Journal of Information Systems Engineering and Management
2025, 10(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Fatty Liver Disease Stage Classification Using Ultra Sound Images by ResNet Deep Learning Model

M. Swapna[1], Dr. B. Sujatha[2]

[1]Research scholar, Department of Computer Science and Engineering, Osmania University, Hyderabad, Telangana, India – 500007. maragoniswapna761@gmail.com

[2]Assistant professor, Department of Computer Science and Engineering, Osmania University, Hyderabad, Telangana, India - 500007. sujatabanothu@gmail.com

ARTICLE INFO

Received: 12 Oct 2024
Revised: 11 Dec 2024
Accepted: 24 Dec 2024

ABSTRACT

Fatty liver disease (FLD) is a growing global health concern, with its prevalence on the rise due to factors such as sedentary lifestyles and poor dietary choices. The non-invasive classification of FLD stages using ultrasound images is a vital diagnostic endeavor that aids in early detection and intervention, thus mitigating the potential for severe liver-related complications. In this research, a ResNet-based deep learning framework is proposed for ultrasound images based automatic staging of FLD. The model uses features obtained through ultrasound scans to classify the stage of FLD, i.e Grade 1, Grade 2 and Grade 3. They are trained, validated and tested on a dataset of standardized ultrasound images. This indicates the model's ability to classify FLD at a stage correctly and proposes a potential avenue for improved efficacy in diagnostics in hepatology.

Keywords: Fatty Liver Disease, Ultrasound Images, Deep Learning, ResNet, Classification.

INTRODUCTION

The condition known as fatty liver disease, which is characterized by the buildup of extra fat within the liver cells, has become a global health concern in recent decades. This syndrome encompasses a variety of liver diseases, from the relatively benign non-alcoholic fatty liver (NAFL) [1] to more severe forms like non-alcoholic steatohepatitis (NASH) [2], which, if ignored, can lead to cirrhosis and liver failure. Alongside the rise in obesity along with metabolic syndrome, the threat of fatty liver illness has been steadily rising. Therefore, successful intervention in the healing process of fatty liver disease requires early and accurate detection.

The severity of fatty liver disease cannot be overstated since it not only causes morbidity related to the liver but is also related to an increased threat of cardiovascular disease, diabetes, and hepatocellular cancer. Therefore, in order to make educated treatment choices and enhance patient outcomes, it is essential to detect and precisely identify the stage of fatty liver disease in a timely way [3].

Ultrasound imaging has established itself as a useful instrument for evaluating liver health [4], making it one of the many diagnostic modalities that are now accessible. Because it is not intrusive, is widely available, and has a low cost, it is a good option for initial screening as well as monitoring. The ability to view and measure hepatic fat accumulation is made possible by ultrasonography for medical professionals [5]. However, correct interpretation of ultrasound pictures may be difficult at times, and it depends significantly on the experience of the physician doing the interpreting.

In this regard, the application of deep learning techniques to the ultrasound image-based diagnosis along with staging of fatty liver illness [6] has gained significant traction. Image analysis and pattern identification are two areas in which deep learning models [7], and in particular convolutional neural networks (CNNs) [8], have shown very impressive capabilities. These models have the potential to completely transform the sector by automating the categorization of the phases of fatty liver disease. As a result, the subjectivity that is associated with human interpretation will be significantly reduced, and prompt intervention will be made possible.

This study investigates the critical function that deep learning plays in the categorization of the phases of fatty liver disease using ultrasound pictures. This article dives into the severity of the condition, discusses the usefulness of ultrasound imaging, and highlights the essential need for automated, precise, and objective diagnostic methods in order to tackle this growing public health crisis. Deep learning advancements could lead to more accurate and timely diagnoses of fatty liver illness, which would be a significant step toward earlier treatment and better patient care.

LITERATURE

Hui Che et al [9] proposed a new deep learning model for nonalcoholic fatty liver disease classification using US data. The authors design directed multi-scale residual CNN with multiple features to gather the feature that is supports for each receptive field. The B-mode ultrasound images combine with the corresponding radial symmetry modified images and local phase filtered images as the multi-feature inputs of the network. Fusion processes are studied to improve forecast accuracy.

Amit Das et al [10] Create a classification model that is based on machine learning (ML) that is capable of distinguishing NAFLD from healthy liver tissue. Evaluate the effectiveness of this model in comparison to indices that are based on pixel intensity. In order to construct an image database, de-identified ultrasound pictures of the liver that were taken as part of a cross-sectional research assessing the prevalence of pediatric NAFLD were employed. Image texture features were extracted from one representative region of interest (ROI) of the ultrasound images of subjects with verified NAFLD and subjects with normal livers using ImageJ and MAZDA image processing and analysis tools. The ROI was selected from ultrasound pictures of participants with NAFLD and those with normal livers. A number of machine learning classification techniques were investigated.

Wen Cao MD et al [11] used the established criteria and ultrasound scoring system for NAFLD, 240 subjects were enrolled and classified into four groups (Normal, mild, moderate, and severe NAFLD groups). The two-dimensional hepatic chemistry was studied using three image-processing techniques—the envelope signal, grayscale signal, and deep-learning index. The return values for each of them was different ranging from 0–4, 0–255 and 0–65,535. Comparison between the four groups were made to determine the best image-processing strategy, plotting receiver operating characteristic curves and comparing area under the curve (AUC) values.

