Original

Classification of Nonalcoholic Fatty Liver Grades using Pre-Trained Convolutional Neural Networks and a Random Forest Classifier on B-Mode Ultrasound Images

Amir Reza Naderi Yaghouti (MSc)[1], Ahmad Shalbaf (PhD)[2]*

[1]Department of Biomedical Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran

[2]Department of Biomedical Engineering and Medical Physics, School of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran

ABSTRACT

Background: Nonalcoholic Fatty Liver Disease (NAFLD) as a prevalent condition can significantly have health implications. Early detection and accurate grading of NAFLD are essential for effective management and treatment of the disease.

Objective: The current study aimed to develop an advanced hybrid machine-learning model to classify NAFLD grades using ultrasound images.

Material and Methods: In this analytical study, ultrasound images were obtained from 55 highly obese individuals, who had undergone bariatric surgery and used histological results from liver biopsies as a reference for NAFLD grading. The features were extracted from the ultrasound images using popular pretrained Convolutional Neural Network (CNN) models, including VGG19, MobileNet, Xception, Inception-V3, ResNet-101, DenseNet-121, and EfficientNet-B7. The fully connected layers were removed from the CNN models and also used the remaining structure as a feature extractor. The most relevant features were then selected using the minimum Redundancy Maximum Relevance (mRMR) method. We then used four classification algorithms: Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), Multilayer Perceptron (MLP) neural network, and Random Forest (RF) classifiers, to categorize the ultrasound images into four groups based on liver fat level (healthy liver, low fat liver, moderate fat liver, and high-fat liver).

Results: Among the different CNN models and classification methods, EfficientNet-B7 and RF achieved the highest accuracy. The average accuracies of the LDA, MLP, SVM, and RF classifiers for the feature extraction method with EfficientNet-B7 were 88.48%, 93.15%, 95.47%, and 96.83%, respectively. The proposed automatic model can classify NAFLD grades with a remarkable accuracy of 96.83%.

Conclusion: The proposed automatic classification model using EfficientNet-B7 for feature extraction and a Random Forest classifier can improve NAFLD diagnosis, especially in regions, in which access to professional and experienced medical experts is limited.

Keywords

Fatty Liver; Ultrasound Images; Liver Diseases; Convolutional Neural Networks; Random Forest

Introduction

Nonalcoholic Fatty Liver (NAFL) disease affects individuals worldwide, with an estimated prevalence of around 20-30% in developed countries [1]. The liver encounters challenges increased in metabolizing fats, resulting in the accumulation of fat within

*Corresponding author:
Ahmad Shalbaf
Department of Biomedical Engineering and Medical Physics, School of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran
E-mail: shalbaf@sbmu.ac.ir

Received: 22 July 2023
Accepted: 11 December 2023

liver tissues and the subsequent development of a fatty liver. NAFL disease often presents without noticeable symptoms and is frequently detected in its advanced and potentially dangerous stages; it occurs when the liver struggles to metabolize fats, leading to the accumulation of fat within liver tissues. Early and accurate diagnosis of NAFL disease and its severity is crucial to prevent disease progression and facilitate timely and appropriate treatment. Liver biopsy and pathological laboratory results are widely regarded as the gold standards for diagnosing and assessing the severity of liver conditions. However, it is important to note that these methods are invasive with potential risks, such as pain and bleeding [2].

Ultrasound imaging, as a powerful and universal diagnostic tool for physicians and radiologists, is widely used to diagnose NAFL disease in most imaging methods. Ultrasound, as a diagnostic tool for patients with NAFL disease, has some advantages, primarily due to its non-invasive nature; additionally, it is a cost-effective and widely accessible imaging technique that provides real-time imaging of the liver, leading to dynamic monitoring of changes over time. This non-invasive approach is particularly valuable for patients with NAFL disease, as it eliminates the need for invasive procedures while enabling long-term monitoring and management. Furthermore, ultrasound is relatively safe without any exposure of patients to ionizing radiation, resulting in a preferred imaging modality for certain patients, such as pregnant women and children [3, 4]. NAFL disease and its severity can be diagnosed based on the assessment of ultrasound images by a highly expert radiologist, visually, which is tedious and subjective. Advanced artificial intelligence tools to quantitatively analyze ultrasound images can automate, improve reliability, and provide objective estimation of the NAFL disease grade, helping physicians and radiologists achieve higher accuracy and efficiency in diagnosis [5].

Ribeiro and Sanches et al. [6] extracted the features of spectral images and radio frequency to detect the NAFL grade based on the Bayesian method as a classifier. Kyriakou et al. [7] extracted texture features from ultrasound images and used the K-Nearest Neighbor (KNN) classifier to categorize the images based on NAFL disease grades. Wan and Zhou [8] extracted features using wavelet packet transform and a Support Vector Machine (SVM) classifier for this task. Acharya et al. [9] extracted image features using three methods: texture features, wavelet transform, and higher-order spectrum properties, and then classified NAFL diseases using a decision tree. Kopili et al. [10] extracted the texture features of ultrasound images using a Gray-Level Co-Occurrence Matrix (GLCM) and classified these features using an SVM classifier. Naderi et al. [11] extracted texture features from ultrasound images using GLCM, employing the minimum Redundancy and Maximum Relevance (mRMR) technique for feature selection, and then categorized NAFL disease into four groups via the AdaBoost classifier. Hassan et al. [12] classified NAFL disease using image features using the stacked scattered automatic encoder method and the Softmax classifier and compared their proposed method with multiclass SVM, KNN, and simple Bayesian methods. Saba et al. [13] extracted features from ultrasound images utilizing five feature extraction methods—Harlick, Gupta, Fourier, Basic geometric, DCT, and Gabor—and subsequently classified them through backpropagation neural network classification. However, these methods have problems such as the inability to extract all the effective features of the image, high computational complexity, the dependence of the result of this algorithm on the segmentation method, and determination of the area to be evaluated by an expert [14].

