Precision and Robust Models on Healthcare Institution Federated Learning for Predicting HCC on Portal Venous CT Images

Chiu-Han Hsiao, *Member, IEEE*, Frank Yeong-Sung Lin, Tzu-Lung Sun, Yen-Yen Liao, Chih-Horng Wu, Yu-Chun Lai, Hung-Pei Wu, Pin-Ruei Liu, Bo-Ren Xiao, Chien-Hung Chen, and Yennun Huang, *Fellow, IEEE*

***Abstract*—Hepatocellular carcinoma (HCC), the most common type of liver cancer, poses significant challenges in detection and diagnosis. Medical imaging, especially computed tomography (CT), is pivotal in non-invasively identifying this disease, requiring substantial expertise for interpretation. This research introduces an innovative strategy that integrates two-dimensional (2D) and three-dimensional (3D) deep learning models within a federated learning (FL) framework for precise segmentation of liver and tumor regions in medical images. The study utilized 131 CT scans from the Liver Tumor Segmentation (LiTS) challenge and demonstrated the superior efficiency and accuracy of the proposed Hybrid-ResUNet model with a Dice score of 0.9433 and an AUC of 0.9965 compared to ResNet and EfficientNet models. This FL approach is beneficial for conducting large-scale clinical trials while safeguarding patient privacy across healthcare settings. It facilitates active engagement in problem-solving, data collection, model development, and refinement. The study also addresses data imbalances in the FL context, showing resilience and highlighting local models' robust performance. Future research will concentrate on refining federated learning algorithms and their incorporation into the continuous implementation and deployment (CI/CD) processes in AI system operations, emphasizing the dynamic involvement of clients. We recommend a collaborative human-AI endeavor to enhance feature extraction and knowledge transfer. These improvements are intended to boost equitable and efficient data collaboration across various sectors in practical scenarios, offering a crucial guide for forthcoming research in medical AI.**

***Index Terms*—Deep learning, federated learning, hepatocellular carcinoma, image segmentation, transfer learning.**

I. Introduction

HEPATOCELLULAR carcinoma (HCC) poses a major threat to public health and will affect over 1 million individuals globally by 2025. HCC and colon metastasis are the predominant forms of primary and metastatic liver cancer [1]. Sub-Saharan Africa and Eastern Asia have a higher incidence of HCCs than other regions due to hepatitis B and C virus infection. Dynamic computed tomography (CT) is one of the main methods for diagnosing HCC according to many practice guidelines [2], [3]. However, interpreting CT images slice-by-slice is an arduous and time-consuming task that demands the expertise of hepatologists, surgeons, oncologists, or radiologists [4]. Doctors can reduce their workload and minimize subjective errors by using computer vision techniques for feature extraction and pattern recognition [5]. Using computer vision techniques to automate liver tumor identification enhances liver cancer diagnosis efficiency and accuracy. It facilitates early detection for optimal treatment and increases survival rates [6].

Segmenting liver tumors from CT images is challenging (as illustrated in Fig. 1). Contrast-enhanced studies typically exhibit hyperdense contrast enhancement during the arterial phase and hypodense contrast washout during the portal venous or delayed phase. Consequently, identifying small tumors becomes more complicated since they have similar grayscale intensities to the surrounding liver tissue [1]. Moreover, accurately delineating tumors from adjacent tissue poses a challenge due to the complex and indistinct boundaries tumors often exhibit [7].

Deep learning (DL) techniques, such as voxel-wise segmentation models, have recently gained prominence as primary approaches for clinical data analysis and have demonstrated effectiveness in identifying liver contours and calculating volume [8]. Notably, UNet models and their variations have proven valuable in segmenting the liver and tumors from medical images

Manuscript received 13 June 2023; revised 6 February 2024 and 1 April 2024; accepted 8 May 2024. Date of publication 13 May 2024; date of current version 7 August 2024. This work was supported in part by the Academia Sinica and National Science and Technology Council, Taiwan under Grant 3012-73C3803, Grant NSTC 111-2221-E-001-020, Grant NSTC 111-2321-B-075-004, and Grant NSTC 112-2221-E-001 -024 -MY2. *(Corresponding author: Chih-Horng Wu.)*

Chiu-Han Hsiao, Yen-Yen Liao, Yu-Chun Lai, and Yennun Huang are with the Research Center for Information Technology Innovation, Academia Sinica, Taipei 115, Taiwan (e-mail: chiuhanhsiao@citi.sinica.edu.tw; tiffany115040@gmail.com; wayne910139@gmail.com; yennunhuang@gmail.com).

Frank Yeong-Sung Lin, Tzu-Lung Sun, Hung-Pei Wu, Pin-Ruei Liu, and Bo-Ren Xiao are with the Department of Information Management, National Taiwan University, Taipei 10617, Taiwan (e-mail: yeongsunglin@gmail.com; suntlung@gmail.com; juliawu0717@gmail.com; B06705058@ntu.edu.tw; kogktt187@gmail.com).

Chih-Horng Wu is with the Center of Minimal-Invasive Interventional Radiology and Department of Medical Imaging, National Taiwan University Hospital, Taipei 100, Taiwan (e-mail: chw1020@ntuh.gov.tw).

Chien-Hung Chen is with the Department of Internal Medicine, National Taiwan University Hospital, Taipei 100, Taiwan (e-mail: chenhcc@ntuh.gov.tw).

Digital Object Identifier 10.1109/JBHI.2024.3400599

due to their convolution and deconvolution processes [9]. UNet models, based on the dimensions of their input training data, can be categorized into two-dimensional (2D) [10], two-and-a-half dimensional (2.5D) [11], [12], [13], [14], or three-dimensional (3D) [15], [16] models. The 2.5D model bears similarity to the 2D model, with the 2D model's input data being either one channel or three channels, depending on whether it is a grayscale or RGB image. The 2.5D model is characterized by its use of a stack of adjacent slices (more than three slices) as input and generates a segmentation map for the central slice. This approach of using 2.5D input effectively reduces the model's size. In the studies by Zhou et al. [12] and Qin et al. [13], varying numbers of slices (3, 5, 7, etc.) from a scan were used to create a segmentation map for the middle slice. In contrast, 3D models involve inputting all slices into the deep learning model for prediction.

Furthermore, these models can be organized based on their intended use and processing methodology. For instance, one model may be more suitable for liver segmentation, while another may be better suited to tumor segmentation. Additionally, preprocessing can improve accuracy and reduce computational resources. It is also a flexible way to customize the UNet model autoencoder architecture [17].

However, the challenge of insufficient data for training deep learning models in medical imaging is substantial. Federated learning (FL) presents a viable solution to this challenge [18], [19]. FL enables the development of models using a decentralized architecture, eliminating the need for a centralized database or the exchange of images among multiple hospitals, thereby addressing privacy and regulatory compliance concerns [19]. Consequently, this study focused on employing global FL algorithms to improve segmentation results without the need to share local datasets. This approach not only maintained data privacy but also achieved high levels of accuracy and efficiency. Additionally, the study underscored the clinical significance of CT imaging in the diagnosis of HCCs. These enhancements strengthen the foundation of this paper and ensure it accurately reflects current advancements in our research field.

This study aimed to enhance the accuracy of liver tumor segmentation models for hepatocellular carcinoma detection using deep learning techniques. However, regulatory and privacy concerns restrict cross-hospital sharing of medical data for centralized learning [20]. Hence, we sought to design appropriate liver tumor segmentation models and an integrated federated learning framework to address the cross-hospital data-sharing problem. To our knowledge, this study is the first research to use a federated learning algorithm to address and solve the cross-hospital data-sharing problem for HCC detection. The key contributions are as follows:

- We have developed an advanced hybrid network architecture with technological innovations aimed at accurately defining the borders and locations of the liver and tumors. Our model employs both 2D and 3D cascading approaches to focus on specific image areas, thereby enhancing tumor detection and reducing the workload for clinicians. This approach also offers significant clinical value by providing spatial information, which improves image readability for subsequent applications. For instance, by stacking segmented slices of the liver, our model is capable of calculating liver volume. This feature is particularly useful in surgical planning as it helps in determining the remaining liver volume.
- We also demonstrated that our proposed hybrid model excelled in liver tumor detection and segmentation. It seamlessly integrated into a federated learning framework for clinical assessment, upholding data privacy through local training and enabling its application across various clinics.
- In conclusion, our study employed an FL framework that trained the model on distributed datasets, accommodating both balanced and imbalanced scenarios. This framework supported the development of a model that is generalizable for all participants, making it suitable for use across medical clinics. Moreover, it enabled the gathering of more training data, resulting in more effective models with improved accuracy. In this way, we can exchange parameters for modeling rather than patients' images for training among institutes. This result proposed a way to improve accuracy and keep data safety and confidentiality.