Pezhman Pasyar et al [12] presented a novel deep classifier made up of deep CNNs that have already been trained. Fully connected networks (FCNs) are utilized in conjunction with other networks, such as ResNeXt, ResNet18, ResNet34, ResNet50, along with AlexNet. Significant classification information could be obtained by extracting deep features using transfer learning. A fully convolutional network (FCN) can categorize images into three disease states: cirrhosis, liver hepatitis, along with normal liver. Classifiers that could distinguish across two classes (normal/cirrhosis, normal/hepatitis, as well as cirrhosis/hepatitis) along with three classes (normal/cirrhosis/hepatitis) were trained using the liver images. Given that two-class classifiers have shown superior performance than three-class classifiers, it is advised to develop a hybrid classifier that incorporates the weighted probability of the classes generated by each separate classifier. A majority vote is then used to choose the class with the highest score.

Trong N. Nguyen et al [13] examined how to assess hepatic steatosis in a live rabbit framework for fatty liver using attenuation and the B-mode ultrasonic scanning technique (BSC). A high-fat diet was given to a group of rabbits for 0, 1, 2, 3, or 6 weeks at different times. There were three rabbits in each diet group, for a total sample size of fifteen rabbits. In this work, radio frequency (RF) backscattered data from live rabbits was collected using a SonixOne scanner and an array transducer (L9-4) running at a center frequency of 4.5 MHz. The estimate of the average attenuation as well as BSC for each individual rabbit was made easier by the use of RF signals. The Backscattering Strength Coefficient (BSC) was parameterized using two different approaches. In particular, a spherical Gaussian model was used to calculate the effective scatterer diameter along with the efficient acoustic concentration. Additionally, principal component analysis (PCA) was used to build a model-free technique. The two main components that explained 96% of the variance in the converted data were found using principal

component analysis (PCA) on the balanced scorecards (BSCs). After that, a support vector machine (SVM) architecture was trained using these elements as input characteristics for classification.

Haley Schoenberger et al [14] conducted a retrospective cohort analysis on a cohort of individuals diagnosed with cirrhosis who had undergone ultrasonography examinations at two major healthcare systems from July 2016 to July 2019. Radiologists assigned LI-RADS Visualization Score (A, B, C) for the adequacy of each exam. This study aimed to assess changes in visualization across the various ultrasound exams in a cohort of patients. We performed a multivariable logistic regression analysis to identify the features associated with inadequate ultrasound image quality, which were scores B or C.

Monica Lupsor-Platon et al [15] evaluated the diagnostic performance of ultrasonic elastography for non-invasive evaluation of nonalcoholic fatty liver disease and nonalcoholic fatty liver disease-related HCC. Liver elastography adds a new axis to the original ultrasound examination (both classic ultrasound and elastography) through liver stiffness measurements. Although the two most effective elastographic methods of assessing liver fibrosis in NAFLD are, vibration controlled transient elastography (VCTE) combined with 2D-Shear wave elastography (2D-SWE), VCTE can also measure steatosis with the controlled attenuation parameter. Two of the most popular elastographic methods for quantitatively describing focal liver lesions (FLLs), with an emphasis on hepatocellular carcinoma, are point shear wave elastography (pSWE) and two-dimensional shear wave elastography (2D-SWE). For the reasons listed in the research, elastography shouldn't be utilized in place of FLL biopsy due to the overlap in stiffness readings.

Jeongin Yoo et al [16] employed a commercial ultrasound equipment, the purpose of this study was to test both the intra-observer and inter-observer repeatability of shear wave dispersion imaging (SWDI) for the assessment of non-alcoholic fatty liver disease in asymptomatic volunteers. The institutional review board gave its permission to this prospective trial, and the authors made sure to get the patients' informed consent before moving forward. One radiologist carried out two sessions of SWDI on the patients that made up group I (n = 71), all of whom were suspected or reported to have non-alcoholic fatty liver disease. This was done in order to evaluate the repeatability of the test within the same observer. Within the second group, which consisted of symptom-free volunteers (n = 19), three separate sessions led by three different radiologists were carried out.

PROPOSED METHOD

Residual Network (ResNet) is type of deep learning architecture used for image classification. Because of its remarkable performance and the capacity to train extremely deep networks without the problem of vanishing gradient, this architecture is widely used in computer vision applications, notably in image classification, object recognition, and image segmentation. The most significant advancement made to the ResNet design was the incorporation of residual blocks. The vanishing gradient issue may cause a typical CNN's performance to suffer when more layers are added to an existing network. This can lead to a decline in the network's overall accuracy. When additional layers are added, it becomes a lot more difficult to train the network such that it works well.

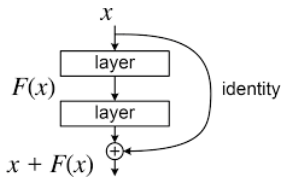

Figure 1: residual block

This problem is solved by incorporating skip connections or shortcuts into the network, which gives it the ability to learn residual functions. Residual blocks. By using these skip connections, the network is able to concentrate on learning the difference, or residual, between the input and the output rather than attempting to learn the complete transformation. This frees up the network to better perform its

intended function of learning. Because of this, training extremely deep networks is simplified, which ultimately results in enhanced performance.

