Çukurova Üniversitesi
Mühendislik Fakültesi Dergisi
Çukurova University
Journal of the Faculty of Engineering
ISSN: 2757-9255 CİLT/VOLUME: 40 SAYI/ISSUE: 3 EYLÜL/SEPTEMBER 2025

Diagnosis of Hepatocellular Carcinoma - HCC Liver Cancer Using Federated Learning on MR Images

Burak UZDUR [1,a], **Erkut TEKELİ** [1,b], **Turgay İBRİKÇİ** [1,c], **Harun Ur RASHID** [2,d], **Geetha RAMACHANDRAN** [3,e]

[1]*Adana Alparslan Turkes Science and Technology University, Software Engineering Department, Adana, Türkiye*
[2]*Hankuk University of Foreign Studies, Department of Information and Communications Engineering, Seoul, South Korea*
[3]*S.A. Engineering College, Department of Computer Science and Engineering, Thiruverkadu, Chennai, India*

[a]***ORCID**: 0009-0008-8093-1097; [b]**ORCID**: 0000-0001-9468-5378; [c]**ORCID**: 0000-0003-1321-2523; [d]**ORCID**: 0000-0003-0874-7590; [e]**ORCID**: 0000-0002-4541-3314*

Article Info
Received : 05.07.2025
Accepted : 12.08.2025
DOI: 10.21605/cukurovaumfd.1735231

Corresponding Author
Erkut TEKELİ
etekeli@atu.edu.tr

Keywords
Federated learning
FedAvg
Liver tumor classification
Convolutional neural networks

***How to cite:** UZDUR, B., TEKELİ, E., İBRİKÇİ, T., RASHID, H.U., GEETHA, R. (2025). Diagnosis of Hepatocellular Carcinoma - HCC Liver Cancer Using Federated Learning on MR Images. Cukurova University, Journal of the Faculty of Engineering, 40(3), 531-544.*

ABSTRACT

In recent years, Federated Learning (FL) has emerged as a powerful paradigm for training machine learning models across decentralized data sources while preserving data privacy. This study proposes an FL framework for the classification of liver tumors from the ATLAS dataset, which provides images of hepatocellular carcinoma cases. A comparative evaluation was performed utilizing CNN, EfficientNet, MobileNetV3, ResNet50, and VGG16 architectures within the federated environment. Among these models, the FL implementation based on EfficientNet achieved superior performance, reaching an accuracy of 93.75% and a ROC-AUC score of 99.19%. The results demonstrate that federated approaches can attain performance levels comparable to centralized learning while ensuring patient data confidentiality. The potential of developing a privacy-preserving collaborative model using the FL method has been demonstrated.

MR Görüntülerinde Federasyonlu Öğrenme Kullanılarak Hepatosit Karsinomu - HCC Karaciğer Kanseri Tanısı

Makale Bilgileri
Geliş : 05.07.2025
Kabul : 12.08.2025
DOI: 10.21605/cukurovaumfd.1735231

Sorumlu Yazar
Erkut TEKELİ
etekeli@atu.edu.tr

Anahtar Kelimeler
Federasyonlu öğrenme
FedAvg
Karaciğer tümörü sınıflandırması
Evrişimsel sinir ağları

***Atıf şekli:** UZDUR, B., TEKELİ, E., İBRİKÇİ, T., RASHID, H.U., GEETHA, R. (2025). MR Görüntülerinde Federasyonlu Öğrenme Kullanılarak Hepatosit Karsinomu - HCC Karaciğer Kanseri Tanısı. Çukurova Üniversitesi, Mühendislik Fakültesi Dergisi, 40(3), 531-544.*

ÖZ

Son yıllarda, Federasyonlu Öğrenme (FÖ), veri gizliliğini korurken merkezi olmayan veri kaynakları arasında makine öğrenimi modellerini eğitmek için güçlü bir paradigma olarak ortaya çıkmıştır. Bu çalışma, hepatosellüler karsinom vakalarının görüntülerini sağlayan ATLAS veri setinden elde edilen Manyetik Rezonans Görüntülerini kullanılarak karaciğer tümörlerinin sınıflandırılması için bir FÖ çerçevesi önermektedir. Federasyonlu ortamda Evrişimli Sinir Ağı, EfficientNet, MobileNetV3, ResNet50 ve VGG16 mimarileri kullanılarak karşılaştırmalı bir değerlendirme yapılmıştır. Bu modeller arasında, EfficientNet tabanlı FÖ uygulaması, %93,75'lik bir doğruluk ve %99,19'luk bir ROC-AUC puanına ulaşarak üstün bir performans elde etmiştir. Sonuçlar, federasyonlu yaklaşımların hasta verilerinin gizliliğini sağlarken merkezi öğrenmeye benzer performans seviyelerine ulaşabileceğini göstermektedir. FÖ ile gizliliği koruyan işbirlikçi model geliştirme potansiyeli olduğu gösterilmiştir.

1. INTRODUCTION

The liver, as the largest internal organ in the human body, plays a vital role in numerous physiological functions, including detoxification, enzyme production, blood clotting, and metabolic regulation [1]. Given its critical importance, timely and accurate diagnosis of liver-related disorders is crucial for effective treatment and improved patient outcomes.

Magnetic Resonance Imaging (MRI) has become a cornerstone in liver disease diagnostics due to its non-invasive nature and its ability to generate high-resolution anatomical and functional images. The evolution of MRI techniques has significantly improved the early detection and characterization of hepatic abnormalities [2]. Alongside these advancements, Deep Learning (DL) techniques have increasingly gained attention for their ability to analyze complex medical imaging data with high precision.

DL, a subset of Machine Learning (ML), has shown transformative potential in healthcare, especially in radiology [3]. It allows for automated detection, classification, and segmentation of medical images while minimizing diagnostic variability and aiding clinicians in decision-making processes [4]. These advances are particularly important in diagnosing critical diseases like Hepatocellular Carcinoma (HCC).

HCC remains one of the most prevalent and fatal forms of liver cancer, with projections estimating over one million cases globally by 2025 [5]. It is commonly observed in regions with high hepatitis B and C virus prevalence, such as sub-Saharan Africa and Eastern Asia. Although dynamic Computed Tomography (CT) is a standard imaging technique for HCC diagnosis [6,7], its interpretation requires expert knowledge and is both time- and labor-intensive [8]. Recent research has demonstrated that automated methods, such as computer vision, can enhance diagnostic efficiency and reduce human error [9,10].

Pre-trained Convolutional Neural Networks (CNN), including MobileNetV3, EfficientNet, ResNet50, and VGG16, have proven effective in medical image classification tasks. However, their performance is highly dependent on large and diverse training datasets. In medical domains, data acquisition and sharing are constrained by strict privacy laws and ethical considerations.

