Kasu Journal of Computer Science Vol.3 No.1 [March, 2026], pp. 015-031
https://doi.org//10.47514/kjcs/2026.3.1.002

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

Quantization-Aware Pruned MobileNetV2 for Real-Time Liver Tumor Classification on CT Images

Amina Bala Jaafaru[1], A. F. D. Kana[2], M. Abdullahi[2]
[1]Department of Informatics, Kaduna State University
[2]Department of Computer Science, Ahmadu Bello University, Zaria
aminabala@kasu.edu.ng[1], donfackkana@abu.edu.ng[2], abdullahilwafu@abu.edu.ng[2]

Abstract

Early diagnosis of liver cancer is critical due to its high mortality rate and late clinical presentation. Computed tomography (CT) remains the standard modality for liver lesion assessment; however, interpretation requires experts that are skilled and prone to diagnostic variability. Deep learning has enhanced liver lesion classification accuracy, though large models come with computational burdens that hinder deployment in clinical and resource-limited setting. This study introduces an optimized MobileNetV2 architecture for liver tumor classification, integrating unstructured pruning and quantization-aware training (QAT) to reduce the model size and latency while maintaining diagnostic performance. Using LiTS17 CT dataset, 17,506 slices were extracted and labeled as tumor or non-tumor. A Baseline finetuned MobileNetV2 achieved 99.6% accuracy but required 26.10 MB and 8 FPS processing. After pruning and weight stripping, size reduced to 8.91 MB with a 31.57 FPS inference rate and 94.8% accuracy. Further QAT yielded an 8-bit TensorFlow Lite model with 8.46 MB size and 94.5% accuracy. These results demonstrate that lightweight, compressed models can support real-time liver tumor diagnosis.

Keywords: Liver cancer, Computed tomography, MobileNetV2, Quantization-aware training, Pruning, LiTS17.

1. Introduction

Liver cancer remains one of the most prevalent malignancies worldwide, contributing significantly to global cancer mortality (Bray et al., 2021). CT scans play a crucial role in liver cancer diagnostics due to their ability to capture abnormal vascular contrast variations and lesion morphology (Hassan et al., 2023). However, manual interpretation is time-intensive and subject to inter-observer variability, particularly in early-stage tumors where visual differences may be subtle (Tang et al., 2020). Consequently, computer-aided diagnosis systems based on deep learning have increasingly been adopted to support radiological decision-making. Deep learning, especially Convolutional Neural Networks (CNNs), has demonstrated outstanding performance in image classification and medical imaging tasks. CNN-based models have been successfully applied to tumor detection, segmentation, and classification. Addressing Challenges with Deep Learning effective diagnosis and treatment of liver cancer are hindered by challenges such as the difficulty in accurately identifying tumors in medical imaging scans due to poor contrast and complex liver structure (McMahon et al.,2023). Utilizing advanced image processing techniques, particularly CNNs, can aid in improving the accuracy of liver cancer diagnosis, potentially reducing errors and biases in medical assessments. CNNs can learn subtle features in CT scans that might be missed by humans, potentially

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

leading to more accurate diagnoses. CNNs are less susceptible to bias compared to human interpretation (Wei et al.,2023). CNNs have achieved notable success in liver lesion recognition, yet larger architectures such as VGG and ResNet require significant computational resources, limiting real-time clinical deployment (Chen et al., 2022). Lightweight CNNs like MobileNet are designed for low-resource environments through depth-wise separable convolutions, but further optimization is needed for low-latency diagnostic applications (Howard et al., 2019). Model compression techniques such as pruning and quantization enable faster inference and reduced storage cost with minimal performance loss, particularly for medical imaging tasks (Xu et al., 2023; Kim et al., 2022). These limitations highlight the need for efficient, lightweight, and robust CNN architectures that can operate effectively on limited hardware while maintaining high diagnostic accuracy, particularly for tasks such as liver cancer detection in low resource clinical environments. This study presents an optimized, deployment efficient MobileNetV2 model for liver cancer classification using unstructured pruning and quantization-aware training (QAT). The approach aims to reduce model size and inference time while maintaining clinically relevant diagnostic accuracy, offering a viable solution for real-time medical decision support systems.

2. Related Works

Deploying convolutional neural networks (CNNs) in clinical practice requires balancing accuracy, computational efficiency, and hardware constraints. While CNNs have shown strong performance in medical image analysis, their clinical translation is often limited by high computational costs, particularly for high-resolution images like CT and MRI data, which demand significant memory and processing power (Tian et al., 2025). These challenges are even more pronounced in resource-limited settings and on edge devices, where processing capacity and energy budgets are restricted (Kumar et al., 2025).

Deep learning applications to liver cancer have predominantly focused on segmentation rather than classification, with models such as UNet derivatives demonstrating improved lesion localization on CT (Zhou et al., 2021). Classification studies using VGG16, ResNet50, or DenseNet have shown promising accuracy but often incur high inference cost unsuitable for rapid deployment in hospital workflows (Chen et al., 2022). Lightweight alternatives such as MobileNet and ShuffleNet have reduced complexity, though many implementations lack compression strategies and remain suboptimal for embedded medical devices (Howard et al., 2019; Wong et al., 2020). To address these barriers, researchers have investigated model compression techniques such as pruning and quantization and explainable methods including Grad-CAM and integrated gradients. These approaches can yield lightweight, transparent models with minimal loss in diagnostic performance, improving suitability for real-world, resource-constrained deployment (Yang et al., 2022; Malik et al., 2025). Lightweight architectures such as MobileNetV2 further enhance model efficiency by reducing inference time and memory usage while preserving diagnostic utility. Several recent studies highlight the effectiveness of lightweight models across medical imaging tasks. (Ahmad et al., 2024) developed SkipGateNet, a compact CNN-LSTM hybrid achieving 99.91% accuracy with an 8 ms inference time, demonstrating feasibility for real-time use on low-power devices. In liver imaging, (Alawneh et al.,2022) introduced LiverNet, an eight-layer CNN achieving 95.6% accuracy, 100% when paired with an SVM though limited architectural detail may impede replication. (Sindhura et al., 2022) proposed Efficient-PSP-Net for CT liver segmentation, combining PSP modules with EfficientNet to balance context understanding and computational efficiency. Similarly, (Alici Karaca et al., 2022) designed a lightweight CNN for classifying radiation induced liver disease, outperforming deeper models, though without pruning

Kasu Journal of Computer Science Vol.3 No.1 [March, 2026], pp. 015-031
https://doi.org//10.47514/kjcs/2026.3.1.002
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

or quantization to further enhance deployability. MobileNet based models have been increasingly explored for liver and lung imaging tasks due to their efficiency. (Raghava et al., 2023) and (Valeri et al., 2023) found MobileNet to provide competitive accuracy for CT protocol optimization, although hybrid models introduce higher computational cost. (Hameed et al., 2023) similarly reported MobileNet as a practical alternative to more accurate but resource intensive models like ResNet and VGG16. For segmentation, (Riaz et al., 2023) integrated MobileNetV2 with U-Net, achieving a Dice score of 89.75% while maintaining computational efficiency. Overall, lightweight architectures particularly MobileNetV2 offer a promising pathway toward clinically viable deep learning systems by enabling rapid inference, reduced memory consumption, and broader use on portable or low power devices. These advantages are especially relevant for CT based liver cancer classification and other applications where real-time or point-of-care deployment is critical (Sandler et al., 2018) Pruning has been widely explored to reduce model redundancy by removing less significant parameters, with structured pruning offering hardware friendly improvements for compressed CNNs (Xu et al., 2023). Quantization-aware training (QAT) allows models to operate at low numeric precision (e.g., 8-bit) with minimal accuracy degradation, showing promise for diagnostic imaging tasks (Kim et al., 2022). Yet, few studies combine pruning and QAT for liver CT classification, and fewer evaluate deployment level performance such as inference speed and memory footprint. This work addresses these gaps by integrating pruning and QAT into MobileNetV2 to produce a lightweight, accurate, and real-time capable liver tumor classification model.