In contrast to the conventional network models, the equation (1) can be used to characterize the residual connection that exists inside residual blocks.

$$x_{l+1} = x_l + f(x_l, w_l) \tag{1}$$

Where x_{l+1} represents the leftover block of the $(l + 1)^{th}$ layer. The leftover block of the l^{th} layer is denoted by x_l. The residual section of the block is indicated by $f(x_l, w_l)$. When the feature map contours in x_{l+1} and x_l are present. Due to the difference between x_{l+1} and x_l, the network needs the dimension procedure. The residual connection block has the following definition:

$$x_{l+1} = h(x_l) + f(x_l, w_l) \tag{2}$$

In this case, the 1X1 convolution operation represents $h(x_l)$. Two block groups are created from the remaining links in this study.

The ResNet-50 neural network requires an input image that is 224 by 224 pixels and has three color channels. The structure of ResNet-50 consists of four stages, each of which has a number of residual blocks. The quantity of residual blocks that are present in each stage is variable, with a greater number of blocks present in later stages.

Typically, there are three convolutional layers included inside each residual block that is part of a stage. The convolutional layers that are included inside a block make use of 1x1, 3x3, and 1x1 filters. The 1x1 filters are used to decrease or expand the number of channels, respectively. A batch normalization step is performed after every convolutional layer, and a ReLU activation function is then used. Each residual block's output consists of the sum of the block's inputs and the residual, or third convolutional layer's output. Stated differently, the residual is the result of every residual block.

Because of these remaining chunks of information, the network can learn ever-more complex features in deeper layers. To reduce the spatial dimensions to 1 x 1, a global average pooling layer is used after the final residual block is applied. A completely connected layer with a thousand units then receives the output. The quantity of classes in the ImageNet dataset, which served as the main dataset for ResNet-50 training, is equivalent to this amount of units.

Input layer
Convolution layer
Max pooling layer
Block 1
Block 2
Block 3
Block 1
Block 2 X 2
Block 3
Block 1
Block 2 X 4
Block 3
Block 1
Block 2 X 2
Batch Normalization layer
ReLu layer
Max pooling layer
Dense layer
Dropout layer
Dense layer
Dropout layer
Dense layer

Figure 2: Modified Residual network:

The portion of the network that takes the data you supply—typically an image—is known as the input layer. A set of learnable filters convolves, or slides, across the input image to extract features. This process is known as a convolution layer. Each filter computes dot products with a specific region of the

input image to produce a feature map. The feature map is then created using these dot products. The max pooling layer selects the largest value from a tiny portion of the feature map to downsample the feature maps following the convolution step. This comes after the phase of convolution. This helps to decrease the spatial dimensions and extract the elements that are most important.

Blocks 1, 2 and 3 are residual blocks made up of convolutional layers, batch normalization, and activation functions like ReLU. The batch normalization layer is responsible for standardizing the output of the layer that came before it. By lowering the amount of internal covariate change, it makes training more stable and speeds up the convergence process. Rectified Linear Unit (ReLU) activation functions are what cause non-linearity to be introduced into the network at this layer. They do this by applying an element-wise activation function, in which case all negative values are generally replaced with zero. Similar to the previous max pooling layer, this layer also downsamples the feature maps in order to further minimize the spatial dimensions of the maps.

Dense layers are those in which all of the neurons in the layer above are linked to all of the neurons in the layer below. These layers contribute to the process of learning high-level characteristics and producing conclusive predictions. Dropout is a regularization approach that involves arbitrarily removing a portion of the neurons in the network as it is being trained. By pushing the network to learn characteristics that are more robust, this helps to avoid overfitting from occurring.

Pseudocode: Block 1

```
Input: x (input feature map)
Output: y (output feature map)

shortcut = x
x = BatchNormalization(x)
x = ReLU(x)
x = Convolution(x)
x = BatchNormalization(x)
x = ReLU(x)
x = Convolution(x)
x = BatchNormalization(x)
x = ReLU(x)
x = Convolution(x)
x = Convolution(x)
y = x + shortcut
```

This pseudocode is an illustration of a Residual Block, also known as Block 1, which is used in proposed residual model. The 'x' feature map that was provided as input is first saved to a'shortcut' variable. After that, it does a Batch Normalization, then a ReLU activation, and finally a convolution operation three times in a row, with each iteration contributing to an improvement in the feature representation. After the last convolution, it adds the 'shortcut' to the output 'x', which makes the remaining connection easier to make. Because of this approach, the model is able to learn both refined features from the convolutional layers and identity mappings as appropriate. This makes it much simpler for the network to train extremely deep architectures while minimizing difficulties related to vanishing gradients. The output feature map of the Residual Block is denoted by the letter 'y' as the end result.

