

RadDenseNet: An Explainable AI Model for Detection of Hepatic Steatosis from Abdominal CT Images

Rahul L.R.^{}, and Satya Prasad Gaddamedi^{}

Department of ECE, Koneru Lakshmaiah Education Foundation, Hyderabad, India

Corresponding author: Rahul LR (e-mail: dr.rahul@klh.edu.in), **Author(s) Email:** Satya Prasad Gaddamedi (e-mail: satyaprasad@klh.edu.in)

Abstract Hepatic steatosis affects approximately 30.2% of the global population and, if undetected, can silently progress to non-alcoholic steatohepatitis, cirrhosis, and hepatocellular carcinoma. Manual CT interpretation for fatty liver assessment is time-consuming and prone to inter-observer variability, motivating research into automated screening tools. We propose RadDenseNet, a DenseNet121 model initialized with RadImageNet weights pretrained on 1.35 million medical images spanning CT, MRI, and ultrasound, and evaluate it as a research prototype. The first 397 of 427 backbone layers were frozen, and the remaining 30 layers were fine-tuned alongside a classification head comprising global average pooling, a dense layer of 256 units, batch normalization, and dropout of 0.5. The model was trained on the Kaggle Liver CT Image Dataset - 3,557 abdominal CT images across Healthy and Hepatic Steatosis classes - using a stratified 70/20/10 split with 5-fold cross-validation. RadDenseNet was compared against six baseline models spanning ImageNet and RadImageNet pretraining across ResNet50, InceptionV3, and DenseNet121 architectures. The decision threshold was optimized on the validation set, yielding $\tau = 0.48$. On the held-out test set, RadDenseNet achieved 96.35% accuracy, AUC of 0.9966, sensitivity of 0.9538, specificity of 0.9752, precision of 0.9789, and F1-score of 0.9662, with 5-fold cross-validation confirming stable performance (mean accuracy $96.46\% \pm 0.68\%$, mean AUC 0.9962 ± 0.0014). RadImageNet pretraining consistently outperformed ImageNet pretraining across all architectures, with accuracy improvements of 2.81 to 5.90 percentage points. Grad-CAM analysis covering all four prediction categories revealed that false positives arose from extra-hepatic attention and false negatives from weak parenchymal activation - clinically meaningful failure modes. External validation on multi-center datasets, extension to multi-class steatosis grading, and liver ROI segmentation as preprocessing are recommended as future directions.

Keywords Hepatic steatosis; RadImageNet; DenseNet121; Explainable AI; Computed tomography.

1. Introduction

Hepatic steatosis, commonly known as Nonalcoholic Fatty Liver Disease (NAFLD), affects approximately 30.2% of the global population [1], [2] and, if undetected, progresses silently through non-alcoholic steatohepatitis, cirrhosis, and hepatocellular carcinoma [3], [4]. Computed tomography (CT) offers a practical non-invasive route for steatosis screening [5], where steatosis appears as a measurable reduction in liver attenuation relative to the spleen, expressed in Hounsfield Units (HU) [6]. However, manual interpretation is time-consuming and subject to inter-observer variability, creating a need for automated detection tools. Deep learning and transfer learning have delivered strong performance across CT-based applications including liver cancer detection [7], [8], lung cancer detection [10], brain tumor classification [11], and kidney tumor detection [12] by adapting feature representations to new target tasks [9]. Despite

this progress, most CT-based steatosis studies rely on segmentation pipelines [13], [15], handcrafted radiomics [14], or 3D volumetric classifiers [16] rather than end-to-end image classification. The few studies that have attempted direct CT image classification used ImageNet-pretrained weights - a domain mismatch since natural photograph features bear little resemblance to grayscale CT tissue density patterns where steatosis-relevant HU differences are as little as 5-10 HU [6]. We hypothesize that RadImageNet [17] - 1.35 million annotated CT, MRI, and ultrasound images curated by radiologists - provides more relevant feature initialization than ImageNet for CT-based steatosis classification [25]. A second unaddressed gap exists: nearly all published CT-based liver models provide no insight into which image regions drove their predictions [31]. This study addresses both gaps. The key contributions are: (1) to the best of our knowledge, the first application of RadImageNet pretraining to direct

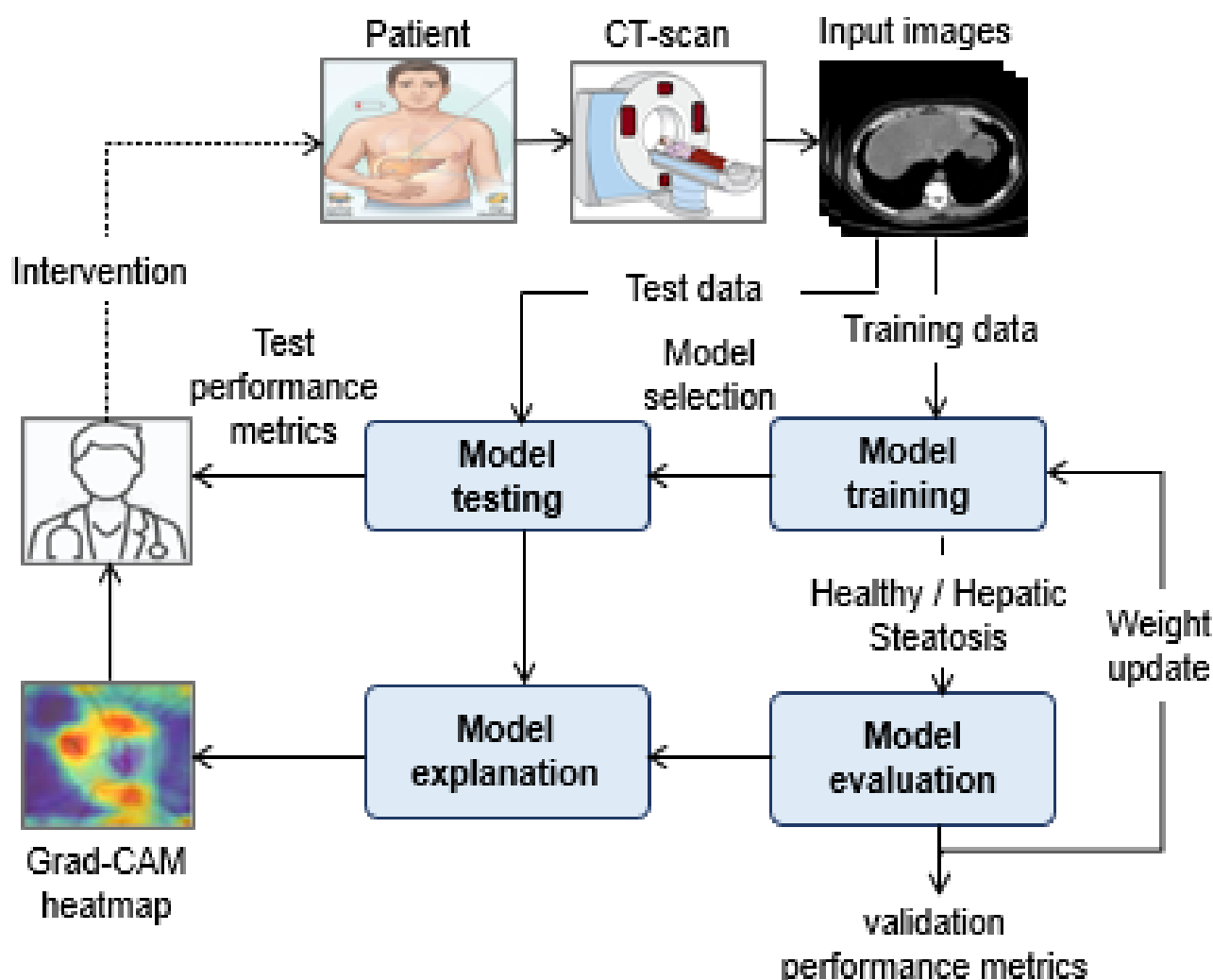


Fig. 1. Overall workflow of the proposed RadDenseNet framework for automated hepatic steatosis detection from abdominal CT images, including data preparation, model training, testing, performance evaluation, and Grad-CAM explainability.

end-to-end binary hepatic steatosis classification from 2D CT images, distinct from segmentation-based [13], [15], radiomics-based [14], and 3D volumetric [16] approaches; (2) systematic comparison against six baseline models spanning two pretraining domains; (3) stratified 70/20/10 split with 5-fold cross-validation; and (4) Grad-CAM explainability covering all four prediction categories - true positives, true negatives, false positives, and false negatives.