3. Methodology

This study follows an **experimental research design.** The approach is to train a deep learning model (MobileNetV2) to classify CT image slices as either containing a tumor or not. To improve efficiency, unstructured pruning was applied to remove redundant weights based on magnitude, introducing sparsity while preserving the overall model structure, and **quantization-aware training** to reduce memory usage and speed up inference, without significantly affecting accuracy. The model is evaluated using multiple performance metrics including accuracy, precision, recall, F1-score, inference time, and model size.

3.1 Dataset description

This study utilized the Liver Tumor Segmentation Challenge dataset (LiTS17), which contains contrast enhanced abdominal CT scans from multiple hospitals and imaging protocols (Bilic et al., 2019). Each case includes a full 3D volume with expert provided liver and tumor masks. From the 130 available volumes, relevant 2D slices were extracted from the 3D CT volumes, a total of **17,506 2D slices** were extracted and labeled into tumor and non-tumor classes based on pixel level annotations. Slices containing ≥1 pixel labeled as tumor were classified as tumor slices, while others were labeled non tumor. A slice-wise split was applied: 70% training, 15% validation, and 15% testing. All slices were resized to a uniform resolution and normalized to standardize pixel intensity.

Kasu Journal of Computer Science Vol.3 No.1 [March, 2026], pp. 015-031
https://doi.org//10.47514/kjcs/2026.3.1.002

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

Figure 1: Sample of the LiTS17 Dataset (Bilic et al., 2023).

3.2 Data Preparation and Pre-processing

Prior to model training, as shown in Figure 2, the CT slices and their corresponding masks were loaded using Python libraries such as `os`, TensorFlow, Keras preprocessing utilities, and image-handling functions. All images were resized to **224 × 224 pixels**, matching the input requirement of MobileNetV2. Pixel intensities were normalized to a **[0,1] range** to ensure stable network convergence. Slice labels were assigned automatically based on the segmentation masks, where any slice containing tumor-labelled pixels was categorized as **"tumor",** while all others were labeled as **"no-tumor."** Labeling was automated using the LiTS17 segmentation masks: a binary tumor label was assigned to any CT slice where the corresponding mask contained at least one tumor pixel. This yields a simple binary classification target. A total of 17506 CT slices were extracted and categorized into tumor slices of 10770 and no-tumor slices of 6736 classes. (Bilic et al., 2023). To improve model generalization and minimize overfitting, data augmentation was applied, including random horizontal and vertical flips, rotational transforms, and brightness/contrast adjustments, consistent with medical augmentation standards. Finally, the dataset was split using a **slice-wise partitioning strategy**, allocating **70%** of the images for training, **15%** for validation, and **15%** for testing.

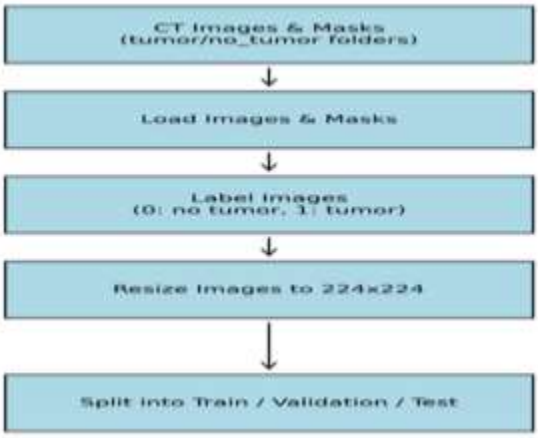

Figure 2: **Dataset Splitting Process**

Kasu Journal of Computer Science Vol.3 No.1 [March, 2026], pp. 015-031
https://doi.org//10.47514/kjcs/2026.3.1.002

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

3.3 Model Architecture

Figure 3 illustrates the overall workflow of the proposed liver tumor classification framework based on a Quantization-Aware MobileNetV2 architecture. The process begins with CT image dataset loading followed by several preprocessing steps, including image resizing, normalization, noise removal, filtering, and contrast enhancement to improve image quality and feature visibility. The dataset is then divided into training, validation, and testing sets, after which image augmentation techniques such as rotation, flipping, and scaling are applied to increase data variability and improve model generalization. A baseline MobileNetV2 model is first trained to establish initial classification performance and learn relevant image features. The trained model is subsequently optimized using unstructured weight pruning, followed by retraining to maintain model accuracy. Next, quantization-aware training (QAT) is performed to simulate low-precision computations during training. Finally, the optimized model is converted to TensorFlow Lite (INT8) format for efficient inference, producing the final classification outputs of tumor and no-tumor from CT images.