Pseudocode: Block 2

```
Input: x (input feature map)
Output: y (output feature map)

shortcut = x
x = BatchNormalization(x)
x = ReLU(x)
x = Convolution(x)
```

```
x = BatchNormalization(x)
x = ReLU(x)
x = Convolution(x)
x = BatchNormalization(x)
x = ReLU(x)
x = Convolution(x)
y = x + shortcut
```

Pseudocode: Block 3

```
Input: x (input feature map)
Output: y (output feature map)

shortcut = x
x = BatchNormalization(x)
x = ReLU(x)
x = Convolution(x)
x = BatchNormalization(x)
x = ReLU(x)
x = Convolution(x)
x = BatchNormalization(x)
x = ReLU(x)
x = Maxpooling(x)
x = Convolution(x)
y = x + shortcut
```

EXPERIMENTAL RESULTS

The dataset contains ultrasound pictures of the liver, separated into three unique grades: Grade 1, Grade 2, and Grade 3. It is likely that these groups reflect various phases or circumstances of either healthy liver function or liver disease. The dataset has been broken up into three sets: training, validation, and testing. Each of these sets serves a distinct function in the process of constructing and assessing machine learning or deep learning models. The information that follows offers a complete summary of this dataset including liver ultrasounds:

The total number of images: There are a total of 583 ultrasound pictures of the liver that are included in the collection. More specifically, it comprises the following:

- 140 pictures are needed for training.
- 322 pictures are waiting to be validated.
- 121 pictures for testing purposes.

Class Distribution: The dataset is divided into three groups, each of which represents a different level of liver disease severity:

- One class is considered to be in Grade 1.
- The second grade is equivalent to the second class.
- The third grade is equivalent to the third class.

Image Size: All of the ultrasound pictures that are included in this collection have been scaled to the same dimensions, which are 224 pixels on each side and 224 pixels high. This uniform picture size makes processing and analysis much simpler, and it guarantees that all of the photos may be used with the machine learning or deep learning models that have been specifically created for use with this dataset.

Training Set: This collection of 140 photos is used for training machine learning and deep learning models. It is referred to as the "training set." The training method teaches the model to detect patterns

and characteristics in the ultrasound pictures that separate the three distinct liver diseases that are represented by the classes. These patterns and features are seen in the liver scans.

Validation Set: The validation set consists of 322 different pictures and is used to evaluate the generalization capacity of the model as well as to fine-tune the model's hyperparameters. Researchers are able to improve the performance of the model and prevent it from being overfit with the help of this collection.

Testing Set: The testing set is an independent assessment data set comprising of 121 photos that is utilized for testing only. It is used to objectively evaluate the effectiveness of the model in classifying ultrasound images of the liver to the right categories. This set replicates real-world events in which the model comes into contact with data that it has not before seen.

When it comes to diagnosing and keeping track of liver diseases, pictures obtained from ultrasounds of the liver are quite helpful. This dataset, with its varied classes and consistent picture size, may be used to construct and test machine learning or deep learning models that aid medical practitioners in the categorization of liver ultrasound images. These models can be used to help diagnose liver conditions and diagnose liver diseases. This dataset is an important resource for research and clinical applications in the realm of medical imaging and healthcare since accurate categorization may help in the early diagnosis of illness and increase patient care.