In summary, this paper presents a method to enhance clinical decision-making by employing a deep learning and federated learning model based on radiologist-annotated CT scans. The goal is to improve the assessment of patient conditions. Currently, our proposed Hybrid-ResUNet method combined with FedAvg demonstrates improved efficiency and accuracy, achieving a Dice score of 0.9433 and AUC of 0.9965. Future objectives include developing HCC-specific deep learning models for malignancy risk prediction and treatment planning, optimizing algorithms for HCC analysis in CT images within a federated learning framework, and commercializing the predictive model as a software as medical device (SaMD) for practical use and ongoing optimization. This SaMD, having been validated across multiple hospitals, will deliver essential information to doctors and patients, aiding in the prediction and prevention of liver diseases. This underscores our dedication to advancing medical AI.

II. Related Work

The following subsections provide an overview of two related topics: (1) hybrid deep learning models and (2) federated learning algorithms.

A. Hybrid Deep Learning Models

The minimization of computing resources is crucial for cost-effective liver tumor segmentation. While 2D models require fewer resources, their accuracy is lower [21]. On the other hand, 3D models demand more resources and time but provide more accurate results [16]. Practical computing constraints should be considered when selecting models for clinical settings. Various studies have focused on semantically segmenting medical images. For instance, the H-DenseUNet, a hybrid densely connected UNet, was introduced for liver and tumor segmentation [22]. It integrates a 2D DenseUNet for slice-based

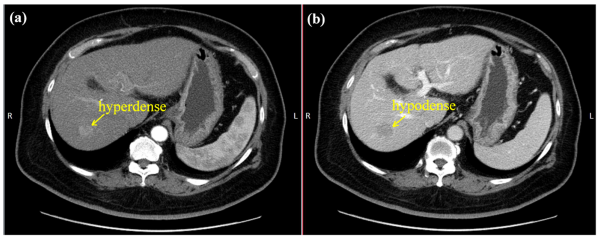

Fig. 1. Hepatocellular carcinoma in CT images. (a) Arterial phase. (b) Portal venous phase.

feature extraction and a 3D DenseUNet for hierarchical context aggregation. Another example is the hybrid CNN combined with a 3D V-Net model for automatic lung tumor delineation in CT images [21]. This model utilizes an encoder-decoder structure with dense connections and combines 2D and 3D features into a single module. A two-stage approach to liver tumor segmentation was proposed, leveraging 3D models and employing a cascaded fully convolutional network (FCN) based on the improved 3D-ResUNet [23]. The first FCN segments the liver as regions of interest (ROIs), while the second FCN focuses on tumor segmentation. In this paper, the overall processing flow of the proposed model can be modified by employing 2D, 3D, or hybrid models. It can be done by narrowing down the ROI in stages and conserving computing resources. The cascaded approach (Hybrid-ResUNet) can also be split into two phases; the first stage is a 2D model and the second is a 3D model for tumor detection.

B. Federated Learning Algorithms

Federated Learning involves multiple clients exchanging model parameters with a server while keeping their datasets local and not shared [20]. Each client trains a model on its data, and a central system aggregates the model parameters. The server processes these parameters to update an aggregated model, that is then shared with clients. This approach ensures data privacy and avoids sensitive data sharing. However, FL models have limitations, prompting the development of various methods such as data partitioning techniques, advanced privacy protection mechanisms, communication architectures, and system heterogeneity [24]. FL models can be updated between clients using sequential or cyclic parameter updating. The federated averaging (FedAvg) method is commonly used in FL to optimize model weights [25]. The server gathers and averages gradients from each local client, distributing the results to clients for model updating and further training. FedAvg has performed well in multi-tasking learning studies [26]. FL has been applied to various tasks, including mining industrial data, secure image analysis, and training models on decentralized medical datasets. During the COVID-19 pandemic, FL trained a global model for distributed disease detection from CT images [27]. FL can be implemented as cross-silo or cross-device learning. Cross-silo learning in multihospital systems requires stable connections and reliable client computing environments. Based on previous studies, FL offers a practical and cost-effective framework for decentralized learning using private medical data to optimize clinical outcomes systematically.

In the field of medical applications, researchers led by Houda introduced a novel framework named HealthFed, which employs FL and blockchain technology [19]. This framework enables multiple clinical practitioners to engage in privacy-protected and decentralized learning. Extensive experiments conducted using a publicly available breast cancer dataset demonstrated that HealthFed not only ensures the privacy of each collaborator's sensitive data but also delivers accurate learning models. These results position HealthFed as a promising framework for medical systems.

C. Summary

This paper introduces Hybrid-ResUNet, a cost-effective hybrid model that combines 2D and 3D segmentation techniques for liver and tumor detection in stages. It balances accuracy and performance, making it suitable for affordable AI applications and integration into clinical FL architectures. FedAvg improves performance without exchanging medical information. Added more hospitals increased model complexity due to increased dataset access. The FL framework ensures all participants receive the optimized model. The hybrid model within the FedAvg framework was evaluated for balanced and imbalanced data scenarios, yielding promising results. The proposed approach for HCC detection can be applied to medical system alliances, enhancing performance without data sharing. This can potentially improve diagnostic accuracy in small hospitals at a reasonable cost [28].

III. Proposed Methods

A. Ethical Approval

The protocols and waiver requests for retrospective data collection and existing biosamples (REC No. 202109100RINC) have been approved by the 148th meeting of the Research Ethics Committee C of the National Taiwan University Hospital. Government regulations and Good Clinical Practice guidelines are followed in the conduct of the experiments. Written informed consent is not required from study participants. The liver tumor segmentation study utilizes the liver tumor segmentation (LiTS) dataset, a publicly available dataset without patient identification, hence the waiver of consent [29].

B. Model Architecture

This study proposed a two-stage approach for liver tumor detection, as shown in Fig. 2. The UNet model serves as the baseline in both stages. In the first stage, medical images are subjected to windowing and preprocessing to amplify the liver region's intensity. The ResNet-50 model is then utilized to isolate liver slices from the original dataset. Subsequently, in the second stage, the isolated liver data is inputted into a 2D-UNet model to assess the influence of various encoders on liver and tumor detection accuracy. Ultimately, a 3D-UNet model is deployed for the segmentation of liver tumors. Conversely, 3D models process by inputting the entire stack of slices into

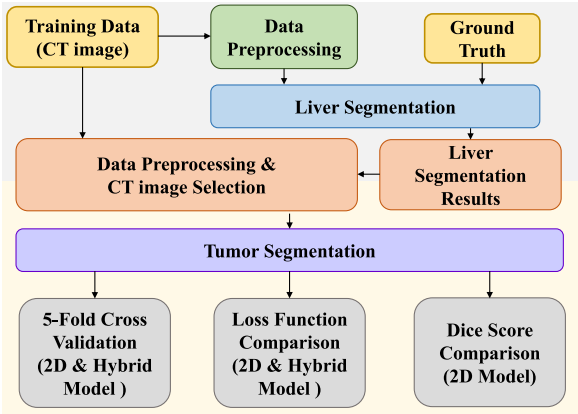

Fig. 2. Flowchart of the two-stage liver segmentation and tumor identification approach.

the deep learning model for comprehensive prediction. This creates a hybrid model that incorporates 2D and 3D methods for liver tumor identification. Utilizing the 2D and 3D models for liver tumor detection enhances the hybrid model's performance and accuracy. Consequently, by combining the data from the 2D model, the 3D-UNet model achieves better tumor detection within the liver region. The study also investigated the influence of various loss functions on model performance during training.

1) UNet Model: UNet is widely recognized for its effectiveness in biomedical image segmentation [9], [30]. Its encoder-decoder structure and skip connections enable efficient training on small labeled datasets. The compact encoder extracts high-level features, which are then resampled and combined with low-level features by the decoder to segment small targets accurately. Unlike traditional segmentation methods, UNet avoids error propagation by directly incorporating input and output feature through skip connections. Additionally, UNet's convolution and pooling operations allow it to process images of varying sizes. This study employed ResNet, DenseNet [31], and EfficientNet [32] as encoders for liver area segmentation. The performance of each encoder model was evaluated in the first-stage 2D model.

2) 2D-UNet With Distinct Encoders: He et al. [33] developed a residual learning framework to tackle the degradation issue encountered with the increase of layers in ResNet architectures. ResNet is structured into stages, each comprising multiple building blocks. Variants such as ResNet-18, ResNet-50, and ResNet-121 utilize distinct layers of building blocks to improve accuracy. Notably, ResNet-50 incorporates bottleneck blocks, which enable more efficient parameter reduction compared to the ResNet-18 or ResNet-34 models that use standard residual blocks.

DenseNet [31] distinguishes itself from ResNet by employing a dense concatenation strategy that connects all layers linearly, facilitating a direct link between the initial and subsequent layers. With $\frac{n(n+1)}{2}$ connections among its n layers, DenseNet promotes feature reuse by concatenating feature maps from various layers. This methodology not only boosts performance but also significantly reduces the number of parameters and computational expenses. The dense connection technique ensures uniformity in feature maps. Additionally, transition layers that function as pooling layers are placed between consecutive dense blocks, effectively reducing the dimensions of feature maps. In our study, we selected the DenseNet-121 model, pre-trained with ImageNet weights, to serve as the encoder within the UNet architecture.