As data privacy and security remain top concerns, Federated Learning (FL) has emerged as a viable alternative to centralized training approaches. FL allows multiple institutions to collaboratively train ML models locally, without sharing sensitive data. Only model parameters are exchanged and aggregated, preserving patient confidentiality while enabling large-scale collaborative research.

In this study, an FL-based classification framework is applied by using EfficientNetB5 for liver tumor identification in MRI images. To assess its effectiveness, we compared the FL model's performance with centralized implementations of widely used DL architectures, including CNN, EfficientNet, MobileNetV3, ResNet50, and VGG16. This comparison aims to evaluate FL's advantages not only in classification accuracy but also in addressing data privacy concerns.

The limitations of centralized data usage in medical imaging—particularly regarding privacy, ethics, and regulatory compliance—have led to increased interest in decentralized learning frameworks. FL provides a robust solution by enabling collaborative model training across institutions without transferring raw patient data. This privacy-preserving approach ensures compliance with legal standards while maintaining diagnostic performance. In the context of this study, FL was applied to liver tumor classification using MRI data. Figure 1 illustrates the architecture of the proposed FL framework and its operational flow in a healthcare setting, highlighting how the model aggregates local updates while preserving data confidentiality (adapted from [11]).

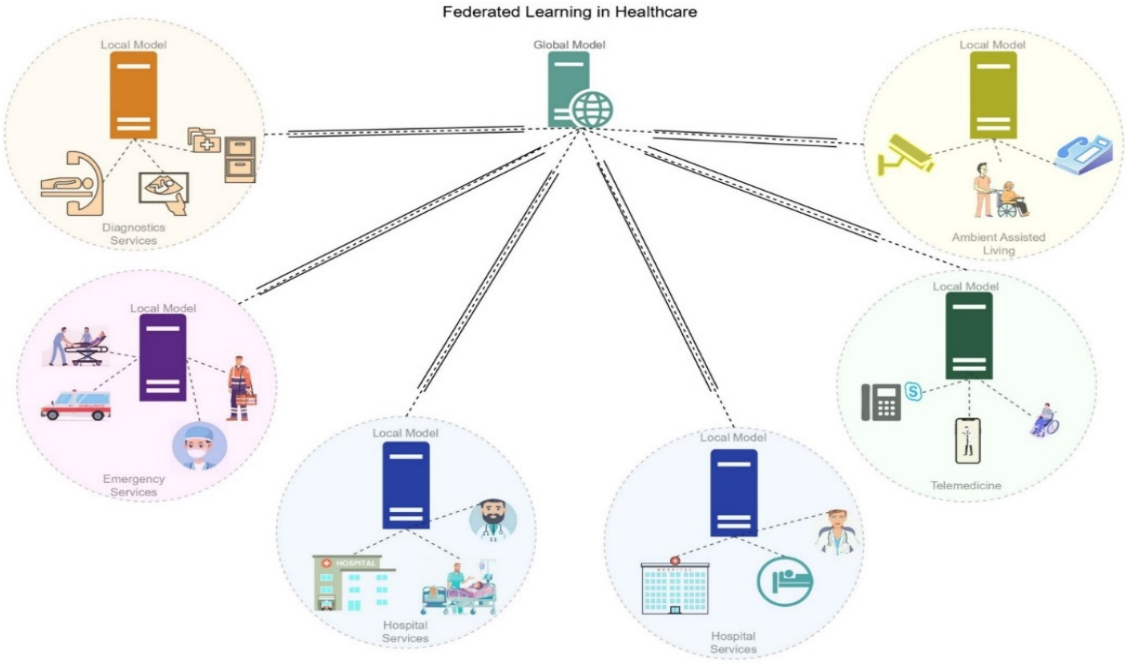

Figure 1. FL technology in the healthcare domain [11]

In conclusion, the study addresses the detection of HCC, a significant issue in medical imaging, by integrating FL to protect data privacy, which is of considerable importance in light of current concerns regarding patient data security. The study also observes that the combination of EfficientNetB5 with FL represents a novel approach and that the comparative analysis with traditional DL models adds significant value.

2. LITERATURE REVIEW

In the realm of medical imaging, the application of ML and DL techniques has led to significant progress, particularly in the classification and diagnosis of liver tumors using MRI images. This section examines relevant studies in the field, emphasizing key approaches, findings, and their connection to the present research (Table 1).

Roth et al. [12] demonstrated the effectiveness of FL in breast density classification by involving seven clinical institutions in a collaborative effort that ensured confidentiality. Their FL-trained model outperformed single-institution models with a 6.3% performance gain and showed a 45.8% improvement in generalizability across external datasets. Despite its success, the study lacks analysis of computational costs and overlooks challenges in large-scale FL deployment, such as infrastructure heterogeneity and data standardization.

Bernecker et al. [13] applied two FL methods, FedNorm and FedNorm+, to perform liver disease segmentation using CT and MRI data from 428 patients across six datasets. FedNorm+ demonstrated superior performance in comparison to local models and equaled the performance of centralized models by attaining a high Dice score of up to 0.961. However, the study is not accompanied by metrics that would allow for an assessment of its accuracy. These limitations, despite the study's initial success in demonstrating the efficacy of the segmentation approach, render it difficult to apply to actual clinical settings.

Mahlool et al. [14] proposed a novel classification model that integrates DL with the FL algorithm. The model was evaluated using the CT-small 2c and CT-large 3c datasets, yielding classification accuracies of 0.82 and 0.96, respectively. The researchers' findings indicate that classification systems developed in FL can offer high reliability and prove effective in clinical decision support systems.

Table 1. Summarization for related works

Authors	Year	Number of patients / images	Data set description	Accuracy / dice (%)	Methodology
Roth et al.	2020	715.000	Mammography (BIRADS, multi-institutional)	6.3 (avg.), 45.8 (gen.)	FL for breast density classification across 7 clinical sites
Bernecker et al.	2022	428	CT and MRI images from 6 different public datasets	Dice: 96.1	FedNorm and FedNorm+ algorithms for FL-based segmentation
Mahlool et al.	2022	253 (BT-small-2c), 3,264 (BT-large-3c)	BT-small-2c: MRI. BT-large-3c: MRI (3 tumor types, 500 healthy)	BT-small-2c Acc: 82, BT-large-3c Acc: 96	FL for brain tumor diagnosis
Trivedi et al.	2023	576	MRI	Acc: 99.59	AlexNet-based FL tested on IID Liver dataset with lightweight CNNs
Chai et al.	2024	733	Gene expression data from TCGA and GEO (cross-institutional, anonymized)	Acc: 54.2	FL with AdFed, DeepSurv-based survival prediction
Lusnig et al.	2024	41 patients / 4,400 images	41 whole-slide images (JPEG2000), divided into 1024×1024-pixel patches; balanced dataset with 1100 images per stage	Centralized Acc: 97%; Federated Acc: ~90%	FL applied to high-resolution histopathology images; images categorized into transplant-suitable vs. unsuitable; used HQNN for classification
Shankar et al.	2025		Multi-modal dataset (CT, MRI, ultrasound + lab values)	Acc: 79.05	Prediction of liver disease using FL from imaging and clinical data

Trivedi et al. [15] evaluated lightweight FL strategies for HCC classification using several pre-trained CNNs, with AlexNet achieving a peak accuracy of 99.59%. A distinguishing feature of their study is its emphasis on system efficiency and deployable architectures. However, the study lacks a detailed discussion on real-world application challenges such as dataset size and communication costs, or training latency.

