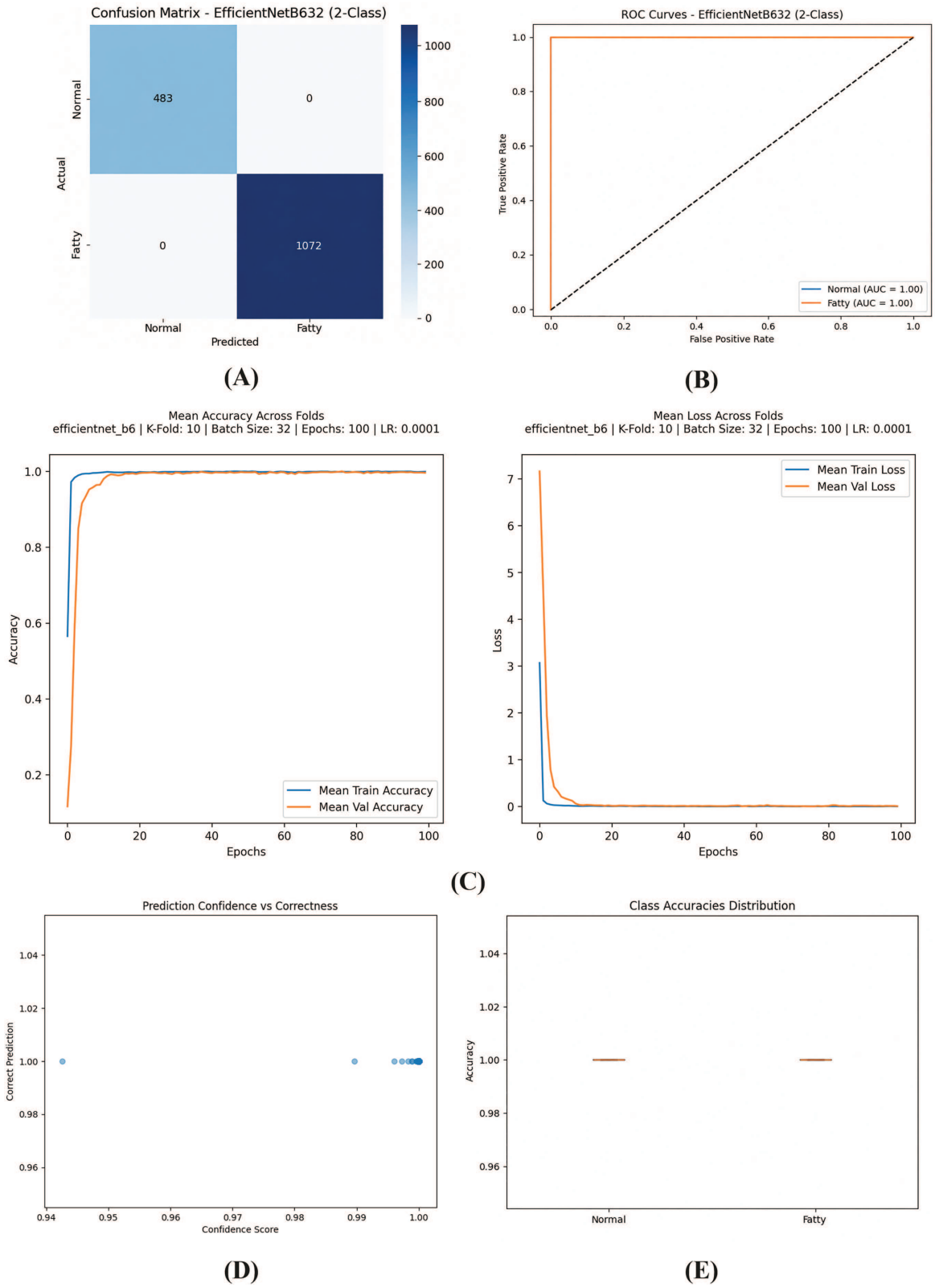

Figure 2 Performance evaluation of the EfficientNet-B6 model for 2-class classification, including (A) confusion matrix (Normal, Fatty), (D) ROC curve for the Fatty class, (C) accuracy and loss trends, (D) confidence vs. correctness, and (E) class-wise accuracy distribution.

