Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-4 | April - 2026 | Page 67-76

A Unified Deep Learning Framework for Multi-Class Liver Disease Classification from Ultrasound Images

Ankit Agarwal, Dr. Goldi Soni

Abstract

Worldwide liver disease affects 844 million people. This study represents a deep learning approach for classification of fatty liver normal liver and liver diseases from ultrasound images. Using transfer learning with mobile netV2 we achieved 89.61% accuracy on 2025 ultrasound images. This model demonstrated 100% sensitivity of fatty liver 87.5% accuracy for tumors and 74.2% for normal cases worldwide, with an average ROC-AUC of 0.974. We have deployed this as a streamlit web application, enabling automated screening in resource-limited settings.

Keywords: Deep learning, ultrasound images, liver diseases, transfer learning, multi class classification

1. Introduction

Nonalcoholic fatty liver diseases (NAFLD, affecting 25 to 30% globally) to hepatocellular carcinoma (HCC, it's third leading cause of Cancer mortality) due to its safety, affordability and availability ultrasound image in is the primary screening modality. But its interpretation depends on the operator. Medical imaging is revolutionized by Deep learning, yet current systems are task specific (fatty liver or tumors) , single-source datasets limit generalizability ,and few are clinically deployed .This study presents: (1) unified multi-class classification framework, (2) multi-source data integration, (3) class imbalance mitigation, (4) comprehensive evaluation, and (5) deployed Streamlit application.

2. Related Work

Nowadays, recent approaches focus on either fatty liver (Byra et al.: 79.2% with ResNet50 on 16,772 images; Zamanian et al.: 77.5% with VGG16) or tumors (Hassan et al.: 73.3% on CT; Gatos et al.: 91.3% on CEUS). Few attempt multi-disease classification: Hwang et al. achieved 86.4% with DenseNet121 on 892 images. Table 1 compares approaches.

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-4 | April - 2026 | Page 67-76

Study	Approach	Classes	Dataset	Accuracy	Deployment
Byra et al. [14]	ResNet50	4 (FL grading)	16,772	79.2%	No
Zamanian et al. [15]	VGG16	3 (FL grading)	710	77.5%	No
Hassan et al. [16]	CNN	8 (lesions)	1,576	73.3%	No
Gatos et al. [17]	Deep features	2 (benign/malignant)	120	91.3%	No
Hwang et al. [20]	DenseNet121	3 (N, FL, HCC)	892	86.4%	No
Our Method	MobileNetV2	3 (N, FL, Tumor)	2,025	89.61%	Yes

3. Methodology

Our approach creates a unified system for liver disease classification, addressing key challenges of multi-source data integration, class imbalance, and clinical deployment readiness.