Chai et al. [16] proposed AdFed, an FL framework for survival prediction in multiple cancers, including liver cancer. Using 733 genetic profiles, the model achieved an AUC of 0.605 for liver cancer, surpassing comparable FL approaches. A major strength is its biological interpretability, with half of the top genes already known to be associated with liver cancer. However, the limited sample size and lack of discussion on real-world FL challenges, such as scalability and communication overhead, weaken its practical applicability.

Lusnig et al. [17] introduced an FL framework using hybrid quantum neural networks (HQNNs) for classifying Non-Alcoholic Fatty Liver Disease (NAFLD) from histopathological biopsy images. The model achieved 97% accuracy under centralized training and around 90% under FL, demonstrating strong performance while preserving data privacy. Despite the promising results, the study is limited by a small dataset (41 patients) and potential scalability issues due to quantum infrastructure requirements.

Shankar et al [18] developed an FL framework that combines CT, MRI, and ultrasound medical imaging data with clinical test data (e.g., bilirubin, ALT, AST) for liver disease prediction. The approach is to achieve an acceptable accuracy of 79.05% while maintaining data confidentiality. However, this study lacks transparency regarding the dataset size and patient numbers. The objective of this study is to identify cases of general liver disease as opposed to tumor-specific classification.

Fofanah et al. [19] presented a comprehensive study that employed CNN and DL techniques for the purpose of detecting skin cancer. The proposed method in this study achieved an 84.3% accuracy rate in skin cancer detection.

Fırat and Üzen [20] proposed a DL method for the classification of Alzheimer's disease on MRI. This method was based on Inception and CNNs. Utilising this approach, the researchers attained a classification accuracy of 98% on a four-class dataset.

3. METHOD

3.1. Dataset Description

This study utilizes a publicly available dataset from the ATLAS (A Tumour and Liver Automatic Segmentation) challenge, comprising contrast-enhanced T1-weighted MRI scans from 90 patients diagnosed with inoperable HCC [21]. The dataset includes segmentation masks for both liver and tumor regions, allowing for binary classification into "normal" and "tumor" categories. A total of 3.623 training and 1.553 test images were balanced across both classes, as illustrated in Figure 2. All images were resized to 224×224 pixels and preprocessed accordingly. The data was randomly split into 70% training and 30% testing sets to ensure generalizability. Additionally, Figure 3 presents representative examples that highlight variations in tumor morphology and anatomical structures.

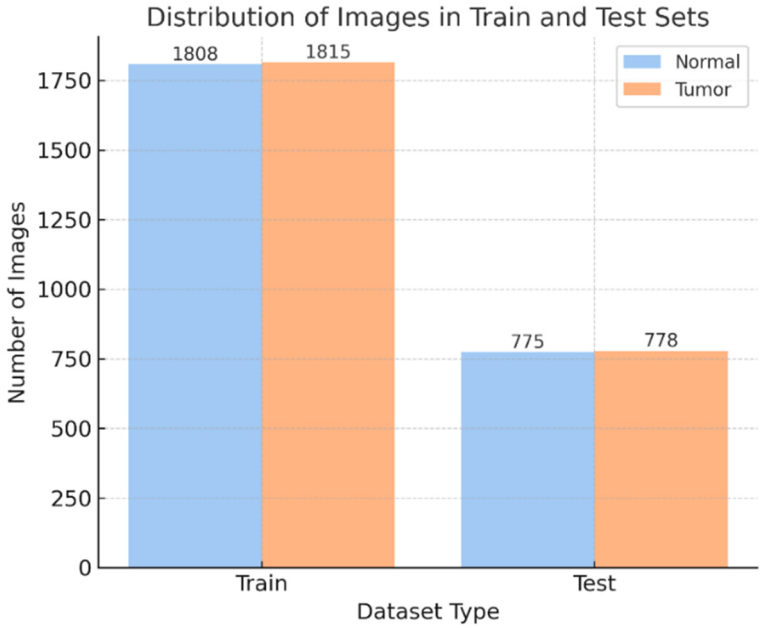

Figure 2. The numerical distribution of normal and tumor images in the training and test datasets