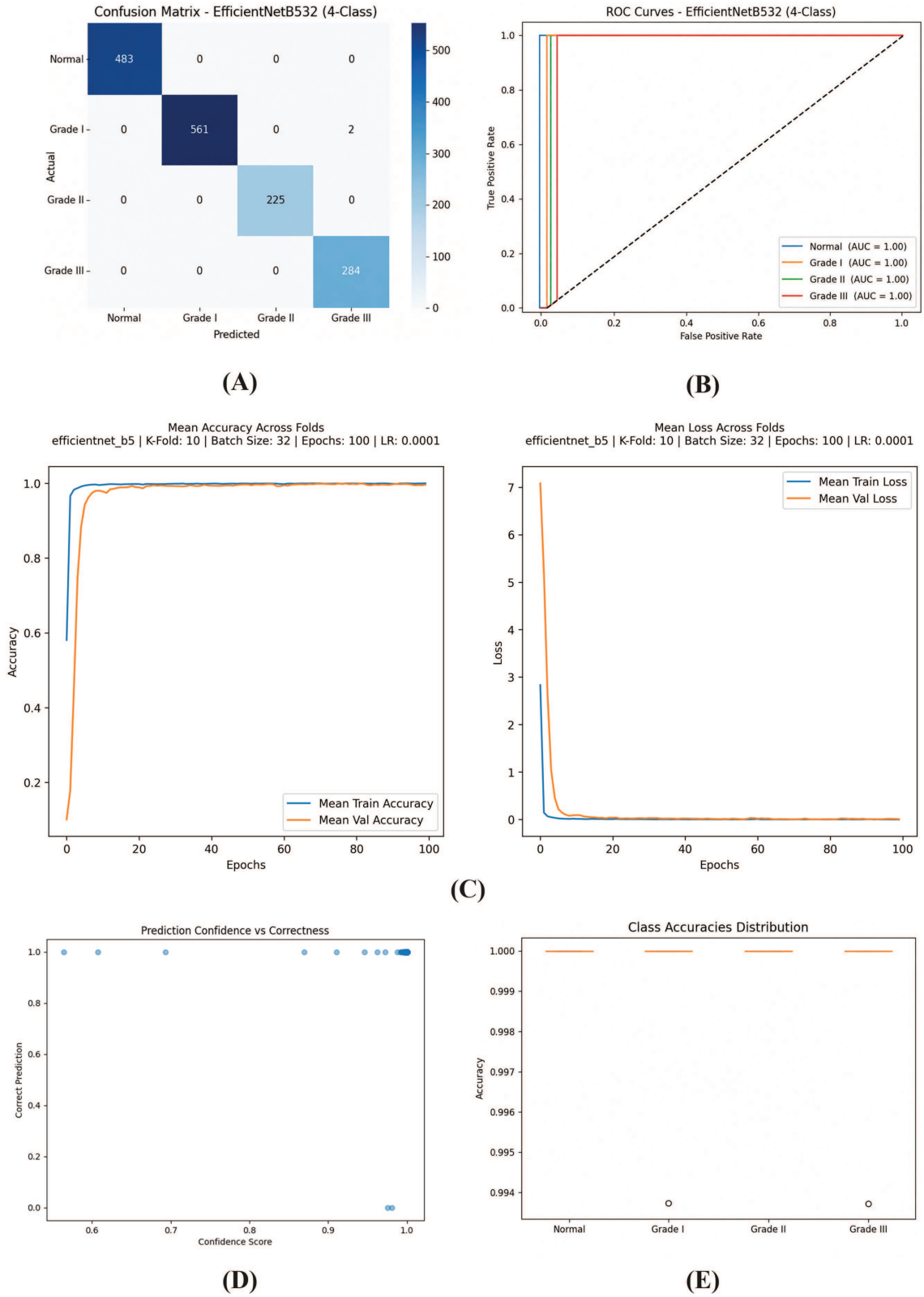

Figure 3 Performance evaluation of the EfficientNet-B5 model for 4-class classification, including (A) confusion matrix, (B) ROC curve (Normal, Grade I, Grade II, Grade III), (C) accuracy and loss trends, (D) confidence vs. correctness, and (E) class-wise accuracy distribution.

Table 4 Summary of state-of-the-art deep learning methods for liver ultrasound image analysis