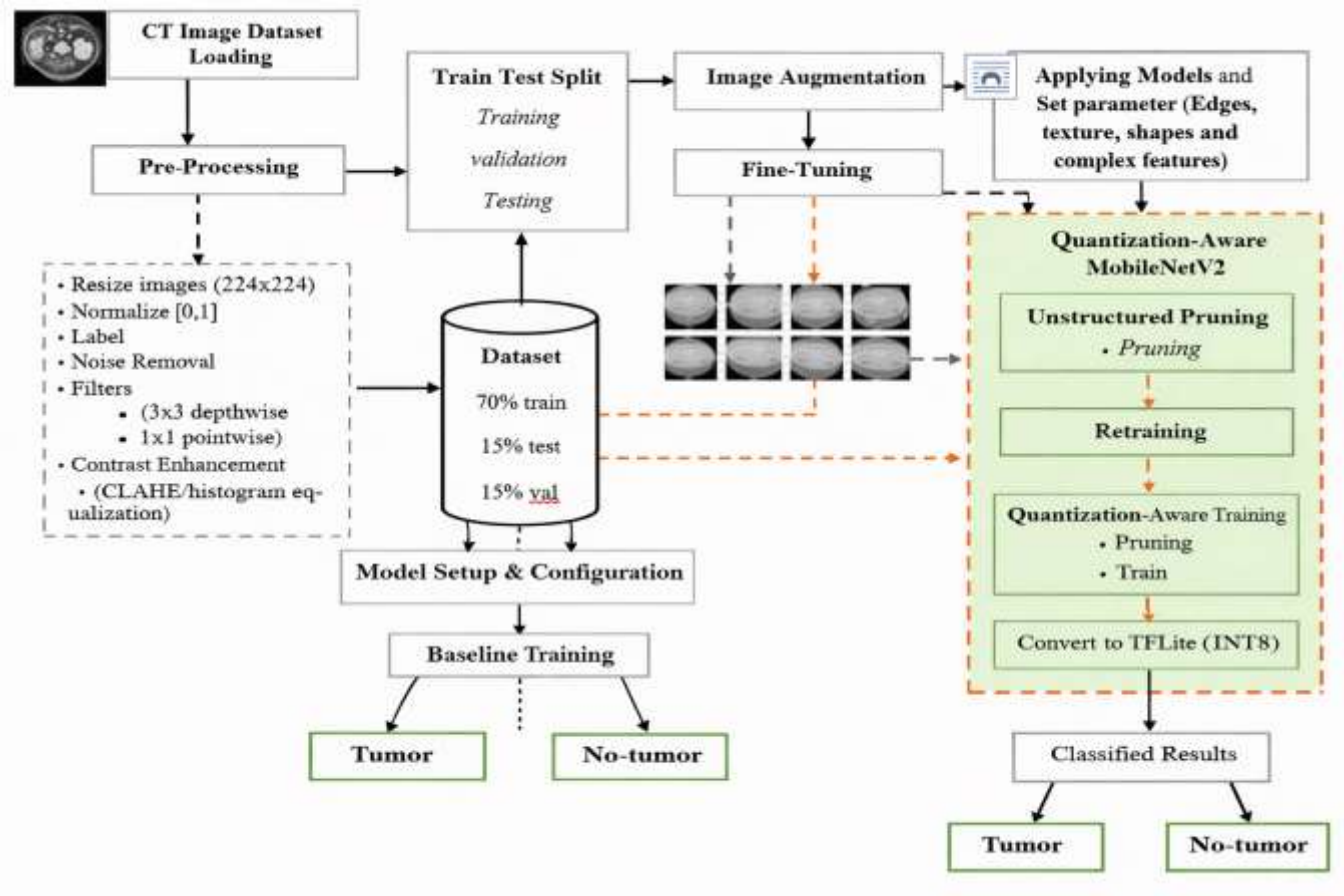

Figure 3: Research Model Architecture

Kasu Journal of Computer Science Vol.3 No.1 [March, 2026], pp. 015-031
https://doi.org//10.47514/kjcs/2026.3.1.002
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

3.4 Pruning method used (unstructured pruning)

Unstructured pruning (weight-level sparsification) was applied using TensorFlow Model Optimization pruning callbacks. Pruning followed a **gradual sparsity schedule** during training, low-magnitude weights were progressively set to zero until a final target sparsity of 20% was reached. After the pruning schedule completed, **post-pruning fine-tuning was performed** to recover any lost accuracy, and finally executed **weight stripping** to physically remove zeroed weights from the saved model so that model-size reductions were realized on disk. Note that unstructured sparsity does not automatically provide hardware acceleration unless the sparse model is stripped or executed on sparse-aware kernels; weight-stripping plus quantization were used to obtain actual size and inference improvements.

3.5 Quantization-Aware Training (QAT)

Following pruning and fine-tuning, quantization-aware training (QAT) was used to prepare the model for 8-bit integer inference. Fake-quantization nodes were inserted into weights and activations during retraining to simulate low-precision arithmetic. After a short QAT fine-tune to stabilize scale parameters, the model was converted to an 8-bit TensorFlow Lite (TFLite) representation with representative calibration data to preserve activation ranges. This produced the final TFLite model used for deployment benchmarking.

3.6 Training Procedure

The model was trained using the **Sparse Categorical Cross-Entropy loss**, which is appropriate for binary image classification tasks with integer labels. Optimization was performed using the **Adam optimizer** with an initial learning rate of **0.0001(**1e-4) and a **batch size of 32**. Base fine-tuning training was conducted for up to **30 epochs**, depending on convergence. A **ReduceLROnPlateau** scheduler was applied to automatically decrease the learning rate whenever the validation loss stopped improving, promoting more stable convergence. To prevent overfitting, **early stopping was implemented based on validation loss rather than validation accuracy**, as loss provides a more sensitive reflection of model learning by incorporating prediction confidence, whereas accuracy may remain unchanged even if the model is overfitting. Monitoring validation loss also aligns directly with the optimization objective of minimizing cross-entropy. Additionally, **checkpointing was used to save the best-performing model based on validation accuracy**, while saving weights at each epoch to facilitate further fine-tuning and compression experiments. This training configuration ensured both optimal convergence and reliable selection of the most generalizable model prior to pruning and quantization steps. Pruning phase followed the same optimizer settings, pruning used a gradual schedule embedded within training and was followed by a fine-tuning phase until validation performance converged.

3.7 Standard MobileNetV2 Convolutional Block

The diagram in Figure 4 explains MobileNetV2 which is based on inverted residual blocks with linear bottlenecks, designed to provide high accuracy with low computational cost. Each convolutional block consists of three stages: a 1×1 expansion convolution that increases feature dimensionality, a 3×3 depthwise convolution that performs efficient spatial filtering, and a 1×1 linear projection that reduces the channels to

Kasu Journal of Computer Science Vol.3 No.1 [March, 2026], pp. 015-031
https://doi.org//10.47514/kjcs/2026.3.1.002
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

the desired output size. Batch normalization and ReLU6 activation are applied after the expansion and depthwise layers to maintain stability and efficiency. When input and output dimensions match, a residual connection is used to improve gradient flow and convergence. This design enables an effective balance between model efficiency and representational power, making MobileNetV2 well-suited for resource-constrained applications.

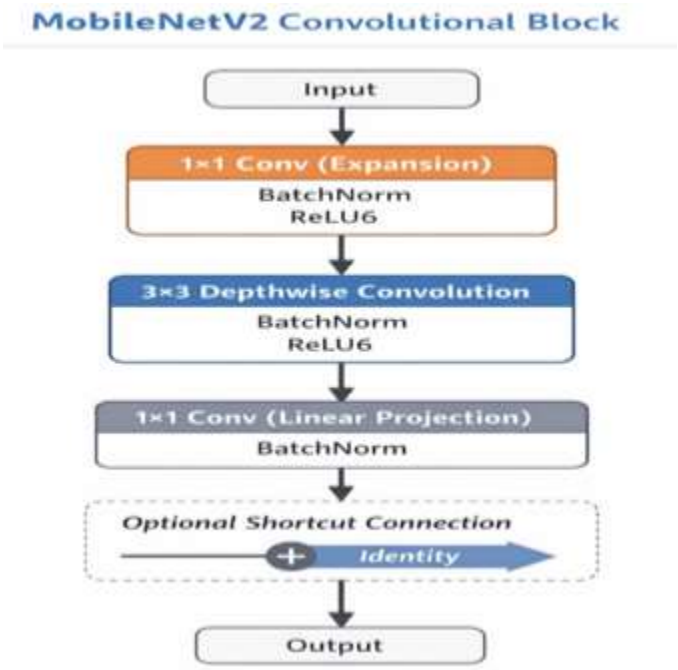

Figure 4: Standard MobileNetV2 Inverted Residual Block (Mark Sandler et al., 2018)

3.8 Modified MobileNetV2 Convolutional Neural Network Algorithm

Although MobileNetV2 is inherently efficient, additional optimization is necessary for deployment in resource-constrained medical imaging environments. This study proposes a modified MobileNetV2 architecture that integrates **unstructured weight pruning** and **quantization-aware training (QAT)** within its convolutional blocks as shown in figure 5. Unstructured pruning removes less significant weights to reduce parameter redundancy and computational cost without altering the network structure. Additionally, QAT simulates low-precision operations during training, enabling robust conversion to low-bit representations for efficient inference. The resulting model preserves the core MobileNetV2 design while achieving reduced memory usage and latency. In contrast to the standard MobileNetV2 block shown in Figure 4, which operates using full-precision weights and dense connections, the modified architecture presented in Figure 5 introduces unstructured pruning and quantization-aware training to further enhance computational efficiency. These modifications enable significant reductions in model size, memory footprint, and inference time, making the proposed approach suitable for deployment in real-world clinical environments.