II. Method

The overall pipeline is illustrated in Fig. 1. The following subsections describe each stage in detail.

A. Dataset

The Kaggle Liver CT Image Dataset [18] comprises 3,557 non-contrast abdominal CT images across two classes - Healthy (1,611 images) and Hepatic Steatosis (1,946 images) - stored in RGB format at 256×256

pixels. The mild class imbalance (1:1.2) was addressed using class weights during training. The diagnostic labels were assigned by the dataset curators based on CT imaging criteria; the exact reference standard is undocumented, and the model is trained on Kaggle-provided annotations rather than independently verified clinical ground truth [18], [49].

B. Data Split

The dataset was divided using stratified random splitting (70/20/10), yielding 2,490 training, 711 validation, and 356 test images, preserving the class ratio across all subsets [19]. The test set was held out throughout and used only once for final evaluation. A fixed random seed of 42 ensured full reproducibility. 5-fold stratified cross-validation was additionally performed on the combined training and validation set of 3,201 images, with the test set kept completely separate. The Kaggle dataset does not provide patient-

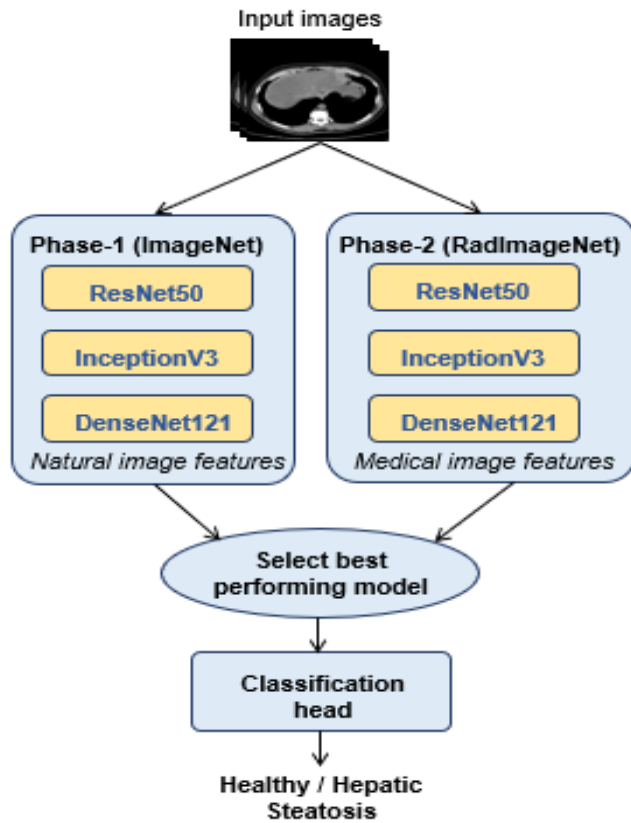


Fig. 2. Two-phase experimental design comparing ImageNet and RadImageNet-pretrained CNN architectures for hepatic steatosis classification.

level identifiers, preventing patient-level splitting. The image-level split preserves class proportions, but slice-level correlation across subsets cannot be excluded, representing an acknowledged limitation.

C. Data Processing

The preprocessing pipeline comprises three sequential operations: spatial resizing, pixel normalization, and data augmentation. The Kaggle dataset provides images as pre-converted 8-bit PNG files rather than raw DICOM files with native HU values; therefore, the original windowing parameters are inaccessible. The 8-bit PNG pixel values cannot be treated as calibrated HU measurements, and the degree to which the PNG representation preserves the subtle liver-to-spleen attenuation differences - as little as 5–10 HU - that carry diagnostic significance for steatosis assessment [6] is uncertain. This represents a meaningful methodological limitation to consider when interpreting the reported results.

1. Image resizing: All images were resized from their original resolution of 256×256 pixels to 224×224 pixels to match the input dimensions required by the proposed backbone architecture. Bilinear interpolation was used to maintain smooth intensity gradients during resizing. As defined in [47], the output spatial dimension after resizing is computed using Eq. (1):

$$O = \text{floor}\left(\frac{l-1}{s}\right) + 1 \quad (1)$$

where O is the output spatial dimension after resizing ($O = 224$ pixels), l is the original input spatial dimension ($l = 256$ pixels), s is the down-sampling stride factor that determines the sampling interval between input and output pixel positions, and $\text{floor}(\cdot)$ denotes the floor function that rounds the result down to the nearest integer to ensure a whole number of output pixels.

2. Pixel Normalization: Raw CT images are stored as 8-bit unsigned integers with pixel intensity values in the range $[0, 255]$. To ensure numerical stability during gradient-based optimization and to bring the input distribution into a range compatible with the pretrained RadImageNet weights, each pixel value was normalized to $[0, 1]$ using min-max scaling [27] as defined in Eq. (2):

$$x_{\text{norm}} = \frac{x}{255} \quad (2)$$

where x is the original pixel intensity value ($x \in [0, 255]$) and x_{norm} is the normalized pixel value ($x_{\text{norm}} \in [0, 1]$). This linear normalization does not change relative pixel-intensity differences within the PNG representation. However, since the original DICOM files and windowing parameters are unavailable, the 8-bit PNG pixel values cannot be confirmed to retain a calibrated linear relationship with native Hounsfield Unit measurements.

3. Data Augmentation: Data augmentation was applied exclusively to the training set to improve generalization and reduce overfitting; validation and test images underwent only resizing and normalization. The pipeline comprised four stochastic transformations - horizontal flip, vertical flip, rotation, and brightness adjustment - each applied independently with a fixed probability, selected to simulate realistic CT acquisition variability without introducing anatomically implausible distortions.

4. Random Horizontal Flip: As defined in [20], each training image was horizontally flipped with probability $p = 0.5$, and computed using Eq. (3):

$$x_{\text{flip}} = \begin{cases} \text{flip}(x) & \text{for } u \leq p \\ x & \text{otherwise} \end{cases} \quad (3)$$

where x is the input image, $u \sim \text{Uniform}(0,1)$ is a random variable drawn at runtime, $p = 0.5$ is the flip probability, and $\text{flip}(x)$ denotes the left-right mirror transformation. Horizontal flipping simulates variability in patient positioning within the CT scanner gantry.

5. Random Vertical Flip: A vertical flip was applied with probability $p = 0.5$, [20] as given in Eq. (4):

$$x_{\text{flip}} = \begin{cases} \text{flip}_y(x) & \text{for } u \leq p \\ x & \text{otherwise} \end{cases} \quad (4)$$

where $\text{flip}_y(x)$ denotes the top-to-bottom mirror transformation. While vertical flipping generates

orientations not encountered in standard axial CT acquisition, it was applied conservatively to increase training set diversity and reduce orientation-specific bias.

6. Random Rotation: Each image was rotated by a random angle θ sampled uniformly from $[-15^\circ, +15^\circ]$ using a 2D affine rotation matrix, [20] as defined in Eq. (5):

$$x_{rot} = R(\theta) \cdot x = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \cdot x \quad (5)$$

where x is the input image, $R(\theta)$ is the 2D rotation matrix, $\theta \sim \text{Uniform}(-15^\circ, +15^\circ)$ is the sampled rotation angle, and $\cdot$ denotes the spatial affine transformation applied to the image coordinates. This transformation simulates variability in patient tilt and gantry angle during CT acquisition.

7. Random Brightness Adjustment: A random brightness scaling factor β was applied to each training image, [20] as given in Eq. (6):

$$x_{bright} = \beta \cdot x_{norm} \quad (6)$$

where x_{norm} is the normalized pixel value from Eq. (2), $\beta \sim \text{Uniform}(0.85, 1.15)$ is a brightness scaling factor sampled uniformly from $[0.85, 1.15]$, and x_{bright} is the brightness-adjusted output, clipped to $[0, 1]$ after scaling. This augmentation simulates variations in CT tube current and scanner exposure settings across different acquisition protocols [6]. Since the Kaggle dataset provides images as pre-converted 8-bit PNG files rather than raw DICOM files, the original HU-level intensity information has already been transformed during DICOM-to-PNG conversion. Within the PNG representation, modest brightness variation simulates acquisition variability. However, since the PNG pixel values cannot be confirmed as calibrated HU measurements, the precise effect of brightness scaling on diagnostically relevant intensity information is uncertain. The conservative $\pm 15\%$ range was selected to minimize this risk. Together, these four augmentation operations expand the effective diversity of the training set without collecting additional data, reducing the tendency of the model to overfit to specific imaging conditions present in the training subset.