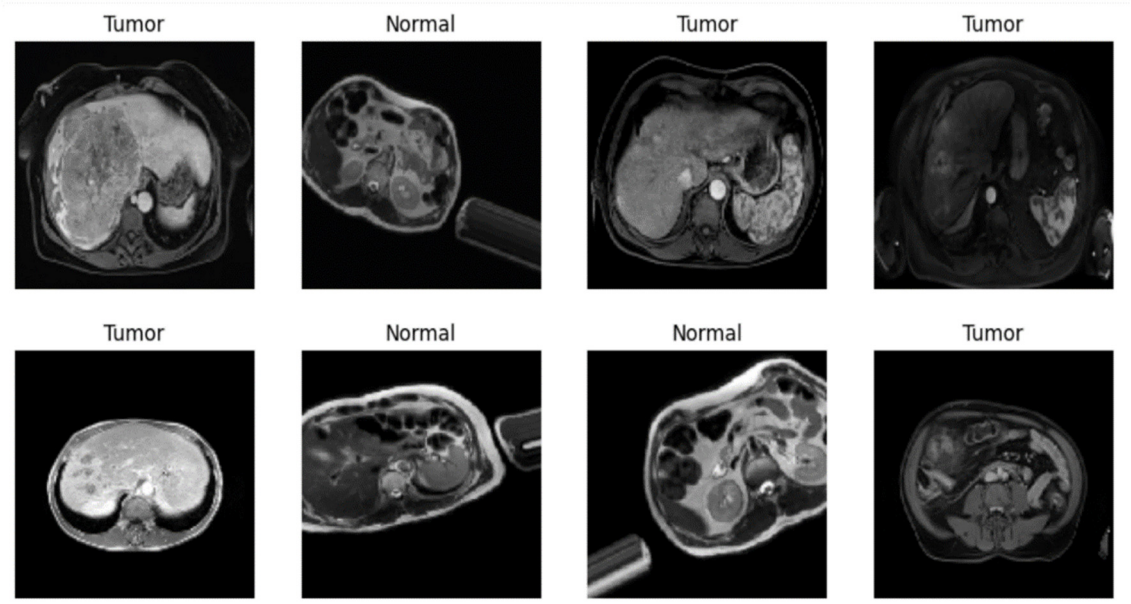

Figure 3. Sample images from the dataset

3.1.1. Data Preprocessing

In this study, a comprehensive data preprocessing pipeline was implemented to enhance the performance of the model and ensure consistency across the dataset. Initially, all MRI images were resized to 224 × 224 pixels, ensuring uniform input dimensions for all models. To improve the generalization capability of the model and mitigate overfitting, several data augmentation techniques were employed. These included rotation (±40°), horizontal flipping, zooming (±30%), brightness adjustments (range: 0.8–1.2), width and height shifting (±30%), and shear transformations (±30%). These augmentations not only expanded the diversity of the dataset but also contributed to improving the model's robustness. Additionally, all images were normalized by scaling pixel values to the [0,1] range, which ensured numerical stability throughout the training process. To avoid class imbalance and reduce potential bias, a class-balancing strategy was applied, ensuring an equal representation of tumor and non-tumor images. Finally, the dataset was partitioned into training and test sets, with 70% of the images allocated for training and the remaining 30% used for evaluation. These preprocessing steps were essential for optimizing the performance of the FL-based classification model.

3.2. Federated Learning

AI has significantly transformed the field of medical imaging by enabling the automation of diagnostics and enhancing accuracy across diverse healthcare settings. However, developing robust AI models requires access to large and varied labeled datasets—a challenge in clinical environments due to stringent data privacy regulations, inter-institutional differences, and the limited availability of annotated medical data. To address these challenges, collaborative and privacy-preserving approaches have been introduced, with FL emerging as a leading decentralized technique. FL enables multiple medical institutions to train models locally on their private datasets and contribute to a shared global model by sending only model updates, thereby safeguarding sensitive patient information within the original institution while facilitating knowledge aggregation from varied data sources [22]. In this study, a strategy of full client participation was employed to maximize the collaborative potential of FL.

In this study, all three clients actively participated in the model training process during each federated round. Each client trained its local model for 3 epochs on its respective dataset, with this process being repeated across 10 federated rounds. The strategy of full participation guarantees that model updates from all clients are integrated into the global model in every round, thereby fostering more efficient and balanced learning.

The training dataset, comprising 90 ATLAS patients and 3,623 images, was partitioned equally and independently among three FL clients to simulate an Independent and Identically Distributed (IID) data setting. Each client received approximately one-third of the total data with balanced class distributions, ensuring statistical similarity across clients. This configuration represents an ideal FL scenario, where data is evenly and randomly distributed among clients, facilitating fair evaluation of model performance.

Federated Averaging (FedAvg) is one of the most widely used optimization algorithms in horizontal FL, where client data remains decentralized and is not shared with a central server. This method has been extensively studied for its convergence properties under various conditions, including data heterogeneity and variations in loss functions, demonstrating its robustness in collaborative learning environments [23]. In this study, FedAvg was employed as the aggregation mechanism to synchronize the local models of the three clients into a unified global model. After each client trained its local model over three epochs, the learned weights were sent to the central server, where layer-wise averaging was conducted. The resulting global weights were then redistributed to all clients for the subsequent training round. This iterative process was repeated across 10 federated rounds, allowing the global model to progressively improve by incorporating knowledge from all clients, all while maintaining data privacy.

While FedAvg facilitates collaborative model training without direct data sharing, it does not inherently guarantee complete privacy. FedAvg remains vulnerable to privacy attacks such as gradient inversion and membership inference, which can potentially expose sensitive client information. This study acknowledges these limitations and considers the integration of advanced privacy-preserving mechanisms as important directions for future work to enhance the robustness of the FL framework.

For the training of the FL, the binary cross-entropy loss function was employed, which is well-suited for the liver tumor classification task due to its binary nature. This loss function effectively measures the difference between predicted probabilities and actual class labels in binary classification tasks, making it a standard and reliable choice for such applications. The model was optimized using the Adam optimizer, renowned for its adaptive learning rate and strong performance in DL tasks. Alongside the minimization of the loss function, various performance metrics were monitored during both training and evaluation, including accuracy, precision, recall, F1-score, and ROC-AUC. These metrics provided a comprehensive evaluation of the model's classification capabilities, particularly in a clinical context where both sensitivity and specificity are crucial for decision-making.

Figure 4 illustrates the training and testing process implemented in this study, following the FL framework. In this setup, separate models are trained locally on distinct clients using their respective data. Following each training round, the local model weights are sent to a central server, where the FedAvg algorithm is applied to update the global model. This iterative process continues until the model achieves optimal accuracy. Once the final model is obtained, its performance is assessed using the central test dataset.

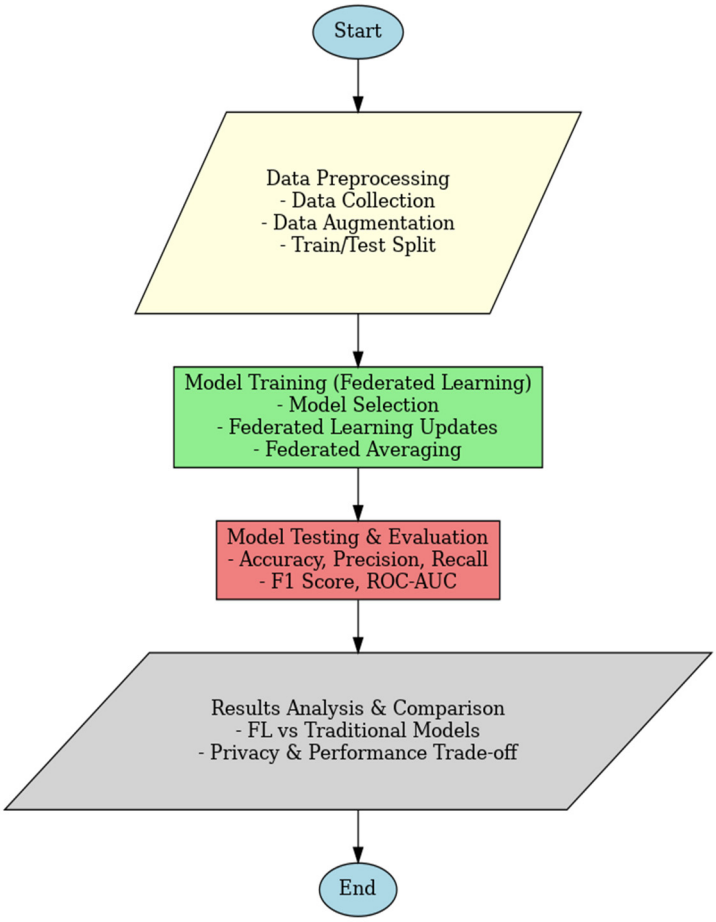

Figure 4. The flowchart of the proposed training and testing process for liver tumor classification