Kasu Journal of Computer Science Vol.3 No.1 [March, 2026], pp. 015-031
https://doi.org//10.47514/kjcs/2026.3.1.002
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

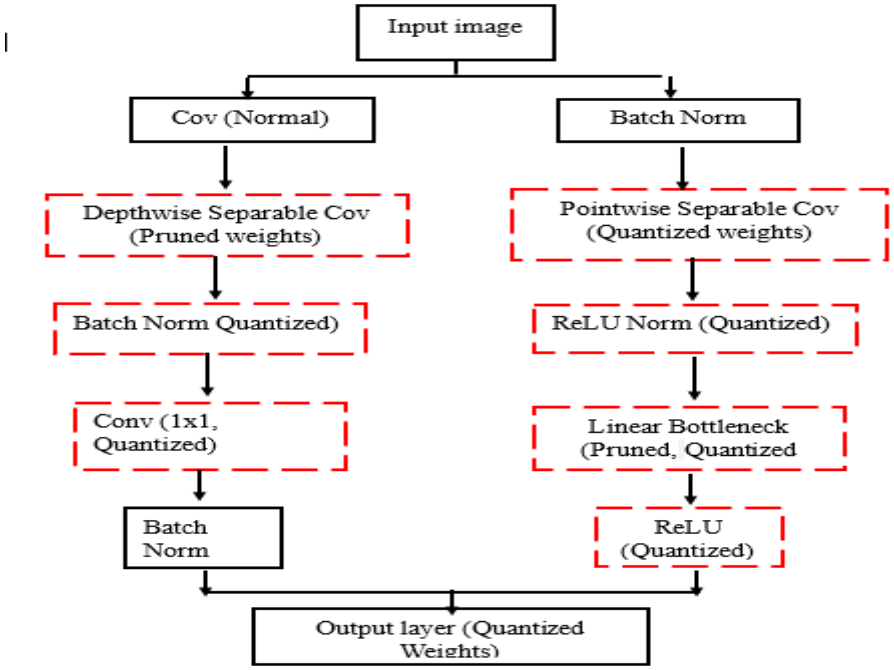

Figure 5: Modified MobileNetV2 CNN Algorithm for liver cancer classification in CT scan

4. Results and discussion

The model was trained using early stopping based on validation loss, which helped prevent overfitting.

4.1 Classification Results for optimized Mobilenetv2 model

Table 1 presents the performance of the fine-tuned, pruned, and quantized MobileNetV2 models developed for liver cancer classification using CT slices. The fine-tuned model achieved the highest performance, with an accuracy of 99.6% and near-perfect precision, recall, and F1-score of 0.99 each. After applying unstructured pruning, performance slightly decreased to 94.8% accuracy, but remained clinically relevant while retaining strong precision and recall values. The quantization-aware trained TFLite model showed comparable results of 94.5% accuracy, demonstrating that compression preserved discriminative capability. These findings confirm that model compression substantially reduces size while maintaining high classification quality.

Table 1. Classification Performance of Optimized MobileNetV2 Models

Model	Accuracy (%)	Precision	Recall	F1score	AUC
Finetuned mobilenetv2	99.6%	0.99	0.99	0.99	0.9999
Pruned and stripped	94.8%	0.96	0.95	0.96	0.9887
QAT TFLite	94.5%	0.96	0.95	0.96	0.9859

Kasu Journal of Computer Science Vol.3 No.1 [March, 2026], pp. 015-031
https://doi.org//10.47514/kjcs/2026.3.1.002
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

4.2 Model Efficiency Results of Optimized Mobilenetv2 Model

The deployment efficiency of MobileNetV2 was significantly improved through unstructured pruning and 8-bit quantization, as shown in Table 2. The fine-tuned model had a size of 26.10 MB and an inference speed of 8 FPS. Applying unstructured pruning by removing less important weights (20% weight sparsity) while largely preserving the network's representational capacity, reduced the model size by 65.9% to 8.91 MB and increased inference speed to 31.57 FPS. Further quantization to 8-bit TFLite reduced the model size slightly to 8.46 MB (67.6% reduction from the original) with minimal impact on speed which remained high at 31.42 FPS. It should be noted that the dataset was **split slice-wise**, meaning slices from the same patient could appear in multiple splits. While this allows more training samples, it may slightly inflate reported metrics. Overall, these results demonstrate that unstructured pruning and quantization effectively optimize MobileNetV2 for deployment on resource-constrained devices, providing a practical balance between model size, inference speed, and predictive performance.

Table 2. Deployment performance of the fine-tuned, pruned, and quantized MobileNetV2 models.

Model	Model Size (MB)	Size Reduction (%)	Inference Speed (FPS)
Mobilenetv2 (fine-tuned)	26.10	-	8.00
Pruned Mobilenetv2	8.91	65.9	31.57
Quantized Mobilenetv2 (8-Bit TFLite)	8.46	67.6	31.42

4.3 Combined Bar Chart of Model Evaluation Metrics

Figure 6 presents a combined visualization of all evaluation metrics for the three models: the fine-tuned MobileNetV2, the pruned and stripped model, and the QAT-TFLite model. Panel A compares the core performance metrics (accuracy, precision, recall, and F1-score), showing that the fine-tuned model achieved the highest values across all metrics, largely due to its dense architecture and slice-wise data separation. Panels B and C highlight the improvements from model optimization. Pruning reduced the model size from 26.10 MB to 8.91 MB, while quantization further reduced it to 8.46 MB. This reduction directly translated into faster inference speeds, increasing from 8 FPS in the baseline model to over 31 FPS in both optimized versions. Panel D displays the sparsity levels applied, showing that only the pruned and QAT models used 20% weight sparsification. Overall, the combined visualization clearly demonstrates the trade-off between predictive performance and computational efficiency, emphasizing that pruning and QAT significantly enhanced deployability with only a modest reduction in classification accuracy.