Ref.	Method	Architecture	Performance metrics	Dataset	Device	Image Type
[20]	Feature learning (texture + gray-scale) with CNN	CNN	Sensitivity = 83% Accuracy = 90% Specificity = 95%	500 images	GE Vive E9	B-Mode (DICOM)
[21]	Transfer learning	Inception v3, VGG-16	AUC = 0.93 Sensitivity = 89.90% Accuracy = 93.23% Specificity = 96.60%	60 patients (629 images)	Siemens Acuson NX3 & S2000	B-Mode
[22]	CNN	ResNet v2, GoogleNet, AlexNet and ResNet101	Accuracy = 98.64% Sensitivity = 97.20% Specificity = 100%	55 patients (55 images)	GE Vive E9	B-Mode (DICOM)
[23]	Training on raw + Shannon entropy images	VGG-16	AUC_{Mild} = 0.71 $AUC_{Moderate}$ = 0.75 AUC_{Severe} = 0.88	205 patients (820 images)	Burlington Terason 3000	B-Mode
[24]	CNN	Inception ResNet-v2	AUC = 0.992 Accuracy = 98.50% Specificity = 92%	55 patients (55 images)	GE Vive E9	B-Mode (DICOM)
[25]	CNN	VGG16 + Fully Connected Layers	AUC = 0.95 $Accuracy_{Average}$ = 77.5%	710 images	Mindray M9	B-Mode + Doppler
Proposed Method	Transfer learning + GWO-optimized denoising + Mask R-CNN anonymization	EfficientNet (B0–B7, V2-S/M/L)	$Accuracy_{binary}$ = 100% AUC_{binary} = 1 $Accuracy_{grading}$ = 99.87% $AUC_{grading}$ = 0.9994	55 patients (55 images)	GE Vivid E9	B-Mode (DICOM)

The accuracy results in Figure 4(A) show a gradual improvement from EfficientNetB0 to the larger models, with the performance reaching 99.87–100% in the range of B5–B7. Interestingly, no improvement was observed for EfficientNetV2-M and V2-L, despite the much larger number of parameters. This means that increasing the model capacity beyond the range of B5–B7 did not yield any significant benefit, at least for this dataset. On the other hand, the distribution of the batch size in the experiments also provides a clue: since models with the same batch size have different accuracies, the difference in performance cannot be attributed solely to the choice of batch size. These differences seem to be more related to the backbone architecture itself.

However, accuracy alone does not tell the whole story, as the model's resolution and error distribution are different things. In Figure 4(B), we see that in binary classification, the correlation between AUC and EBI/CBI is significant for EfficientNetB4. Although B4 achieved 99.68% accuracy, its EBI/CBI value was 4.0000 (the highest imbalance among all architectures) and its AUC was also the lowest (0.9954). This combination means that the small number of classification errors was spread asymmetrically between the two classes; therefore, a high overall accuracy does not necessarily mean good class-level resolution. A similar pattern was observed in the four-class scenario, although with a greater dispersion between architectures. EfficientNetB2 had the highest imbalance based on EBI/CBI, while EfficientNetB0 had the lowest AUC. This is interesting, because B0 is the smallest architecture in terms of capacity, meaning that perhaps insufficient representation capacity can also limit class discrimination when the classification becomes more detailed. These findings suggest that accuracy and AUC should be read alongside class balance metrics, as overall metrics can sometimes mask asymmetric error distributions.

Another point is made by examining the accuracy in terms of the AI (Figure 4(C)). EfficientNetB5 had one of the highest accuracy levels among all models, but at the same time it showed the lowest AI in both scenarios. According to the definition of AI, this means that the largest difference between training and validation performance was precisely for this model. This combination (high validation accuracy combined with relatively low AI) is worth considering when trying to assess the reliability of B5. Although the cross-validation results show high predictive performance, this discrepancy between training and validation could be an indication of reduced generalizability stability. This finding supports the recommendation made in Section 4: before considering EfficientNetB5 for clinical use, external or independent validation at the patient level should be performed.

4. Discussion

In this study, we evaluated the performance of different EfficientNet models (B0 to B7, as well as V2S, V2M, and V2L) in a 2-class scenario (normal and fatty) and a 4-class scenario (normal, grade I, grade II, and grade III). The initial dataset consisted of 55 images, whose distribution in the 2-class scenario included 17 patients with normal liver and 38 patients with fatty liver, and in the 4-class scenario included 17, 20, 8, and 10 patients for normal, grade I, grade II, and grade III, respectively. This initial distribution showed class imbalance, especially for grade II and grade III,