EfficientNet models, spanning from B0 to B7, employ the mobile inverted bottleneck (MBConv) as their fundamental building block [32]. These models utilize a compound scaling method that harmonizes network width, depth, and resolution, scaling them based on a fixed constant. By adjusting this constant, the network can be expanded, leading to variations from EfficientNet-B1 through EfficientNet-B7. In this experiment, the noisy student technique alongside pre-trained weights of EfficientNet was employed to deploy EfficientNet-B0 and EfficientNet-B5.

Li et al. [22] introduced H-DenseUNet, a method that starts by dividing original CT images into adjacent slices for input into a 2D model. Following this, 3D features extracted by a 3D model are combined with the 2D and 3D inputs through hybrid feature fusion, resulting in the output of the 3D model. This integration of 2D and 3D models preserves the 3D features that are otherwise lost in purely 2D modeling, while also alleviating the computational burden associated with fully 3D models.

Therefore, due to the significant memory usage caused by the extensive concatenation in the DenseUNet approach, we chose to adapt the 2D, 3D, and hybrid models into 2D-ResUNet, 3D-UNet, and Hybrid-ResUNet in stage 2, respectively, with the goal of reducing computational requirements.

The 2D-ResUNet architecture features several stages, each incorporating a variety of building blocks. This model resembles the 2.5D model with the input image size set to 160x160x3, where the input comprises three channels from a stack of adjacent slices, creating a segmentation map for the central slice, thereby effectively minimizing the model's size. Notably, the ResNet-50 encoder utilizes bottleneck blocks, enabling a more efficient reduction of parameters compared to ResNet-18 or ResNet-34 models, which employ general residual blocks. The network architecture of the 2D-ResUNet is illustrated in Fig. 3. The 2D-ResUNet model integrates the skip connections from ResNet into the UNet architecture to improve the performance of semantic segmentation. It begins with initial processing steps, including convolution with 64 filters and batch normalization, followed by max-pooling for downsampling. Residual blocks in stages 2 to 5 maintain spatial dimensions and perform spatial downsampling as needed. During upsampling, skip connections are integrated, enriching contextual understanding and preserving details. Upsampling involves convolutional layers with decreasing filter sizes, followed by batch normalization and ReLU activation. Dropout regularization can be applied for overfitting control. For multi-class classification, the last layer lowers the channels to three classes using softmax activation.

The 3D-UNet model includes stages in the downsampling path, each with two 3D convolutional layers (3x3x3 kernel size), batch normalization, ReLU activation, and max-pooling after the first two stages. The image size is 160x160x64. Feature maps

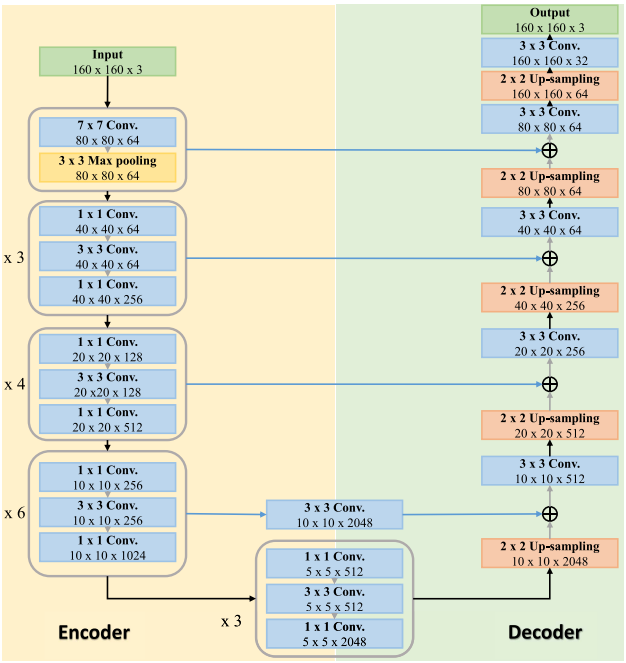

Fig. 3. Depiction of the 2D-ResUNet model architecture.

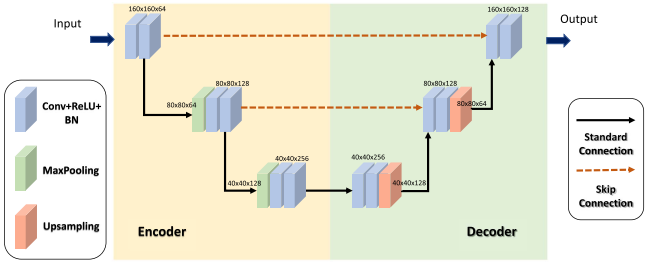

Fig. 4. Depiction of the 3D-UNet model architecture.

are stored for concatenation in the upsampling path. The upsampling path also consists of stages with 3D transposed convolutional layers, concatenating with corresponding downsampling features. Each upsampling stage includes two 3D convolutional layers, batch normalization, and ReLU activation. The final layer is a 3D convolutional layer (1x1x1 kernel size) for multi-class segmentation. The architecture is illustrated in Fig. 4. Compared with the 2D model, the convolution layers of the 3D model include filters with 3D size. The 3D convolution layers can capture spatial information, meaning the 3D model needs more parameters and more computing resources. The batch size when the training stage cannot be set to a sufficient size. In addition, the 3D model has another problem. That is, the 3D model architecture lacks the weights of pre-trained models to load. Therefore, many types of research that use the 3D model are not satisfactory, even worse than some outstanding 2D models.

This modified model architecture, referred to as Hybrid-ResUNet, is depicted in Fig. 5. The hybrid model combines 2D-ResUNet and 3D-UNet branches to leverage both 2D and 3D

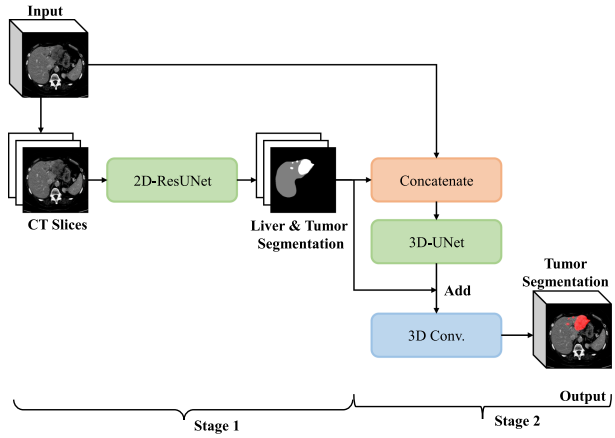

Fig. 5. Proposed Hybrid-ResUNet model architecture as a fusion of the 2D-ResUNet and the 3D-UNet.

information for enhanced feature extraction and classification. The model's core consists of an input layer for 3D volumetric data (160x160x64), which is processed in batches through the 2D-ResUNet branch. This branch extracts features from each 2D slice of the input volume and combines adjacent slices using concatenation operations, creating a comprehensive representation of the 2D information.

C. Transfer Learning

Due to the scarcity of annotated training data for liver segmentation, this paper leverages transfer learning as an effective solution. Specifically, the kidney tumor segmentation 2019 (KiTS19) dataset [34] was employed as the source data and the LiTS Challenge dataset as the target data for HCC segmentation. Both datasets are formatted in NIfTI (.nii). A notable challenge in transfer learning is the potential mismatch of features when the target and source datasets significantly differ. To mitigate this, the coordinate system of the LiTS dataset was adjusted to match KiTS19. The approach involved initializing the proposed Hybrid-ResUNet model with weights from a pre-trained kidney tumor detection model based on the KiTS19 dataset. This strategy enabled the effective use of pre-trained weights in the retraining and prediction phases for HCC detection, thereby accelerating training and improving accuracy with a model initially trained on an extensive dataset.

D. Loss Function

Deep learning utilizes a loss function to gauge the difference between predicted and actual outcomes, aiming to minimize this function during training to improve accuracy [35]. The objective is to understand the relationship between features and labels in the training data to accurately predict outcomes in testing data. Choosing the right loss function is pivotal for the performance of medical image segmentation, as different loss functions may be more effective depending on the data characteristics and the model's architecture. In this research, four distinct loss functions were evaluated for their effectiveness with the 2D-ResUNet and

Hybrid-ResUNet models to identify the most suitable one for our application. A particular challenge was the misclassification of small-sized liver tumors in CT images to background, prompting the adjustment of weights to 8.57 for tumors and 0.7 for liver regions, respectively.

1) Categorical Cross-Entropy Loss: Categorical cross-entropy loss has advantages in medical image segmentation: it is simple to implement and commonly used in deep learning models, making it convenient for practical applications. Gradients can be calculated with predicted probabilities and proper labels, improving segmentation accuracy. It is also suitable for various classification tasks, including binary and multiclass classification, making it widely used in medical image segmentation. However, it has limitations in handling imbalanced classes and voxel correlation, which must be considered for accurate and stable segmentation results. In multiclass tasks, categorical cross-entropy is often called binary cross-entropy and utilizes a softmax activation function at the final neural network layer.