D. Baseline Models and Proposed Architecture

This study follows a two-phase experimental design to evaluate the role of pretraining domain. In both phases, an identical classification head - consisting of Global Average Pooling, a Dense layer of 256 units with ReLU activation, Batch Normalization, Dropout of 0.5, and a final sigmoid output node - is attached to each backbone, ensuring performance differences are attributable solely to backbone architecture and pretraining source (Fig. 2). The last 30 layers of each backbone were kept trainable, while, all preceding

layers were frozen, with all other training settings identical across all six models.

1. Phase 1: In the Phase 1 three widely used CNN architectures are evaluated with weights pretrained on ImageNet - a large-scale dataset of over 1.2 million natural photographs spanning 1,000 object categories [21]. The architectures selected are ResNet50, InceptionV3, and DenseNet121, each representing a distinct design philosophy.

ResNet50: ResNet50 introduces residual (skip) connections to address the vanishing gradient problem that limits training of very deep networks. The output of each residual block is computed as given in Eq. (7):

$$x_{l+1} = F(x_l, \{W_l\}) + x_l \quad (7)$$

where x_l is the input to residual block l , $F(x_l, \{W_l\})$ is the residual mapping learned by two or three stacked convolutional layers with weight set $\{W_l\}$, and x_l is the identity shortcut connection added element-wise to the residual output. This formulation allows gradients to flow directly through the identity path during backpropagation, enabling effective training of networks up to hundreds of layers deep [22].

InceptionV3: InceptionV3 employs parallel convolutional branches of different kernel sizes within each Inception module, enabling the model to capture features at multiple spatial scales simultaneously. As in [23], the output of an Inception module is the channel-wise concatenation of all parallel branches, as defined in Eq. (8):

$$y = \text{Concat}(\text{Conv1} \times 1(x), \text{Conv3} \times 3(x), \text{Conv5} \times 5(x), \text{MaxPool}(x)) \quad (8)$$

where x is the input feature map, $\text{Conv1} \times 1$, $\text{Conv3} \times 3$, and $\text{Conv5} \times 5$ represent convolutions with kernel sizes 1×1 , 3×3 , and 5×5 respectively applied in parallel branches, $\text{MaxPool}(x)$ is a 3×3 max-pooling branch, and $\text{Concat}(\cdot)$ denotes channel-wise concatenation of all branch outputs. In InceptionV3, larger convolutions are factorized into sequences of smaller convolutions to reduce computational cost while maintaining receptive field coverage [23].

DenseNet121: DenseNet121 uses dense connectivity, where each layer receives concatenated feature maps from all preceding layers within a dense block. This is in contrast to ResNet's element-wise summation and enables full feature reuse throughout the network. As in [24] the output of layer l within a dense block is defined in Eq. (9):

$$x_l = H_l([x_0, x_1, \dots, x_{l-1}]) \quad (9)$$

where x_l is the output of layer l , $H_l(\cdot)$ is the composite function consisting of Batch Normalization ReLU

Conv2D (3×3) applied to the input, and $[x_0, x_1, \dots, x_{l-1}]$ denotes channel-wise concatenation of the feature maps from all preceding layers 0 through $l - 1$. This dense connectivity preserves low-level features throughout the network, which is particularly advantageous for tasks requiring fine-grained texture discrimination, such as hepatic steatosis classification from CT images [24].

The number of output feature maps at each dense layer is controlled by the growth rate k , [24] as given in Eq. (10):

$$\text{input channels}_l = k_0 + k * (l - 1) \quad (10)$$

where k_0 is the number of channels in the initial input to the dense block, k is the growth rate ($k = 32$ in DenseNet121), and l is the layer index within the block. Between dense blocks, transition layers reduce spatial dimensions and compress feature channels by a compression factor $\theta = 0.5$, [24] as given in Eq. (11):

$$m_{out} = \lfloor \varphi \cdot m_{in} \rfloor \quad (11)$$

where m_{in} is the number of input feature maps to the transition layer, m_{out} is the number of output feature maps, and $\varphi = 0.5$ is the compression ratio used in DenseNet121.

While these architectures have proven effective across a wide range of natural image tasks, their suitability for CT-based medical imaging is inherently limited by the domain gap between natural photographs and grayscale tissue density maps [25].

2. Phase 2: Phase 2 addresses this limitation directly. The same three architectures are re-evaluated using weights pretrained on RadImageNet [17], a large-scale medical imaging dataset containing 1.35 million annotated CT, MRI, and ultrasound images across 165 pathological categories, curated by trained radiologists. By replacing ImageNet initialization with RadImageNet initialization while keeping everything else identical, this phase isolates the effect of pretraining domain on classification performance. The best-performing model from this phase is selected as the proposed model for this study (see Fig. 2). An identical classification head is appended to each backbone in both phases. The head consists of four sequential operations: Global Average Pooling (GAP), a Dense layer with ReLU activation, Batch Normalization, Dropout, and a final sigmoid output layer. GAP reduces the $7 \times 7 \times 1024$ spatial feature map output of the backbone to a 1024-dimensional vector by computing the spatial average of each channel [24]. This operation discards spatial position information while retaining channel-wise feature strength, and is defined as in Eq. (12):

$$z_c = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W f_c(i, j) \quad (12)$$

where z_c is the pooled scalar output for channel c , H and W are the spatial height and width of the feature map, and $f_c(i, j)$ is the activation value at spatial position (i, j) in channel c . GAP produces a compact feature vector while retaining channel-wise information and is invariant to spatial input size. Batch Normalization: Batch Normalization is applied after the dense layer to stabilize the distribution of activations and accelerate training convergence, [44] as defined in Eq. (13):

$$\hat{x} = \frac{x - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}, y = \Gamma \hat{x} + \beta \quad (13)$$

where x is the input activation, μ_B and σ_B^2 are the mini-batch mean and variance, $\epsilon = 1 \times 10^{-5}$ is a small constant added for numerical stability, and Γ and β are learnable affine transformation parameters (scale and shift). During inference, population statistics accumulated during training replace the mini-batch statistics [27]. The fully connected Dense layer transforms the 1024-dimensional GAP output to a 256-dimensional representation, followed by ReLU activation, [27] as given in Eq. (14):

$$h = \text{ReLU}(W \cdot z + b) = \max(0, W \cdot z + b) \quad (14)$$

where z is the 1024-dimensional input vector from GAP, $W \in \mathbb{R}^{(256 \times 1024)}$ is the learnable weight matrix, $b \in \mathbb{R}^{256}$ is the bias vector, $h \in \mathbb{R}^{256}$ is the output activation, and $\max(0, \cdot)$ is the ReLU non-linearity that introduces non-linear feature transformation while avoiding the vanishing gradient problem [27]. Dropout is applied after Batch Normalization with a rate of $p = 0.5$ during training to prevent co-adaptation of neurons and reduce overfitting, [45] as defined in Eq. (15):

$$\tilde{x}_i = \begin{cases} 0 & \text{with probability } p \\ \frac{x_i}{1-p} & \text{with probability } 1-p \end{cases} \quad (15)$$

where x_i is the i -th neuron activation, $p = 0.5$ is the dropout probability, and $\tilde{x}_i$ is the masked output. The factor $\frac{1}{1-p}$ scales the surviving activations to preserve the expected output magnitude. During inference, dropout is disabled, and all neurons contribute with their full activations [27].

Sigmoid Output Layer: The final classification layer applies sigmoid activation to produce a probability output [27]. The sigmoid function maps any real-valued input to a continuous output in the range (0, 1), making it well-suited for binary classification tasks. The probability output $\hat{y}$ is defined as Eq. (16):

$$\hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}} \quad (16)$$

Algorithm 1. Training and evaluation procedure for proposed algorithm

Input: CT image dataset D with binary class labels; medical-domain pretrained backbone weights W_{med} ; hyperparameters: learning rate lr ; batch size B ; max epochs E ; early stopping patience p_{es} ; LR reduction patience p_{lr}

Output: Trained model M^* , classification predictions, Grad-CAM heatmaps

Initialization:

1. Split $D \rightarrow D_{train}, D_{val}, D_{test}$ using stratified splitting (seed = 42)
2. Compute class weights for each class c using Eq. (18)
3. Apply augmentation to D_{train} only Eq. (3) – (6); normalize all subsets Eq. (2)
4. Load backbone with W_{med} ; freeze early layers; keep last N layers trainable
5. Attach classification head: GAP $\rightarrow$ Dense($ReLU$) $\rightarrow$ BN $\rightarrow$ Dropout $\rightarrow$ Sigmoid Eq. (12) – (16)

Training:

6. Initialize Adam optimizer with lr (Eq. 19); set $best_score \leftarrow 0$, $no_improve \leftarrow 0$
7. **for** epoch = 1 **to** E **do**
8. **for** each mini-batch b in D_{train} **do**
9. Compute $\leftarrow \text{sigmoid}(M(b))$ using Eq. (16)
10. Compute weighted loss using Eq. (17)
11. Update via backpropagation using Eq. (19)
12. **end for**

Validation:

13. Evaluate M on $D_{val} \rightarrow$ compute val_score (AUC)
14. **if** $val_score > best_score$ **then** $best_score \leftarrow val_score$; save M^* ; $no_improve \leftarrow 0$
15. **else** $no_improve \leftarrow no_improve + 1$
16. **if** $no_improve \geq p_{es}$ **then** restore M^* ; **break** (Early Stopping)
17. **if** plateau(val_score, p_{lr}) **then** $lr \leftarrow lr \times 0.5$ (ReduceLROnPlateau)
18. **end for**

Testing:

19. Load best checkpoint M^*
20. **for** each image x in D_{test} **do**
21. Compute $\leftarrow \text{sigmoid}(M^*(x))$ using Eq. (16)
22. Apply threshold: $\hat{c} \leftarrow$ Steatosis **if** $> \tau$, **else** Healthy Eq. (20)
23. **end for**
24. Compute Accuracy, Se, Sp, Precision, F1, AUC using Eq. (21) – (27)

Grad-CAM Explainability:

25. **for** each representative case in $\{TP, TN, FP, FN\}$ **do**
26. Compute gradient weights
27. Generate class activation heatmaps
28. Normalize and overlay heatmap on original CT image
29. **end for**

where z is the pre-activation scalar output of the final dense layer and $\hat{y} \in (0, 1)$ is the predicted probability of Hepatic Steatosis. The classification decision is applied using a threshold of τ : images with $\hat{y} \geq \tau$ are classified as Hepatic Steatosis (positive class), and those with $\hat{y} < \tau$ are classified as Healthy (negative class).

3. Training Configuration

All models were trained using the Adam optimizer with an initial learning rate of 1×10^{-4} [26] and weighted binary cross-entropy loss [27]. Class weights were computed inversely proportional to class frequency - 1.1037 for Healthy and 0.9141 for Hepatic Steatosis - to handle the mild class imbalance [28]. Three callbacks guided training: EarlyStopping monitored validation AUC with patience of 8 epochs;

ReduceLROnPlateau halved the learning rate when validation AUC plateaued for 4 consecutive epochs with a floor of 1×10^{-7} [29]; and ModelCheckpoint saved weights corresponding to the highest validation AUC. Together, these callbacks ensured that training terminated at the optimal point, avoided overfitting through adaptive learning rate reduction, and preserved the best-performing model weights for final evaluation.

E. Loss Function, Optimization Objective, and Training Algorithm

As explained in [27], the weighted binary cross-entropy loss function assigns higher penalty to misclassified minority-class samples and is used for training, as defined in Eq. (17):

Table 1. Training and validation performance of the ImageNet-pretrained DenseNet121 model across training epochs.

Epoch	Train Loss	Val Loss	Val Acc.	Val AUC	Val Prec.	Val Recall
1	0.8643	0.7821	58.37%	0.7214	0.5831	0.9124
10	0.5231	0.5312	74.56%	0.8241	0.7321	0.8512
20	0.4124	0.4312	81.23%	0.8734	0.8124	0.8512
30	0.3512	0.3621	85.43%	0.9043	0.8512	0.8712
40	0.2943	0.3124	88.56%	0.9287	0.8843	0.8912
50	0.2534	0.2812	91.01%	0.9512	0.9124	0.9187

Table 2. Training and validation performance of the RadImageNet-pretrained DenseNet121 (RadDenseNet) model across training epochs.

Epoch	Train Loss	Val Loss	Val Acc.	Val AUC	Val Prec.	Val Recall
1	0.7991	0.6508	64.98%	0.8135	0.6148	0.9640
10	0.4533	0.3912	84.11%	0.9359	0.8026	0.9409
20	0.3341	0.2748	90.01%	0.9750	0.9760	0.8380
30	0.2612	0.1871	93.81%	0.9860	0.9726	0.9126
40	0.2145	0.1438	95.08%	0.9919	0.9836	0.9254
50	0.1798	0.1065	96.34%	0.9958	0.9814	0.9512

$$L_w = -\frac{1}{N} \sum_{i=1}^N [w_i y_i (\hat{y}_i) + w_o (1 - y_i) \log \log (1 - \hat{y}_i)] \tag{17}$$

where N is the number of training samples, $y_i \in \{0,1\}$ is the true label, $\hat{y}_i$ is the predicted probability from Eq. (16), $w_1 = 0.9141$ is the class weight for Hepatic Steatosis, and $w_0 = 1.1037$ is the class weight for Healthy. As in [28], class weights are computed inversely proportional to class frequency as defined in Eq. (18):

$$w_c = \frac{N}{C.N_c} \tag{18}$$

where N is the total number of training samples, $C = 2$ is the number of classes, and N_c is the number of samples in class C . The fine-tuning optimization minimizes the weighted loss over the trainable parameters θ_c - comprising the last 30 backbone layers and the classification head - while keeping the frozen RadImageNet pretrained parameters θ_{frozen} fixed, [34] as defined in Eq. (19):

$$\theta_{ft} = \arg \arg \min L_w(f(x; \theta_{frozen}, \theta_{ft}) y) \tag{19}$$

where $f(x; \theta_{frozen}, \theta_{ft})$ is the model output, θ_{frozen} , are the RadImageNet pretrained weights held fixed throughout training, and θ_{ft} are the parameters updated by the Adam optimizer [26]. As in [33], the binary classification decision is applied using an optimized threshold τ , as defined in Eq. (20):

$$\hat{c} = \begin{cases} \text{Hepatic steatosis} & \hat{y} \geq \tau \\ \text{Healthy} & \hat{y} < \tau \end{cases} \tag{20}$$

where $\hat{y}$ is the sigmoid output from Eq. (16) and τ is the decision threshold optimized on the validation set by maximizing F1-score across all candidate threshold values. The optimal threshold identified was $\tau = 0.48$, which was applied to the held-out test set without further adjustment to avoid any form of test-set leakage. Algorithm 1 defines the complete pseudocode for training and evaluation.

F. Grad-CAM Explainability

Gradient-weighted Class Activation Mapping (Grad-CAM) was applied to test images from the proposed model to support clinical interpretability [30]. Grad-CAM computes the gradient of the predicted class score with respect to the last convolutional layer feature maps, producing a spatial heatmap where warmer regions indicate higher model attention. This allows clinicians to verify whether the model attends to the hepatic parenchyma rather than unrelated background structures, providing a level of transparency important for clinical trust [31].

G. Performance metrics

All models were evaluated on the held-out test set using six standard metrics [32]. Accuracy measures the overall proportion of correctly classified images across both classes. Sensitivity (Se) measures the model's ability to correctly identify true Hepatic Steatosis cases - a low Se means steatosis cases are missed, losing

Table 3. Performance comparison of ImageNet- and RadImageNet-pretrained models on the test set.

Model	Weights	Acc (%)	Se	Sp	Precision	F1-score	AUROC
ResNet50	ImageNet	87.36	0.8654	0.8812	0.8923	0.8787	0.9287
InceptionV3	ImageNet	90.45	0.8976	0.9124	0.9187	0.9080	0.9523
DenseNet121	ImageNet	91.57	0.9089	0.9231	0.9276	0.9182	0.9612
ResNet50	RadImageNet	93.26	0.9278	0.9371	0.9412	0.9344	0.9741
InceptionV3	RadImageNet	94.38	0.9401	0.9472	0.9518	0.9459	0.9821
DenseNet121 (RadDenseNet)	RadImageNet	96.35	0.9538	0.9752	0.9789	0.9662	0.9966

the opportunity for early intervention. Specificity (Sp) measures the ability to identify truly Healthy cases, minimizing unnecessary follow-up investigations correctly. Precision reflects the reliability of positive predictions - the proportion of predicted steatosis cases that are truly steatotic. F1-score provides a single balanced measure combining precision and sensitivity via their harmonic mean. AUROC evaluates discriminative ability across all classification thresholds and is threshold-independent, making it particularly informative in clinical deployment scenarios [33]. Model calibration was assessed using the Expected Calibration Error (ECE), which quantifies the weighted average absolute difference between predicted confidence and observed accuracy across ten equal-width bins [46]. An ECE close to zero indicates well-calibrated predictions.