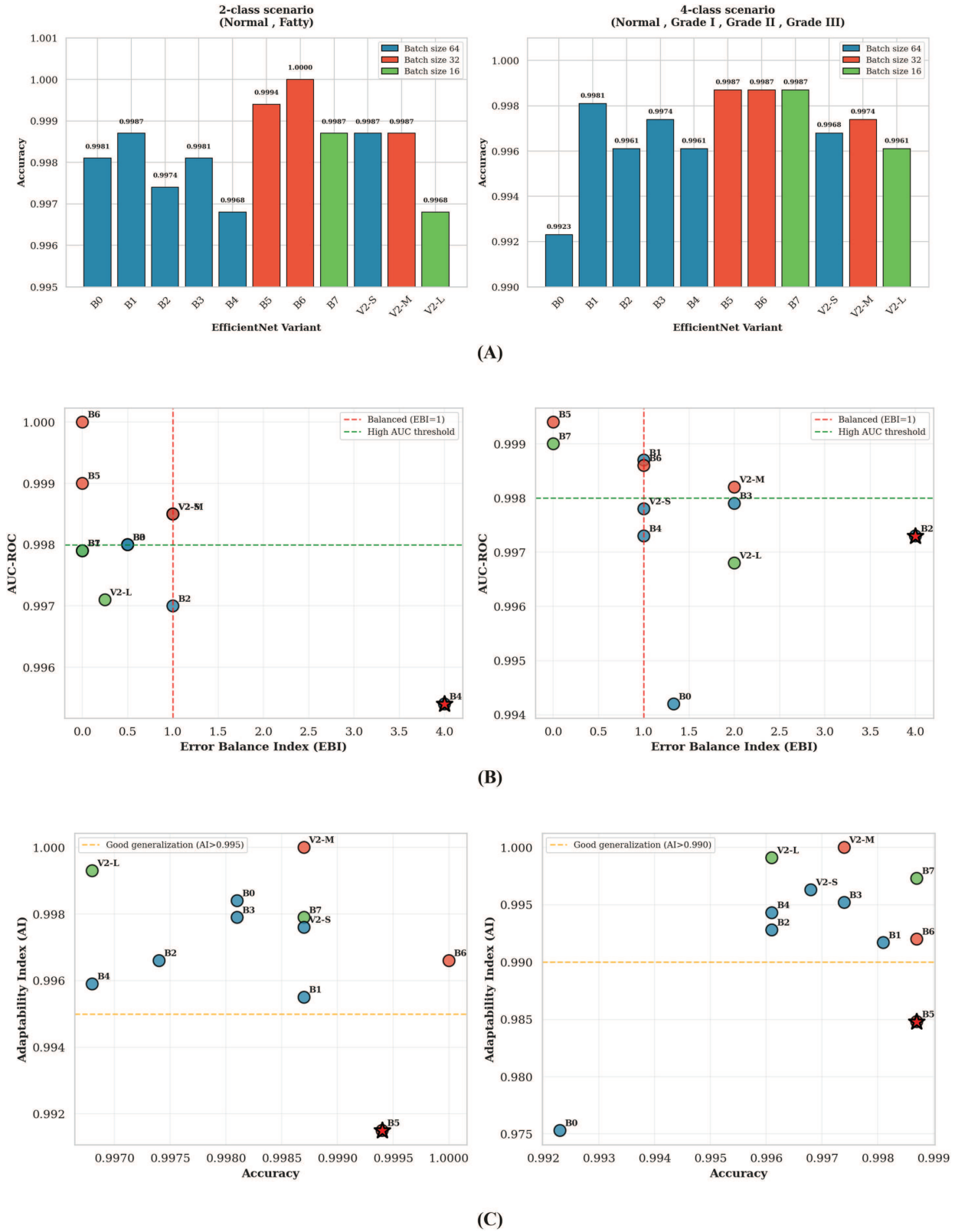

Figure 4 Performance evaluation of EfficientNet architectures in binary and 4-class scenarios: (A) classification accuracy grouped by batch size, (B) AUC versus Error Balance Index and Cost-Balance Index, and (C) accuracy versus Adaptability Index.

which had fewer samples. To address this problem, we applied the data augmentation techniques detailed in Table 2 and increased the dataset to 1550 images. All experiments and evaluations were performed on this augmented dataset of 1550 images.

In the 2-class scenario, the performance of the models was at a very high level; their accuracy fluctuated in the range of 0.9968 to 1. Meanwhile, EfficientNetB6 showed a perfect performance, so that its accuracy, AUC, as well as Precision, Recall, and F1-Score all reached a value of 1. This result indicates the high ability of this model to distinguish normal and fatty classes. Examination of the CIS showed that most of the models (including B1, B5, B6, and B7) recorded a value of 0, indicating a perfect balance in performance between the two classes. However, two models, EfficientNetB4 and EfficientNetV2S, were exceptions to this pattern, showing CIS values of 4 and 1, respectively. A closer examination of these cases showed that the main reason for this increase was the greater sensitivity of these two models to changes applied in the data augmentation process. At the same time, the same data augmentation techniques played an effective role in improving the robustness and overall performance of the models in this scenario by simulating conditions such as mirrored medical structures and optical diversity.

In the 4-class scenario, the overall accuracy of the models ranged from 0.9923 (EfficientNetB0) to 0.9987 (EfficientNetB5, B6, and B7). EfficientNetB5, B6, and B7 achieved the highest accuracy of 0.9987, with AUC values between 0.9994 and 0.9995, demonstrating their strong capability to classify all 4 classes. The Precision, Recall, and F1-Score for these models were also 0.9987, indicating high reliability in identifying each class. The CIS for EfficientNetB5 and B7 was 0, signifying perfect balance in performance across the classes. In contrast, models such as EfficientNetB2 and EfficientNetB0 recorded higher CIS values of 4 and 1.3333, respectively. We analyzed this outcome and found that these higher CIS values result from the increased complexity of the data for these classes within the augmented dataset.