Deep learning approaches based on Convolutional Neural Network (CNN) models have received considerable attention in the medical field [15]. Compared with traditional image recognition and classification approaches, which require manual feature extraction and optimal selection, CNN models can automatically extract useful image features. However, the training of a CNN model conventionally demands a substantial volume of input data, posing challenges for researchers engaged in medical image processing via CNN methodologies. Due to the shortage of adequately labeled medical images, training a CNN model becomes challenging, and in some cases, impossible. As a solution, transfer learning methods have been employed, which involves utilizing the knowledge acquired from a pre-trained CNN model to address the problem, rather than constructing an entirely new CNN model. The usefulness of a pretrained CNN model depends on its ability to adapt to images outside the main educational dataset [16]. Byra et al. [17] pioneered the integration of transfer learning in fatty liver classification, employing Inception-ResNet-V2 for feature extraction from ultrasound imaging sequences and conducting comparative analyses against the hepatorenal index and GLCM algorithms. Constantinescu et al. [18] classified ultrasound images using two pre-trained networks, VGG19 and Inception-V3, using the Softmax classifier.

This study aimed to develop an advanced hybrid model integrating deep transfer learning through CNN models and diverse machine-learning methods to classify the grade of NAFL disease. The classification relies on ultrasound images obtained from 55 individuals with severe obesity who underwent bariatric surgery. The proposed methodology involves employing transfer learning across distinct CNN models to extract comprehensive features from liver ultrasound images, subsequently refining and selecting the most discriminative features. Finally, these images were classified using a Random Forest (RF) classification method into four groups. In other words, the current study aimed to use the capabilities of deep learning and traditional machine learning methods concurrently. In the present study, the proposed novel approach aimed to achieve high accuracy and robustness in medical image analysis. Additionally, this study can significantly contribute to various clinical applications.

Material and Methods

Dataset

In this analytical study, liver ultrasound images of 55 highly obese patients, who had undergone obesity surgery, were obtained. The datasets, obtained one or two days preoperatively, were obtained from the Department of Internal Medicine, Warsaw University of Medical Sciences, Poland. As part of the university's routine protocol, every patient who underwent obesity surgery underwent a biopsy, followed by a histological evaluation of the liver by a pathologist. Fatty liver levels were defined based on the percentage of hepatocytes with fatty infiltration [17]. Figure 1 illustrates the distribution of fatty liver levels derived from biopsy results obtained by pathologists from the patient dataset. Data [19] were classified into four classes according to fatty liver stages: healthy liver (steatosis level <5%, n=17), low-fat liver (steatosis level 5-30%, n=20), moderate fat liver (steatosis level 30-70%, n=8), and high-fat liver (steatosis level >70%, n=10).

Ultrasound images were acquired using a GE Vivid E9 ultrasound device with a 2.5 MHz probe with a resolution of 434×636 pixels. A sequence of images corresponding to a heartbeat was obtained and saved in DICOM format for each patient. Due to the movement, speckle pattern, and relative position of the liver and kidney, the images in each sequence varied slightly for each patient. In addition, the number of images per sequence was not fixed and depended on the frame rate of the device-scanner probe. From each image sequence, ten images were selected for further processing, which increased the total number of images

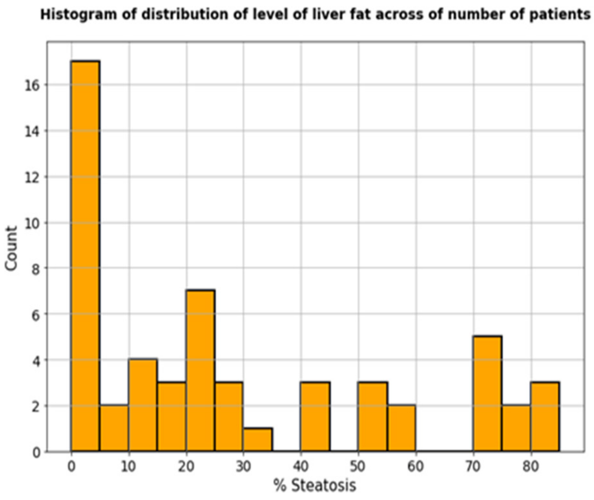

Figure 1: Distribution of liver fat levels in the study population based on biopsy results obtained from 55 highly obese patients who underwent obesity surgery. The liver fat levels were classified into four groups: healthy liver (steatosis level <5%), low fat (5%≤ steatosis level <30%), moderate fat (30%≤ steatosis level <70%), and high fat (steatosis level ≥70%). The dataset was obtained from the Department of Internal Medicine, Warsaw University of Medical Sciences, Poland and was divided based on the percentage of hepatocytes with fatty infiltration.

in our dataset and provided our models with more diverse data. The final dataset consisted of 550 ultrasound images (10 images per individual×55 individual). To ensure a robust evaluation and address potential correlations due to shared patient data, the dataset was randomly split into five folds. For each fold, one-fifth of the patients (11 patients) were designated as the test set, with the remaining used for training. This random splitting and testing process was repeated 50 times to mitigate dataset-specific biases and provide a comprehensive evaluation. The dataset was divided into four groups based on fatty liver levels: healthy liver, low-, moderate-, and high-fat liver.

Preprocessing

During the preprocessing step, the numbers and signals of the ultrasound images were removed. Thereafter, the image margins and any extraneous data points were cropped to isolate the central Region of Interest (ROI), in which encompasses all diagnostically relevant features while reducing image dimensionality for computational efficiency. As a result, the initial image dimensions were reduced from 434×636 pixels to 399×399 pixels. Figure 2 exhibits a sample image, alongside the outcomes post-preprocessing. For the