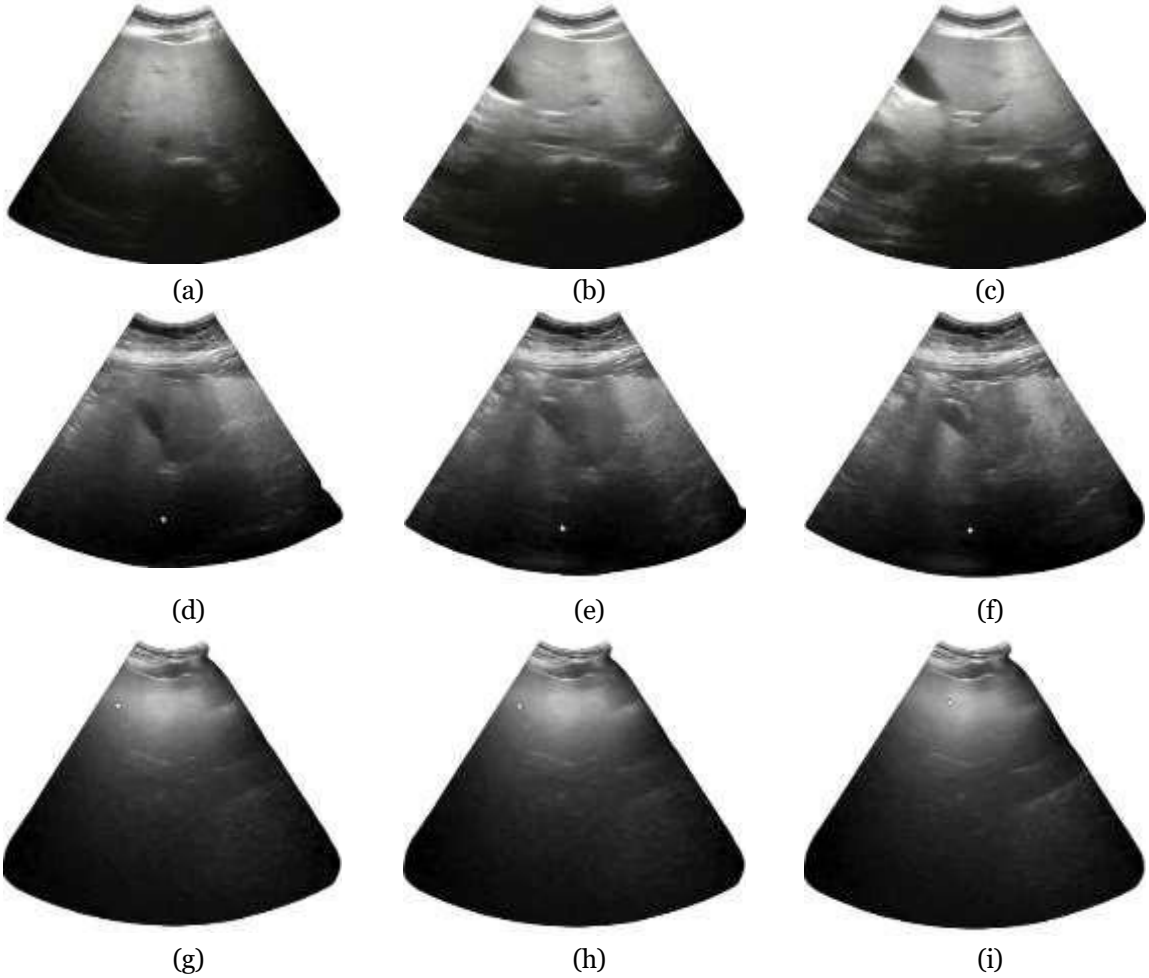

Figure 3: Input database images

Figure 3 shows the input images used in the database. Images (a) to (c) show images of grade 1, images (d) to (f) show grade 2 images, figures (g) to (i) show the grade 3 images.

In machine learning, a graphical representation called the training and validation loss plot, which is depicted in figure 4, is frequently used to assess how well a model is trained and how well it progresses. An understanding of how well a machine learning model is learning from its training data and extrapolating its results to data it hasn't seen before is provided by this chart.

Typically, the plot will include two separate curves, which are:

- This curve shows how well the model performed on the training data throughout the course of epochs or iterations during the training phase. We refer to it as the Training Loss Curve. It illustrates how the loss, a metric representing the discrepancy between the target values and the model's predictions, varies as the model learns new information. The training loss often has a tendency to be large early on in the process of training, which is an indication that the model is generating significant mistakes. On the other hand, as the training goes on, the loss usually goes down, which is a sign that the model is becoming better at its prediction powers and fitting the training data better. A steady improvement in the model's training loss over time is suggestive that the model is learning correctly.
- Validation Loss CurveThe validation loss curve, on the other hand, shows how well the model performs on a unique validation dataset that was not used for training. The training data does not contain this dataset. This dataset could serve as an independent gauge of the model's ability to generalize to novel circumstances. The validation loss curve is a highly useful tool for figuring out whether the model is "overfitting" the training data. Overfitting is a problem that happens when a model becomes too specialized in capturing noise and nuances in the training data. This leads to poor performance on fresh data that has not been seen before. The validation loss first starts to drop in the plot as the model learns and starts to generalize from the training data. However, if the model begins to overfit, the validation loss curve can begin to increase, indicating that the model's performance is declining when used with previously unseen data..

Figure 4: Training and validation loss

A graphical representation that is frequently used in machine learning to assess a model's performance and learning trajectory is the training and validation accuracy plot, which is displayed in figure 5. This figure provides important information about how well a model that uses machine learning is learning from the training data and how well it can generalize to new kinds of data that it hasn't encountered before.

In most cases, this graph will have two separate curves, which are:

- This curve depicts the model's accuracy on the training data throughout epochs or iterations while it is being trained. The term "training accuracy curve" refers to this particular curve. It illustrates the model's capacity to produce accurate predictions based on the dataset that it was trained on. It's possible that the training accuracy will be poor at first since the model's predictions won't be spot on. On the other hand, the accuracy tends to become better as training goes on, which is a sign that the model is picking up new information and becoming more

adaptable to the training data. The fact that the training accuracy has been steadily improving provides evidence that the model is successfully acquiring new knowledge from the data.

- The effectiveness of the model on an independent validation dataset, which it did not come across during the training stage of its development, is shown by the validation accuracy curve. This dataset is used to assess how well the model generalizes to new situations. Important details regarding whether the model is overfitting the training set are provided by the validation accuracy curve. "Overfitting" is a phenomena that happens when a model becomes very specialized in capturing both noise and complexity in the training data. This results in subpar performance on previously unseen data. When the model first begins to learn to generalize from the training data, the validation accuracy initially shows an upward trend in the plot. On the other hand, if there is overfitting, the validation accuracy curve can reach a plateau or start to drop, which would imply that the model's capacity to produce correct predictions based on fresh data is deteriorating.

Figure 5: Training and validation accuracy

Table 1: Output parameters of training

Metric	Value
Loss	9.04181
Accuracy	92.56%

The table that was presented comprises important performance indicators for a classification model, in particular when considering the evaluation of liver disorders that may be connected with fatty liver disease. The following is a full explanation of the metrics and the values that correlate to them:

- Loss: The value of 9.04181 for loss represents the average error or difference between the model's projected values and the actual ground truth labels in the dataset. This number was calculated using the data. In an issue involving classification, having loss values that are smaller is preferable since this indicates that the model's predictions are closer to the actual labels.
- Accuracy: Out of the overall amount of cases in the test set, the accuracy value, 0.92562, represents the proportion of occurrences that were correctly classified. The model has achieved an accuracy of approximately 92.56% for this dataset, meaning that it has nearly 93% of the time correctly diagnosed liver illnesses.