The boxplot for EfficientNetB5 in the 4-class scenario shows that the Normal class achieved the highest median accuracy of 0.998 with the lowest variance, indicating a strong performance for this class. We determined that this performance is due to the distinctive features of the Normal class, which are further enhanced by data augmentation techniques such as CLAHE to increase contrast and translation to resist positional variations. In contrast, the Grade I class showed the lowest median accuracy of 0.994 and the highest variance. We investigated this result and found that the high variance is due to the similarity of the features of Class I with those of other classes, such as Class II or Normal. However, the CIS of 0 for EfficientNetB5 highlights the overall balance of the model across classes, which is a significant strength.

These results should nonetheless be interpreted with appropriate caution given the size and validation strategy of the underlying dataset. The original cohort comprised only 55 patients and was expanded to 1,550 images through augmentation. In medical imaging studies, however, a large number of derived or augmented images does not necessarily compensate for a limited number of independent subjects, and improper handling of augmented or derived samples during dataset splitting has been shown to substantially inflate reported classification performance [34, 35]. Because augmentation was applied before the 10-fold split, augmented copies from the same source patient could appear in both the training and validation folds within a given run. Similar methodological studies have demonstrated that when augmented samples originating from the same subject are distributed across training and validation folds, data leakage may substantially inflate estimated diagnostic performance, emphasizing the importance of subject-wise rather than sample-wise cross-validation in medical AI research [17]. This within-subject overlap can therefore inflate cross-validated accuracy relative to what would be observed on genuinely unseen patients, and is the most likely explanation for the near-ceiling accuracies (up to 1.0) reported in Table 3.

Consequently, the reported figures should be read as upper-bound estimates of in-distribution performance rather than generalization estimates. The methodologically preferable alternative, patient-level (group-wise) cross-validation, in which all augmented copies of a given patient remain within a single fold, together with evaluation on an independent, patient-disjoint test set, is identified as a priority for future work rather than a claim we make on the present data. We similarly note that the varying batch sizes across models (16–64, dictated by GPU memory rather than tuning) introduce a potential confound when comparing architectures head-to-head, and a fixed-batch-size ablation (using gradient accumulation for the larger models) is planned to disentangle this effect from genuine architectural differences.

Regarding interpretability, the present study does not yet include a pixel-level explanation method such as Grad-CAM. Qualitatively, we expect that such an analysis would be particularly informative for the Grade I misclassifications noted above, where the model's attention may fall on peri-hepatic tissue boundaries rather than the parenchymal texture changes that radiologists rely on; confirming or refuting this is left as a concrete direction for future work. Likewise, a like-for-like comparison against the six prior methods summarized in Table 4, re-implemented on our dataset under identical experimental settings, would more directly validate the advantages of the present approach, but was outside the scope of the current experiments since several of those methods rely on proprietary or non-public data and are not yet available for re-implementation.

5. Conclusion

In this study, we evaluated the performance of EfficientNet models (including B0 to B7, V2S, V2M, and V2L) for two classification scenarios: a 2-class scenario (normal and fatty) and a 4-class scenario (normal, grade I, grade II, and grade III). As announced in previous sessions, the initial dataset was expanded to 1550 images using data augmentation techniques, and all experiments were performed on this augmented dataset. In the 2-class scenario, all models achieved high accuracy (0.9968 to 1), among which EfficientNetB6 showed the best performance with an accuracy of 1. Also, in the 4-class scenario, EfficientNetB5, B6, and B7 models achieved the highest accuracy with a value of 0.9987 and area under the curve (AUC) values in the range of 0.9994 to 0.9995. The boxplot of EfficientNetB5 in the 4-class scenario showed that the Normal class had the highest median accuracy (0.998) and the lowest variance, while the Grade I class had the lowest median accuracy (0.994) and the highest variance, and was identified as the most challenging class. However, the CIS index of zero for the EfficientNetB5 and B7 models indicates an excellent overall balance among the classes. This study shows that EfficientNet is an efficient model for diagnosing fatty liver disease and its stages, although improvements are needed for challenging classes such as Grade I.

As a follow-up, we plan to add multi-task learning (MTL) alongside the STL approach used here. MTL would let the model handle related tasks at once which should help with harder classes

like Grade I. We'll also bring in Grad-CAM to make the model's decisions easier to read: by showing which parts of an image actually drive a prediction, it tells us what the model is picking up on and where it's slipping, particularly for Grade I. We also want a closer look at accuracy, sensitivity, and F1 per class rather than just overall, since that's usually where the interesting problems hide. Testing on a completely new dataset comes next, to see if performance holds up outside this data. Worth adding: none of this is tied to MASLD specifically. The same anonymization pipeline plus EfficientNet should work just as well for other ultrasound tasks, like thyroid nodules or renal cortex grading, and that's a direction we want to take separately.