Kasu Journal of Computer Science Vol.3 No.1 [March, 2026], pp. 015-031
https://doi.org//10.47514/kjcs/2026.3.1.002

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

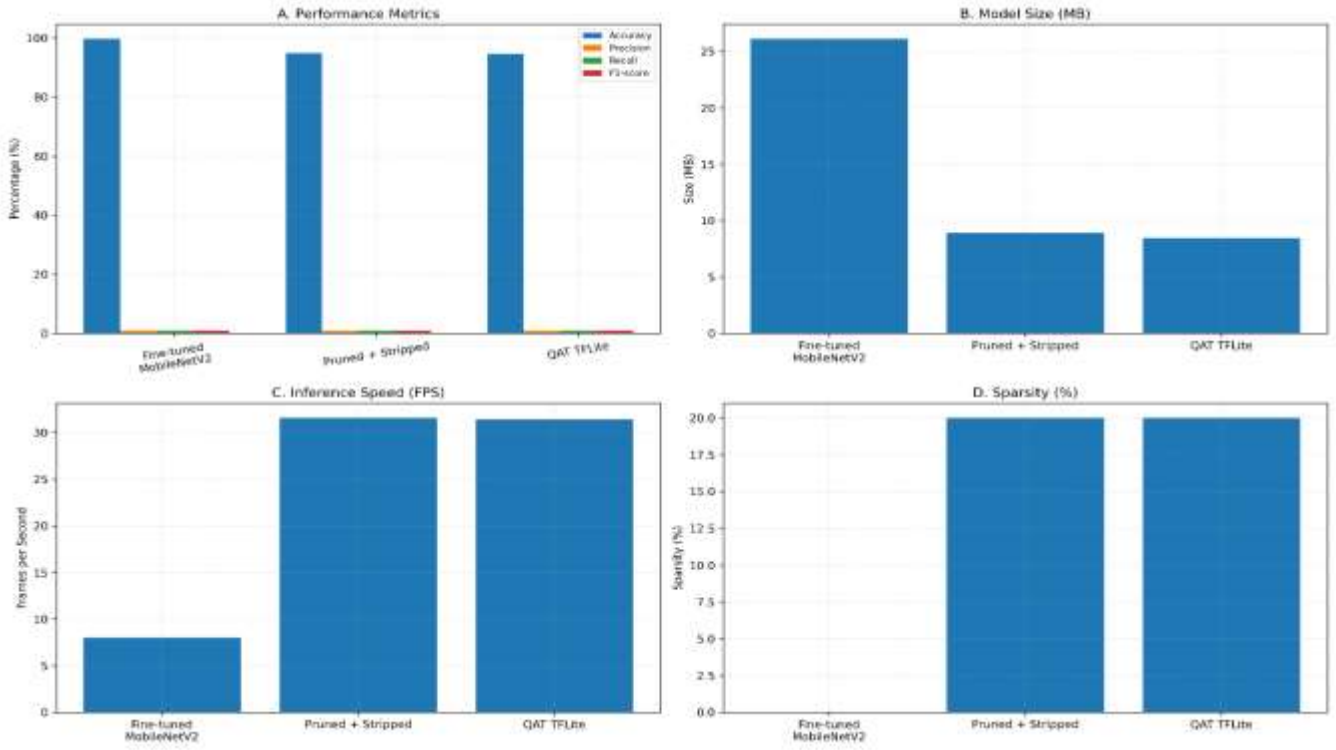

Figure 6: Performance Overview of the Optimized Liver Cancer Classification Model

4.4 Confusion Matrix Results *of Optimized Mobilenetv2 Model*

The confusion matrix results in Table 3 provide a detailed view of the classification performance of the MobileNetV2 models. The fine-tuned model achieved 2,040 true positives and 1,281 true negatives, with only 9 false positives and 6 false negatives, reflecting its very high accuracy of 99.6%, with precision, recall, and F1-score of 0.99 each. After unstructured pruning, there is a slight decrease in true positives and true negatives, results include TP = 1,025 and TN = 634, with FP = 40 and FN = 52, while the 8-bit quantized TFLite model achieved similar values of TP = 1,019, TN = 636, FP = 38, and FN = 58. These small increases in misclassifications are consistent with the modest reduction in accuracy of the pruned and TFLite quantized models with 94.8 and 94.5% respectively with F1-score of 0.96 after compression. These results demonstrate that MobileNetV2 can maintain robust predictive performance for liver cancer classification while being highly optimized for deployment on resource-constrained devices.

Table 3: Confusion matrix values for the fine-tuned, pruned, and quantized MobileNetV2 models.

Model	TP	TN	FP	FN
Mobilenetv2 (fine-tuned)	2040	1281	9	6
Pruned Mobilenetv2	1025	634	40	52
Quantized Mobilenetv2 (8-Bit TFLite)	1019	636	38	58

Kasu Journal of Computer Science Vol.3 No.1 [March, 2026], pp. 015-031
https://doi.org//10.47514/kjcs/2026.3.1.002

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

4.5 Auc/Roc Curve and Confusion Matrix for the Fine-Tuned MobileNetV2

Figure 7 shows the evaluation of the fine-tuned MobileNetV2 model on the test set. Having ROC curve with **AUC = 0.9999**, indicate near-perfect discriminative ability between tumor and non-tumor slices. Confusion matrix showing true positives (TP) = 2,040, true negatives (TN) = 1,281, false positives (FP) = 9, and false negatives (FN) = 6, highlighting minimal misclassifications.

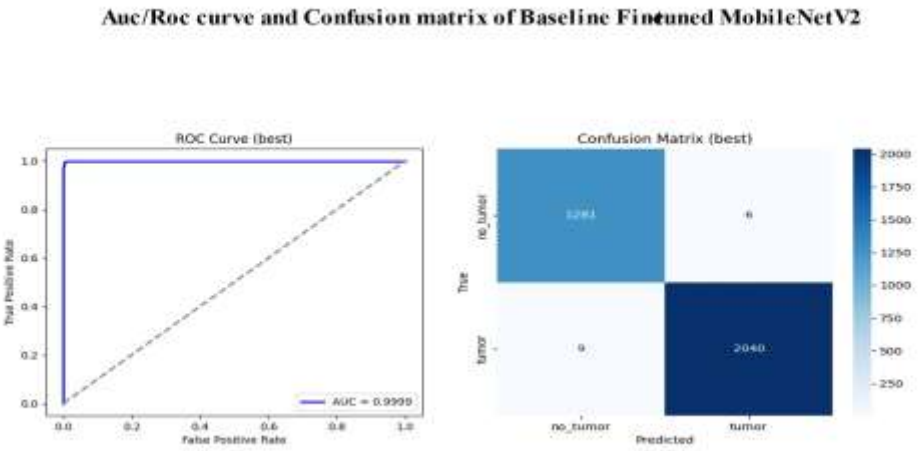

Figure 7: Auc/Roc Curve and Confusion Matrix for the Fine-Tuned MobileNetV2

4.6 Auc/Roc Curve and Confusion Matrix for the Pruned MobileNetV2

Evaluation of the pruned MobileNetV2 model (20% unstructured weight sparsity) on the test set. Shows in figure 8 the ROC curve with AUC = 0.9887, showing a modest reduction in discriminative ability after pruning. Confusion matrix showing TP = 1,025, TN = 634, FP = 40, and FN = 52, reflecting slightly higher misclassification compared with the fine-tuned model.