3.3. Comparison Models for Liver Tumor Classification

To assess the performance of the proposed FL model for liver tumor classification, five well-established DL architectures—CNN, EfficientNet, MobileNetV3, ResNet50, and VGG16—were implemented as baseline models. These models were selected for comparison and were trained under identical hyperparameter settings to ensure consistency and fairness in the evaluation process. Specifically, each model underwent training for 10 epochs with a batch size of 32, utilizing the Adaptive Moment Estimation -Adam optimizer and binary cross-entropy as the loss function. ReLU activation was applied to the hidden layers, while the output layer employed the sigmoid activation function, which is suitable for binary classification tasks. Table 2 provides an overview of the training configurations for these models.

Table 2. Training configurations for models

	Number of epochs	Batch size	Optimizer	Loss function	Activation functions	Output layer
CNN	10	32	Adam	Binary Crossentropy	Relu	Sigmoid
EfficientNet	10	32	Adam	Binary Crossentropy	Relu	Sigmoid
MobileNetV3	10	32	Adam	Binary Crossentropy	Relu	Sigmoid
ResNet50	10	32	Adam	Binary Crossentropy	Relu	Sigmoid
VGG16	10	32	Adam	Binary Crossentropy	Relu	Sigmoid

3.3.1. Convolutional Neural Networks

In this study, a custom CNN was developed as a baseline model for liver tumor classification. The architecture includes three convolutional layers with progressively larger filter sizes (32, 64, and 128), each followed by a MaxPooling2D layer to reduce spatial dimensions and mitigate overfitting. After the convolutional operations, the feature maps are flattened and passed through a fully connected dense layer containing 128 neurons, activated by the ReLU function. Finally, a sigmoid-activated output layer is employed to perform binary classification. This relatively straightforward yet efficient architecture serves as a foundational model for comparing the performance of more advanced, pre-trained models in liver tumor classification.

3.3.2. EfficientNet

In the comparative analysis, EfficientNet was employed to evaluate the performance of liver tumor classification. The model was fine-tuned by unfreezing the last 50 layers of the pre-trained base and retraining them on the target dataset. To tailor the model for binary classification, additional layers were incorporated on top of the base, including a GlobalAveragePooling2D layer, a dropout layer with a 0.3 rate to mitigate overfitting, and two dense layers—one with 128 neurons activated by ReLU, and a final output layer with a sigmoid activation function. This architecture enabled the model to capture domain-specific features while leveraging the benefits of transfer learning for improved performance.

3.3.3. ResNet50

The ResNet50 architecture with 50 layers deep was developed by Microsoft Research in 2015, considered to be among the most popular CNN architectures around, which employs skip connections to address the vanishing gradient problem, enabling effective training of deeper networks. This model has shown robust performance in medical imaging classification, particularly in identifying complex patterns within tumor images. However, its computationally intensive nature may pose challenges for deployment in real-world medical settings, where resources may be limited.

3.3.4. MobileNetV3

In this study, MobileNetV3 Small was adapted using a transfer learning approach specifically designed for medical image classification tasks. The model utilized pre-trained weights, with the initial layers frozen to retain low-level feature extraction capabilities. The MobileNetV3 small architecture was employed, followed by the addition of a Global Average Pooling layer, a fully connected dense layer with 128 neurons and ReLU activation, a Dropout layer with a 0.5 rate to mitigate overfitting, and a final output layer with a single neuron using sigmoid activation for binary classification. The training process utilized the Adam optimizer along with the binary cross-entropy loss function. Additionally, the learning rate was dynamically adjusted using the ReduceLROnPlateau technique, which monitored the validation loss and reduced the learning rate when further improvements plateaued.

3.3.5. VGG16

One of the comparison models employed in this study was the VGG16 architecture. The base of the model consisted of the VGG16 network, which was pre-trained on ImageNet, with its top layers removed. To minimize overfitting and allow the model to focus on training the newly added layers, the weights of the pre-trained layers were frozen. A Global Average Pooling layer was incorporated to flatten the feature

maps, followed by a fully connected dense layer comprising 128 neurons and ReLU activation, and a Dropout layer with a rate of 0.5 to reduce overfitting. The final output layer incorporated a sigmoid activation function to facilitate binary classification. The model was constructed with the Adam optimiser, with a learning rate of 0.001 and binary cross-entropy as the loss function.

4. RESULTS

To evaluate the effectiveness of the proposed FL model in liver tumor classification, five distinct DL architectures were implemented for comparison. The custom-designed CNN employed a simplified architecture consisting of three convolutional layers, each followed by max-pooling and fully connected layers. EfficientNet was fine-tuned by unfreezing its last 50 layers and enhanced with a Global Average Pooling layer, dropout regularization, and dense layers. Similarly, MobileNetV3 and ResNet50—both initialized with ImageNet pre-trained weights—were adapted with additional layers including Global Average Pooling, ReLU-activated dense layers, and dropout to improve generalization. The VGG16 model was also employed by freezing its convolutional base and appending dense layers for the binary classification task. All models were trained under consistent settings using the binary cross-entropy loss function and the Adam optimizer, which is an optimisation algorithm that is frequently employed in the training of DL models. This approach synthesises the merits of two other extensions of stochastic gradient descent (SGD): AdaGrad and RMSProp. The architectural and training configurations of these models are summarized in Table 3, ensuring a fair and systematic comparison of model performances.

Table 3. Architectural and training configurations of DL models used for image classification

Model	Pretrained weights	Trainable layers	Pooling type	Dropout rate	Dense layer (units)
CNN	No	All	MaxPooling2D	None	128
EfficientNet	ImageNet	Last 50 layers	GlobalAveragePooling2D	0.3	128
MobileNetV3	ImageNet	Frozen (feature extractor)	GlobalAveragePooling2D	0.5	128
ResNet50	ImageNet	Frozen (feature extractor)	GlobalAveragePooling2D	0.5	128
VGG16	ImageNet	Frozen (feature extractor)	GlobalAveragePooling2D	0.5	128

To comprehensively assess the performance of the proposed liver tumor classification model, five key evaluation metrics—accuracy, precision, recall, F1-score, and ROC-AUC score—were utilized. These metrics offer valuable insights into the model's ability to correctly classify MRI images, differentiate between normal and tumor cases, and perform effectively in a real-world diagnostic setting.