III. Result

A. Experimental Setup

All experiments were implemented in TensorFlow 2.19 with Keras using Python 3.12 on the Kaggle platform with two NVIDIA T4 GPUs (16 GB VRAM each) using MirroredStrategy, giving an effective batch size of 64. RadImageNet-pretrained DenseNet121 weights were sourced from <https://github.com/BMEII-AI/RadImageNet>. RadDenseNet contains 7,301,185 total parameters - 904,577 trainable and 6,396,608 frozen. All augmentation transformations were applied with $p = 0.5$. EarlyStopping used patience = 8; ReduceLROnPlateau used patience = 4, factor = 0.5, and minimum lr = 1×10^{-7} . Total training time was approximately 35 minutes across 50 epochs. The learning rate was reduced from 1×10^{-4} to 5×10^{-5} at epoch 43. All random operations used seed 42 for reproducibility. Source code and trained weights are available upon request to the corresponding author.

B. Training and Validation Results

All six models were trained under identical conditions across 50 epochs. In Phase 1 ImageNet-pretrained

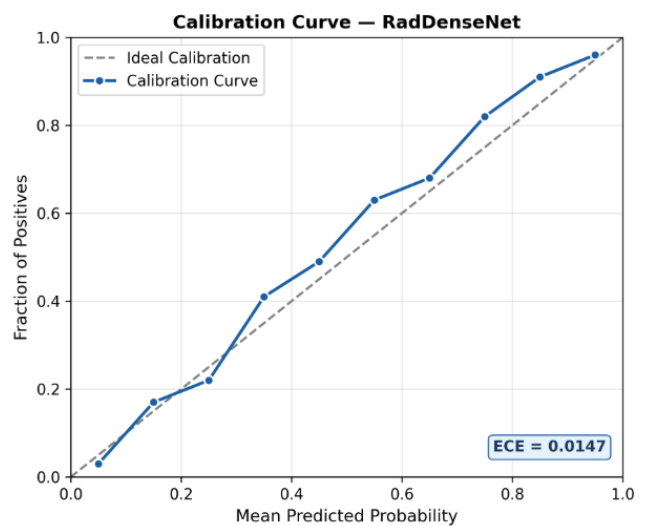


Fig. 3. Calibration curve of the proposed RadDenseNet model on the held-out test set, with an expected calibration error of 0.0147.

models, DenseNet121 demonstrated the strongest convergence, reaching a validation AUC of 0.9512 and accuracy of 91.01% at epoch 50 (Table 1), compared with the other two models. In Phase 2, replacing ImageNet with RadImageNet, weights, i.e., RadDenseNet produced consistently faster convergence and higher validation performance across all three architectures with a validation AUC of 0.9958 and a validation loss of 0.1065 at epoch 50 (Table 2).

C. Test Set Results

The best checkpoint for each of the six trained models was evaluated on the completely held-out test set of 356 images. Hepatic Steatosis is defined as the positive class (label = 1) and Healthy as the negative class (label = 0). The decision threshold τ was optimized on the validation set by maximizing F1-score, yielding $\tau = 0.48$, applied to the test set without further adjustment.

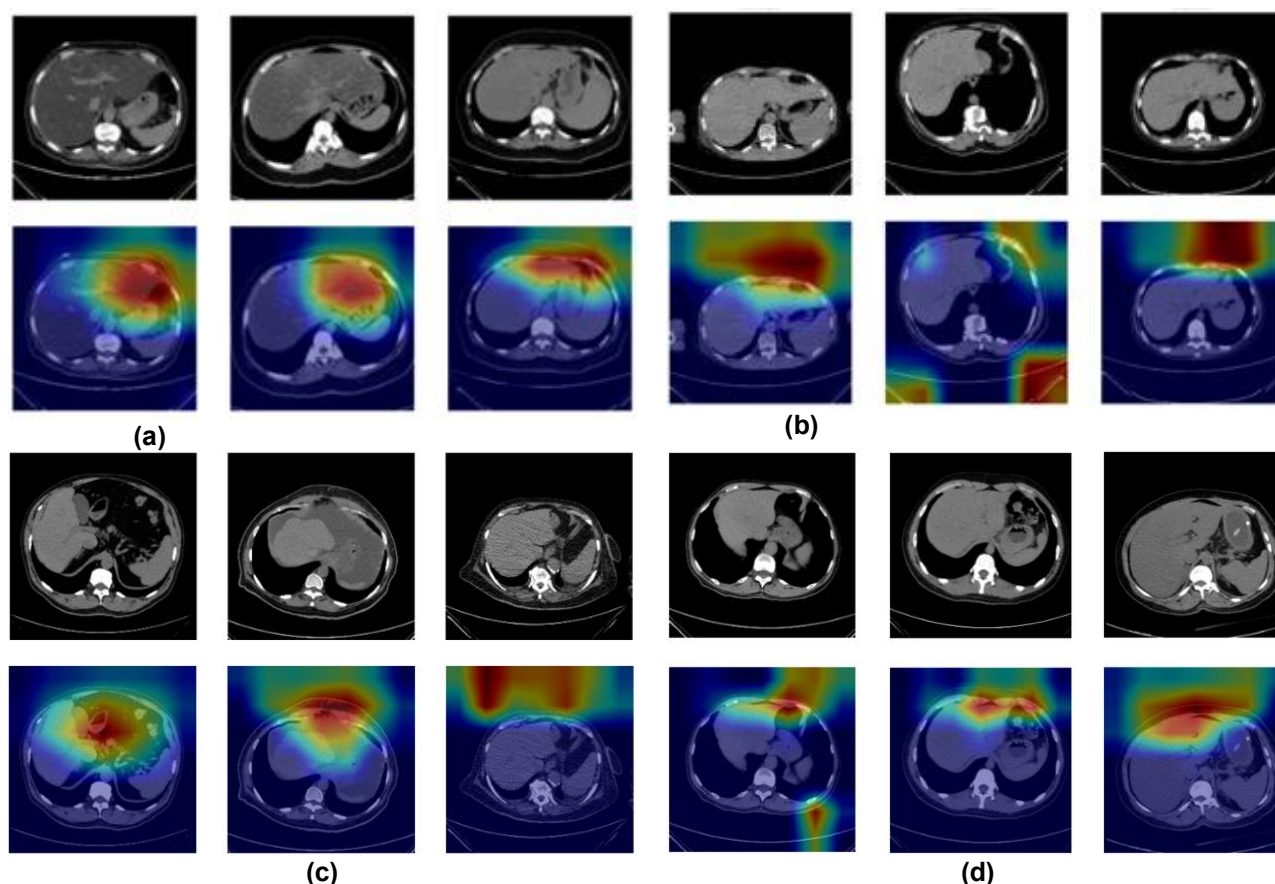


Fig. 4. Grad-CAM visualizations of representative test cases: (a) true positives, (b) true negatives, (c) false positives, and (d) false negatives. The upper row shows the original CT images and the lower row shows the corresponding Grad-CAM activation maps.

Phase 1 results showed a clear performance hierarchy among ImageNet-pretrained models (Table 3). ResNet50 achieved 87.36% accuracy (AUC 0.9287), InceptionV3 achieved 90.45% (AUC 0.9523), and DenseNet121 achieved 91.57% accuracy, sensitivity 0.9089, specificity 0.9231, and AUC 0.9612.

Phase 2 demonstrated consistent improvement across all architectures when ImageNet was replaced with RadImageNet weights (Table 3). ResNet50 improved from 87.36% to 93.26% (AUC 0.9741) and InceptionV3 from 90.45% to 94.38% (AUC 0.9821), confirming that medical-domain pretraining provides more relevant feature initialization than natural image pretraining. RadDenseNet achieved the best performance across all metrics - 96.35% accuracy, sensitivity 0.9538, specificity 0.9752, precision 0.9789, F1-score 0.9662, and AUC 0.9966. The low false negative count is clinically significant in a screening context [37], [36]. Two clear trends emerge: RadImageNet pretraining consistently outperforms ImageNet pretraining with accuracy gains of 2.81–5.90 percentage points, and DenseNet121 is the most effective backbone under both pretraining conditions. Model calibration was assessed using ECE on the test

set (Fig. 3) with ECE = 0.0147, confirming well-calibrated probability estimates.