Recommendations

The results of this work open new avenues, but more effort is needed before real clinical use. Below are points to guide the next steps. This dataset is small. It includes 55 patients, and the model also depends on data augmentation. Because of that, later research should test the same framework on more subjects. It should also include multicenter and multiplatform trials. For ultrasound, that means using scanners from different brands. Also, patient-level splitting is needed so the same person does not appear in both training and validation, which would cause leakage. Before any clinical rollout, the model should be checked on data it has not seen. Use held-out sets, and if you can, add external cohorts. This matters because you want the accuracy seen here to hold up even when the data changes. To build trust and make the output easier to interpret, use explainable AI tools. Grad-CAM is one option. It can show which parts of the image drive the model's decision. This should be useful for hard cases, like Grade I. For clinical integration, keep the Mask R-CNN anonymization step. Treat it as a routine part of the workflow. In that way, MASLD screening tools can be privacy-friendly by default. Finally, consider improving the system with a multi-task learning setup. Predict steatosis grading in a way that runs at the same time as the binary label. Doing both together may help performance, especially for less common groups, such as grades II and III.

Acknowledgements

This work is supported in part by Shahid Chamran University of Ahvaz under grant number **SCU.EC1404.450**.

Funding Support

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethical Statement

This study utilized publicly available human/animal datasets from established repositories and previously published research. The research does not involve the collection, generation, or primary analysis of new human or animal subject data by the authors.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this study.

Data Availability Statement

The liver ultrasound image dataset used in this study is publicly available on the Zenodo archive (https://doi.org/10.5281/zenodo.1009146) under the CC BY 4.0 license. The data can be accessed through the reference provided in the bibliography.

Author Contribution Statement

A'adel Sayyahi: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing-original draft, Writing-review & editing, Visualization. **Seyed Enayatallah Alavi:** Conceptualization, Methodology, Validation, Resources, Supervision, Project administration. **Morteza Jaderyan:** Conceptualization, Methodology,Validation, Resources, Supervision.

Supplementary Information

The supplementary file is available at https://doi.org/10.47852/bonviewMEDIN620210151.

References

[1] Fritsche, K., Ziková-Kloas, A., Marx-Stoelting, P., & Braeuning, A. (2023). Metabolism-disrupting chemicals affecting the liver: Screening, testing, and molecular pathway identification. *International Journal of Molecular Sciences*, *24*(3), 2686. https://doi.org/10.3390/ijms24032686

[2] Li, Y., Yang, P., Ye, J., Xu, Q., Wu, J., & Wang, Y. (2024). Updated mechanisms of MASLD pathogenesis. *Lipids in Health and Disease*, *23*(1), 117. https://doi.org/10.1186/s12944-024-02108-x

[3] Nivethitha, V., Daniel, R. A., Debnath, A., Dwarakanathan, V., Jeer, G., & Kavipriya, G. (2025). Diagnostic accuracy of artificial intelligence models for imaging detection of hepatic steatosis through systematic review and meta analysis. *Scientific Reports*, *15*(1), 34408. https://doi.org/10.1038/s41598-025-17386-3

[4] Teng, M. L., Ng, C. H., Huang, D. Q., Chan, K. E., Tan, D. J., Lim, W. H., . . . , & Muthiah, M. D. (2023). Global incidence and prevalence of nonalcoholic fatty liver disease. *Clinical and Molecular Hepatology*, *29*(Suppl), S32–S42. https://doi.org/10.3350/cmh.2022.0365

[5] Miao, L., Targher, G., Byrne, C. D., Cao, Y. Y., & Zheng, M. H. (2024). Current status and future trends of the global burden of MASLD. *Trends in Endocrinology & Metabolism*, *35*(8), 697–707. https://doi.org/10.1016/j.tem.2024.02.007

[6] Wild, S. H., Lazarus, J. V., Spearman, C. W., Ocama, P., Bilson, J., Zhou, X.-D., . . . , & Byrne, C. D. (2026). MASLD prevalence, incidence and global aspects. 3034–3047. https://doi.org/10.1007/s00125-026-06726-1 *Diabetologia*,*69*

[7] Indre, M. G., Stefanini, B., Boe, M., Capelli, R., Chen, R., Abbati, C., . . . , & Ravaioli, F. (2026). HCC is the predominant liver-related event in MASLD: 2-step non-invasive algorithms to stratify risk in non-cirrhotic patients. *Journal of Hepatocellular Carcinoma*, *13*, 561956. https://doi.org/10.2147/JHC.S561956