Table 2: Classification report

	Precision	Recall	F1-score
grade-1	1	0.95	0.97
grade-2	0.87	0.96	0.91
grade-3	0.94	0.86	0.9

The classification report in Table 2 displays the performance metrics for a classification with multiple classes model. The F1-score balances precision and recall, recall gauges the model's capacity to collect genuine positives, and precision indicates the accuracy of positive predictions. With a high recall of

0.95, which suggests it successfully captures the majority of genuine grade-1 cases, and an exceptional F1-score of 0.97, the framework achieves perfect precision (1.0) for grade-1 in this table, meaning all positive predictions are accurate. With an F1-score of 0.91 for grade-2, the model shows a little lower precision of 0.87 while maintaining a high recall of 0.96. Last but not least, the model's precision of 0.94 for grade 3 indicates high accuracy in positive predictions, while its recall is lower at 0.86, yielding an F1-score of 0.90. The model functions effectively overall, with grade 1 attaining the greatest overall performance indicators, closely followed by grades 2 and 3.

Table 3: Proposed output paper

Accuracy	0.92562
Precision	0.929724
Recall	0.92562
F1 score	0.926046
Cohens kappa	0.88708

Table 3 presents the performance metrics of the proposed output paper's classification model. The accuracy score, which measures the overall correctness of predictions, is impressive at 92.56%, indicating that the model accurately classifies the data. The precision score of 92.97% demonstrates the model's ability to make precise positive predictions, minimizing false positives. The recall score, identical to accuracy in this case, also stands at 92.56%, highlighting the model's capacity to capture true positives effectively. The F1 score, a balanced measure considering both precision and recall, is notably high at 92.60%, indicating a harmonious trade-off between precision and recall. Lastly, the Cohen's kappa coefficient, a statistic that assesses inter-rater agreement, is at 88.71%, suggesting substantial agreement between the model's predictions and the actual data, reinforcing the model's reliability. Overall, these metrics collectively illustrate the robust performance of the proposed output paper's classification model.

Table 4: Confusion matrix

	Grade 1	Grade 2	Grade 3
Grade 1	35	2	0
Grade 2	0	45	2
Grade 3	0	5	32

The confusion matrix presented in Table 4 summarizes the classification performance of a model across three different grades For Grade 1, the model correctly classified 35 instances, with 2 cases incorrectly classified as Grade 1 instead of Grade 2. There were no instances of Grade 1 being misclassified as other grades. For Grade 2, the model correctly classified 45 instances, with 2 cases mistakenly categorized as Grade 2 instead of Grade 3. Again, no instances of Grade 2 were erroneously classified as other grades. In the case of Grade 3, the model correctly classified 32 instances, but 5 instances were incorrectly labeled as Grade 3 when they were actually Grade 2. In summary, this confusion matrix offers an overview of the performance of the model, showcasing its accuracy in classifying each grade while revealing specific instances of misclassification.

Table 5: Comparative analysis

	Precision	**Recall**	**F1 score**	**Cohens kappa**	**Accuracy**
MobilenetV2	0.40086	0.32231	0.17641	0.02381	0.32231
NASNetMobile	0.81587	0.53719	0.50709	0.31858	0.53719
VGG19	0.8492	0.70248	0.65883	0.56674	0.70248
Proposed model	0.92972	0.92562	0.92605	0.88708	0.92562

The table 5 presents performance metrics for four different models, namely MobilenetV2, NASNetMobile, VGG19, and the Proposed model, in a classification task. The metrics include Precision, Recall, F1 score, Cohen's kappa, and Accuracy, which collectively assess the models' effectiveness in classifying data. With an astounding Precision of 0.92972, which indicates a low rate of false positives, a Recall of 0.92562, which indicates the capacity to correctly identify true positives, along with an

impressive F1 score of 0.92605, which indicates a balance between precision as well as recall, the suggested approach notably performs the best across all metrics. Furthermore, the high Accuracy of 0.92562 indicates that the model's classifications are generally accurate, while the Cohen's kappa of 0.88708 indicates a significant agreement between the model's predictions along with the actual data. The Proposed model, on the other hand, performs the best across these important evaluation parameters, while other designs, such as MobilenetV2, NASNetMobile, along with VGG19, show differing levels of performance.