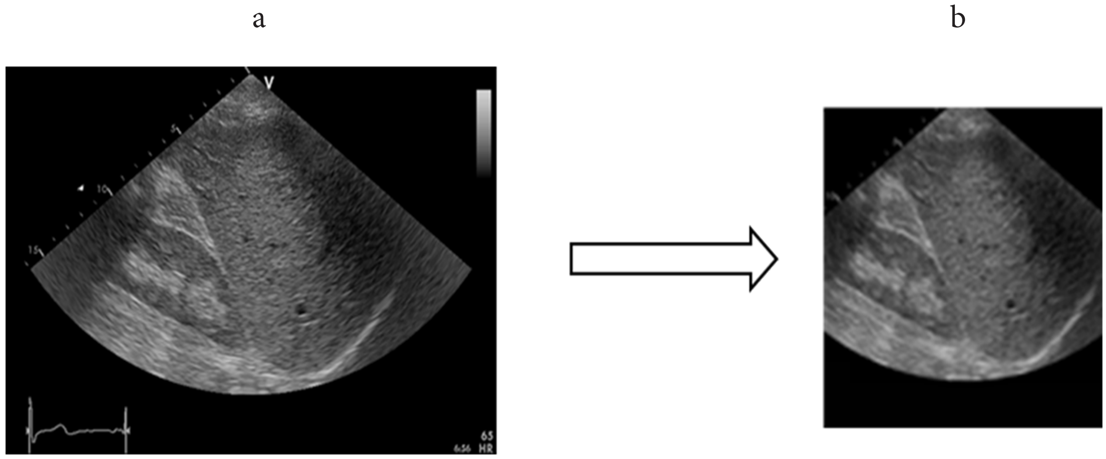

Figure 2: (**a**) sample the original image (**b**) the result after preprocessing

enhancement of model robustness and data heterogeneity, horizontal flipping was employed as a data augmentation strategy. This simple yet effective method involved creating mirrored versions of the images horizontally, effectively doubling the number of images for training. Unlike more complex augmentations like rotation or scaling, horizontal flipping maintains the anatomical consistency crucial for medical imaging, ensuring that no misleading features are introduced. Data augmentation plays a crucial role in machine learning, particularly when working with limited datasets. By augmenting the existing dataset, we can introduce variations and expand the training samples, causing the model to learn from a broader set of examples, leading to enhanced performance and better adaptation to real-world scenarios.

Feature extraction using the pretrained CNN

Feature extraction is an essential step in any pattern recognition task and particularly important for classifying NAFL disease in ultrasound images, due to the low quality of the images and the variability of fatty liver grades [20]. This study utilized seven popular pretrained CNN models trained on the ImageNet dataset to extract features: VGG19, MobileNet, Xception, Inception V3, ResNet-101, DenseNet-121, and EfficientNet-B7. To extract features using the pretrained CNN models, the last fully connected layers (output layers) were removed, and the last layer of the CNN structure was considered as the feature extractor. Specifically, the weights of the last layer were extracted in the network as image features; these weights are typically learned during the training process and capture high-level representations of the input data. By using the last layer of the CNN as the feature extractor, the learned features were used to classify NAFL disease in ultrasound images without requiring extensive training on the small NAFL dataset. Post-feature extraction, the data were normalized, and any features with zero variance were excised, resulting in ensuring that the extracted features were comparable across different models and that any noisy or uninformative features were removed.

The VGG model, a deep-learning network developed by Oxford University in 2014 [21], is known for its simplicity, practicality, and exceptional performance in image classification and target recognition tasks with CNN models. VGG19 with 193 million parameters is a VGG network with 19 layers. Another notable pre-trained CNN model, the inception network [22], introduced by Google in 2014, features 22 layers employing 1×1, 3×3, and 5×5 filter sizes for effective feature extraction, housing five million parameters. The innovation of the inception network is the use of multiple parallel convolutional branches with different kernel sizes, which capture features at different scales and reduce the computational cost. One year later, Google introduced the Inception V3 model from the inception family. Unlike previous versions of inception, this version replaced 5×5 filters with two 3×3 filters to reduce the number of required parameters and calculations without affecting the network performance [23]. Inception V3 has 43 layers. The ResNet, which was introduced in 2015, demonstrated outstanding performance with significantly fewer parameters than VGG. In this study, we utilized the ResNet-101 model, which consists of 101 layers, leading to a deeper and more complex network architecture that can extract intricate features and potentially improve the model's performance on the given task. Another CNN model used in this research was the Xception architecture, introduced in 2017 by Chollet [24]. The Xception architecture comprises 14 modules and 36 convolutional layers. Except for the initial and final modules, linear residual connections were used to connect the remaining modules. MobileNet introduced a new type of convolution called depth-wise separable convolution to reduce the number of parameters. In this architecture, the max-pooling layer was omitted, and stride 2 convolution was employed for the reduction of spatial dimensions. Although the size and computational cost of MobileNet are 1/30 the size of VGG16, it can achieve similar accuracy [25]. Another pre-trained CNN model used in this study was DenseNet with a 7×7 convolution layer after the input layer, followed by a 3×3 max-pooling layer. The grid has four dense blocks, each comprising at least six consecutive 1×1 convolution layers, followed by a 3×3 convolution layer. The DenseNet types have 121, 169, 201, or 264 layers, but all of these networks have four dense blocks, differing only in the number of consecutive convolution layers in each dense block [26]. In this study, DenseNet-121 we also used. In 2019, the EfficientNet Network was introduced by Tan and Le [27]. EfficientNet has significantly fewer parameters than those of other existing CNN models, with almost the same accuracy. Moreover, its accuracy and efficiency generally surpass those of the Inception-V3, VGG19, and MobileNet models. EfficientNet is based on the concept of hybrid scaling, which forms its main principle and comprises a series of eight models, namely EfficientNet-B0 to EfficientNet-B7, with varying parameter sizes ranging from 5.3 million to 66 million. In this study, the EfficientNet-B7 model was specifically employed, consisting of 66 million parameters. By utilizing this model, features were extracted from images with a high level of complexity and detail, potentially leading to improved performance and accuracy in our analysis.