The evaluation of performance criteria constitutes a pivotal step in both ML and DL. The True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) outcomes are pivotal for the evaluation of the performance of classification models. These metrics facilitate comprehension of the extent to which their models can generate precise predictions and subsequently optimise them for discrete or particular decisions, thereby enhancing decision-making processes across diverse domains.

TP: The number of times the model correctly predicted the positive class.
TN: The number of cases where the model correctly predicted the negative class.
FP: The number of instances where the model incorrectly predicted the positive class.
FN: The number of cases where the model incorrectly predicted the negative class.

Accuracy is one of the most fundamental metrics for evaluating the overall performance of a classification model. It is defined as the ratio of correctly classified instances—both positive and negative—to the total number of predictions made, as shown in Equation 1. While accuracy can serve as a reliable indicator in balanced datasets, it may become misleading in imbalanced scenarios.

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \quad (1)$$

Precision is defined as the proportion of predicted positive cases that are actually positive and is formally expressed in equation 2. A high precision score indicates that the model effectively minimizes false positives, meaning its positive classifications are more reliable.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{2}$$

Recall assesses the model's ability to identify positive cases correctly and is mathematically defined in equation 3. This metric is particularly important in applications where failing to detect a positive case can have serious consequences, such as in medical diagnosis

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

F1-Score is the harmonic mean of precision and recall, as defined in Equation 4. It is particularly effective in evaluating model performance on imbalanced datasets, as it considers both false positives and false negatives. A high F1-score reflects the model's ability to strike a balance between precision and recall, effectively minimizing both types of errors. The F1-Score serves as a crucial metric, ensuring that the model not only avoids false alarms but also captures the majority of true positive cases.

$$F1 - score = 2 * \frac{\mathrm{Precision} * \mathrm{Recall}}{Precision + Recall} \tag{4}$$

ROC-AUC Score is a graphical representation that is used to evaluate the performance of a classification model at various threshold values. The Area Under the Curve (AUC) represents the area beneath the Receiver-Operating Characteristic Curve (ROC) and measures the model's overall discriminative ability. The ROC-AUC metric is particularly important for evaluating classification models on imbalanced datasets. As illustrated in Figure 5, the ROC curves of the models under investigation are displayed.

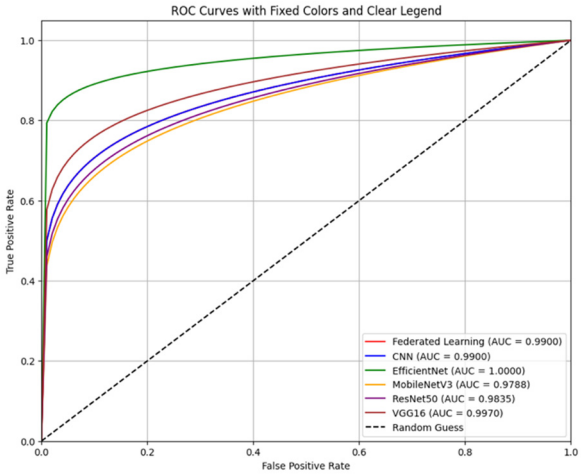

Figure 5. ROC curves

The classification performance of the models studied is shown in Table 4.

Table 4. Performance of models

	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC (%)
FL	93.75	99.71	87.79	93.37	99.19
CNN	98.58	99.74	97.43	98.57	98.59
EfficientNet	99.55	99.61	99.49	99.55	99.55
MobileNetV3	95.43	97.20	93.57	95.35	97.88
ResNet50	94.01	94.54	93.44	93.99	98.35
VGG16	98.71	99.74	97.69	98.70	99.70

Figure 6 presents a graphical representation that compares the classification performance of the models.

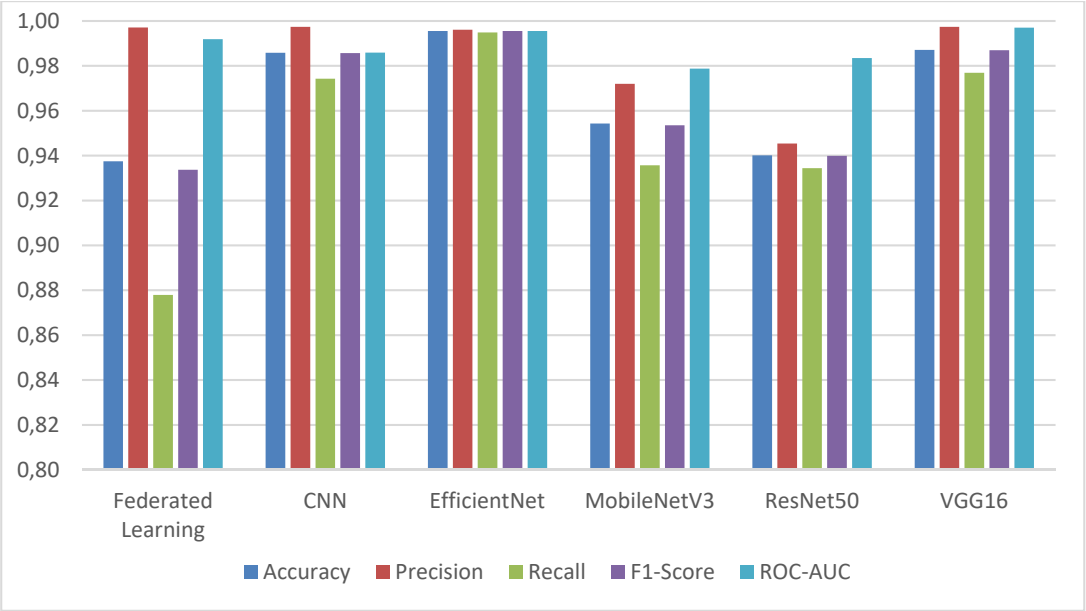

Figure 6. Comparison of model performances

To empirically justify our choice of hyper parameters in the FL setup, we conducted an ablation study by varying local epochs and global rounds. Table 5 summarizes the performance metrics across four different configurations.

Table 5. Ablation study of FL configuration

Local epochs	Rounds	Accuracy (%)	Precision (%)	Recall (%)	F1–Score (%)	ROC-AUC (%)
3	10	93,75	99,71	87,79	93,37	99,19
1	5	92,14	90,80	93,83	92,29	97,40
1	10	92,98	96,65	89,07	92,71	98,49
1	20	82,68	75,02	98,07	85,01	97,62

As shown in Table 5, the configuration with 3 clients, 3 local epochs, and 10 rounds yielded superior results in terms of accuracy and ROC-AUC compared to the other evaluated configurations. Based on these findings, this setup was selected for use in the subsequent experiments.