[8] Tincopa, M. A., & Loomba, R. (2023). Non-invasive diagnosis and monitoring of non-alcoholic fatty liver disease and non-alcoholic steatohepatitis. *The Lancet Gastroenterology & Hepatology*, *8*(7), 660–670. https://doi.org/10.1016/S2468-1253(23)00066-3

[9] Lai, J. C., Liang, L. Y., & Wong, G. L. (2024). Noninvasive tests for liver fibrosis in 2024: Are there different scales for different diseases? *Gastroenterology Report*, *12*, goae024. https://doi.org/10.1093/gastro/goae024

[10] Graupera, I., Thiele, M., Serra-Burriel, M., Caballeria, L., Roulot, D., . . . , & Wong, G. L. H. (2022). & Investigators of the LiverScreen Consortium. Low accuracy of FIB-4 and NAFLD fibrosis scores for screening for liver fibrosis in the population. *Clinical Gastroenterology and Hepatology*, *20*(11), 2567–2576. https://doi.org/10.1016/j.cgh.2021.12.034

[11] Harrison, A. P., Li, B., Hsu, T.-H., Chen, C.-J., Yu, W.-T., Tai, J., . . . , & Tai, D.-I. (2023). Steatosis quantification on ultrasound images by a deep learning algorithm on patients undergoing weight changes. *Diagnostics*, *13*(20), 3225. https://doi.org/10.3390/diagnostics13203225

[12] Li, J., Chen, J., Zeng, X., Lyu, G., Lin, S., & He, S. (2025). Update of machine learning for ultrasound diagnosis of metabolic dysfunction-associated steatotic liver disease: A bright future for deep learning. *PeerJ*, *13*, e19645. https://doi.org/10.7717/peerj.19645

[13] Ma, Z., Liu, B., Zhou, S., Yang, S., Sun, X., Jiang, X., & Sun, X. (2026). Advances in quantitative ultrasound for metabolic dysfunction-associated steatotic liver disease diagnosis. *Frontiers in Physiology*, *17*, 1802284. https://doi.org/10.3389/fphys.2026.1802284

[14] Lepcha, D. C., Goyal, B., Dogra, A., Alkhayyat, A., Sahu, P. K., Ali, A., & Kukreja, V. (2025). Deep learning in medical image analysis: A comprehensive review of algorithms, trends, applications, and challenges. *Computer Modeling in Engineering & Sciences*, *145*(2), 1487–1573. https://doi.org/10.32604/cmes.2025.070964

[15] Rempe, M., Heine, L., Seibold, C., Hörst, F., & Kleesiek, J. (2025). De-identification of medical imaging data: A comprehensive tool for ensuring patient privacy. *European Radiology*, *35*(12), 7809–7818. https://doi.org/10.1007/s00330-025-11695-x

[16] Chiu, I. M., Chen, T. Y., Zheng, Y. C., Lin, X. H., Cheng, F. J., Ouyang, D., & Cheng, C. Y. (2024). Prospective clinical evaluation of deep learning for ultrasonographic screening of abdominal aortic aneurysms. *npj Digital Medicine*, *7*(1), 282. https://doi.org/10.1038/s41746-024-01269-4

[17] Santoro, S., Khalil, M., Abdallah, H., Farella, I., Noto, A., Dipalo, G. M., . . . , & Portincasa, P. (2024). Early and accurate diagnosis of steatotic liver by artificial intelligence (AI)-supported ultrasonography. *European Journal of Internal Medicine*, *125*, 57–66. https://doi.org/10.1016/j.ejim.2024.03.004

[18] Cai, L., & Pfob, A. (2025). Artificial intelligence in abdominal and pelvic ultrasound imaging: Current applications. *Abdominal Radiology*, *50*(4), 1775–1789. https://doi.org/10.1007/s00261-024-04640-x

[19] Wang, X., Wang, L., Hou, X., Li, J., Li, J., & Ma, X. (2024). Comparison and analysis of deep learning models for discriminating longitudinal and oblique vaginal septa based on ultrasound imaging. *BMC Medical Imaging*, *24*(1), 347. https://doi.org/10.1186/s12880-024-01507-x

[20] Zhang, L., Zhu, H., & Yang, T. (2019). Deep neural networks for fatty liver ultrasound images classification. In *2019 Chinese Control and Decision Conference (CCDC)*, 4641–4646. https://doi.org/10.1109/CCDC.2019.8833364