Feature selection

Feature selection is a crucial step in the development of machine learning and pattern recognition models, as it helps to identify the most relevant features for the target classification task while alleviating computational complexity and mitigating overfitting. In this study, the minimum redundancy maximum relevance (mRMR) technique was utilized to select the most discriminative features obtained from several pre-trained CNN architectures. The mRMR method is a supervised feature selection approach that considers both the relevance and redundancy of the features. Specifically, the mRMR technique employs Spearman's rank correlation coefficient to assess the relevance of individual features to the target variable and their redundancy relative to other features. By ranking the features based on their relevance and redundancy scores, the mRMR method can identify a subset of features that are both highly relevant to the classification task and minimally redundant with each other [28, 29]. The mRMR method [28, 29], possesses the advantage of handling high-dimensional data, which proves particularly valuable in image classification, where the number of features can be exceedingly large. By effectively selecting a subset of the most pertinent and informative features, the mRMR method mitigates the issue of dimensionality while retaining critical information. Consequently, this capability enhances the efficiency of image classification tasks. In the present study, the mRMR method we implemented to the features extracted from seven popular pre-trained CNN models, including VGG19, MobileNet, Xception, Inception-V3, ResNet-101, DenseNet-121, and EfficientNet-B7.

Classification

This study used four classification methods: Linear Discriminant Analysis (LDA), SVM, Multilayer Perceptron (MLP) neural network, and RF. LDA is a well-known statistical method that maximizes the ratio of interclass to intra-class dispersion and exhibits good classification accuracy in compromising the processing time and processing. In other words, compared to other classifiers, the LDA classifier is relatively simple to implement and quick to train [30]. MLP is another machine-learning method inspired by the learning process and information processing in the human brain and can do complex analyses, such as those involving nonlinear models. MLP is a network of artificial neurons of feedforward layers and consists of input, output, and hidden layers [31, 32]. For this classification, two hidden layers were used. The ReLU function was used as the activation function for the hidden layers of the MLP classifier, while the SoftMax function was used for its output layer. Another classifier was SVM, selecting a hyperplane with a more reliable margin between two classes [33]. The SVM algorithm utilizes nonlinear mapping to transform the training data space into a higher dimension. In this new dimensional space, the SVM identifies a hyperplane that effectively separates instances of one class from those of other classes and acts as a decision boundary, classifying the data points into their respective classes. Through appropriate nonlinear mapping, even two-class datasets that are not linearly separable in the original space can be successfully separated using a hyperplane in the transformed dimension [34]. The SVM classifier from scikit-learn was utilized in this study, with the default RBF kernel function. The SVM is a binary classifier. A multiclass problem can be solved by combining a two-class SVM. The strategy used was one class versus the remaining classes to categorize each class. The output of the SVM binary classifiers was then combined to solve the multiclass problem [35]. The last classification method utilized in this study was RF, as a supervised learning classification and an ensemble learning method, can be used for pattern recognition and machine learning for high-accuracy classification [36]. This algorithm also combines multiple decision trees. The RF generates decision trees using a subset of randomly selected placement training datasets, showing decision trees in this classifier used a slightly different training dataset that is called bootstrapping. The size of the dataset used for each tree was the same as that of the training dataset. The final result of this classifier entails voting on those from the decision trees. Increasing the number of decision trees in this algorithm does not compromise the model's performance but does slow down its execution [37, 38].

Evaluation methods

In this study, the performance of our proposed method was evaluated using several metrics, including accuracy, sensitivity, specificity, and F1-score. Sensitivity, the true positive rate, is calculated as the proportion of correctly predicted positive cases out of all positive cases. Specificity, the true negative rate, refers to the proportion of correctly predicted negative cases out of all negative cases. Accuracy is defined as the ratio of correct predictions to the total number of predictions. The F1-score is calculated as the harmonic mean of precision and recall.

Proposed method

Figure 3 illustrates a block diagram of the proposed method for classifying the severity of NAFL disease from ultrasound images. After collecting liver ultrasound images, preprocessing was performed, including removing numbers, symbols, and redundant information. Images were initially 434×636 pixels but were cropped to 399×399 pixels. We addressed dataset correlations by employing repeated 5-fold cross-validation, ensuring comprehensive evaluation while mitigating potential biases from shared patient data. Horizontal flipping was also used for data augmentation, enhancing model robustness. The pretrained CNN models VGG19, MobileNet, Xception, Inception-V3, Resent-101, DanseNet-121, and EfficientNet-B7 were used to extract the features. In feature extraction using pretrained CNN models, we removed the last fully connected layer (output layer) and considered the remaining CNN structure as the feature extractor. We extracted the weights of the last layer of the CNN model as image features. We then normalized these features and removed those with zero variance. Next, mRMR was used for feature selection. The best features of the pretrained CNN models were selected and applied to the classifier input. Finally, images were classified into four groups: no steatosis, mild steatosis, moderate, and severe steatosis levels from ultrasound images using different classifiers: LDA, MLP, multiclass SVM, and RF.

Results

The results of NAFL disease classification into four classes with different pretrained CNN models as the feature extractor, mRMR method as the feature selection, and all classifiers (LDA, MLP, SVM, and RF) are listed in Table 1. The mRMR method can select a small subset of highly informative features from each CNN model, with an average reduction of more than 90% in the number of features, which not only improved computational efficiency but also enhanced the accuracy and generalization performance of the subsequent classification models. The highest accuracy values among the different

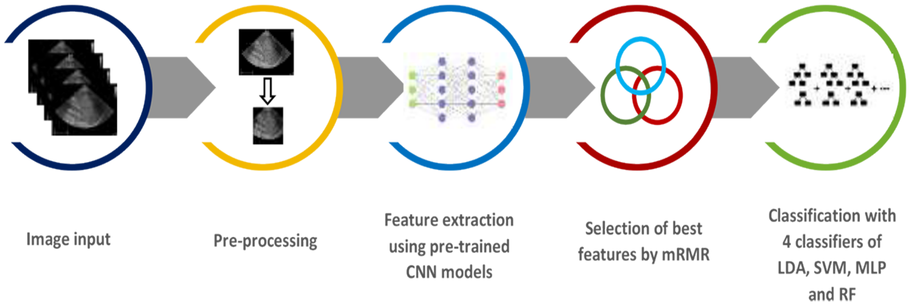
Image input
Pre-processing
Feature extraction using pre-trained CNN models
Selection of best features by mRMR
Classification with 4 classifiers of LDA, SVM, MLP and RF

Figure 3: Block diagram illustrating the five-step process implemented to classify the liver fat quantity from ultrasound images. These steps include Image input, Pre-processing, Feature extraction using pre-trained CNN models, Selection of best features by mRMR, and Classification with 4 classifiers: LDA (Linear Discriminant Analysis), SVM (Support Vector Machine), MLP (Multi-Layer Perceptron), and RF (Random Forest).