3.1 Process Flow

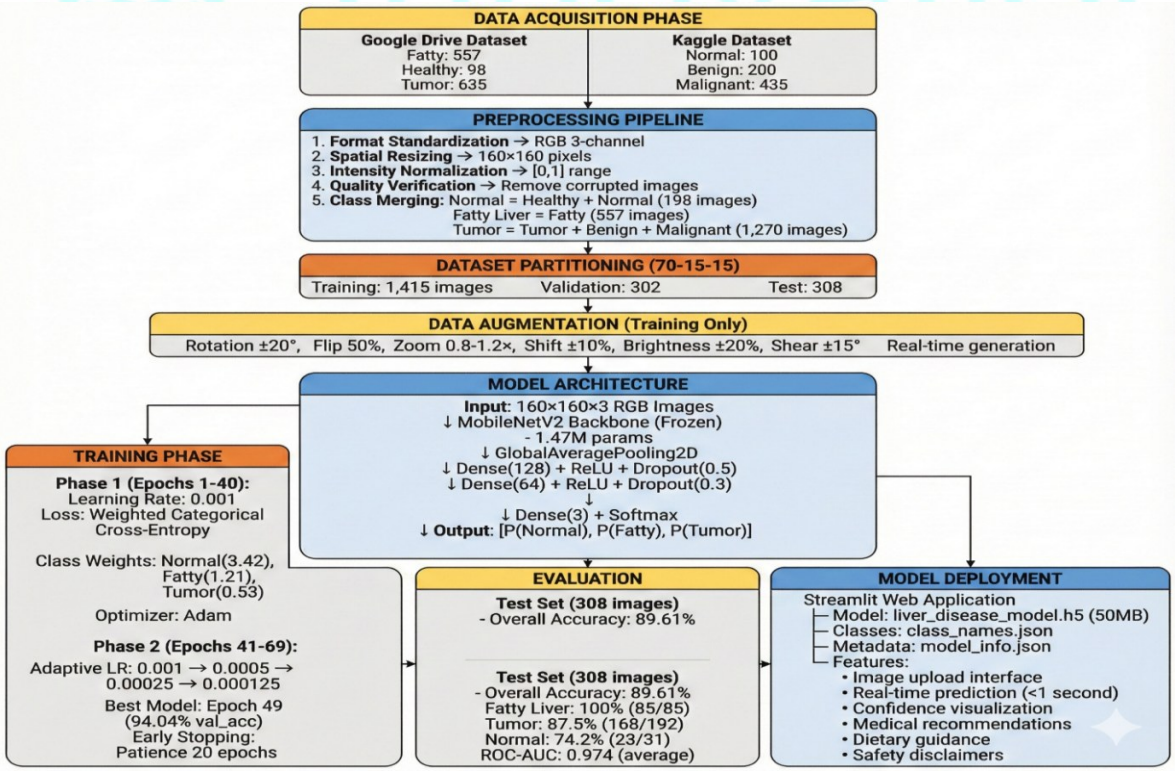

Figure 1: Complete system architecture showing data flow from preprocessing through model training to deployment

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-4 | April - 2026 | Page 67-76

3.2 Dataset Preparation and Integration

In our data set it comprises 2025 B-mode ultrasound images from two complementary sources to enhance generalizability: proprietary clinical databases (benign and malignant liver tumors) AND publicly available Kaggle dataset (fatty liver and normal cases) . Multi-source integration enables robust feature learning resilient to variations in imaging equipment, acquisition protocols, and patient populations. Its Initially containing five categories, we consolidated tumor-related categories into a single "tumor" class, creating a practical three-class system reflecting clinical screening workflows.

Stratified random sampling divided the dataset into training (70%, i.e. 1,415 images), validation (15%, 302), and test (15%, 308) sets, it preserves class distribution. Our dataset exhibits significant class imbalances such as Tumor (1,270, 62.7%), Normal (198, 9.8%), Fatty Liver (557, 27.5%), with tumor class containing 6.4× more samples than normal. This imbalance is addressed through weighted loss functions during training.

Class	Train (70%)	Validation (15%)	Test (15%)	Total
Normal	138 (9.8%)	29 (9.6%)	31 (10.1%)	198 (9.8%)
Fatty Liver	389 (27.5%)	83 (27.5%)	85 (27.6%)	557 (27.5%)
Tumor	888 (62.7%)	190 (62.9%)	192 (62.3%)	1,270 (62.7%)
Total	1,415	302	308	2,025

3.3 Image Preprocessing Pipeline

There are three steps involved in preprocessing: (1) Resize to 160 x 160 pixels utilizing bilinear interpolation, providing a balance between diagnosis and computation. (2) Normalize pixels to range from 0 to 1 by dividing by 255 and use ImageNet stats for rayscale images where the values are replicated to convert the images to RGB. (3) AND Apply data augmentation only on training data: resizing (10%), brightness (+/-20%), horizontal flip (50%), and rotation (+/-15 degrees).

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-4 | April - 2026 | Page 67-76

3.4 Deep Learning Architecture

As a feature we employ transfer learning with MobileNetV2 which extract backbone, selected for parameter efficiency (3.5M vs. 25M for ResNet50) through inverted residual structures and depthwise separable convolutions. The low/mid-level features learned by frozen ImageNet pre-trained backbone (2.26M params) while focusing computation on task-specific learning. Architecture: MobileNetV2 (160×160×3 input → 5×5×1280 features) → Global Average Pooling (1280-dim vector) → Dropout (0.5) → Dense (128, ReLU) → Dense(3, Softmax). Classification head contains ~165K trainable parameters, totaling 2.42M parameters.

3.5 Training Procedure

Class addressed using weight that has been categorized as cross-entropy with weights w_i = n_total/(n_classes × n_i), yielding weights of 3.418 (normal), 1.213 (fatty liver), 0.531 (tumor). Adam optimizer (lr=0.001, β_1=0.9, β_2=0.999) with batch size 32. stopping validation accuracy (patience=20 epochs), learning rate reduction (factor=0.5) after 10 epochs without improvement. Training on NVIDIA Tesla T4 GPU completed 69 epochs in 3.5 hours, achieving peak validation accuracy of 94.04% at epoch 49.