D. Grad-CAM Visualization

Grad-CAM heatmaps were generated for all four prediction categories - true positives, true negatives, false positives, and false negatives - with representative cases shown in Fig. 4. For correctly classified Steatosis cases (Fig. 4(a)), activation concentrated predominantly over the hepatic parenchyma, confirming that RadDenseNet responds to clinically relevant regions rather than background structures. For correctly classified Healthy cases (Fig. 4(b)), activation was more diffuse and lower in intensity, consistent with the absence of focal steatosis-related texture changes. For false-positive cases (Fig. 4(c)), attention extended beyond the hepatic parenchyma into periabdominal adipose tissue and the liver-diaphragm interface, suggesting the model responded to extra-hepatic low-attenuation tissue sharing HU characteristics with steatotic parenchyma. This highlights a key limitation of processing full CT slices without liver ROI segmentation. For false-negative cases (Fig. 4(d)), two failure patterns were identified - weak dispersed parenchymal activation in borderline

steatosis cases, and misdirected attention toward extra-hepatic structures such as the diaphragm and lung base. Both failure modes point to the absence of explicit liver segmentation as a preprocessing step and could be addressed through liver ROI extraction and a lower decision threshold tuned for clinical screening sensitivity.

IV. Discussion

This study proposed RadDenseNet for binary classification of hepatic steatosis from abdominal CT images. On the held-out test set of 356 images, the model achieved 96.35% accuracy, sensitivity of 0.9538, and F1-score of 0.9662 - confirming the effectiveness of the proposed approach.

architectures (Table 3). Switching from ImageNet to RadImageNet initialization alone - without any other change - improved DenseNet121 accuracy from 91.57% to 96.35% and AUC from 0.9612 to 0.9966. This consistent improvement across three architecturally distinct backbones confirms that the benefit originates from the pretraining domain itself rather than any architecture-specific property [17], [25]. ImageNet models learn color gradients, object edges, and shape outlines - features largely absent in CT imaging [25]. CT images encode tissue density in HU, and the diagnostically relevant signal for hepatic steatosis lies in subtle parenchymal attenuation differences [6]. RadImageNet, trained on 1.35 million annotated medical images, produces feature

Table 4. Comparison of the proposed RadDenseNet model with previously reported CT-based hepatic steatosis detection methods.

Reference	Year	Datasize	Modality	Task	Val. design	method	Acc (%)	AUC	Se
Graffy et al. [15]	2019	Population-level VA CT; exact N not reported	Non-contrast abdominal CT	Binary (steatosis present/absent)	Single-site; no patient-level CV	segmentation + HU thresholding	90.10	-	0.870
Zhang et al. [13]	2024	1,014 images; 8 countries; multinational	Non-contrast chest CT	Binary (steatosis present/absent)	Multi-site validation; no patient-level CV reported	DL parenchyma segmentation (nnU-Net) + liver attenuation measurement	91.20	0.943	-
Zhang et al. [14]	2025	840 scans; single-center	Non-contrast abdominal CT	4-class grading	Single-site; train/test split; no CV	2D/3D radiomics + RF / DL	86.40	0.973	0.873
Leiby et al. [16]	2024	Private clinical dataset; exact N not reported	3D unenhanced abdominal CT	Binary (fatty/non-fatty)	Single-site; private dataset; no CV reported	3D data-driven end-to-end DL classification	88.50	0.931	0.871
Kim et al. [50]	2025	906 patients; contrast-enhanced CT; Asan Medical Center	Contrast-enhanced CT	Binary (moderate-severe)	Single-center; independent validation cohort	DL automated liver-spleen attenuation measurement	-	0.868	0.827
Nivethitha et al. [51]	2025	Meta-analysis; 29 AI studies; multiple modalities	CT + MRI + US	Binary (pooled)	Systematic review; pooled estimates	Pooled AI diagnostic accuracy	-	-	0.860
RadDenseNet (Proposed)	2026	3,557 images; Kaggle public dataset	Non-contrast abdominal CT	Binary (Healthy/Steatosis)	Stratified 70/20/10 split + 5-fold CV + repeated runs (3 seeds)	RadImageNet pretrained DenseNet121; end-to-end 2D classification	96.35	0.9966	0.9538

A. Impact of Pretraining Domain

The most revealing finding is the consistent performance advantage of RadImageNet pretraining over ImageNet pretraining across all three

representations already sensitive to these patterns [17], requiring far less adaptation during fine-tuning [35].

B. Architecture Selection

Among the three tested architectures, DenseNet121 consistently outperformed ResNet50 and InceptionV3 under both pretraining conditions. We offer a motivated hypothesis for this observation, though it is not experimentally demonstrated. DenseNet121 employs dense connectivity where each layer receives concatenated feature maps from all preceding layers [24] - plausibly advantageous for hepatic steatosis classification where the diagnostic signal lies in subtle HU-level attenuation differences rather than high-level semantic features. ResNet50 uses element-wise summation [22], which does not explicitly preserve earlier feature maps, and InceptionV3's parallel multi-scale convolutions [23] are designed for spatial scale diversity rather than fine-grained texture preservation. Other factors - parameter count, gradient flow dynamics, and architecture-RadImageNet compatibility - may equally contribute to the observed hierarchy. The architectural interpretation presented here should be read as a plausible hypothesis motivated by established literature [24], [35] rather than an experimentally verified conclusion, and formal ablation studies are identified as a direction for future work.

C. Comparison with State-of-the-Art

Table 4 compares RadDenseNet against six published hepatic steatosis detection studies. Graffy et al. [15] achieved 90.10% accuracy using a segmentation-based liver attenuation pipeline. Kim et al. [50] reported an AUC of 0.868 on contrast-enhanced CT. Zhang et al. [13] achieved an AUC of 0.943 using nnU-Net segmentation on 1,014 multinational chest CT images. Zhang et al. [14] reported an AUC of 0.973 using radiomics for four-category grading - a more complex task than binary classification. Leiby et al. [16] reported 88.50% accuracy using 3D end-to-end classification on private CT data. A 2025 systematic review [51] reported pooled sensitivity of 0.860 and specificity of 0.890 across multiple modalities. Against all of these, RadDenseNet achieved the highest accuracy (96.35%), sensitivity (0.9538), specificity (0.9752), and AUC (0.9966) among CT-based binary classification approaches. However, direct comparison must be interpreted with caution, as these studies differ substantially in dataset source, imaging protocol, task complexity, and validation design. Nevertheless, RadDenseNet is the only method among those listed to combine direct end-to-end 2D CT classification, medical-domain pretraining, validation-set threshold optimization, and Grad-CAM explainability covering all four prediction categories - addressing methodological gaps consistently absent from the existing literature.

D. Limitations

The limitations are organized into three categories for clarity. Dataset Limitations: The Kaggle Liver CT Image Dataset does not include clinical metadata such as patient demographics, BMI, liver function tests, or

histological confirmation of steatosis grade [42]. The labelling methodology is undocumented - the absence of a verified reference standard means the model may reproduce uncertain annotations rather than clinically confirmed diagnoses. Patient-level identifiers are unavailable, preventing patient-level splitting; the image-level split may allow slice-level correlation, potentially inflating reported metrics [48]. The dataset originates from a single source with no documented scanner specifications, limiting generalizability. Images are provided as pre-converted 8-bit PNG files without DICOM metadata; PNG pixel values cannot be confirmed as calibrated HU measurements, limiting applicability to raw DICOM-based clinical workflows. Methodological Limitations: The binary classification setup does not capture steatosis severity grading, which carries direct implications for treatment planning [41]. The augmentation pipeline was not validated through a formal ablation study; vertical flipping introduces anatomically implausible CT orientations and brightness adjustment may alter diagnostically relevant pixel-intensity ratios. Grad-CAM heatmaps were generated without prior liver segmentation, occasionally producing activation beyond the liver boundary [30], and the absence of pixel-level segmentation masks prevented quantitative localization analysis such as Dice coefficient or IoU overlap [39]. Grad-CAM analysis revealed two failure modes - extra-hepatic attention in false positives and weak parenchymal activation in false negatives - both attributable to the absence of liver ROI segmentation. A fixed layer count of 30 trainable layers represents different proportions across architectures - 17% for ResNet50, 10% for InceptionV3, and 7% for DenseNet121 - limiting the comparison design [35]. Clinical Translation Limitations: RadDenseNet is a research prototype evaluated on a single publicly available dataset, and results should be interpreted as preliminary. Dataset shift poses a well-recognized risk - CT attenuation values vary across scanner models, tube voltage settings, reconstruction kernels, and slice thickness, and the model's robustness to such variability has not been assessed. The dataset consists exclusively of non-contrast CT images; performance on contrast-enhanced acquisitions cannot be assumed. Population differences in demographics, BMI, ethnicity, and steatosis aetiology may further affect generalizability. Model calibration, while acceptable on the current test set, may degrade under domain shift. Prospective validation on multi-center data with radiologist-confirmed ground truth and formal ethical oversight is a necessary prerequisite before any clinical application. In future, we plan to incorporate a pre-trained liver segmentation model as preprocessing to constrain model attention to the hepatic parenchyma; extend the framework to multi-class steatosis severity grading for clinically actionable output; conduct

a formal ablation study to evaluate the contribution of each augmentation transformation; architecture-specific fine-tuning optimization through independent hyperparameter search [35]; external validation on multi-center clinical datasets spanning diverse scanner types, contrast protocols, and patient demographics; and a formal reader study comparing RadDenseNet predictions with expert radiologist assessments alongside radiologist evaluation of Grad-CAM heatmap quality.