Table 1: Accuracy of classifying NAFL disease into four grades using different pretrained CNN models as the feature extractor and using all selected classifiers (LDA, MLP, SVM, and RF). The accuracy is the percentage of images that are correctly classified by the model. (NAFL: Nonalcoholic Fatty Liver, LDA: Linear Discriminant Analysis; MLP: Multilayer Perceptron; SVM: Support Vector Machine; RF: Random Forest; Pre-trained CNN models: Convolutional Neural Networks that are pre-trained on large datasets and can be used as feature extractors.

Pre-trained CNN Models	LDA	MLP	SVM	RF
Mobile Net	87.26±2.51	76.48±6.51	93.78±1.48	91.14±2.57
VGG 19	74.98±2.72	74.59±7.04	79.76± 2.81	85.70±2.05
Xception	78.99±2.69	76.12±5.87	92.58±2.44	87.64±2.32
DensNet121	76.50± 2.67	79.28±6.02	84.12±2.69	84.20±2.70
Inception V3	74.87±3.04	77.56±5.82	90.87±2.13	86.67±2.32
ResNet101	86.93±2.62	82.76±3.96	94.03±1.76	94.36±1.98
EfficientNet B7	88.48±2.26	93.15±2.46	95.47±1.81	96.83±1.59

LDA: Linear Discriminant Analysis, MLP: Multilayer Perceptron, SVM: Support Vector Machine, RF: Random Forest pretrained CNN models as feature extraction methods were obtained for EfficientNet-B7 for different classifiers. The accuracies of the LDA, MLP, SVM, and RF classifiers for extracted features using EfficientNet-B7 were 88.48, 93.15, 95.47, and 96.83%, respectively. The accuracy of all classifiers in EfficientNet-B7 was significantly higher than that of the other pretrained CNN models. In addition, the highest accuracy among the different classification methods was obtained for the RF classifier. Therefore, the pretrained CNN model named EfficientNet-B7 as the feature extraction method, along with the feature selection using the mRMR and RF classifier, has the highest accuracy of 96.83%. Table 2 presents the results of classifying NAFL disease into four categories (healthy liver, low, medium, and high fatty liver levels) using Efficient Net-B7 as the feature extractor and four classifiers (LDA, MLP, SVM, and RF), evaluated by accuracy, sensitivity, specificity, and F1-score. Figure 4 shows the ratio of the number of features selected by the mRMR method for feature selection from EfficientNet-B7 as a feature extraction method to the accuracy obtained by RF classification. The best 40 features of the EfficientNet-B7 model were selected and applied to the classifier input. In addition, Figure 5 shows the ratio of the number of decision trees in the RF classifier to the obtained accuracy. The RF method with 90 decision trees achieved the highest accuracy. Table 3 compares the results of the proposed

Table 2: Results of classifying NAFL disease into four grades (no steatosis, mild steatosis, moderate, and severe steatosis) with the EfficientNet B7 as the feature extractor and all classifiers (LDA, MLP, SVM, RF) based on accuracy, sensitivity, specificity, and F1-scores (NAFL: Nonalcoholic Fatty Liver, LDA: Linear Discriminant Analysis, MLP: Multilayer Perceptron, SVM: Support Vector Machine, RF: Random Forest)

		Accuracy	Sensitivity	Specificity	F1 score
SVM	No steatosis	93.33±3.59	94.66±5.47	98.67±1.33	95.73±2.57
	Mild steatosis	78.79±2.88	96.22±3.66	95.86±2.14	94.59±2.34
	Moderate steatosis	88.48±5.54	93.29±14.22	99.30±0.78	94.42±3.28
	Severe steatosis	92.12±3.39	97.55±11.15	99.52±0.63	97.66±2.28
LDA	No steatosis	92.25±1.81	92.16±3.76	96.49±1.98	92.16±2.63
	Mild steatosis	83.09±3.05	87.10±4.18	91.26±2.54	86.40±3.59
	Moderate steatosis	91.46±2.55	80.77±6.01	97.84±1.31	84.00±6.48
	Severe steatosis	94.81±1.52	92.10±4.99	97.16±1.62	90.57±3.77
MLP	No steatosis	96.56±1.92	93.36±4.07	98.02±1.81	94.39±3.02
	Mild steatosis	94.49±2.18	91.60±4.26	96.22±1.95	92.35±2.96
	Moderate steatosis	96.57±1.65	95.62±3.91	96.73±1.75	89.08±4.77
	Severe steatosis	98.68±0.98	94.25±4.45	99.64±0.53	96.17±2.70
RF	No steatosis	98.47±1.22	97.20±3.18	99.06±0.93	97.47±2.03
	Mild steatosis	97.32±1.33	98.84±1.41	96.49±1.96	96.39±1.79
	Moderate steatosis	98.69±0.76	92.34±4.43	99.78±0.40	95.32±2.76
	Severe steatosis	99.18±0.67	96.15±3.16	99.89±0.29	97.78±1.77

SVM: Support Vector Machine, LDA: Linear Discriminant Analysis, MLP: Multilayer Perceptron, RF: Random Forest

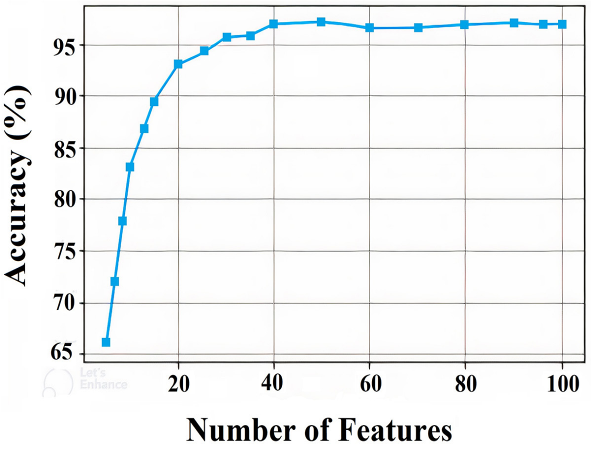

Figure 4: Ratio of the number of features selected by the mRMR method from the EfficientNet B7 model to the accuracy of the data obtained via RF (Random Forest) classification. The 40 best features of the EfficientNet B7 model were selected and applied to the classifier input