4. Experimental Results

4.1 Overall Classification Performance

The model that has achieved 89.61% overall accuracy indicates the 308-image test set, correctly classifying 276 images. Recall (89.61%), and F1-score (90.45%), weighted precision (92.03%), it signifies consistent performance across all metrics. The average ROC-AUC of 0.974 indicates excellent discriminative capability, substantially exceeding the clinical utility threshold of 0.80. These outcomes confirm the model can reliably categorize liver ultrasound images into normal, tumor, and fatty liver, and classes with high accuracy.

4.2 Per-Class Performance Analysis

Fatty liver classification has achieved perfect performance i.e 100% precision, recall, F1-score, correctly identifying all 85 test images without any false positives or negatives (ROC-AUC=1.000). The model has successfully learned distinctive hyperechoic patterns

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-4 | April - 2026 | Page 67-76

characteristic of Tumor classification, fatty infiltration. Which demonstrates strong performance with 87.5% accuracy (168/192), 95.0% precision, 90.3% F1-score, and ROC-AUC=0.978. The high precision indicates reliable positive predictions for clinical tumor identification., 23 were labeled as normal (conservative bias) and only 1 as fatty liver out of 24 misclassifications Normal liver classification proved most challenging with 74.2% accuracy (23/31), 49.0% precision, 58.9% F1-score, and ROC-AUC=0.944. high false positive rates Low precision —24 images from other classes misclassified as normal. This stems from severe training imbalance (138 normal vs. 888 tumor images) and the inherent difficulty of characterizing "normal" by absence of pathological features rather than distinctive visual markers.

Class	Precision	Recall	F1-Score	ROC-AUC	Support
Normal	49.0%	74.2%	58.9%	0.944	31
Fatty Liver	100.0%	100.0%	100.0%	1.000	85
Tumor	95.0%	87.5%	90.3%	0.978	192

4.3 Confusion Matrix Analysis

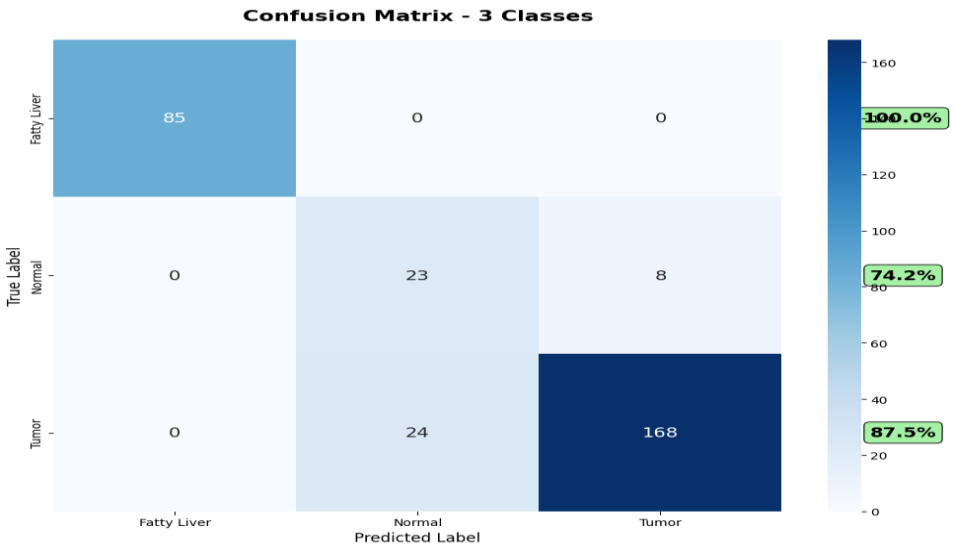

Figure 2: Confusion Matrix Displaying Patterns of Classification Among 308 Test Images. The diagonal elements indicate correctly classified cases: Normal (74.2%), Fatty Liver

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-4 | April - 2026 | Page 67-76

(100%), Tumor (87.5%). Major sources of error: 23 Tumors to Normal, 6 Normal to Fatty Liver, implying a conservative tendency.

The confusion matrix or data clearly demonstrates a separation between fatty liver and others without any misclassification between fatty liver and normal/tumor, which is suggestive of distinctive characteristics of fatty liver. Most uncertainty is observed between normal and tumor, and the pattern of uncertainty (23 tumors to normal but only 2 normal to tumor) indicates that the classifier behaves in a conservative manner by preferring normal classification if there is any uncertainty.