V. Conclusion

This study proposed RadDenseNet, a DenseNet121 model initialized with RadImageNet weights pretrained on 1.35 million medical images, for automated binary classification of hepatic steatosis from abdominal CT images. The proposed model achieved superior performance with an accuracy of 96.35%, AUC-ROC of 0.9966, sensitivity of 0.9538, specificity of 0.9752, precision of 0.9789, and F1-score of 0.9662 on the held-out test set of 356 images. Notably, replacing ImageNet pretraining with RadImageNet weights in the same DenseNet121 architecture improved accuracy from 91.57% to 96.35% - a gain of 4.78 percentage points - and AUC from 0.9612 to 0.9966, demonstrating that medical-domain pretraining alone yields a substantial and statistically significant performance improvement. Beyond its strong predictive performance, Grad-CAM analysis confirmed that model attention was concentrated over the hepatic parenchyma in correctly classified cases, lending clinical credibility to the predictions. For future work, the model should be validated on larger and more diverse multi-center clinical datasets spanning diverse scanner types and contrast protocols. Enhancements such as liver ROI segmentation as preprocessing, extension to multi-class steatosis severity grading, formal ablation studies, and integration into clinical decision support systems are also recommended to ensure practical deployment in radiology workflows.

Acknowledgment

The authors acknowledge the use of the Kaggle Liver CT Image Dataset and the Kaggle platform for providing the computational resources used in this study.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

The publicly available dataset can be accessed at <https://www.kaggle.com/datasets/ucimachinelearning/liver-ct-image-dataset>. Source code and trained model

weights will be made available upon reasonable request to the corresponding author.

Author Contribution

Author contributions are reported using the CRediT (Contributor Roles Taxonomy) standard:

Rahul LR: Conceptualization, Visualization, methodology design, algorithm development, and manuscript writing.

SP Gaddamedi: Data analysis, Validation, preprocessing, manuscript review, and critical revision.

Declarations

Ethical Approval

This study uses the publicly available Kaggle Liver CT Image Dataset, which contains de-identified and anonymized abdominal CT images. No new patient data were collected, no clinical interventions were performed, and no direct patient contact was involved. Consequently, no additional institutional ethics committee approval was required for this study. The dataset is used strictly for research purposes in accordance with its publicly available terms of access.

Consent for Publication Participants.

Consent for publication was given by all participants.

Dataset License and Ethical Provenance

The Kaggle Liver CT Image Dataset is publicly available under the Kaggle community data sharing terms. The dataset was published by UCI Machine Learning on the Kaggle platform and is accessible without restriction for non-commercial research use. The original ethical approval for data collection, de-identification, and public release was the responsibility of the dataset curators; no independent ethical clearance was sought or required for this secondary analysis study. The dataset is used in this work solely for the purpose of developing and evaluating a research prototype for hepatic steatosis classification, consistent with its intended reuse conditions for academic research.

Competing Interests

The authors declare no competing interests.

Declaration of Generative Artificial Intelligence and AI-Assisted Technologies

During the development and preparation of this study, the authors used AI tool to assist with software development, code debugging, documentation, and language refinement. All generated code and manuscript content were reviewed and validated by the authors. Experimental design, data processing, statistical analysis, interpretation of findings, and final scientific conclusions remain the responsibility of the

authors. The authors take full responsibility for the integrity and reproducibility of the reported research.

References

- [1] Z. M. Younossi, P. Golabi, J. M. Paik, A. Henry, C. Van Natta, and L. Younossi, "The global epidemiology of nonalcoholic fatty liver disease (NAFLD) and nonalcoholic steatohepatitis (NASH): a systematic review," *Hepatology*, vol. 77, no. 4, pp. 1335–1347, 2023, doi: 10.1097/HEP.0000000000000004.
- [2] K. Riazzi, H. Azhari, J. H. Charette, F. E. Underwood, J. A. King, E. E. Afshar, M. G. Swain, S. S. Congly, G. G. Kaplan, and F. Shaheen, "The prevalence and incidence of NAFLD worldwide: a systematic review and meta-analysis," *Lancet Gastroenterol. Hepatol.*, vol. 7, no. 9, pp. 851–861, 2022, doi: 10.1016/S2468-1253(22)00165-0.
- [3] D. Georgescu, D. F. Lighezan, C. I. Rosca, O. M. Bratu, and C. Diaconu, "NASH/NAFLD-related hepatocellular carcinoma: an added burden," *Life (Basel)*, vol. 14, no. 1, p. 25, 2023, doi: 10.3390/life14010025.
- [4] M. J. Zhu, X. H. Sun, and X. M. Li, "Pathogenesis of NAFLD-related hepatocellular carcinoma: an up-to-date review," *J. Hepatocell. Carcinoma*, vol. 10, pp. 347–356, 2023, doi: 10.2147/JHC.S400231.
- [5] J. Starekova, D. Hernando, P. J. Pickhardt, and S. B. Reeder, "Quantification of liver fat content with CT and MRI: state of the art," *Radiology*, vol. 301, no. 2, pp. 250–262, 2021, doi: 10.1148/radiol.2021204288.
- [6] H. N. Kim, H. J. Jeon, H. G. Choi, I. S. Kwon, W. S. Ryu, T. H. Lee, S. H. Kim, and B. S. Lee, "CT-based Hounsfield unit values reflect the degree of steatohepatitis in patients with low-grade fatty liver disease," *BMC Gastroenterol.*, vol. 23, no. 1, p. 77, 2023, doi: 10.1186/s12876-023-02706-4.
- [7] S. Survarachakan, R. Prasad, Z. S. A. Rahman, J. de Frutos, F. Lindseth, and T. L. A. van Walsum, "Deep learning for image-based liver analysis — a comprehensive review focusing on malignant lesions," *Artif. Intell. Med.*, vol. 130, p. 102331, 2022, doi: 10.1016/j.artmed.2022.102331.
- [8] K. Xiang, B. Jiang, and D. Shang, "The overview of the deep learning integrated into the medical imaging of liver," *Hepatol. Int.*, vol. 15, no. 4, pp. 868–880, 2021, doi: 10.1007/s12072-021-10229-z.
- [9] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017, doi: 10.1016/j.media.2017.07.005.
- [10] D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, D. P. Naidich, and S. Shetty, "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," *Nat. Med.*, vol. 25, no. 6, pp. 954–961, 2019, doi: 10.1038/s41591-019-0447-x.
- [11] S. Deepak and P. M. Ameer, "Brain tumor classification using deep CNN features via transfer learning," *Comput. Biol. Med.*, vol. 111, p. 103345, 2019, doi: 10.1016/j.combiomed.2019.103345.
- [12] D. Alzu'bi, A. Abdullah, M. Hmeidi, R. AlAzab, M. Gharaibeh, A. El-Heis, A. Almotairi, A. Forestiero, C. Yuan, and G. Merigó, "Kidney tumor detection and classification based on deep learning approaches: a new dataset in CT scans," *J. Healthc. Eng.*, vol. 2022, p. 3861161, 2022, doi: 10.1155/2022/3861161.
- [13] Z. Zhang, G. Li, Z. Wang, F. Xia, N. Zhao, H. Nie, Z. Ye, J. S. Lin, Y. Hui, and X. Liu, "Deep-learning segmentation to select liver parenchyma for categorizing hepatic steatosis on multinational chest CT," *Sci. Rep.*, vol. 14, p. 11987, 2024, doi: 10.1038/s41598-024-62887-2.
- [14] H. Zhang, J. Liu, D. Su, Z. Bai, Y. Wu, Y. Ma, Q. Miao, M. Wang, and X. Yang, "Diagnostic of fatty liver using radiomics and deep learning models on non-contrast abdominal CT," *PLoS ONE*, vol. 20, no. 2, p. e0310938, 2025, doi: 10.1371/journal.pone.0310938.
- [15] P. M. Graffy, V. Sandfort, R. M. Summers, and P. J. Pickhardt, "Automated liver fat quantification at nonenhanced abdominal CT for population-based steatosis assessment," *Radiology*, vol. 293, no. 2, pp. 334–342, 2019, doi: 10.1148/radiol.2019190512.
- [16] J. S. Leiby, M. E. Lee, E. K. Choe, and D. Kim, "Data-driven classification of fatty liver from 3D unenhanced abdominal CT scans," in *Proc. Machine Learning in Medical Imaging (MLMI 2023)*, Lecture Notes in Computer Science, vol. 14349, Springer, Cham, pp. 264–272, 2024, doi: 10.1007/978-3-031-45676-3_27.
- [17] X. Mei, Z. Liu, P. M. Robson, B. Marinelli, M. Huang, A. Doshi, A. Jacobi, C. Cao, K. E. Link, T. Yang, Y. Wang, H. Greenspan, T. Deyer, Z. A. Fayad, and Y. Yang, "RadImageNet: an open radiologic deep learning research dataset for effective transfer learning," *Radiol. Artif. Intell.*,