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{1}$$

where $\mathbf{y} = [y_1, y_2, \ldots, y_C]$ is the true one-hot encoded label vector with C categories, and $\hat{\mathbf{y}} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_C]$ is the predicted probability vector over the same C categories in (1).

2) Dice Loss: The Dice loss function is ideal for medical image segmentation, especially in imbalanced class distributions. It effectively addresses situations where the number of voxels for different classes varies significantly. This loss function prioritizes the minority class, improving segmentation accuracy. Moreover, it captures intervoxel relationships by calculating the Dice coefficient between predicted and actual segmentation results, resulting in contiguous segmentation regions. The Dice loss function also provides clear gradient signals, facilitating better model training and faster convergence for enhanced segmentation accuracy. It was specifically chosen to improve liver tumor segmentation.

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum_{i=1}^{N} y_i \hat{y}_i + \epsilon}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i + \epsilon} \tag{2}$$

where y_i and $\hat{y}_i$ are the actual label and predicted probability of the ith voxel, N is the total number of pixels or voxels in the image, and ϵ is a small constant (either 1 or $1e^{-5}$) added to the denominator to avoid division by zero in (2). The Dice Loss is used in image segmentation tasks to optimize the overlap between predicted and actual segmentation masks. This measure mitigates overfitting and ensures numerical stability.

3) F-Score Loss: The F-score loss function effectively addresses imbalanced class distributions in medical image segmentation. It improves the segmentation accuracy of the minority class by assigning it more weight while maintaining a balance between precision and recall. The F-score loss also provides clear gradient signals, aiding model training and convergence for superior segmentation results. F-score loss parameters can be fine-tuned to meet specific practical requirements for optimal segmentation outcomes. In (3), the coefficient β balances the precision-recall ratio, with a higher value emphasizing precision.

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta \cdot precision) + recall)} \tag{3}$$

F-scores can be used for micro-averaging or macro-averaging. In micro-averaging, the average is calculated across all samples, regardless of their class. The macro-averaging method calculates the average for each class. F-score values range from 0 to 1, with higher values indicating better performance. The loss function can be optimized as (1 − F-score).

4) Combo Loss: The combo loss combines the cross-entropy and Dice loss functions, merging their strengths to enhance model performance. It leverages Dice loss's ability to handle class imbalance, voxel correlation, and cross-entropy gradient signals. By overcoming the limitations of each loss function, such as the local minima trap of Dice loss and poor performance with imbalanced classes of cross-entropy, the combo loss aims to improve recall while maintaining accuracy for superior segmentation outcomes. It provides clear gradient signals, aiding in model training and convergence and improving segmentation results. This versatile function can be applied to both binary and multiclass segmentation tasks. See (4) for the combo loss description.

$$\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{Dice} \tag{4}$$

E. Federated Learning Algorithm

Federated Learning enables model training on distributed devices while preserving user privacy. It is particularly suitable for medical image segmentation, where centralized training is not feasible due to data privacy concerns. Training data remains on local devices in FL, and only model parameters are aggregated on a central server. FL benefits from diverse data sources, enhancing model generalization. Medical image segmentation datasets often suffer from class imbalance and high annotation costs, which can be addressed through local training in FL.

Our study employed an FL framework for liver tumor segmentation in CT images using distributed datasets and Hybrid-ResUNet models. The framework involved three clients exchange model weights with a central server through communication rounds. The FedAvg algorithm, proposed by McMahan et al. [36], was used for FL training. The algorithm consists of local and global training steps, with computational requirements determined by the number of communication rounds ($\boldsymbol{R}$) and local training epochs ($\boldsymbol{E}$).

During the local training step, the server broadcasts the current model parameters (w_0) to all participants. Each client initializes their local model with these parameters (w_j) and trains it on their respective dataset D_j. After training, the updated model parameters (w_j^{r+1}) are sent back to the server, which aggregates them to generate a revised server model (w_G^{r+1}). This updated model is then distributed to all clients for further training. This process is repeated until a stopping criterion is met, resulting in a global model ($\boldsymbol{W_G}$) for the distributed dataset. Early stopping can be implemented to monitor model performance during

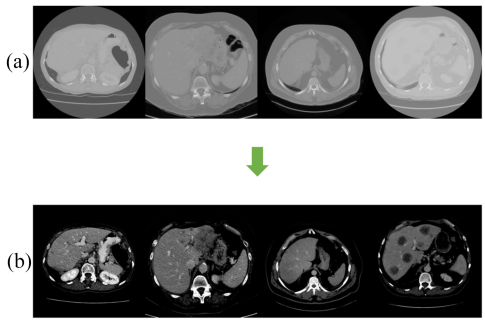

Fig. 6. Comparison of images with and without the windowing method. (a) No windowing. (b) HU values set to −200 to 250.

Algorithm 1: FedAvg [36].

Input: Number of communication rounds R,
Number of local epochs E,
Local minibatch size B,
Learning rate η
Output: Global Model W_G

Function SERVERPROCESS ▷ Run on the server
  initialize w_0
  $S \leftarrow$ (a set of N clients)
  for each round r from 1 to R **do**
    for each client $j \in S$ **Parallel do**
      $w_j^{r+1} \leftarrow \text{CLIENTUPDATE}(j, w, E)$
    end for
    $w_G^{r+1} \leftarrow \sum_{j \in S} \frac{n_j}{n} w_j^{r+1}$
  end for
end function

function CLIENTUPDATEj, w, E ▷ Run on client j
  $\mathcal{B} \leftarrow$ (split D_j into batches of size B)
  for each local epoch e from 1 to E **do**
    for batch $b \in \mathcal{B}$ **do**
      $w \leftarrow w - \eta \nabla l(w; b)$
    end for
  end for
  return w to server
end function

training. The complete pseudo-code for the FedAvg algorithm can be found in Algorithm 1.

IV. Experiments

A. Dataset

This study used the LiTS dataset, which includes 131 cases [29]. All cases include contrast-enhanced CT volumes in the portal venous phase, with ground truth data available for the training cases. Each slice in the dataset is in NIfTI format, with a resolution of 512 pixels x 512 pixels. The number of slices in each volume ranges from 42 to 1026. The ground-truth data includes three labels: 0 for background, 1 for liver regions, and 2 for liver tumor regions. A few non-tumor cases are also included in the dataset (e.g., cases 32, 34, 38, 41, 47, 87, 89, 91, 105, 106, 114, 115, and 119). The dataset was collected from multiple clinics and research institutions and manually labeled by three independent radiologists. Liver tumors exhibit significant variations in contrast, size, and shape due to individual physiological differences. The dataset's diverse sources include different machines and measurement protocols. Some tumors may resemble healthy liver regions or be too small to detect. Training deep learning models with these divergent features is challenging.

B. Data Preprocessing

1) Windowing Process: CT volumes are quantified in Hounsfield units (HUs), which span from −1024 to 3071, varying with the organs or tissues under examination. The broad spectrum of voxel values in CT images necessitates complex processing. Preprocessing CT images is crucial for minimizing irrelevant voxel values and converting the image into a grayscale format, with values between 0 and 255. Insufficient preprocessing can skew the distribution of HU values for the liver, resulting in diminished contrast, increased noise, and challenges in differentiating the liver on a grayscale image. Windowing is a technique applied to improve contrast by focusing on a specific HU range. In this research, the min-max normalization method was utilized, transforming CT scans into grayscale images by adjusting values outside the selected HU range to the range's minimum and maximum values. The HU value ranges were established using (5) for windowing.

$$HU_{\max} = WindowLevel + \frac{WindowWidth}{2}$$
$$HU_{\min} = WindowLevel - \frac{WindowWidth}{2} \quad (5)$$

The min-max normalization method may cause information loss, and it can remove excessive image details by restricting the wide range of HUs to a specific interval. In this study, the final preprocessing step did not utilize min-max normalization. Instead, a HU window range of -200 to 250 was chosen, demonstrating optimal performance for tumor segmentation but deviating from the conventional clinical window range of -79 to 304. Fig. 6 illustrates that more information was preserved using this windowing technique, enabling visualization of the liver and tumor.

2) Data Augmentation: Due to the limited number of samples in the LiTS dataset, data augmentation techniques were implemented prior to training. These techniques were essential in addressing the scarcity of training data, aimed at increasing data diversity and preventing overfitting. The data augmentation primarily involved two types: geometric transformation and contrast adjustment. Geometric transformation involved rotation (90° or 180° counterclockwise), flipping (up, down, left, or right), and scaling (random zoom of 0.8%–1.2%). Contrast adjustment was achieved by applying gamma correction using a specific formula ((6)). These augmentation methods were

randomly applied to the data.

$$I' = aI^{\gamma} \tag{6}$$

The gamma correction technique processes the input image I to an output image I'. a is a fine-tuning constant, and γ indicates the contrast adjustment value, a random value between 0.4 and 2.5. The image values are initially normalized from 0 to 1, and the γ value is then adjusted. If $\gamma > 1$, the image becomes darker, whereas a value closer to 0 produces a brighter image. Some gamma configurations were used for data augmentation in the experiments.