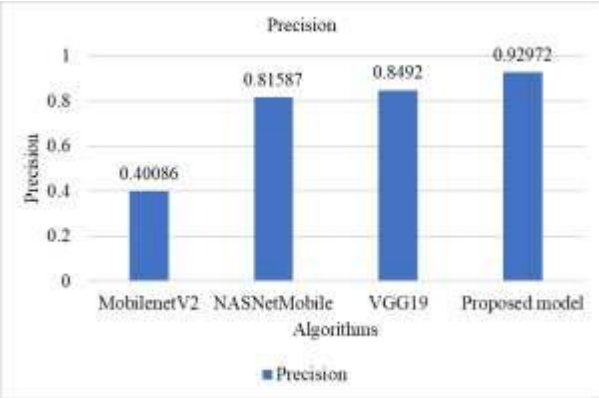

Figure 6: Precision

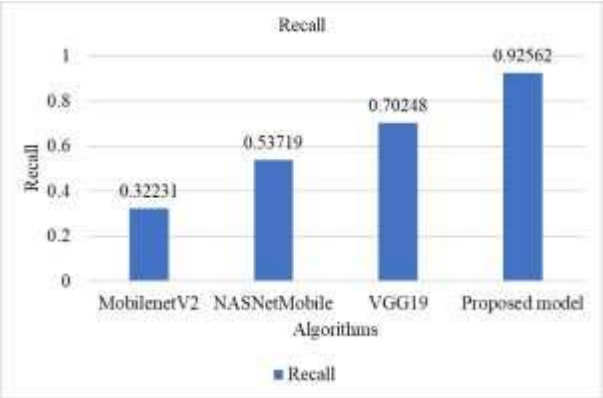

Figure 7: Recall

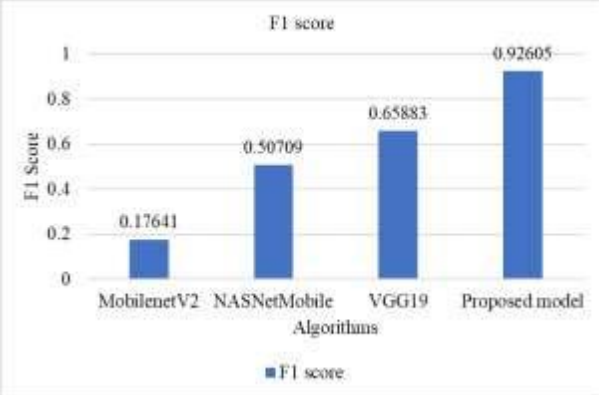

Figure 8: F1 score

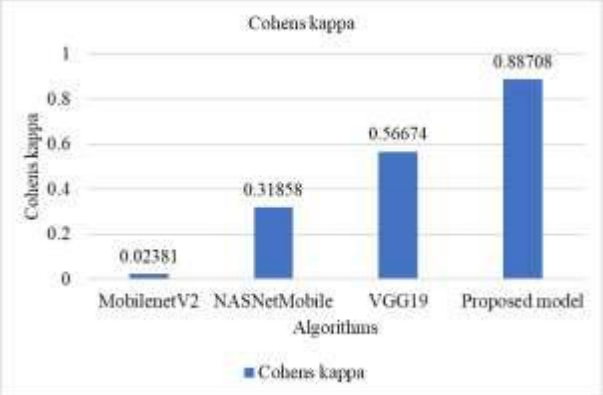

Figure 9: Cohens kappa

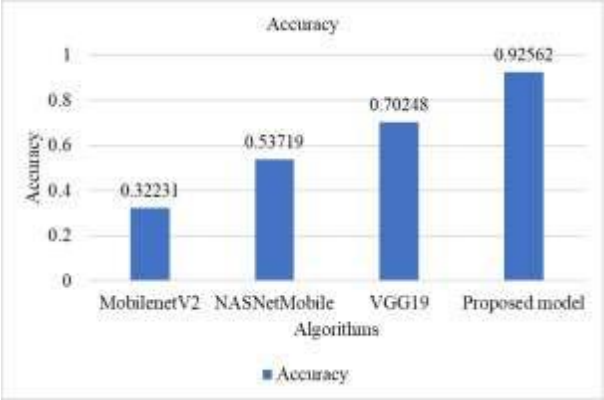

Figure 10: Accuracy

CONCLUSION

In conclusion, this study addresses a pressing healthcare challenge by developing a deep learning solution for the non-invasive classification of Fatty Liver Disease (FLD) stages using ultrasound images. FLD, a widespread condition with potentially severe consequences, demands early detection and monitoring. The deep learning model presented here, based on the ResNet architecture, has demonstrated its potential to provide a reliable and automated means of classifying FLD stages. The necessity for such a model arises from the escalating prevalence of FLD and the need for efficient

diagnostic tools. Using ultrasound pictures to manually diagnose FLD phases can be laborious and prone to interobserver variability. As a result, deep learning's automation of this procedure is a useful addition to the fields of medical imaging along with hepatology. The results obtained from the model underscore its effectiveness in accurately classifying FLD stages, thus facilitating timely interventions and personalized treatment plans. This automation enhances diagnostic efficiency and reduces the risk of overlooking critical conditions. It is imperative to acknowledge the collaborative effort between the medical community and machine learning practitioners, as this fusion of expertise yields tangible benefits in patient care.

REFERENCES

[1] Yang, Yang, Jing Liu, Changxuan Sun, Yuwei Shi, Julianna C. Hsing, Aya Kamya, Cody Auston Keller et al. "Nonalcoholic fatty liver disease (NAFLD) detection and deep learning in a Chinese community-based population." *European Radiology* (2023): 1-13.