- vol. 4, no. 5, p. e210315, 2022, doi: 10.1148/ryai.210315.
- [18] UCI Machine Learning, "Liver CT Image Dataset," *Kaggle*, 2024. [Online]. Available: <https://www.kaggle.com/datasets/ucimachinelearning/liver-ct-image-dataset>.
- [19] J. D. Kelleher, B. Mac Namee, and A. D'Arcy, *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*, 2nd ed. MIT Press, Cambridge, MA, USA, 2020, ISBN: 978-0262044691.
- [20] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, p. 60, 2019, doi: 10.1186/s40537-019-0197-0.
- [21] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "ImageNet: a large-scale hierarchical image database," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Miami, FL, USA, pp. 248–255, 2009, doi: 10.1109/CVPR.2009.5206848.
- [22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.
- [23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 2818–2826, 2016, doi: 10.1109/CVPR.2016.308.
- [24] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 4700–4708, 2017, doi: 10.1109/CVPR.2017.243.
- [25] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: understanding transfer learning for medical imaging," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, Vancouver, Canada, pp. 3347–3357, 2019. [Online]. Available: <https://proceedings.neurips.cc/paper/2019/hash/eb1e78328c46506b46a4ac4a1e378b91-Abstract.html>.
- [26] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *Proc. Int. Conf. Learning Representations (ICLR)*, San Diego, CA, USA, 2015. [Online]. Available: <https://openreview.net/forum?id=8gmWwjFyLj>.
- [27] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016. [Online]. Available: <https://www.deeplearningbook.org>.
- [28] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Netw.*, vol. 106, pp. 249–259, 2018, doi: 10.1016/j.neunet.2018.07.011.
- [29] L. N. Smith, "Cyclical learning rates for training neural networks," in *Proc. IEEE Winter Conf. Applications of Computer Vision (WACV)*, Santa Rosa, CA, USA, pp. 464–472, 2017, doi: 10.1109/WACV.2017.45.
- [30] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, 2020, doi: 10.1007/s11263-019-01228-7.
- [31] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.
- [32] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, p. e0118432, 2015, doi: 10.1371/journal.pone.0118432.
- [33] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.
- [34] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010, doi: 10.1109/TKDE.2009.191.
- [35] A. Raza, M. S. Alshehri, S. Almakdi, A. A. Siddique, M. Alsulami, and M. Alhaisoni, "Enhancing brain tumor classification with transfer learning: leveraging DenseNet121 for accurate and efficient detection," *Int. J. Imaging Syst. Technol.*, vol. 34, no. 1, p. e22957, 2024, doi: 10.1002/ima.22957.
- [36] S. A. Harrison and C. P. Day, "Benefits of lifestyle modification in NAFLD," *Gut*, vol. 56, no. 12, pp. 1760–1769, 2007, doi: 10.1136/gut.2006.112094.
- [37] P. J. Pickhardt, P. M. Graffy, M. G. Zea, A. T. Sandfort, and R. M. Summers, "Utilizing fully automated abdominal CT-based biomarkers for opportunistic screening for metabolic syndrome," *AJR Am. J. Roentgenol.*, vol. 215,

- no. 5, pp. 1173–1179, 2020, doi: 10.2214/AJR.19.22537.
- [38] R. Benjamins, P. Dhunoo, and B. Meijboom, "The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database," *NPJ Digit. Med.*, vol. 3, no. 1, p. 118, 2020, doi: 10.1038/s41746-020-00324-0.
- [39] P. Bilic, P. Christ, H. B. Li et al., "The liver tumor segmentation benchmark (LiTS)," *Med. Image Anal.*, vol. 84, p. 102680, 2023, doi: 10.1016/j.media.2022.102680.
- [40] A. L. Rinella, M. E. Lazarus, V. Ratzu, F. M. Francque, A. Sanyal, L. Kanwal, et al., "A multisociety Delphi consensus statement on new fatty liver disease nomenclature," *Hepatology*, vol. 78, no. 6, pp. 1966–1986, 2023, doi: 10.1097/HEP.0000000000000520.
- [41] M. Ekstedt, H. Hagstrom, P. Nasr, M. Fredrikson, P. Stal, R. Kechagias, and S. Hultcrantz, "Fibrosis stage is the strongest predictor for disease-specific mortality in NAFLD after up to 33 years of follow-up," *Hepatology*, vol. 61, no. 5, pp. 1547–1554, 2015, doi: 10.1002/hep.27368.
- [42] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study," *PLOS Med.*, vol. 15, no. 11, p. e1002686, 2018, doi: 10.1371/journal.pmed.1002686.
- [43] M. Q. Pembury Smith and G. D. Ruxton, "Effective use of the McNemar test," *Behav. Ecol. Sociobiol.*, vol. 74, no. 11, p. 133, 2020, doi: 10.1007/s00265-020-02916-z.
- [44] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Machine Learning (ICML)*, Lille, France, vol. 37, pp. 448–456, 2015. [Online]. Available: <http://proceedings.mlr.press/v37/ioffe15.html>.
- [45] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014. [Online]. Available: <https://jmlr.org/papers/v15/srivastava14a.html>.
- [46] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Machine Learning (ICML)*, Sydney, Australia, vol. 70, pp. 1321–1330, 2017. [Online]. Available: <http://proceedings.mlr.press/v70/guo17a.html>.
- [47] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 4th ed. Pearson, New York, USA, 2018, ISBN: 978-0133356724.
- [48] S. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PLOS ONE*, vol. 14, no. 11, p. e0224365, 2019, doi: 10.1371/journal.pone.0224365.
- [49] B. Norgeot et al., "Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist," *Nat. Med.*, vol. 26, no. 9, pp. 1320–1324, 2020, doi: 10.1038/s41591-020-1041-y.
- [50] H. Y. Kim, K. J. Lee, S. S. Lee, S. J. Choi, D. H. Kim, S. Heo, H. J. Jang, and S. H. Choi, "Diagnosis of moderate-to-severe hepatic steatosis using deep learning-based automated attenuation measurements on contrast-enhanced CT," *Abdom. Radiol.*, vol. 50, no. 9, pp. 4139–4147, 2025, doi: 10.1007/s00261-025-04872-5.
- [51] V. Nivethitha, R. A. Daniel, A. Debnath, V. Dwarakanathan, G. Jeer, and G. Kavipriya, "Diagnostic accuracy of artificial intelligence models for imaging detection of hepatic steatosis through systematic review and meta-analysis," *Sci. Rep.*, vol. 15, no. 1, p. 34408, 2025, doi: 10.1038/s41598-025-17386-3.

Author Biography



Rahul LR received the B.Tech. degree in Electronics and Communication Engineering from Kerala University, Kerala, India, in 2013, the M.Tech. degree in Electronics and Communication Engineering from the National Institute of Technology (NIT), Jalandhar, India, in 2015, and the Ph.D. degree in Electrical Engineering from the Indian Institute of Technology (IIT) Hyderabad, India, in 2025. Since 2025, he has been serving as an Assistant Professor in the Department of Electronics and Communication Engineering at Koneru Lakshmaiah Education Foundation (KLEF), Bowrampet, Hyderabad, India. His research interests span biomedical signal and image processing, deep learning, and machine learning applications in healthcare.



Satya Prasad Gaddamedi received the B.E. degree in Electronics and Communication Engineering from Jawaharlal Nehru Technological University, India, in 2015, and the M.Tech. degree in Electrical, Electronics and Communication

Engineering from Osmania University, Hyderabad, India, in 2020. He is currently pursuing the Ph.D. degree in Speech Signal Processing at the Indian Institute of Technology (IIT) Kharagpur, India. Since 2025, he has been serving as an Assistant Professor in the Department of Electronics and Communication Engineering at Koneru Lakshmaiah Education Foundation (KLEF), Bowrampet, Hyderabad, India. His research interests include speech signal processing and enhancement, deep learning, and machine learning applications.