C. Evaluation Metrics

Evaluation metrics are critical for accurately assessing segmentation models' performance. Precision measures the ratio of correctly identified voxels to the total number of identified voxels. Similarly, recall measures the percentage of correctly identified voxels to the total number of actual voxels ((7), 8). TP, FN, and FP indicate true positive, false negative, and false positive results, respectively. The Dice similarity coefficient (DSC, Dice score, or Dice) or F1-score is a widely used evaluation metric in medical image segmentation that combines both precision and recall ((9)). Radiomics pipelines usually use the Dice score as the primary evaluation metric.

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$Dice = \frac{2Precision \times Recall}{(Precision + Recall)} = \frac{2 \times TP}{(TP + FN) + (TP + FP)} \tag{9}$$

D. Experimental Results and Analysis

To optimize model performance, various windowing intervals and arrangements were evaluated for 2D and hybrid models. The model is intended to be usable as the primary model in both centralized and distributed architectures. Centralized and distributed training methods were compared regarding liver and tumor segmentation model accuracy.

1) Centralized Training Process: Implementation Details: One experiment aimed to determine the optimal windowing range for liver and tumor segmentation. Five windowing ranges from previous studies were tested: (−200, 250) [37], (−20, 220) [17], (50, 250) [38], (−200, 400) [39], and (−79, 304) [40]. The (−200, 250) range, slightly wider than the clinical windowing range, was chosen based on liver tumor segmentation studies to preserve more information. The (−20, 220) range was determined statistically using the standard deviation (SD) and mean of HU values in LiTS, assuming that values within three SDs of the mean are representative of some outliers. The (50, 250) range was selected as liver HU values mostly fall within this range. The last two settings were borrowed from kidney segmentation studies [39], [40] to explore the potential use of parameters from other organ segmentation models. The LiTS dataset was partitioned into a training set of 105 cases and a test set of 26 cases. Models with different windowing settings were assessed using K-Fold cross-validation on the training data. The K-Fold cross-validation method splits the data into K equal segments, with K being an adjustable parameter. For example, setting K to 10 implies that the training dataset is divided into ten equal parts. Consequently, the model is trained ten times, with each iteration involving nine of these ten segments as the training data, while the tenth segment serves as the validation set, not used in training. In the upcoming experiments, K is set to 5. The training process involved 500 epochs, each with 500 steps, and utilized the stochastic gradient descent (SGD) optimizer with a learning rate of 0.001. This was conducted on a machine equipped with an NVIDIA GeForce RTX 3090 GPU, boasting 24 GB of RAM. For liver tumor segmentation, the ResUNet architecture was employed as the model. The study concluded by assessing the influence of the windowing range on the identification of liver and tumors.

TABLE I
AVERAGE DICE SCORE FOR 5-FOLD CROSS-VALIDATION OF LIVER AND TUMOR SEGMENTATION WITH VARIOUS WINDOWING RANGES

Windowing range	Liver Dice score	Tumor Dice score
(-200, 250)	0.8984	0.6661
(-20, 220)	0.8956	0.6461
(50, 250)	0.8901	0.6172
(-200, 400)	0.9010	0.6654
(-79, 304)	0.8908	0.6424

Table I shows the average Dice scores of liver and tumor segmentation models achieved through 5-fold cross-validation using different window range settings. The liver segmentation model achieved the highest average Dice score with a windowing range of (−200, 400). The Dice scores for various liver windowing ranges indicate the possibility of transfer learning configurations (see Section III-C and the Discussion for more details). The tumor segmentation model performed best with a windowing range of (−200, 250), as indicated in Table I. Fig. 7 presents tumor segmentation results for different windowing ranges. It shows a minor false negative (FN) area in the prediction with the (−200, 250) range compared to the (50, 250) range. The Dice scores are for various tumor windowing ranges, with the model achieving the highest average Dice score using the (−200, 250) range. This range was selected for data preprocessing.

Several experiments were conducted on modified models: 2D-ResUNet, 2D-DenseUNet, and Hybrid-ResUNet. The 2D-ResUNet and 2D-DenseUNet models used 2D approaches for liver tumor segmentation, while the Hybrid-ResUNet model employed a hybrid approach. Initially, the liver was segmented using 2D models, and liver slices were generated to train a tumor detection model aimed at reducing error rates. The training and validation process involved 5-fold cross-validation. The Hybrid-ResUNet model achieved a higher average Dice score (0.9433) than the 2D-ResUNet model (0.6661) and the 2D-DenseUNet model (0.5736). This improvement was attributed to the hybrid

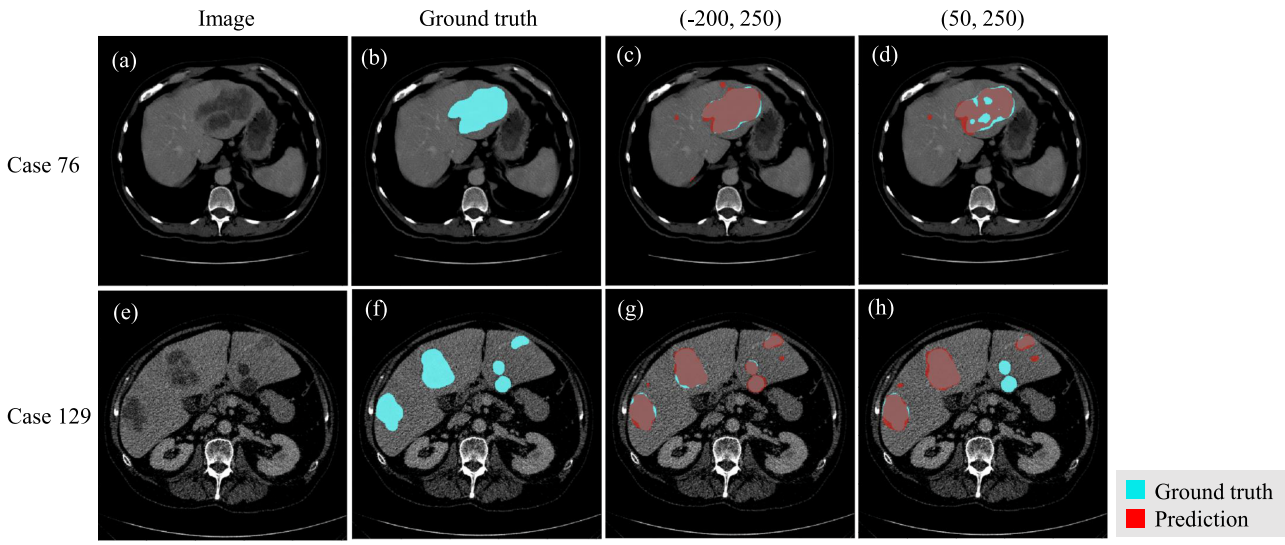

Fig. 7. Tumor segmentation predictions by the 2D-ResUNet model. (a) Original CT image of case 76 (b) overlaid with the ground truth, and overlaid with both the ground truth and the segmented result for the windowing ranges of (c) (−200, 250) and (d) (50, 250). (e) Original CT image of case 129, (f) overlaid with the ground truth, and overlaid with both the ground truth and the segmented result for the windowing ranges of (g) (−200, 250) and (h) (50, 250).

TABLE II
ABLATION STUDY OF MODEL SELECTION, LOSS FUNCTION, AND ARCHITECTURE EVALUATION

Model	Cross Entropy Loss	Dice Loss	F-score Loss	Combo Loss	Stage 1	Stage 2	Dice score	Recall	Precision
EfficientNet-B5	X	X	O	X	O	X	0.49	0.42	0.82
3D-UNet	X	X	O	X	O	X	0.54	0.48	0.85
2D-DenseUNet	X	X	O	X	O	X	0.5748	0.5609	0.709
ResNet-50	X	X	O	X	O	X	0.62	0.57	0.79
2D-DenseUNet	X	X	X	O	O	X	0.5444	0.5103	0.7119
2D-DenseUNet	O	X	X	X	O	X	0.5455	0.5082	0.7127
2D-DenseUNet	X	O	X	X	O	X	0.5736	0.5340	0.7445
2D-ResUNet	O	X	X	X	O	X	0.5866	0.8058	0.5071
2D-ResUNet	X	O	X	X	O	X	0.5919	0.5763	0.7564
EfficientNet-B0	X	X	X	**O**	O	X	0.62	0.65	0.65
ResNet-50	X	X	X	**O**	O	X	0.62	0.57	0.79
DenseNet-121	X	X	X	**O**	O	X	0.64	0.65	0.7
2D-ResUNet	X	X	O	X	O	X	0.6661	0.7058	0.7108
Hybrid-ResUNet	O	X	X	X	O	O	0.8942	0.9909	0.8166
Hybrid-ResUNet	X	X	O	X	O	O	0.9273	0.9795	0.8812
Hybrid-ResUNet	X	O	X	X	O	O	0.9378	0.9372	0.939
Hybrid-ResUNet	X	X	X	O	O	O	0.9433	0.9517	0.9354

model's retention of 3D image features and information. Details can be found in Table II.