As the final results, the FL-based model achieved an accuracy of 93.75% and a precision value of 99.71%. These results demonstrate that the model provides highly reliable and accurate positive results in the diagnosis of liver tumors, while also effectively differentiating between classes. The F1-Score was 93.37%, and the ROC-AUC value was measured at 99.19%, both of which indicate the high overall performance of the model and its potential for use in critical decision-support systems in healthcare.

It is observed that the centralized EfficientNet model achieves a higher accuracy (99.55%) compared to the FL model (93.75%). This performance gap is expected and can be attributed to the inherent challenges of FL, such as data heterogeneity across clients, communication constraints, and the absence of direct access to the full training dataset.

Despite this difference, the FL model still achieves a high level of accuracy and offers substantial privacy advantages by enabling model training without centralized data aggregation. In real-world medical environments, such as inter-hospital collaborations, data sharing is often limited by strict privacy regulations (e.g., HIPAA, GDPR). Consequently, the observed reduction in accuracy is justified by the critical need to maintain data privacy, particularly in healthcare settings where centralized data sharing is not feasible.

In the diagnosis of HCC using MRI, sensitivity (recall) rates are typically reported in the range of 85% to 95% [24, 25, 26]. Deep learning-based models have also achieved recall rates comparable to those of expert radiologists [24]. In our study, the achieved recall rate of 87.79% falls within the lower bound of this clinically accepted range. However, this corresponds to a 12.21% false-negative rate, which could lead to missed diagnoses in practice. According to the Liver Imaging Reporting and Data System (LI-RADS) developed by the American College of Radiology, sensitivity levels above 88% are considered desirable for MRI-based HCC diagnosis [27]. In our study, the recall rate of 87.79% falls within the lower limit of the range accepted as clinically acceptable in the literature.

In comparison, EfficientNet exhibited the highest performance with an accuracy of 99.55%. Its precision and recall values are very close, reflecting the model's success in accurately diagnosing tumors. Both VGG16 and CNN models also showed high accuracy rates (98.71% and 98.58%) and excellent precision values (99.74%). However, when compared to the FL model, these models exhibited higher recall values (97.69% and 97.43%), which is a crucial factor in healthcare applications, where false negatives can have significant consequences. This situation demonstrates that further development of FL is required in order to offer a more balanced and reliable solution.

MobileNetV3 and ResNet50 models, with accuracy rates of 95.43% and 94.01%, respectively, presented the lowest performance. Nevertheless, their precision and F1-score values are still at levels suitable for healthcare applications, providing useful results.

5. CONCLUSION AND DISCUSSIONS

Accurate diagnosis of HCC is essential for guiding life-saving treatments. In this study, it is shown that FL—a privacy-preserving ML approach—can reliably detect HCC from MRI medical imaging. The necessity to classify liver cancer while protecting patient privacy has become increasingly important. This study investigates the feasibility of using FL and DL to categorize liver cancer. The study also examines the comparisons of CNN, EfficientNet, MobileNetV3, ResNet50, and VGG16 with FL.

This study confirms that FL, particularly when integrated with EfficientNetB5, is a viable solution for privacy-preserving liver tumor classification in MRI imaging. The findings demonstrate the system's capacity to generate outcomes that are competitive with those of the baseline system, while utilizing fewer resources and ensuring superior data privacy. These results underscore the practical viability of FL for HCC classification in clinical environments.

Conversely, it is imperative to recognise that the Federated EfficientNet model exhibits a lower classification accuracy (93.75%) in comparison to its central counterpart (99.55%). This performance gap is expected, given the decentralized training constraints and the limited global information available to each client. However, this trade-off is justified in privacy-sensitive environments where direct data sharing is infeasible. Future work will aim to reduce this gap by incorporating more sophisticated aggregation techniques (e.g., FedProx, FedOpt), dynamic client selection, and hybrid learning approaches that strike a better balance between performance and privacy.

Although the proposed FL framework achieved promising accuracy and ROC-AUC scores, the recall value remained relatively lower compared to clinical expectations. This indicates that the model may still miss a portion of true positive cases, which is critical in hepatocellular carcinoma diagnosis. To address this limitation, future work will explore advanced strategies such as cost-sensitive learning, focal loss functions, and ensemble techniques tailored to enhance recall. Additionally, incorporating clinical metadata (e.g., patient history, lab results) alongside imaging data could improve diagnostic robustness. Finally, investigating personalization techniques in FL, such as federated fine-tuning or clustering-based model adaptation, may further reduce false negatives by tailoring models to client-specific distributions.

While FedAvg facilitates privacy-aware training by avoiding direct data sharing, recent studies have shown that it is still vulnerable to privacy leakage through indirect inference attacks. In particular, gradient inversion attacks [28] can reconstruct sensitive input data from shared model updates, while membership inference attacks [29] can expose whether a particular data point was part of a client's training set. These vulnerabilities highlight the need for complementary privacy-preserving mechanisms such as differential privacy, secure aggregation, or homomorphic encryption. Incorporating such techniques in future

implementations will be critical to ensuring stronger privacy guarantees without compromising model performance.

Future work may also involve the deployment of FL frameworks in multi-institutional hospital networks and the integration of multimodal clinical data.