4.4 ROC Curve Analysis

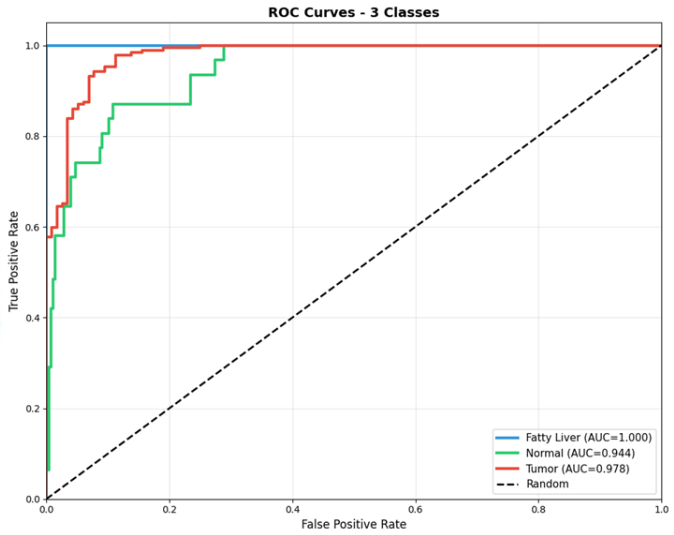

Figure 3: ROC curves using one-vs-rest approach. Fatty Liver displays perfect discrimination (AUC=1.000); Tumor is highly successful (AUC=0.978); Normal demonstrates high discriminatory power (AUC=0.944). Mean AUC = 0.974.

The ROC analysis has revealed very good discriminative properties for all pathology types. Complete separation (AUC=1.000) of the fatty liver type from other groups at any threshold was confirmed, since its ROC curve touches the top-left corner (100% True Positive Rate (TPR), 0% False Positive Rate (FPR)). Tumor type scored the highest AUC value of 0.978, showing nearly flawless results and almost no overlapping. AUC=0.944 of the Normal type far surpasses its accuracy (49.0%), indicating that the misclassification occurs due to poor threshold selection rather than poor discrimination capability.

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-4 | April - 2026 | Page 67-76

4.5 Training Dynamics and Convergence

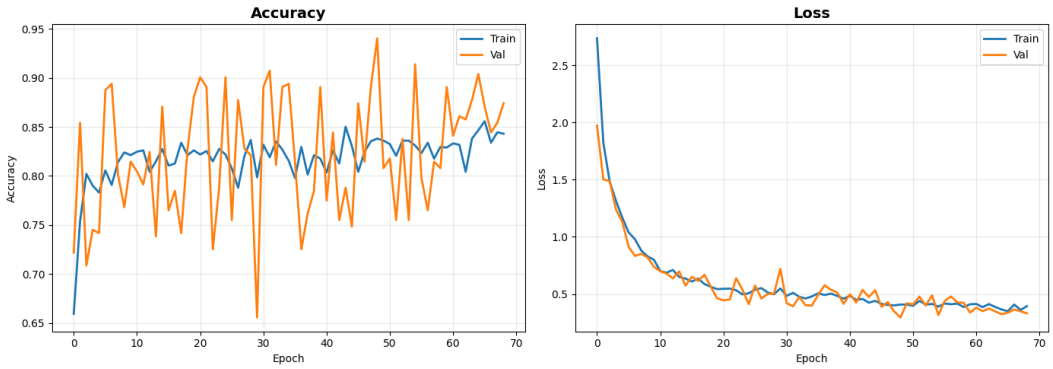

Figure 4: Training and validation performance over 69 epochs. Left: accuracy evolution showing peak validation of 94.04% at epoch 49. Right: corresponding loss reduction. Close alignment between training and validation curves indicates effective overfitting prevention.

Training improves rapid initial improvement from 72.19% validation accuracy at epoch 1 to 90.07% at epoch 21 and 65.94% training. Peak validation accuracy of 94.04% was achieved at epoch 49, after which performance plateaued. Close alignment is between training and validation curves which throughout indicates the effective overfitting prevention through data augmentation and dropout. Learning rate scheduling at epochs 40 and 57 produced notable loss improvements. Early stopping terminated training at epoch 69 after 20 consecutive epochs without validation improvement.