Furthermore, we employed the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) to assess model performance in 26 testing cases. The ROC curve is constructed by plotting the true positive rate (TPR) against the false positive rate (FPR) ((10) and 11) at different threshold settings. The AUC is the area between the ROC curve and the x-axis. Fig. 8 presents the ROC curves and AUCs. The Hybrid-ResUNet model performs superior to the ResUNet model, but their ROC curves and AUCs are similar. Hybrid-ResUNet's AUC is 0.9965, and ResUNet's is 0.9944. Hence, the two models perform similarly. Notably, background pixels (non-liver regions) in images Fig. 9(a) comprise the vast majority of image regions. After removing these background pixels, as with the mask in Fig. 9(b), the model can better focus on the liver. This

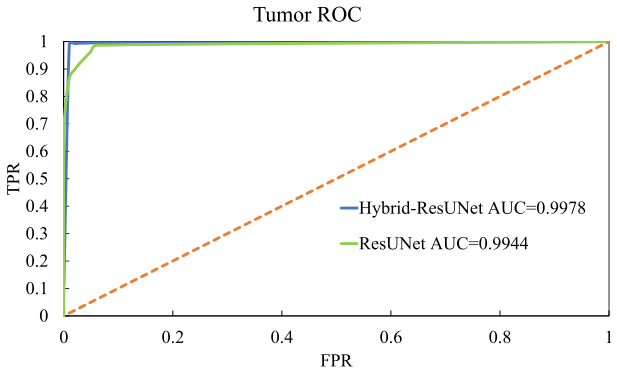

Fig. 8. 2D-ResUNet and Hybrid ResUNet ROC curves.

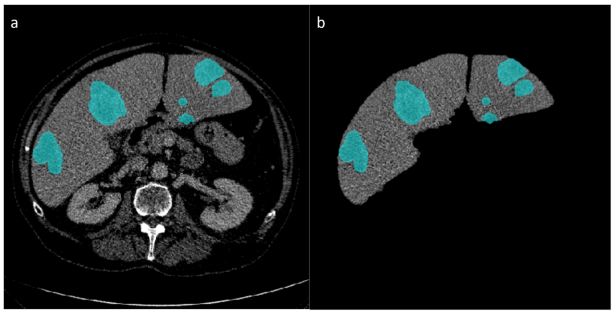

Fig. 9. (a) Abdomen image showing the liver, tumors, other organs, and the background. (b) Image only containing the liver and tumors.

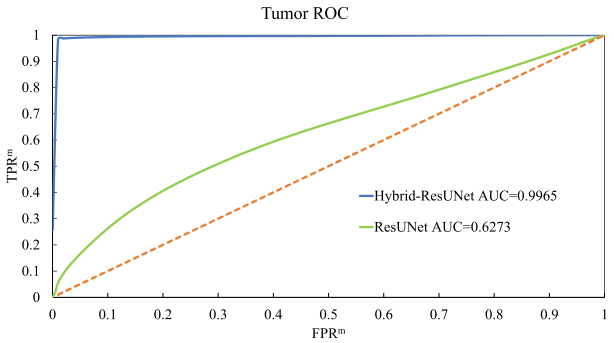

Fig. 10. ROC curves for 2D-ResUNet and Hybrid ResUNet with background removed.

mask is called a m. The (10) and 11 were slightly reformulated to use TPR^m ((12)) and FPR^m ((13)), respectively, to evaluate the model performance on images without backgrounds. Fig. 10 reveals that 2D-ResUNet's background-removed AUC drops sharply to 0.6273. By contrast, Hybrid-ResUNet's AUC is 0.996. Hybrid-ResUNet has the highest tumor segmentation performance.

$$TPR = \frac{TP}{TP + FN} \tag{10}$$

$$FPR = \frac{FP}{FP + TN} \tag{11}$$

$$TPR^m = \frac{TP^m}{TP^m + FN^m} \tag{12}$$

$$FPR^m = \frac{FP^m}{FP^m + TN^m} \tag{13}$$

The performance of the 2D-ResUNet, 2D-DenseUNet, and Hybrid-ResUNet methods with four different loss functions was compared to optimize their performance. Initially, a significantly greater foreground weight was assigned to prevent misclassification, but this weight decreased over time. Table II shows that the F-score loss function achieved the highest Dice scores and recall values for the 2D-ResUNet model. In contrast, the Hybrid-ResUNet model with the Combo loss function obtained the highest Dice scores. A high Dice score indicates impressive precision and recall values that are nearly equal.

Encoder Comparison: In this research, ResNet, DenseNet, and EfficientNet were employed as encoders for segmenting liver and tumor areas. The design of the models, encompassing the number of layers and the count of skip connections, was meticulously evaluated. The efficacy of each encoder was first examined for liver and tumor segmentation within the context of a 2D model framework. A selection of encoders, specifically EfficientNet-B0, EfficientNet-B5, ResNet-50, and DenseNet-121, underwent evaluation to determine the most effective 2D-UNet model, as detailed in Table III. These encoders produced comparable Dice scores, with DenseNet-121 leading with the highest score and demonstrating the least variance. Nonetheless, DenseNet-121 demanded the most significant computation time due to its layered structure. Conversely, the ResNet-50 model, though slightly lower in Dice score, required less computation time, positioning it as a viable alternative to DenseNet-121. The Hybrid-ResUNet model outperformed all 2D models Fig. 11, registering a Dice score of 0.9433 in HCC detection, as indicated in Table II. Further comparative analysis between the Hybrid-ResUNet and 3D-UNet models revealed that the Hybrid-ResUNet's 2D component captured more features and ensured stable prediction times without significantly increasing the computation time. In summary, the Hybrid-ResUNet model proved to be superior to the 3D-UNet model, as indicated in Table III.

2) Federated Learning Framework: Implementation Details: The study employed an FL framework to assess the model's effectiveness. It used 105 training cases divided among three clients and 26 for testing. The training data sets were distributed among clients in balanced (35 cases per client) and imbalanced (53, 31, and 21 cases) manners. The Hybrid-ResUNet model performed better in previous experiments and was applied with the Dice loss function. The models underwent five-fold cross-validation using the LiTS Challenge data set. The training involved 100 epochs with 100 steps per epoch and an SGD optimizer with a learning rate of 0.001. Computation was performed on a machine with NVIDIA GeForce RTX 3090 GPU, AMD Ryzen 5 5600X 6-Core processor, and 32 GB of RAM. Fig.12 depicts the FL framework employed in the study.

FL with Balanced Data Set Distribution: Table IV summarizes the results of local model training, model testing, FL, and global verification for the balanced data set scenario. The

TABLE III
PARAMETERS AND RESULTS OF 2D, 3D, AND HYBRID MODELS FOR LIVER TUMOR SEGMENTATION

Encoder	Layers	Parameters	Average computation time (sec)	Dice score	Recall	Precision
EfficientNet-B0	237	4M	10.3	0.62	0.65	0.65
EfficientNet-B5	576	28.3M	12.5	0.49	0.42	0.82
ResNet-50	202	36M	4.6	0.62	0.57	0.79
DenseNet-121	344	19.8M	11.6	0.64	0.65	0.70
3D-UNet	67	24M	18.1	0.54	0.48	0.85
Hybrid-ResUNet	325	24M	20.6	0.9433	0.9517	0.9354