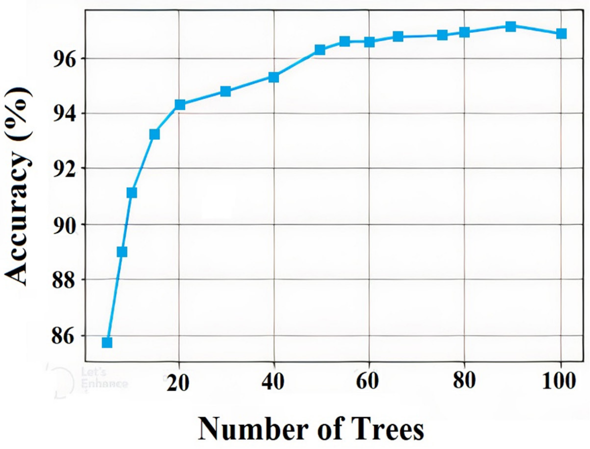

Figure 5: Ratio of the number of decision trees in the RF (Random Forest) classifier to the accuracy obtained. The highest accuracy was obtained with 90 decision trees

Table 3: Comparison of the proposed method with other recent approaches in terms of the classification of NAFL (Nonalcoholic Fatty Liver) disease. The table shows the accuracy of classifying NAFL disease into four grades (healthy, low, moderate, and high) using different classification methods and feature extraction methods. The accuracy is the percentage of images that are correctly classified by the method.

Author's Name	Number of Cases	Feature extraction	Classification method	Accuracy (%)
Wan and Zhou [8]	590	32 wavelet packet transform-based features	SVM	85.8
Acharya et al. [9]	100	Texture, HOS, DWT	Decision Tree	93.3
Kuppili et al. [10]	63	GLCM	SVM	86.42
Naderi et al. [11]	550	GLCM	Adaboost	92.72
Hassan et al. [12]	110	SSAE	Softmax	98
Saba et al. [13]	124	Harlick, Gupta, Fourier, basic geometric, DCT, Gabor	back-propagation neural network	97.6
Byra et al. [17]	550	Inception-ResNet-v2	SVM	96.3
Constantinescu et al. [18]	629	Inception-v3, VGG19	Softmax	93.23
Proposed method	**550**	**EfficientNet-B7**	**RF**	**96.83**

SVM: Support Vector Machine, HOS: Higher Order Statistics, DWT: Discrete Wavelet Transform, GLCM: Gray-Level Co-Occurrence Matrix, SSAE: Stacked Sparse Autoencoder, DCT: Discrete Cosine Transform, RF: Random Forest method with other recent approaches for grading NAFLD disease. Table 3 compares the number of datasets used, feature extraction techniques, classification methods, and accuracies obtained.

Discussion

This study presented an automated medical diagnostic system using advanced AI techniques to grade NAFLD disease severity from ultrasound images. A total of 550 ultrasound images were used in the current study, which was publicly available from the Department of Internal Medicine, Medical University of Warsaw, Poland. It is essential to emphasize that this dataset was obtained through a collaborative effort involving pathologists and physicians to ensure both its quality and relevance within the medical domain. Features were extracted from the ultrasound images leveraging the pretrained CNN architectures VGG19, MobileNet, Xception, Inception-V3, ResNet-101, DenseNet-121, and EfficientNet-B7. Subsequently, 40 of the best features were selected using the mRMR method. Finally, the images were classified into four classes using LDA, multiclass SVM, MLP, and RF classifiers. While the proposed automatic model achieved a high accuracy of 96.83% in the detection of NAFLD grade using EfficientNet-B7 and RF classifier, it is important to note that SVM and RF also demonstrated competitive performance in this study. SVM achieved an accuracy of 95.47% in this task. Therefore, future studies could explore the use of different feature extraction methods and classification algorithms to further improve the accuracy and generalizability of automated NAFLD grading systems.

In this study, EfficientNet-B7 with 66 million parameters achieved the best results among the powerful pretrained CNN models. The EfficientNet-B7 model generally has higher accuracy and performance than other pretrained CNN models. The main idea behind EfficientNet is a hybrid scaling method that uniformly scales the base network based on three factors: network depth (number of layers), network width (number of nodes per layer), and image resolution (input image size) [27]. The main block of this network includes MBConv, to which squeeze-and-excitation optimization has been added. MBConv is similar to the inverted residual blocks in MobileNet-V2. Swish, a novel activation function, is applied to each layer to preserve more information than ReLU [26].

Among the evaluated classification methods, the RF classifier, an ensemble learning technique, demonstrated the most superior results and performance and can learn complex patterns and perform pattern recognition and ML for high-precision classification [36]. The RF generates its decision using the bootstrapping technique and achieves higher accuracy by increasing the number of decision trees. The final result of this classifier is the result of voting on the decision trees. RF offers a distinct advantage in its robustness to outliers and effectiveness in handling nonlinear data.

Briefly, our study has several contributions that distinguish it from previous studies in the literature. Firstly, we applied a larger number of powerful pretrained CNN models for feature extraction and selected the 40 most relevant features for NAFLD grading through a feature selection process. We also evaluated multiple classification algorithms and found that the RF classifier achieved the best performance. Therefore, the contribution of our study lies in the selection and combination of specific powerful pretrained CNN models and classifiers to achieve a high level of accuracy in classifying NAFLD disease grades from ultrasound images. In other words, we develop a novel hybrid model based on deep transfer learning using powerful pretrained CNN models and various machine-learning methods. Our specific selection and combination of models is what sets our approach apart and enables us to achieve a remarkable accuracy of 96.83%. From another point of view, we use the capabilities of powerful deep learning methods as well as traditional ML methods simultaneously. This approach offers several benefits and enhances the model's accuracy. Finally, we conducted a thorough comparison with other state-of-the-art approaches in recent years in terms of accuracies obtained.