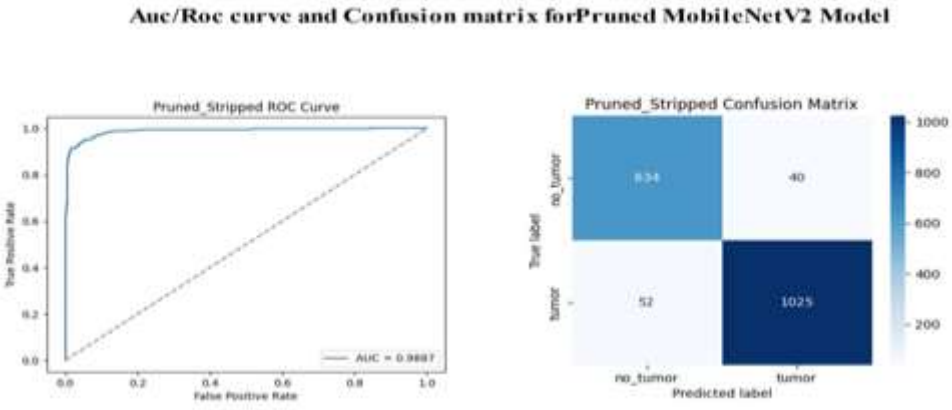

Figure 8: Auc/Roc Curve and Confusion Matrix for the Pruned MobileNetV2

Kasu Journal of Computer Science Vol.3 No.1 [March, 2026], pp. 015-031
https://doi.org//10.47514/kjcs/2026.3.1.002

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

4.7 Auc/Roc Curve and Confusion Matrix for quantized TFLite (8-bit) MobileNetV2

Evaluation of the 8-bit quantized MobileNetV2 model on the test set in Figure 9 shows ROC curve with AUC = 0.9859, demonstrating minimal reduction in discriminative performance after quantization, with Confusion matrix showing TP = 1,019, TN = 636, FP = 38, and FN = 58, indicating robust predictive capability despite model compression.

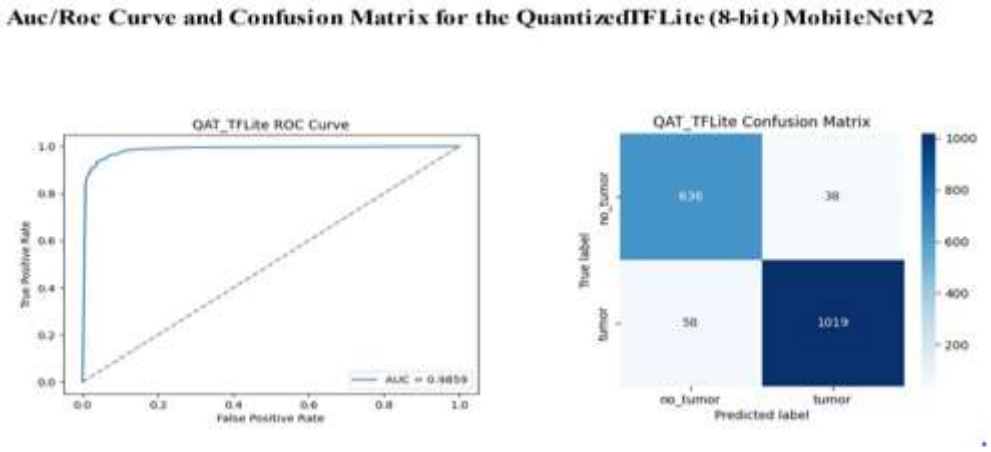

Figure 9: Auc/Roc Curve and Confusion Matrix for TFLite (8-bit) MobileNetV2

4.8 Training and Validation Performance (Accuracy and Loss Curve)

In Figure 10, Accuracy curves showing the rapid increase in both training and validation accuracy during the first five epochs, with validation accuracy closely tracking training accuracy and plateauing near 0.99. Loss curves illustrating the sharp decline in training and validation loss, stabilizing below 0.02. The consistency between training and validation curves indicates that the model learned to distinguish tumor from non-tumor CT slices effectively, without significant overfitting.

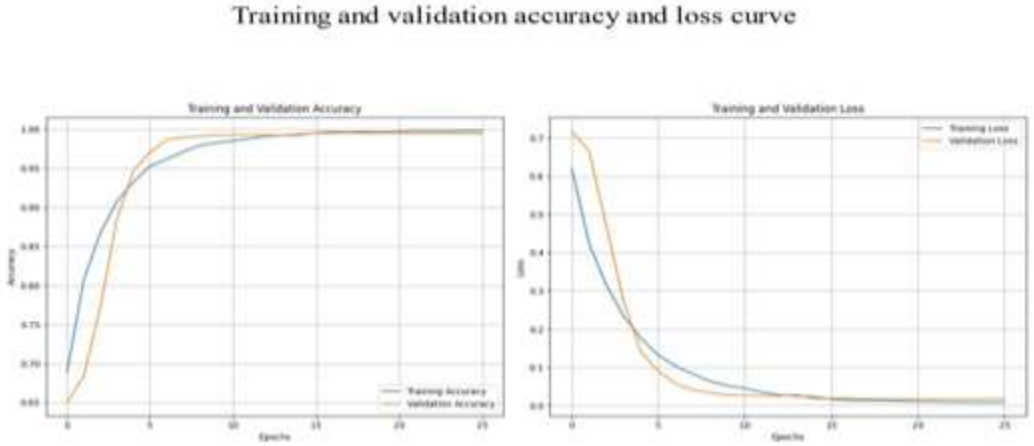

Figure 10. Training and validation performance of MobileNetV2 model.

Kasu Journal of Computer Science Vol.3 No.1 [March, 2026], pp. 015-031
https://doi.org//10.47514/kjcs/2026.3.1.002

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

4.9 Benchmark Comparison Discussion

The benchmark comparison in Table 4 highlights the improvement achieved by the proposed models relative to the baseline architectures previously reported by Hameed et al. Rather than re-implementing the baseline networks, this study adopts their reported performance metrics for ResNet50, VGG16, and MobileNetV1, as they were evaluated on the same LiTS dataset used in this work. According to Hameed et al., ResNet50 achieved only 62% accuracy, which may be attributed to overfitting and its high computational cost on a limited medical dataset. VGG16 performed considerably better, reaching 91.1% accuracy, although its large parameter size makes it less suitable for practical deployment. Meanwhile, MobileNetV1 achieved 94.1% accuracy, confirming the potential of lightweight architectures but without additional optimization. In comparison, the fine-tuned MobileNetV2 developed in this study attained a significantly higher 99.6% accuracy, outperforming all baseline models. More importantly, the pruned version achieved a substantial 65.9% reduction in model size, while maintaining 94.8% accuracy, a performance similar to MobileNetV1 but with markedly greater efficiency. The QAT (TFLite) compressed model further reduced the model size to 8.46 MB and achieved real-time inference speeds of 31 FPS, while sustaining 94.5% accuracy. These results indicate that the optimized MobileNetV2 models not only surpass conventional high-capacity architectures like ResNet50 and VGG16, but also provide a superior balance of accuracy, computational efficiency, and deployability, even when compared to previous lightweight models such as MobileNetV1. With the performance comparison establishing the superiority of the optimized MobileNetV2 models over existing baselines, the next step focuses on evaluating their practicality for real-world deployment. To this end, the effects of pruning and quantization on model size, execution speed, and inference stability are examined to determine suitability for use on resource-limited clinical systems.