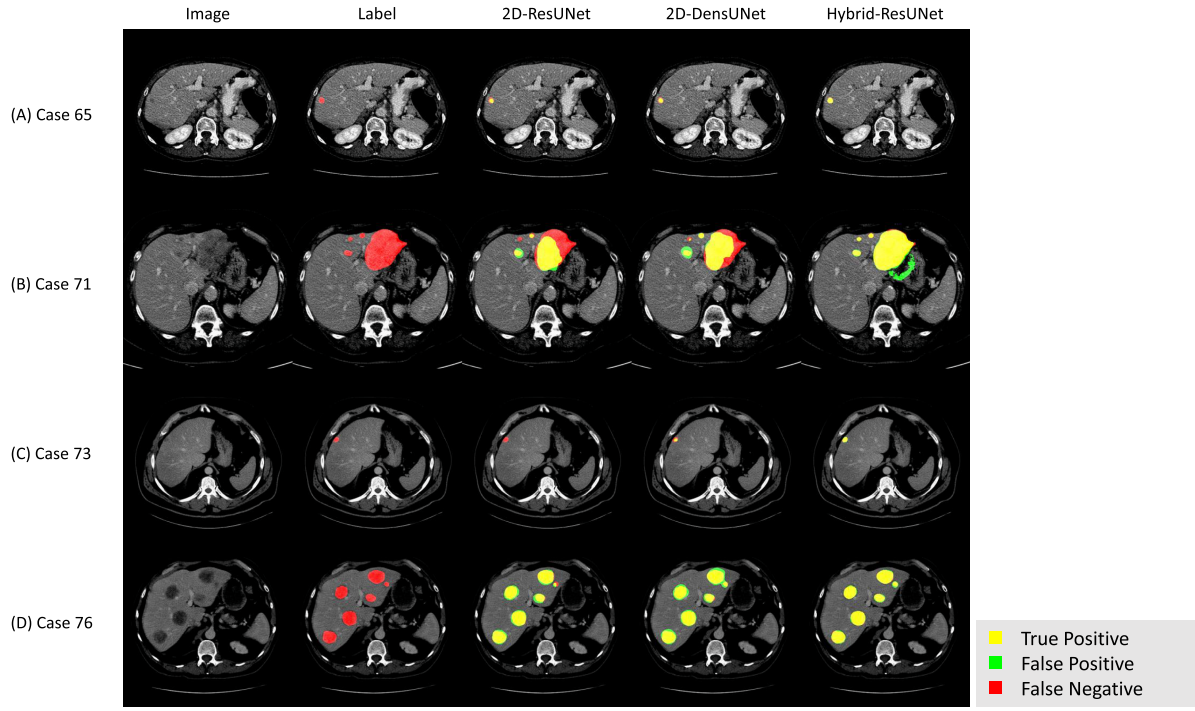

Fig. 11. Results for three liver tumor segmentation models for cases 65, 71, 73, and 76. First column, original images; second column, annotated photographs; the third, fourth, and fifth columns are the segmentation results for the 2D-ResUNet, 2D-DenseUNet, and Hybrid-ResUNet models, respectively. The Hybrid-ResUNet model outperforms the other models in all cases.

TABLE IV
FL WITH A BALANCED DATA SET DISTRIBUTION

	Dataset 1	Dataset 2	Dataset 3	Performance	Global Verification
Client 1 **(35)**	0.8488	0.6324	0.6642	0.7151	0.6256
Client 2 **(35)**	0.4766	0.7861	0.4578	0.5735	0.4942
Client 3 **(35)**	0.5692	0.6624	0.8780	0.7032	0.5449
FedAvg	0.7071	0.7233	0.7119	0.7132	0.7418

first three rows present the performance of each client's model on each local data set. The last row presents the results of the global model iteratively updated by the FedAvg algorithm and tested on all three clients' data sets. The global model has a mean Dice score of 0.7418. Each client performs well on its data set, achieving a Dice score of approximately 0.78–0.87, but performs poorly on other data sets. All clients had improved performance after FL, which achieved an average Dice score of 0.7132. The FedAvg global model has the highest Dice score of 0.7418 in the global testing set. Federated learning enables cross-hospital institutions to improve liver tumor segmentation models by increasing data volume without exchanging data.

Federated Learning with Imbalanced Data Set Distribution: The proposed framework was investigated for resilience by using experiments involving FL with an imbalanced data distribution among local data sets. The results are presented in Table V. Each client performs favorably on its local data set (DSCs of 0.84–0.92). However, they perform poorly on other data sets because

TABLE V
FL WITH AN IMBALANCED DATA SET DISTRIBUTION

	Dataset 1	Dataset 2	Dataset 3	Performance	Global Verification
Client 1 **(53)**	0.8430	0.7154	0.6552	0.7727	0.7511
Client 2 **(31)**	0.5398	0.8916	0.4503	0.6382	0.3752
Client 3 **(21)**	0.5394	0.4851	0.9245	0.6048	0.3500
FedAvg	0.7703	0.7678	0.8201	0.7788	0.7155

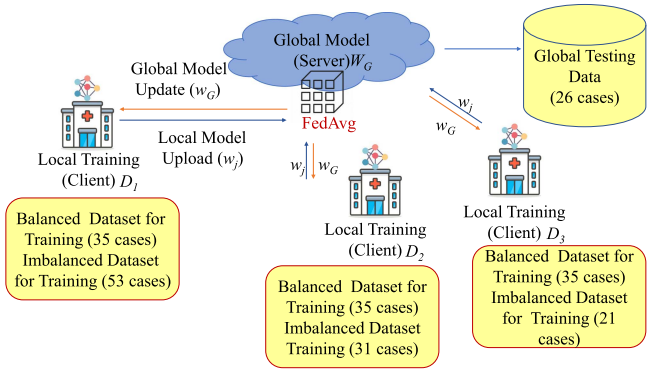

Fig. 12. Proposed FL framework including the workflow, relevant parameters, and balanced and imbalanced data distributions.

of imbalanced data distribution. The global model, which was iteratively updated by the FedAvg algorithm, had a Dice score of 0.7155 for the local data sets. FL enhanced this performance, achieving a Dice score of 0.7788 for the local data sets. It also achieved the highest Dice score of 0.7155 on the global testing data set. In summary, this federated learning method is effective even if the data distribution among local data sets is imbalanced.

Comparison with Centralized and Distributed Training Results: Based on the designed Hybrid-ResUNet model, centralized training outperformed distributed training. Hospitals typically follow strict data exchange and sharing regulations to protect patient privacy. Hence, centralized training cannot be used for medical data. However, the FL framework is suitable for distributed training for clients with small data sets. Training quality also plays a crucial role in improving performance. The results indicate that the proposed FL approach can be applied to real-world medical scenarios, even to unbalanced data. The comparable performance of the proposed approach on both balanced and imbalanced data sets also demonstrates its robustness.

V. DISCUSSION

In this study, an ablation analysis was conducted to evaluate the effectiveness of windowing ranges, loss functions, and training methods across various deep learning models. The aim was to identify the parameters that yielded the highest performance. Transfer learning was implemented by applying pre-trained weights from the KiTS dataset, which facilitated model convergence and enhanced accuracy, as noted in [41]. Windowing techniques were used specifically to improve liver and tumor segmentation methods. Additionally, the study investigated the impact of different loss functions and encoders on the overall performance of the model. The findings indicated that both preprocessing and adjustments to the model architecture had a significant effect on the accuracy of liver and tumor segmentation. In this context, the Hybrid-ResUNet model with a combo loss function employed in this study notably increased segmentation accuracy by initially isolating the liver region, thereby minimizing interference from surrounding organs. The ultimate objective was to develop a high-performing Hybrid-ResUNet model suitable for broad hospital use, aiming to ensure a high-quality and standardized approach to segmentation.

Strict regulations in hospitals often hinder data exchange and sharing, posing challenges in developing consistent and accurate medical models. Federated learning presents a potential solution to this dilemma. However, the varying quantity and quality of data among different clients can affect the accuracy of the global model in FL compared to centralized models. Therefore, the choice of an effective and high-performing FL architectural model is essential. The findings of this study demonstrate the viability of implementing an FL architecture across multiple medical institutions, leveraging cross-silo computation within the secure and stable network environments of hospitals. The Hybrid-ResUNet model, employed in this research, shows high accuracy in tumor segmentation and the assessment of liver and tumor volumes. To our knowledge, the integration of a liver tumor semantic segmentation model with a federated learning framework represents a novel application. This is specifically applied to liver cancer detection scenarios where a systematic evaluation of various parameters, such as encoder and decoder selection, transfer learning, loss function evaluation, and adaptation of federated learning algorithms, is conducted. The innovative combination of these methods and steps highlights the novelty and progressiveness of this paper, where, within the federated learning environment, each client can independently compute using their dataset while collaboratively optimizing through the Federator. Furthermore, utilizing this model, doctors can precisely determine tumor size and location by calculating liver volume and overlaying liver tumor slices on a 3D image, as well as identifying affected tissues. This accurate information is crucial for developing optimal treatment plans, monitoring treatment progression, and predicting the growth rate and prognosis of liver tumors. Such accurate predictions enable healthcare professionals to provide improved patient care and more effective treatment planning.

The study has limitations, including the number of clients in the federated learning framework. The FedAvg algorithm has limitations, such as excessive model parameter increases

when many clients participate in training, potentially leading to communication bandwidth constraints. Slow model convergence requires increased communication frequency, resulting in higher computational costs. Additionally, imbalanced data distribution in distributed learning across multiple clients can affect system performance. Further research is necessary to develop mechanisms or algorithms for balancing user data distribution across medical institutions.

VI. Conclusion

This study utilized deep learning and federated learning techniques for liver cancer detection, specifically hepatocellular carcinoma. The proposed Hybrid-ResUNet model combines 2D and 3D methodologies to achieve precise detection and analysis of liver tumors, reducing doctors' workload. Transfer learning and hybrid models were employed to minimize training time and computational resources and enhance system performance. Optimization of the models included evaluating various data preprocessing strategies, windowing ranges, encoders, and loss functions. This resulted in improved performance for liver tumor detection (Dice score: 0.9433, AUC: 0.9965). Additionally, the model enables liver and tumor volume calculation by overlaying segmented voxels on image slices. This valuable information helps doctors formulate appropriate treatment plans, including surgery, radiation therapy, or chemotherapy. Moreover, it facilitates treatment progress monitoring, effectiveness assessment, and treatment plan adjustment.