Due to the limited dataset size, the omnipresent challenge of overfitting necessitates a thoughtful approach. In our study, we have judiciously employed a tailored set of strategies to combat overfitting. Data augmentation was a key element, where we performed horizontal flipping to effectively double the number of training images, thereby introducing crucial variability into the dataset. Another pivotal strategy was transferring learning, wherein we harnessed the power of pretrained CNNs. Leveraging these models as feature extractors enabled us to capture pertinent image features that are highly relevant to our specific task, effectively mitigating overfitting by leveraging the rich representations they had already learned from extensive datasets. In conjunction with these strategies, we also employed feature selection using mRMR, aiding in the identification of the most informative features for our task. For enhanced robustness, we employed an ensemble of traditional classifiers, including RF. This approach significantly reduced the risk of overfitting and contributed to improved overall model performance due to the ensemble's inherent advantages.

The proposed method offers several key advantages and presents some limitations. This approach is completely automatic without any operators. Additionally, it can be used in rural and remote areas lacking expert radiologists. Radiologists can use this approach to aid their diagnoses. A principal constraint in model development stemmed from the limited sample size inherent to medical imaging datasets, with only 550 images obtained from 55 subjects available for training and evaluation. Future explorations could involve investigating alternative ensemble learning methods to potentially broaden the method's applicability across diverse settings. External validation across heterogeneous ultrasound platforms and clinical environments would further fortify generalizability for widespread NAFLD grading. Data augmentation techniques can also be used to expand labeled training data, curbing overfitting and enhancing model generalizability. The limitations of the present study are listed, as follows: 1) the limited size of the training dataset, affecting the robustness and generalizability of our classification model and 2) all ultrasound images were acquired using a single ultrasound machine, limiting the applicability of results to other imaging systems.

Conclusion

In this study, an automated diagnostic system was developed and validated for the grading of NAFLD in obese patients, who had undergone bariatric surgery. The proposed system used advanced machine learning techniques, including a state-of-the-art EfficientNet-B7 pretrained CNN model, the mRMR feature selection algorithm, and an RF ensemble classifier. Utilizing a dataset of 550 ultrasound images, the proposed system achieved 96.83% accuracy in differentiating normal livers from those with mild, moderate, or severe steatosis. The proposed method can improve the accuracy and accessibility of NAFLD diagnoses, particularly in regions, in which there is a shortage of experienced medical professionals. In subsequent research endeavors, it is needed to explore how advanced feature extraction methodologies, such as texture analysis and vision transformer features, can augment the efficacy of the proposed automated diagnostic framework. Furthermore, the current dataset must be completed with a large collection of clinical data and ultrasound imagery from different diagnostic equipment, leading to cross-validate the system more robustly and, importantly, to speed up its adoption in clinical practice with more empirical evidence.

Authors' Contribution

A. Shalbaf was the supervisor of the project and conceived the original idea. AR. Naderi Yaghouti carried out the study, simulation, and results, and wrote the manuscript. Both authors read, modified, and approved the final version of the manuscript.

Ethical Approval

The study was approved by the Ethical Committee at the Shahid Beheshti University of Medical Sciences (IR.SBMU.MSP.REC.1402.117).

Informed Consent

All patients gave informed consent for echocardiography and abdominal ultrasound examinatio.

Conflict of Interest

None

References

1. Bharath R, Rajalakshmi P. Deep scattering convolution network based features for ultrasonic fatty liver tissue characterization. 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); Jeju, Korea (South): IEEE; 2017. p. 1982-5.
2. Kim KB, Kim CW. Quantification of Hepatorenal Index for Computer-Aided Fatty Liver Classification with Self-Organizing Map and Fuzzy Stretching from Ultrasonography. *Biomed Res Int.* 2015;**2015**:535894. doi: 10.1155/2015/535894. PubMed PMID: 26247023. PubMed PMCID: PMC4515496.
3. Alizadeh Sani Z, Shalbaf A, Behnam H, Shalbaf R. Automatic computation of left ventricular volume changes over a cardiac cycle from echocardiography images by nonlinear dimensionality reduction. *J Digit Imaging.* 2015;**28**(1):91-8. doi: 10.1007/s10278-014-9722-z. PubMed PMID: 25059548. PubMed PMCID: PMC4305052.
4. Shalbaf A, AlizadehSani Z, Behnam H. Echocardiography without electrocardiogram using nonlinear dimensionality reduction methods. *J Med Ultrason.* 2015;**42**(2):137-49. doi: 10.1007/s10396-014-0588-y. PubMed PMID: 26576567.
5. Das A, Connell M, Khetarpal S. Digital image analysis of ultrasound images using machine learning to diagnose pediatric nonalcoholic fatty liver disease. *Clin Imaging.* 2021;**77**:62-8. doi: 10.1016/j.clinimag.2021.02.038. PubMed PMID: 33647632.
6. Ribeiro R, Sanches J. Fatty liver characterization and classification by ultrasound. In: Pattern Recognition and Image Analysis; Berlin, Heidelberg: Springer Berlin Heidelberg; 2009. p. 354-61.
7. Kyriacou E, Pavlopoulos S, Konnis G, Koutsouris D, Zoumpoulis P, Theotokas L. Computer assisted characterization of diffused liver disease using image texture analysis techniques on B-scan images. Nuclear Science Symposium Conference Record; Albuquerque, NM, USA: IEEE; 1997.
8. Wan J, Zhou S. Features extraction based on wavelet packet transform for B-mode ultrasound liver images. 3rd International Congress on Image and Signal Processing; Yantai, China: IEEE; 2010.
9. Acharya UR, Sree SV, Ribeiro R, Krishnamurthi G, Marinho RT, Sanches J, Suri JS. Data mining framework for fatty liver disease classification in ultrasound: a hybrid feature extraction paradigm. *Med Phys.* 2012;**39**(7):4255-64. doi: 10.1118/1.4725759. PubMed PMID: 22830759.
10. Kuppili V, Biswas M, Sreekumar A, Suri HS, Saba L, Edla DR, et al. Extreme Learning Machine Framework for Risk Stratification of Fatty Liver Disease Using Ultrasound Tissue Characterization. *J Med Syst.* 2017;**41**(10):152. doi: 10.1007/s10916-017-0797-1. PubMed PMID: 28836045.
11. Naderi Yaghouti AR, Shalbaf A, Maghsoudi A. Automatic classification of Non-alcoholic fatty liver using texture features from ultrasound images. *Tehran-Univ-Med-J.* 2021;**79**(1):10-7.
12. Hassan TM, Elmogy M, Sallam ES. Diagnosis of focal liver diseases based on deep learning technique for ultrasound images. *Arabian Journal for Science and Engineering.* 2017;**42**(8):3127-40. doi: 10.1007/s13369-016-2387-9.
13. Saba L, Dey N, Ashour AS, Samanta S, Nath SS, Chakraborty S, et al. Automated stratification of liver disease in ultrasound: An online accurate feature classification paradigm. *Comput Methods Programs Biomed.* 2016;**130**:118-34. doi: 10.1016/j.cmpb.2016.03.016. PubMed PMID: 27208527.
14. Song T, Yu X, Yu S, Ren Z, Qu Y. Feature extraction processing method of medical image fusion based on neural network algorithm. *Complexity.* 2021;**2021**:1-10. doi: 10.1155/2021/7523513.
15. Reddy DS, Bharath R, Rajalakshmi P. Classification of nonalcoholic fatty liver texture using convolution neural networks. 20th International Conference on

