FL-PancreasNet: Federated Transformer Approach for Secure Multi-Center Pancreatic Lesion Analysis in CT Imaging

Dr. D. Narmadha [1,2], Prof. Shashi Kant Gupta[3],
[1] Lincoln University College, Malaysia;
[2] Division of AIML, Karunya Institute of Technology and Sciences, Coimbatore, India;
[3] Lincoln University College, Malaysia

pdf.narmadha@lincoln.edu.my
narmadha@karunya.edu
shashigupta@lincoln.edu.my

Abstract

Computed Tomography (CT) is a mainstay for assessing pancreatic lesions. Clear delineation of the pancreas and tumour is essential for diagnosis and treatment planning. Training on data from multiple centres can improve robustness. In practice, however, clinical scans are usually kept within individual hospitals. Privacy rules and institutional policies often prevent central pooling. Differences in data distributions across sites can also undermine model stability. Most existing work relies on centralised three-dimensional U-Net variants or Transformer-based segmentation networks trained on pooled datasets. Such setups are difficult to realise in routine deployments. Federated Learning (FL) offers a practical alternative by enabling collaborative training without sharing raw scans. Yet, with heterogeneous client data, standard aggregation can become unreliable. To address this issue, a Federated Transformer-UNet hybrid, FL-PancreasNet, is introduced to improve cross-site generalisation. Secure training is supported through privacy-aware aggregation and robust handling of client updates. Experiments use the public Medical Segmentation Decathlon (MSD) Task07 Pancreas dataset. It contains 420 portal venous phase CT volumes with pancreas and tumour labels. A total of 282 cases are used for training and 139 for testing. On the MSD test set, FL-PancreasNet achieves a pancreas Dice of 0.892 and a tumour Dice of 0.748. The tumour Hausdorff Distance at the 95th percentile (HD95) is reduced to 15.6 mm. This exceeds the performance of Federated Averaging (FedAvg), which reports a tumour Dice of 0.721 and HD95 of 17.4 mm. With secure aggregation, performance is preserved at a tumour Dice of 0.747 with modest overhead. These results support privacy-preserving collaboration while maintaining strong segmentation accuracy in multi-centre settings.

Keywords: Federated Learning, Pancreatic Tumour Segmentation, Computed Tomography Imaging, Transformer-UNet Architecture, Multi-Centre Medical Imaging, Privacy-Preserving Machine Learning, Medical Segmentation Decathlon

1 INTRODUCTION

Pancreatic cancer is one of the most deadly cancers in clinical oncology, characterised by late stage presentation and a 5-year survival rate that has shown only marginal improvement in spite of advances in diagnostic technologies [1]. The irregular morphology of the pancreas and its proximity to the surrounding vascular structures make it difficult to accurately segment pancreatic parenchyma and tumour regions a particularly challenging computation task, with direct clinical implications for early detection, surgical planning and target delineation for radiotherapy.

Pancreatic tumour segmentation is complicated by severe class imbalance, highly variable tumour morphology, and ambiguous tissue boundaries which together problematize learned representations [2]. Beyond these algorithmic difficulties, assembling sufficiently large annotated training datasets across institutions is infeasible due to strict regulatory constraints relating to patient data sharing, meaning individual centres rarely have sufficient data in isolation to train reliably generalising models.

Recent deep learning approaches, from U-Net to Attention-U-Net to transformer-based approaches like TransUNet and SwinUNETR have provided incremental improvements in segmentation accuracy [3]. However, all of these approaches assume centralised access to training data at optimisation stage, a premise that is fundamentally opposed to the privacy preserving requirements of real world, multi-institutional clinical research.

Federated learning provides a principled way of coordinating model training across distributed institutions using parameter exchange, without requiring raw data transfer [4]. Although algorithms like FedAvg, FedProx, SCAFFOLD and FedBN have tackled the issue of statistical heterogeneity and client drift in general settings, their deployment to pancreatic tumour segmentation is not well studied. This work proposes FL-PancreasNet, a secure federated framework based on a Transformer-UNet hybrid backbone with transformer encoder blocks, cross attention multi-scale fusion, lesion aware loss weighting and secure aggregation that is evaluated on MSD Task07 Pancreas dataset distributed across 5 simulated federated centres.