Table 4: Benchmark Comparison with Baseline Models

Model	Accuracy	Precision	Recall	F1- score	AUC	Size	inference
ResNet50	62%	-	-	-	-	-	-
VGG16	91.1%	-	-	-	-	-	-
MobileNetV1	94.1%	-	-	-	-	-	-
MobileNetV2	99.6%	0.99	0.99	0.99	0.9999	26.10	8.00
Pruned	94.8%	0.96	0.95	0.96	0.9887	8.91	31.57
QAT TFLite	94.5%	0.96	0.95	0.96	0.9859	8.46	31.42

4.10 Comparative Evaluation Report

Figure 11 presents a comprehensive comparison of MobileNetV2 at different stages of optimization, pre-optimization (fine-tuned), pruned, and quantization-aware trained (QAT 8-bit) alongside baseline models including ResNet50, VGG16, and MobileNetV1. The chart visualizes multiple performance dimensions: accuracy, F1-score, model size, and inference speed (FPS). The fine-tuned MobileNetV2 achieves the highest accuracy of 99.6% and F1-score of 0.99, indicating excellent discriminative capability at the slice level. However, it is the largest model at 26.10 MB and exhibits the slowest inference speed of 8 FPS,

Kasu Journal of Computer Science Vol.3 No.1 [March, 2026], pp. 015-031
https://doi.org//10.47514/kjcs/2026.3.1.002
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

which may limit real-time deployment in clinical environments. The figure also highlights a potential metric inflation due to slice-wise splitting, where adjacent slices from the same volume appear across training and test sets. Applying 20% unstructured pruning and model stripping reduces the model size significantly to 8.91 MB, while maintaining strong accuracy of 94.8% and F1-score of 0.96. Inference speed improves dramatically to 31.57 FPS, showing the efficiency gains achievable with model compression. The QAT 8-bit MobileNetV2 model retains the same level of sparsity but further reduces the size to 8.46 MB. Accuracy and F1-score remain robust at 94.5% and 0.96, respectively, while inference speed remains high at 31.42 FPS, demonstrating that quantization-aware training can preserve predictive performance under lower numerical precision. By contrast, baseline models ResNet50, VGG16, and MobileNetV1 achieve lower accuracy of 62%, 91.1% and 94.1% and provide no reported F1-score, model size, or inference metrics, highlighting the advantages of the proposed MobileNetV2 optimization pipeline. Overall, the chart clearly demonstrates that pruning and quantization effectively balance predictive performance with model efficiency, making the optimized MobileNetV2 suitable for real-time clinical deployment.

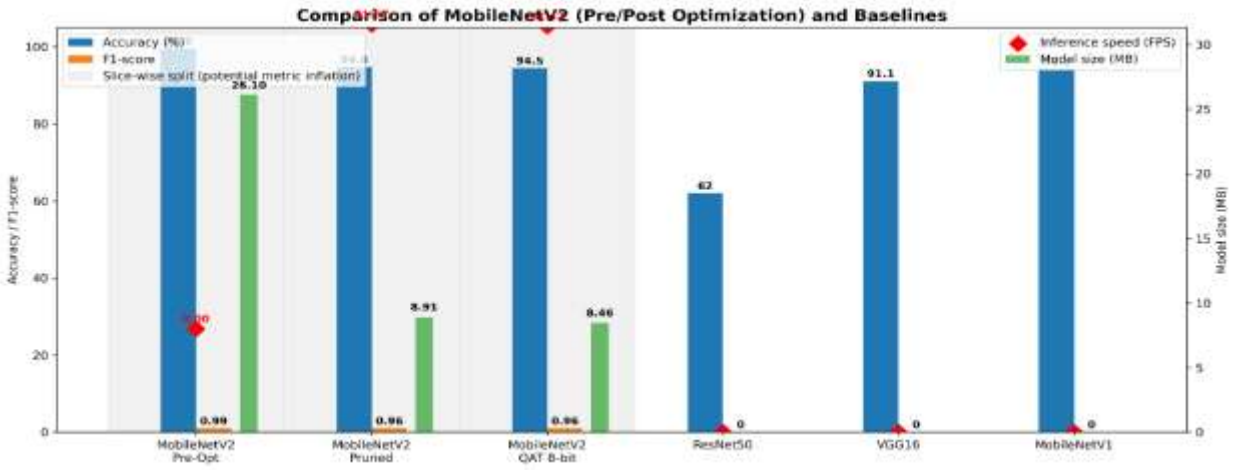

Figure 11: Comparative Evaluation of MobileNetV2 (Pre/Post Optimization) and Baseline Models

4.11 Summary of Findings

The fine-tuned MobileNetV2 achieved the highest predictive performance, recording an accuracy of **99.6%** and an AUC of **0.9999,** confirming its strong capacity for liver tumor classification. However, the model required **26.10 MB** of storage and an inference time of **~125 ms per image**, limiting its suitability for deployment on lightweight or embedded clinical systems. Applying unstructured pruning reduced the model size to **8.91 MB** and improved inference speed to **31.57 FPS**, with only a moderate reduction in accuracy to **94.8%** and AUC to **0.9887**, demonstrating a substantial gain in computational efficiency while maintaining clinically reliable performance. Further compression through 8-bit Quantization-Aware Training (QAT TFLite) produced the smallest model (**8.46 MB**) with a similarly high inference speed (**31.42 FPS**) and a slightly reduced accuracy of **94.5%**, alongside an AUC of **0.9859**. Overall, the results emphasize the **trade-off between maximum accuracy and deployment efficiency**, showing that while the fine-tuned model performs best in prediction, the pruned and quantized models offer a practical balance of speed, size, and accuracy for real-world medical use, especially in resource-limited environments.

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

5. Conclusion

This study demonstrates that MobileNetV2, when enhanced with **unstructured pruning and quantization-aware training**, provides an effective and deployable solution for liver cancer classification using CT images. Although the fine-tuned model delivers the highest accuracy, the compressed models particularly the QAT TFLite model offer a more balanced option for clinical settings where computational resources and real-time performance are essential. These optimization strategies show that it is feasible to **retain high diagnostic reliability while significantly reducing model size and accelerating inference**, thereby improving the accessibility of deep learning solutions in real-world medical workflows. Although the optimized MobileNetV2 models demonstrated strong performance for liver cancer classification, further work is required to support full clinical deployment. Future studies should focus on validating the model using larger, multi-institutional datasets to ensure generalizability across different patient populations, scanners, and imaging conditions. Additionally, the current work utilized a slice-wise dataset split, which, although common in medical imaging research, may lead to data leakage; therefore, future experiments should adopt a strict **patient-wise splitting strategy**, ensuring that slices from the same patient do not appear across training, validation, and test sets. To further enhance diagnostic accuracy, especially after pruning and quantization, advanced optimization strategies such as knowledge distillation, hybrid quantization, or structured pruning can be explored to recover accuracy losses while maintaining efficiency. Finally, integrating the optimized model into clinical imaging workflows such as PACS systems, radiology workstations, edge devices, or point-of-care diagnostic tools would allow real-time decision support for radiologists and could accelerate the clinical translation of machine-learning-assisted cancer diagnosis.