[2] Gao, Feng, De-Chan Lu, Tian-Lei Zheng, Shi Geng, Jun-Cheng Sha, Ou-Yang Huang, Liang-Jie Tang et al. "Fully connected neural network-based serum surface-enhanced Raman spectroscopy accurately identifies non-alcoholic steatohepatitis." *Hepatology International* 17, no. 2 (2023): 339-349.

[3] Tahmasebi, Aylin, Shuo Wang, Corinne E. Wessner, Trang Vu, Ji-Bin Liu, Flemming Forsberg, Jesse Civan, Flavius F. Guglielmo, and John R. Eisenbrey. "Ultrasound-Based Machine Learning Approach for Detection of Nonalcoholic Fatty Liver Disease." *Journal of Ultrasound in Medicine* (2023).

[4] Nguyen, Trong N., Anthony S. Podkowa, Trevor H. Park, Rita J. Miller, Minh N. Do, and Michael L. Oelze. "Use of a convolutional neural network and quantitative ultrasound for diagnosis of fatty liver." *Ultrasound in medicine & biology* 47, no. 3 (2021): 556-568.

[5] Kresevic, Simone, Milos Ajcevic, Mauro Giuffrè, Pierpaolo Pupa, Lory Saveria Crocè, and Agostino Accardo. "Semi-automatic Approach to Estimate the Degree of Non-alcoholic Fatty Liver Disease (NAFLD) from Ultrasound Images." In *Nordic-Baltic Conference on Biomedical Engineering and Medical Physics*, pp. 227-235. Cham: Springer Nature Switzerland, 2023.

[6] Pandit, Harshul, Joseph P. Tinney, Yan Li, Guozhen Cui, Suping Li, Bradley B. Keller, and Robert CG Martin II. "Utilizing contrast-enhanced ultrasound imaging for evaluating fatty liver disease progression in pre-clinical mouse models." *Ultrasound in Medicine & Biology* 45, no. 2 (2019): 549-557.

[7] Colantonio, Sara, Antonio Salvati, Claudia Caudai, Ferruccio Bonino, Laura De Rosa, Maria Antonietta Pascali, Danila Germanese, Maurizia Rossana Brunetto, and Francesco Faita. "A deep learning approach for hepatic steatosis estimation from ultrasound imaging." In *International Conference on Computational Collective Intelligence*, pp. 703-714. Cham: Springer International Publishing, 2021.

[8] Antony Asir Daniel, V., and J. Jeha. "An Optimal Modified Faster Region CNN Model for Diagnosis of Liver Diseases from Ultrasound Images." *IETE Journal of Research* (2023): 1-18.

[9] Che, Hui, Lloyd G. Brown, David J. Foran, John L. Nosher, and Ilker Hacihaliloglu. "Liver disease classification from ultrasound using multi-scale CNN." *International Journal of Computer Assisted Radiology and Surgery* 16, no. 9 (2021): 1537-1548.

[10] Das, Amit, Mary Connell, and Shailesh Khetarpal. "Digital image analysis of ultrasound images using machine learning to diagnose pediatric nonalcoholic fatty liver disease." *Clinical Imaging* 77 (2021): 62-68.

[11] Cao, Wen, Xing An, Longfei Cong, Chaoyang Lyu, Qian Zhou, and Ruijun Guo. "Application of deep learning in quantitative analysis of 2-dimensional ultrasound imaging of nonalcoholic fatty liver disease." *Journal of Ultrasound in Medicine* 39, no. 1 (2020): 51-59.

[12] Pasyar, Pezhman, Tahereh Mahmoudi, Seyedeh-Zahra Mousavi Kouzehkanan, Alireza Ahmadian, Hossein Arabalibeik, Nahid Soltanian, and Amir Reza Radmard. "Hybrid classification of diffuse liver diseases in ultrasound images using deep convolutional neural networks." *Informatics in Medicine Unlocked* 22 (2021): 100496.

[13] Nguyen, Trong N., Anthony S. Podkowa, Alex Y. Tam, Eben C. Arnold, Rita J. Miller, Trevor H. Park, Minh N. Do, and Michael L. Oelze. "Characterizing fatty liver in vivo in rabbits, using quantitative ultrasound." *Ultrasound in medicine & biology* 45, no. 8 (2019): 2049-2062.

[14] Schoenberger, Haley, Nicolas Chong, David T. Fetzer, Nicole E. Rich, Takeshi Yokoo, Gaurav Khatri, Jocelyn Olivares et al. "Dynamic changes in ultrasound quality for hepatocellular carcinoma screening in patients with cirrhosis." *Clinical Gastroenterology and Hepatology* 20, no. 7 (2022): 1561-1569.

[15] Lupsor-Platon, Monica, Teodora Serban, Alexandra-Iulia Silion, Alexandru Tirpe, and Mira Florea. "Hepatocellular carcinoma and non-alcoholic fatty liver disease: A step forward for better evaluation using ultrasound elastography." *Cancers* 12, no. 10 (2020): 2778.

[16] Yoo, Jeongin, Jeong Min Lee, Ijin Joo, Dong Ho Lee, Jeong Hee Yoon, Hyo-Jin Kang, and Su Joa Ahn. "Prospective validation of repeatability of shear wave dispersion imaging for evaluation of non-alcoholic fatty liver disease." *Ultrasound in Medicine & Biology* 45, no. 10 (2019): 2688-2696.