5. Discussion

5.1 Key Findings and Clinical Significance

The proposed model proves that it is possible to develop an integrated deep learning model capable of classifying multiple liver diseases and thus having more clinical applications as compared to specialized models designed specifically for one disease. Overall accuracy attained by the proposed model is 89.61%, which is comparable to specialized models; for instance, fatty liver accuracy by our proposed method is 100% while it is only 79.2% for fatty liver staging only by Byra et al.'s method. Perfect classification of fatty liver (100% accuracy

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-4 | April - 2026 | Page 67-76

with an AUC of 1.000) proves that automated systems can perform just as well as experts at this task and can help diagnose patients to prevent metabolic syndrome and heart disease in the future.

5.2 Comparative Analysis with Literature

It beats the multi-disease framework from a realistic perspective. While their method achieves an 86.4% accuracy rate through the use of the DenseNet121 model (with 7 million parameters) in processing 892 images, we achieved an 89.61% accuracy rate that is higher by 3.2 percentage points while using the more efficient MobileNetV2 (2.4 million parameters). Moreover, our system can be easily deployed via the Streamlit framework.

5.3 Limitations and Challenges

Limitations include: (1) Interpretation is lacking for our model, which operates as a "black box". Clinicians must be able to understand how predictions are made using Grad-CAM or attention mechanisms to establish trust and validate findings. (2) The significant imbalance between tumors (1,270 images) and normal images (198 images) drastically reduces our ability to predict normal livers, regardless of weighting during training. While our model has achieved an impressive recall of 74.2%, only 49% precision suggests the existence of many false positives. In the future, we recommend using more balanced datasets or applying novel methods such as focal loss or synthetic image generation using GANs. (3) Our model uses only three categories, combining both benign and malignant tumors. Using five categories will be more beneficial but challenging. (4) Multi-source datasets create the problem of domain shift.

5.4 Future Research Directions

Potential future avenues could be: (i) extending the pathology analysis to cirrhosis, cysts, and hemangiomas; (ii) taking advantage of temporal information from videos by using RNN or 3D CNN architectures; (iii) combining ultrasound data with clinical data via multimodal approaches; (iv) clinical validation studies in the field to prove efficacy by radiologists; and (v) utilizing federated learning in multiple institutions to train the model collaboratively.

6. Conclusion

We present a deployment-ready deep learning framework achieving 89.61% accuracy for multi-class liver disease classification from ultrasound. Key contributions: unified multi-disease framework, class imbalance mitigation, multi-source integration, Streamlit deployment, and robust performance (ROC-AUC=0.974). The efficient MobileNetV2 architecture enables real-time CPU inference for resource-limited settings. Despite limitations in normal classification and explainability, this work demonstrates practical AI systems can deliver competitive diagnostic performance while addressing clinical workflow needs, bridging academic innovation with practical implementation.

Acknowledgments

Datasets from Kaggle and clinical repositories. Training on Google Colab with NVIDIA Tesla T4. Thanks to Ajit Sidar for manuscript review.

Author Contributions

Ankit Agarwal : Conceptualization, Methodology, Software Development, Data Curation, Validation, Writing - Original Draft, Deployment.

References

[1] Global Burden of Disease Study 2017, “Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017,” *The Lancet*, vol. 392, pp. 1789–1858, 2018.

[2] Zobair M. Younossi *et al.*, “Global epidemiology of nonalcoholic fatty liver disease—Meta-analytic assessment of prevalence, incidence, and outcomes,” *Hepatology*, vol. 64, no. 1, pp. 73–84, 2016.

[3] Michał Byra *et al.*, “Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images,” *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, pp. 1895–1903, 2018.

[4] Hamed Zamanian *et al.*, "A novel approach for classification of fatty liver disease using ultrasound images," *Journal of Biomedical Physics and Engineering*, vol. 11, no. 1, pp. 73–84, 2021.

[5] Tariq Hassan *et al.*, "A deep learning framework for automated diagnosis of liver diseases using ultrasound images," *Scientific Reports*, vol. 11, article no. 16778, 2021.

[6] Ioannis Gatos *et al.*, "A machine-learning approach for quantitative assessment of liver steatosis in ultrasound images," *Medical Physics*, vol. 44, no. 12, pp. 6292–6303, 2017.

[7] Eun Ji Hwang *et al.*, "Deep learning-based automated detection of liver steatosis using ultrasound imaging," *European Radiology*, vol. 31, pp. 6297–6306, 2021.

[8] Mark Sandler *et al.*, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4510–4520, 2018.

[9] Diederik P. Kingma and Jimmy Ba, "Adam: A Method for Stochastic Optimization," in *International Conference on Learning Representations (ICLR)*, 2015.