[21] Constantinescu, E. C., Udriștoiu, A. L., Udriștoiu, Ș. C., Iacob, A. V., Gruionu, L. G., Gruionu, G., . . . , & Săftoiu, A. (2021). Transfer learning with pre-trained deep convolutional neural networks for the automatic assessment of liver steatosis in ultrasound images. *Medical Ultrasonography*, *23*(2), 135–139. https://doi.org/10.11152/mu-2746

[22] Zamanian, H., Mostaar, A., Azadeh, P., & Ahmadi, M. (2021). Implementation of combinational deep learning algorithm for non-alcoholic fatty liver classification in ultrasound images. *Journal of Biomedical Physics and Engineering*, *11*(1), 73–84. https://doi.org/10.31661/jbpe.v0i0.2009-1180

[23] Chen, J. R., Chao, Y. P., Tsai, Y. W., Chan, H. J., Wan, Y. L., Tai, D. I., & Tsui, P. H. (2020). Clinical value of information entropy compared with deep learning for ultrasound grading of hepatic steatosis. *Entropy*, *22*(9), 1006. https://doi.org/10.3390/e22091006

[24] Mohammad, U. F., & Almekkawy, M. (2021). Automated detection of liver steatosis in ultrasound images using convolutional neural networks. In *2021 IEEE International Ultrasonics Symposium (IUS)* (pp. 1–4). https://doi.org/10.1109/IUS52206.2021.9593420

[25] Liu, Y., Yu, W., Wang, P., Huang, Y., Li, J., & Li, P. (2024). Deep learning with ultrasound images enhance the diagnosis of nonalcoholic fatty liver. *Ultrasound in Medicine & Biology*, *50*(11), 1724–1730. https://doi.org/10.1016/j.ultrasmedbio.2024.07.014

[26] Byra, M., Styczynski, G., Szmigielski, C., Kalinowski, P., Michałowski, L., Paluszkiewicz, R., . . . , & Nowicki, A. (2018). Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images. *International Journal of Computer Assisted Radiology and Surgery*, *13*(12), 1895–1903. https://doi.org/10.1007/s11548-018-1843-2

[27] Brunt, E. M., Kleiner, D. E., Carpenter, D. H., Rinella, M., Harrison, S. A., . . . , & Loomba, R. (2021). & American Association for the Study of Liver Diseases NASH Task Force. NAFLD: Reporting histologic findings in clinical practice. *Hepatology*, *73*(5), 2028–2038. https://doi.org/10.1002/hep.31599

[28] Tan, Z. X., Mehta, B., Kusel, K., Seow, J., Zelesco, M., Abbott, S., . . . , & Ayonrinde, O. T. (2024). Hepatic steatosis: Qualitative and quantitative sonographic assessment in comparison to histology. *Australasian Journal of Ultrasound in Medicine*, *27*(3), 179–188. https://doi.org/10.1002/ajum.12381

[29] Kantor, C. A., Boussioux, L., Rauby, B., & Talbot, H. (2021). Gradient-based localization and spatial attention for confidence measure in fine-grained recognition using deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(18), 15807–15808. https://doi.org/10.1609/aaai.v35i18.17900

[30] Karaoğlu, O., Bilge, H. Ş., & Uluer, I. (2022). Removal of speckle noises from ultrasound images using five different deep learning networks. *Engineering Science and Technology, an International Journal*, *29*, 101030. https://doi.org/10.1016/j.jestch.2021.06.010

[31] Mirjalili, S., Mirjalili, S. M., & Lewis, A. (2014). Grey wolf optimizer. *Advances in Engineering Software*, *69*, 46–61. https://doi.org/10.1016/j.advengsoft.2013.12.007

[32] Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, *97*, 6105–6114. https://doi.org/10.48550/arXiv.1905.11946

[33] Tan, M., & Le, Q. (2021). EfficientNetV2: Smaller models and faster training. In *International Conference on Machine*

Learning, *139*, 10096–10106. https://doi.org/10.48550/arXiv.2104.00298

[34] Tampu, I. E., Eklund, A., & Haj-Hosseini, N. (2022). Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images. *Scientific Data*, *9*(1), 580. https://doi.org/10.1038/s41597-022-01618-6

[35] Bussola, N., Marcolini, A., Maggio, V., Jurman, G., & Furlanello, C. (2021). AI slipping on tiles: Data leakage in digital pathology. In *Pattern Recognition. ICPR International Workshops and Challenges* (pp. 167–182). https://doi.org/10.1007/978-3-030-68763-2_13

How to Cite: Sayyahi, A., Alavi, S. E., & Jaderyan, M. (2026). Metabolic Dysfunction-Associated Steatotic Liver Disease Diagnosis Using EfficientNet Transfer Learning and Ultrasound Images. *Medinformatics*. https://doi.org/10.47852/bonviewMEDIN620210151