e-Health Networking, Applications and Services (Healthcom); Ostrava, Czech Republic: IEEE; 2018. p. 1-5.

16. Salehi AW, Khan S, Gupta G, Alabduallah BI, Almjally A, Alsolai H, et al. A Study of CNN and Transfer Learning in Medical Imaging: Advantages, Challenges, Future Scope. *Sustainability.* 2023;**15**(7):5930. doi: 10.3390/su15075930.
17. Byra M, Styczynski G, Szmigielski C, Kalinowski P, Michałowski , Paluszkiewicz R, et al. Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images. *Int J Comput Assist Radiol Surg.* 2018;**13**(12):1895-903. doi: 10.1007/s11548-018-1843-2. PubMed PMID: 30094778. PubMed PMCID: PMC6223753.
18. Constantinescu EC, Udriștoiu AL, Udriștoiu ȘC, Iacob AV, Gruionu LG, Gruionu G, et al. Transfer learning with pre-trained deep convolutional neural networks for the automatic assessment of liver steatosis in ultrasound images. *Medical Ultrasonography*. 2021;**23**(2):135-9. doi: 10.11152/mu-2746.
19. Zenoob. Dataset of B-mode fatty liver ultrasound images. 2018. Available from: https://zenodo.org/records/1009146.
20. Wang D, Fang Y, Hu B, Cao H. B-scan image feature extraction of fatty liver. Sixth International Conference on Internet Computing for Science and Engineering; Zhengzhou, China: IEEE; 2012.
21. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [Internet]. arXiv [Preprint]. 2014 [cited 2014 Sep 4]. Available from: https://arxiv.org/abs/1409.1556.
22. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition; CVPR; 2015. p: 1-9.
23. Adegun A, Viriri S. Deep learning techniques for skin lesion analysis and melanoma cancer detection: a survey of state-of-the-art. *Artificial Intelligence Review.* 2021;**54**:811-41. doi: 10.1007/s10462-020-09865-y.
24. Chollet F. Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition; CVPR; 2017. p. 1251-8.
25. Arcos-García Á, Álvarez-García JA, Soria-Morillo LM. Evaluation of deep neural networks for traffic sign detection systems. *Neurocomputing*. 2018;**316**:332-44. doi: 10.1016/j.neucom.2018.08.009.
26. Nugroho BA. An aggregate method for thorax diseases classification. *Sci Rep*. 2021;**11**(1):3242. doi: 10.1038/s41598-021-81765-9. PubMed PMID: 33547338. PubMed PMCID: PMC7864910.
27. Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning; PMLR; 2019. p. 6105-14.
28. Azadi H, Akbarzadeh-T MR, Kobravi HR, Sarcheshmeh AN, Shahsavanpour N, Asgharzade MR. Presentation of a new gender dependent feature selection approach for diagnosis of Parkinson disease using speech signal processing. International congress on technology, communication and knowledge (ICTCK); Mashhad, Iran: IEEE; 2015. p. 371-5.
29. Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell.* 2005;**27**(8):1226-38. doi: 10.1109/TPAMI.2005.159. PubMed PMID: 16119262.
30. Torkkola K. Linear discriminant analysis in document classification. In Proceedings of the IEEE International Conference on Data Mining: Workshop on Text Mining; San Jose, CA, USA: IEEE; 2001.
31. Jakkula V. Tutorial on support vector machine (svm). School of EECS, Washington State University; 2006.
32. Vapnik V. The nature of statistical learning theory. 2nd edn. Berlin: Springer; 1999.
33. Cortes C, Vapnik V. Support-vector networks. *Machine learning.* 1995;**20**:273-97.
34. Berzal F, Matín N. Data mining: concepts and techniques by Jiawei Han and Micheline Kamber. *SIGMOD Rec.* 2002;**31**(2):66-8. doi: 10.1145/565117.565130.
35. Huang GB, Zhou H, Ding X, Zhang R. Extreme learning machine for regression and multiclass classification. *IEEE Trans Syst Man Cybern B Cybern*. 2012;**42**(2):513-29. doi: 10.1109/TSMCB.2011.2168604. PubMed PMID: 21984515.
36. Breiman L. Random forests. *Machine learning.* 2001;**45**:5-32.
37. Hartshorn S. Machine learning with random forests and decision trees: A Visual guide for beginners. Kindle edition; 2016.
38. Alam MS, Vuong ST. Random forest classification for detecting android malware. International conference on green computing and communications and IEEE Internet of Things and IEEE cyber, physical and social computing; Beijing, China: IEEE; 2013. p. 663-9.