References

Ahmad, J., Alshehri, M. S., Almakdi, S., Al Qathrady, M., Ghadi, Y. Y., & Buchanan, W. J. (2024). SkipGateNet: A lightweight CNN-LSTM hybrid model with learnable skip connections for efficient botnet attack detection in IoT. *IEEE Access, 12*, 35521–35538. https://doi.org/10.1109/ACCESS.2024.3371992

Ahmed, H. A., Rahouma, K. H., et al. (2024). Automated detection of primary liver cancer using different deep learning approaches. *Journal of Advanced Engineering Trends, 43*(1), 433–449. https://doi.org/10.21608/jaet.2024.255537.1269

Alici-Karaca, D., Akay, B., Yay, A. H., Suna, P., Nalbantoğlu, Ö. U., Karaboğa, D., Bastürk, A., Balcıoğlu, E., & Baran, M. (2022). A new lightweight convolutional neural network for radiation-induced liver disease classification. *Biomedical Signal Processing and Control, 73*, Article 103463. https://doi.org/10.1016/j.bspc.2021.103463

Alawneh, K. Z., Alquran, H. H., Alsalatie, M., & Badarneh, A. (2022). LiverNet: Diagnosis of liver tumors in human CT images. *Applied Sciences, 12*(11), 5501.

Anwanwan, D., Singh, S. K., Singh, R., Saikam, V., & Singh, R. (2020). Challenges in liver cancer and emerging therapies. *Biochimica et Biophysica Acta (BBA) – Reviews on Cancer, 1873*(1), 188314.

Bera, K., Braman, N., Gupta, A., Velcheti, V., & Madabhushi, A. (2019). Artificial intelligence in radiology: Challenges and opportunities. *Journal of the American College of Radiology, 16*(9), 1239–1246.

Bilic, P., Christ, P. F., Vorontsov, E., Chlebus, G., Chen, H., Dou, Q., Fu, C., Han, X., Heng, P. A., Hesser, J., Kadoury, S., Konopczynski, T., Le, M., Li, C., Li, X., Ma, C., Mamani, G. E. H., Pauli, J., Pereira, S. P.,

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

… Menze, B. H. (2019). The liver tumor segmentation benchmark (LiTS). *Medical Image Analysis, 59*, 101632.

Bray, F., Ferlay, J., Laversanne, M., et al. (2024). Global cancer statistics 2024: GLOBOCAN estimates of incidence and mortality worldwide. *CA: A Cancer Journal for Clinicians*.

Chung, J. W., Park, M. S., Kim, M. J., & Kim, K. A. (2021). Diagnostic pitfalls in liver imaging. *Clinical and Molecular Hepatology, 27*(2), 248–263.

Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., … Dean, J. (2019). Deep learning-enabled medical computer vision. *Nature Communications, 10*(1), 1–9.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.

Glocker, B., Robinson, R., Castro, D. C., Dou, Q., & Konukoglu, E. (2019). Machine learning with multi-site imaging data: An empirical study on the impact of scanner differences. *Medical Image Analysis, 56*, 44–53.

Han, S., Mao, H., & Dally, W. J. (2016). Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hameed, U., Rehman, M. U., Rehman, A., Damaševičius, R., Sattar, A., & Saba, T. (2023). A deep learning approach for liver cancer detection in CT scans. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*. Advance online publication.

Hameed, U., Rehman, M. U., Rehman, A., et al. (2024). A deep learning approach for liver cancer detection in CT scans. *Biomedical Engineering: Imaging & Visualization, 6*(2), 157–169.

Ichikawa, T., Sano, K., & Morisaka, H. (2014). Diagnosis of liver cancer by computed tomography and magnetic resonance imaging: A review. *Liver Cancer, 3*(1), 51–65.

Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, P., Howard, A., … Adam, H. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2704–2713).

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., … Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis, 42*, 60–88.

Llovet, J. M., Kelley, R. K., Villanueva, A., Singal, A. G., Pikarsky, E., Roayaie, S., … Finn, R. S. (2021). Hepatocellular carcinoma. *Nature Reviews Disease Primers, 7*(1), 6.

Lundervold, A. S., & Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik, 29*(2), 102–127.

McGlynn, K. A., Petrick, J. L., & London, W. T. (2021). Global epidemiology of hepatocellular carcinoma: Etiology, trends, and predictions. *Gastroenterology, 160*(1), 223–238.

McMahon, B., Cohen, C., Brown Jr, R. S., El-Serag, H., Ioannou, G. N., Lok, A. S., … Block, T. (2023). Opportunities to address gaps in early detection and improve outcomes of liver cancer. *JNCI Cancer Spectrum, 7*(3), pkad034.

Midya, A., Chakraborty, J., Srouji, R., Narayan, R. R., Boerner, T., Zheng, J., Pak, L. M., Creasy, J. M., Escobar, L. A., Harrington, K. A., Gonen, M., D'Angelica, M. I., Kingham, T. P., Do, R. K. G., Jarnagin, W. R., & Simpson, A. L. (2023). Computerized diagnosis of liver tumors from CT scans using a deep neural network approach. *IEEE Journal of Biomedical and Health Informatics, 27*(5), 2456–2464.

Owolabi, M. O., et al. (2023). Liver cancer burden and hepatitis prevalence in sub-Saharan Africa. *The Lancet Global Health, 11*(2), e250–e260.

Raghava, Y. B., Srinidhi, V. P., et al. (2023). Detection of tumor in the liver using CNN and MobileNet. In *Proceedings of the IEEE International Conference on Vision Towards Future (ICVTF)* (pp. 1–6).

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

Riaz, Z., Khan, B., Abdullah, S., Khan, S., & Islam, M. S. (2023). Lung tumor image segmentation from computer tomography images using MobileNetV2 and transfer learning. *Bioengineering, 10*(8), 981.

Rumgay, H., Arnold, M., Ferlay, J., Lesi, O., Cabasag, C. J., Vignat, J., … Soerjomataram, I. (2022). Global burden of primary liver cancer in 2020 and predictions to 2040. *Journal of Hepatology, 77*(1), 159–168.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4510–4520).

Sammut, C., & Webb, G. I. (Eds.). (2017). *Encyclopedia of machine learning and data mining*. Springer.

Shen, D., Wu, G., & Suk, H.-I. (2017). Deep learning in medical image analysis. *Annual Review of Biomedical Engineering, 19*, 221–248.

Sindhura, M., Reddy, B. E., & Kumar, P. S. (2022). Efficient-PSP-Net: A lightweight deep learning model for liver and tumor segmentation in CT scans. In *Proceedings of the International Conference on Computational Intelligence and Data Science (ICCIDS 2022)* (pp. 3571600–3571642).

Trefts, E., Gannon, M., & Wasserman, D. H. (2017). The liver. *Current Biology, 27*(21), R1147–R1151.

Valeri, F., Bartolucci, M., Cantoni, E., et al. (2023). UNet and MobileNet CNN-based model observers for CT protocol optimization: Comparative performance evaluation by means of phantom CT images. *Journal of Medical Imaging, 14*(2), Article 027001.

Wei, Q., Tan, N., Xiong, S., Luo, W., Xia, H., & Luo, B. (2023). Deep learning methods in medical image-based hepatocellular carcinoma diagnosis: A systematic review and meta-analysis. *Cancers, 15*(23), 5701.

World Health Organization. (2024). *Liver cancer*.

Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., Sun, Y., He, T., Mueller, J., Manmatha, R., Li, M., & Smola, A. (2021). ResNeSt: Split-attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2736–2746).

Zhou, S. K., Greenspan, H., & Shen, D. (2021). Deep learning for medical image analysis: Challenges and future directions. *Medical Image Analysis, 71*, 102105.