6. REFERENCES

1. Singh, A. & Pandey, B. (2016). Diagnosis of liver disease by using least squares support vector machine approach. *International Journal of Healthcare Information Systems and Informatics, 11*(2), 62-75.
2. Çaviş, T. & Arda, K.N. (2024). Advanced magnetic resonance imaging techniques in the diagnosis of liver diseases. *The Turkish Journal of Current Gastroenterology, 26*(3), 130-141.
3. Chan, H.P., Samala, R.K., Hadjiiski, L.M. & Zhou, C. (2020). Deep learning in medical image analysis. *In: Lee, G., Fujita, H. (eds) Deep Learning in Medical Image Analysis. Advances in Experimental Medicine and Biology*, Springer, 181.
4. Kwak, L. & Bai, H. (2023). The role of federated learning models in medical imaging. *Radiology: Artificial Intelligence, 5*(3), 1-2.
5. Llovet, J.M., Kelley, R.K., Villanueva, A., Singal, A.G., Pikarsky, E., Roayaie, S., Lencioni, R., Koike, K., Rossi, J.Z. & Finn, R.S. (2021). Hepatocellular carcinoma. *Nature Rev. Dis. Primers, 7*(1), 6-34.
6. Heimbach, J.K., Kulik, L.M., Finn, R.S., Sirlin, C.B., Abecassis, M.M., Roberts, L.R., Zhu, A.X., Murad, M.H. & Marrero J.A. (2018). AASLD guidelines for the treatment of hepatocellular carcinoma. *Hepatology, 67*(1), 358-380.
7. Shao, Y.Y., Wang, S.Y. & Lin, S.M. (2021). Management consensus guideline for hepatocellular carcinoma: 2020 update on surveillance, diagnosis, and systemic treatment by the Taiwan liver cancer association and the gastroenterological society of Taiwan. *J Formos Med Assoc., 120*(4), 1051-1060.
8. Chen, H., Gomez, C., Huang, C.M. & Unberath, M. (2022). Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. *npj Digital Medicine, 5*(156), 1-15.
9. Song, L., Geoffrey, K. & Kaijian, H. (2020). Bottleneck feature supervised U-net for pixel-wise liver and tumor segmentation. *Expert Syst. Appl., 145*(5), 1-11.
10. Song, D., Wang, Y., Wang, W. & Wang, Y. (2021). Using deep learning to predict microvascular invasion in hepatocellular carcinoma based on dynamic contrast-enhanced MRI combined with clinical parameters. *J. Cancer Res. Clin. Oncol., 147*(12), 3757-3767.
11. Srinivasu, P.N., Lakshmi, G.J., Narahari, S.C., Shafi, J., Choi, J. & Ijaz, M.F. (2024). Enhancing medical image classification via federated learning and pre-trained model. *Egyptian Informatics Journal, 27*(1), 1-16.
12. Roth, H.R., Chang, K., Singh, P., Neumark, N., Li, W., Gupta, V., Gupta, S., Qu, L., Ihsani, A., Bizzo, B.C., Wen, Y., Buch, V., Shah, M., Kitamura, F., Mendonça, M., Lavor, V., Harouni, A., Compas, C., Tetreault, J., Dogra, P., Cheng, Y., Erdal, S., White, R., Hashemian, B., Schultz, T., Zhang, M., McCarthy, A., Yun, B.M., Sharaf, E., Hoebel, K.V., Patel, J.B., Chen, B., Ko, S., Leibovitz, E., Pisano, E.D., Coombs, L., Xu, D., Dreyer, K.J., Dayan, I., Naidu, R.C., Flores, M., Rubin, D. & Cramer, J.K. (2020). Federated learning for breast density classification: a real-world implementation. *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning, DART 2020*, Peru.
13. Bernecker, T., Peters, A., Schlett, C.L., Bamberg, F., Theis, F., Rueckert, D., Weib, J. & Albarqouni, S. (2022). FedNorm: modality-based normalization in federated learning for multi-modal liver segmentation. *ArXiv preprint*, abs/2205.11096, 1-21.
14. Mahlool, D.H. & Abed, M.H. (2022). Distributed brain tumor diagnosis using a federated learning environment. *Bulletin of Electrical Engineering and Informatics, 11*(6), 3313-3321.
15. Triverdi, N.K., Shukla, S., Tiwari, R.G., Agarwal, A.K. & Gautam, V. (2023). Liver cancer diagnosis with lightweight federated learning using identically distributed images. *12th International Conference on System Modeling & Advancement in Research Trends (SMART-2023)*, Moradabad, India.
16. Chai, H., Huang, Y., Xu, L., Song, X., He, M. & Wang, Q. (2024). A decentralized federated learning-based cancer survival prediction method with privacy protection. *Heliyon, 10*(11), 1-11.
17. Lusnig, L., Sagingalieva, A., Surmach, M., Protasevich, T., Michiu, O., McLoughlin, J., Mansell, C., Petris, G.D., Bonazza, D., Zanconati, F., Melnikov, A. & Cavalli, F. (2024). Hybrid quantum image classification and federated learning for hepatic steatosis diagnosis. *Diagnostics, 14*(5), 1-16.

18. Shankar, P.U., Rahul, E.S., Rao, K.D., Satish, K., Kumar, U.D., Ravindra, D. & Subbarao, G. (2025). Liver disease prediction using federated learning. *International Journal of Innovative Science and Research Technology, 10*(4), 880-887.

19. Balla Fofanah, A., Özbilge, E. & Kırsal, Y. (2023). Skin cancer recognition using compact deep convolutional neural network. *Çukurova Üniversitesi Mühendislik Fakültesi Dergisi, 38*(3), 787-797.

20. Fırat, H. & Üzen, H. (2024). MR görüntülerinden alzheimer hastalığının sınıflandırılması için inception ve sıkma-uyarma ağı tabanlı derin öğrenme modeli. *Çukurova Üniversitesi Mühendislik Fakültesi Dergisi, 39*(2), 555-567.

21. Quinton, F., Popoff, R., Presles, B., Leclerc, S., Meriaudeau, F., Nodari, G., Lopez, O., Pellegrinelli, J., Chevallier, O., Gignac, D., Vrigneaud, J.-M. & Alberini, J.-L. (2023). A tumour and liver automatic segmentation (ATLAS) dataset on contrast-enhanced magnetic resonance imaging for hepatocellular carcinoma. *Data, 8*(5), 1-9.

22. Lotfinia, M., Tayebiarasteh, A., Samiei, S., Joodaki, M. & Arasteh, S.T. (2025). Boosting multi-demographic federated learning for chest x-ray analysis using general-purpose self-supervised representations. *European Journal of Radiology Artificial Intelligence, 3*(1), 1-13.

23. Overman, T. & Klabjan, D. (2025). Continuous-time analysis of federated averaging. *ArXiv preprint*, abs/2501.18870, 1-25.

24. Yasaka, K., Akai, H., Abe, O. & Kiryu, S. (2018). Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: A preliminary study. *Radiology, 286*(3), 887-896.

25. Ahn, J.C., Qureshi, T.A., Singal, A.G., Li, D. & Yang, J.D. (2021). Deep learning in hepatocellular carcinoma: current status and future perspectives. *World J Hepatol., 13*(12), 2039-2051.

26. American College of Radiology (ACR), (2021). *Liver imaging reporting and data system (LI-RADS) v2018 manual.* Retrieved from https://www.acr.org/Clinical-Resources/Clinical-Tools-and-Reference/Reporting-and-Data-Systems/LI-RADS, Access date: 16/07/2025.

27. Choi, J.Y., Lee, J.M. & Sirlin, C.B. (2014). CT and MR imaging diagnosis and staging of hepatocellular carcinoma: Part II. Extracellular agents, hepatobiliary agents, and ancillary imaging features. *Radiology, 273*(1), 30-50.

28. Zhu, L., Liu, Z. & Han, S. (2019). Deep leakage from gradients. *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, Canada.

29. Shokri, R., Stronati, M., Song, C. & Shmatikov, V. (2017). Membership inference attacks against machine learning models. *38th IEEE Symposium on Security and Privacy (SP)*, San Jose, CA.