Furthermore, the study implemented the FedAvg algorithm for federated learning. This algorithm allows collaborative model optimization using distributed data sets from multiple hospitals, making it highly suitable for cross-hospital learning applications. Integrating the Hybrid-ResUNet model into a practical FL framework enables clinical assessment while preserving data privacy and facilitating the collection of large training data sets. The proposed model and FL framework offer potential improvements in tumor detection accuracy, making them well-suited for implementation in medical clinics and hospitals.

This paper presents a detailed platform or workflow designed to deeply involve participants in problem-solving, data collection, model development, and refinement activities. Future research will aim to enhance active participation in these areas. The emphasis will be placed on improving federated learning algorithms, especially their incorporation into the continuous implementation and deployment (CI/CD) processes within AI system operations, while ensuring the dynamic involvement of clients. The goal is to create strategies that provide incentives, ensure fairness, and facilitate ongoing optimization across a diverse array of applications. Such developments are expected to foster equitable and efficient data collaboration across various sectors in practical, real-world environments. This addition serves as an essential guide for future research initiatives, particularly focusing on the advancement of the medical AI field.

References

[1] J. M. Llovet et al., "Hepatocellular carcinoma," *Nature Rev. Dis. Primers*, vol. 7, no. 1, pp. 6–34, 2021.

[2] J. K. Heimbach et al., "AASLD guidelines for the treatment of hepatocellular carcinoma," *Hepatology*, vol. 67, no. 1, pp. 358–380, 2018. [Online]. Available: https://aasldpubs.onlinelibrary.wiley.com/doi/abs/10.1002/hep.29086

[3] Y.-Y. Shao et al., "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. Formosan Med. Assoc.*, vol. 120, no. 4, pp. 1051–1060, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0929664620305313

[4] H. Chen, C. Gomez, C.-M. Huang, and M. Unberath, "Explainable medical imaging AI needs human-centered design: Guidelines and evidence from a systematic review," *npj Digit. Med.*, vol. 5, no. 1, pp. 156–171, 2022.

[5] D. Song et al., "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.*, vol. 147, no. 12, pp. 3757–3767, 2021.

[6] L. Song, K. Geoffrey, and H. Kaijian, "Bottleneck feature supervised U-net for pixel-wise liver and tumor segmentation," *Expert Syst. Appl.*, vol. 145, 2020, Art. no. 113131.

[7] J. Mejía, A. Ochoa, and B. Mederos, "Improving segmentation of liver tumors using deep learning," in *Intuitionistic and Type-2 Fuzzy Logic Enhancements in Neural and Optimization Algorithms: Theory and Applications*. Berlin, Germany: Springer, 2020, pp. 771–780.

[8] A. Mojtahed et al., "Repeatability and reproducibility of deep-learning-based liver volume and couinaud segment volume measurement tool," *Abdominal Radiol.*, vol. 47, no. 1, pp. 143–151, 2022.

[9] P. Malhotra, S. Gupta, D. Koundal, A. Zaguia, and W. Enbeyle, "Deep neural networks for medical image segmentation," *J. Healthcare Eng.*, vol. 2022, 2022, Art. no. 9580991.

[10] H. Seo, C. Huang, M. Bassenne, R. Xiao, and L. Xing, "Modified U-net (mu-net) with incorporation of object-dependent high level features for improved liver and liver-tumor segmentation in CT images," *IEEE Trans. Med. Imag.*, vol. 39, no. 5, pp. 1316–1325, May 2020.

[11] L. Han, Y. Chen, J. Li, B. Zhong, Y. Lei, and M. Sun, "Liver segmentation with 2.5D perpendicular UNets," *Comput. Elect. Eng.*, vol. 91, 2021, Art. no. 107118. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0045790621001221

[12] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++ : Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.

[13] W. Qin et al., "Superpixel-based and boundary-sensitive convolutional neural network for automated liver segmentation," *Phys. Med. Biol.*, vol. 63, no. 9, May 2018, Art. no. 095017, doi: 10.1088/1361-6560/aabd19.

[14] C. Zhang, Q. Hua, Y. Chu, and P. Wang, "Liver tumor segmentation using 2.5D UV-net with multi-scale convolution," *Comput. Biol. Med.*, vol. 133, 2021, Art. no. 104424.

[15] N. Alalwan, A. Abozeid, A. A. ElHabshy, and A. Alzahrani, "Efficient 3D deep learning model for medical image semantic segmentation," *Alexandria Eng. J.*, vol. 60, no. 1, pp. 1231–1239, 2021.

[16] W. Zhou et al., "Prediction of microvascular invasion of hepatocellular carcinoma based on contrast-enhanced MR and 3D convolutional neural networks," *Front. Oncol.*, vol. 11, 2021, Art. no. 588010. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fonc.2021.588010

[17] C.-H. Hsiao et al., "A deep learning-based precision volume calculation approach for kidney and tumor segmentation on computed tomography images," *Comput. Methods Programs Biomed.*, vol. 221, 2022, Art. no. 106861.

[18] E. Tacconelli et al., "Challenges of data sharing in European Covid-19 projects: A learning opportunity for advancing pandemic preparedness and response," *Lancet Regional Health - Europe*, vol. 21, 2022, Art. no. 100467. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666776222001636

[19] Z. A. E. Houda, A. S. Hafid, L. Khoukhi, and B. Brik, "When collaborative federated learning meets blockchain to preserve privacy in healthcare," *IEEE Trans. Netw. Sci. Eng.*, vol. 10, no. 5, pp. 2455–2465, Sep./Oct. 2023.

[20] N. Rieke et al., "The future of digital health with federated learning," *npj Digit. Med.*, vol. 3, no. 1, pp. 119–126, 2020.

[21] W. Gan et al., "Automatic segmentation of lung tumors on CT images based on a 2D & 3D hybrid convolutional neural network," *Brit. J. Radiol.*, vol. 94, 2021, Art. no. 20210038.

[22] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-denseunet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018.

[23] C. Zhang, D. Ai, C. Feng, J. Fan, H. Song, and J. Yang, "Dial/hybrid cascade 3DResUNet for liver and tumor segmentation," in *Proc. 4th Int. Conf. Digit. Signal Process.*, 2020, pp. 92–96.
[24] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao, "A survey on federated learning," *Knowl.-Based Syst.*, vol. 216, 2021, Art. no. 106775.
[25] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2017, pp. 1273–1282.
[26] X. Jiang, J. Zhang, and L. Zhang, "Fedradar: Federated multi-task transfer learning for radar-based internet of medical things," *IEEE Trans. Netw. Serv. Manage.*, vol. 20, no. 2, pp. 1459–1469, 2023.
[27] W. Zhang et al., "Dynamic-fusion-based federated learning for COVID-19 detection," *IEEE Internet Things J.*, vol. 8, no. 21, pp. 15884–15891, Nov. 2021.
[28] Q. Li et al., "A survey on federated learning systems: Vision, hype and reality for data privacy and protection," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 4, pp. 3347–3366, Apr. 2023.
[29] Liver Tumor Segmentation Challenge, 2017, Accessed: Mar. 6, 2023. [Online]. Available: https://competitions.codalab.org/competitions/17094
[30] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access*, vol. 9, pp. 82031–82057, 2021.
[31] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.
[32] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[34] KiTS19 Challenge Homepage, 2019, Accessed: Mar. 22, 2022. [Online]. Available: https://kits19.grand-challenge.org/
[35] J. Ma et al., "Loss odyssey in medical image segmentation," *Med. Image Anal.*, vol. 71, 2021, Art. no. 102035.
[36] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Statist.*, 2017, pp. 1273–1282.
[37] Y. Tian et al., "ARR-GCN: Anatomy-relation reasoning graph convolutional network for automatic fine-grained segmentation of organ's surgical anatomy," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 7, pp. 3258–3269, 2023.
[38] R. S. Alomari, S. Kompalli, and V. Chaudhary, "Segmentation of the liver from abdominal CT using Markov random field model and GVF snakes," in *Proc. Int. Conf. Complex Intell. Softw. Intensive Syst.*, 2008, pp. 293–298.
[39] C.-H. Hsiao et al., "Automatic kidney volume estimation system using transfer learning techniques," in *Proc. Int. Conf. Adv. Inf. Netw. Appl.*, 2021, pp. 370–381.
[40] Q. Li et al., "Densely connected u-net with criss-cross attention for automatic liver tumor segmentation in ct images," *IEEE/Assoc. Comput. Machinery Trans. Comput. Biol. Bioinf.*, vol. 20, no. 6, pp. 3399–3410, 2023.
[41] C.-H. Hsiaoet al., "Automatic Kidney Volume Estimation System Using Transfer Learning Techniques," in *Advanced InformationNetworking and Applications*, L. Barolli, I. Woungang, and T. Enokido Eds. Cham, Switzerland: Springer, 2021, pp. 370–381.