The following are the key contributions:

- A federated multi-centre training framework for pancreatic organ and tumour segmentation in computed tomography that eliminates the need for centralised data aggregation and enables privacy-preserving collaboration among participating institutions.
- A hybrid Transformer-UNet segmentation backbone specifically designed in the context of a federated learning application, which utilizes global context modelling and multi-scale feature decoding to improve the accuracy of lesion demarcation.
- A structured evaluation protocol to analyse segmentation fidelity, boundary precision, client-specific generalisation and convergence dynamics, as well as privacy-related overhead under realistic non-identically distributed client scenarios.
- An ablation study-driven approach that measures the impact of architectural components, loss weighting schemes, and privacy controls, and therefore the contribution of each individual component to overall performance.

The rest of this paper is organized as follows. Section 2 provides a review of the related literature. Section 3 outlines the dataset, experimental setup and proposed framework. Quantitative results and ablation studies are presented in Section 4. Section 5 presents the results and Section 6 concludes with a discussion of the limitations and future research directions.

2 Related Work

Multi-center studies of the pancreas have expanded dramatically as investigators attempt to find cohorts that are large enough to provide generalizable results while working around institutional privacy barriers. Kobayashi et al. [5] registered a phase-2 randomized protocol of robotic-assisted versus open pancreaticoduodenectomy on complication endpoints, with outcome data still pending. Hendrikx et al. [6] reported feasibility and acceptable safety for irreversible electroporation in pancreatic cancer in several centers with the usual caveats of bias in patient selection. Sosnowska-Sienkiewicz et al. [7] pooled retrospective indocyanine green use across the entire country in pediatric surgery, yielding practice-level information limited by lack of consistent institutional protocols. Londono Barrientos et al. [8] synthesized the evidence on pediatric trauma of the pancreas and found no difference in mortality between conservative and operative management, with the risk of developing pseudocysts and fistulas depending on the approach used in heterogeneously reported observational studies.

Clinical pharmacology and procedural trials provide additional context. Gao et al. [9] reported real-life surufatinib activity across Chinese centers for neuroendocrine tumors with the retrospective design leaving the issue of confounding from patient selection unresolved. Park et al. [10] conducted a phase II randomized double blind trial of CKD-495 in gastritis that resulted in clean short-term efficacy data that needs confirmatory phase III replication. Wang et al. [11] showed significant triglyceride reduction with highly purified ethyl icosapentate for twelve weeks in a phase III placebo-controlled trial but the reduction in risk of long-term pancreatitis has not been investigated. Wang et al. [12] published a protocol for a randomized trial of early versus late pancreatic stent placement for the prevention of post-ERCP pancreatitis, but no estimates of outcomes are yet available. Hays et al. [13] used a multi-center American registry to link robotic pancreaticoduodenectomy to reduced major morbidity in octogenarians, while patient selection for minimally invasive surgery makes it hard to interpret causally.

Methodological reviews report the ongoing evolution from radiomics to attention-based deep learning, consistently identifying the lack of external validation as the central unresolved weakness in the field. Almufareh et al. [14] traced this architectural progression in pancreatic cancer detection and prognosis, arguing that reproducible multi-center pipelines are a prerequisite for meaningful clinical translation. Bi et al. [15] performed a systematic review and meta-analysis of deep learning methods for pancreatic cancer diagnosis, and found that dataset heterogeneity and publication bias are likely to overestimate performance across modalities. Gupta et al. [16] reviewed the applications of artificial intelligence in abdominal imaging tasks and enumerated the common barriers of demographic dataset bias, site-specific imaging variability and workflow integration constraints. Xu et al. [17] constructed two-center PET/CT radiomics models for histologic grading of pancreatic ductal adenocarcinoma and showed that combined modality features performed better than single-modality variants, but that external validation has not yet been conducted. Wang et al. [18] assessed camrelizumab with apatinib and SOX in a multi-center single-arm phase II study of alpha-fetoprotein producing gastric adenocarcinoma, reporting high response rates that need to be confirmed in a randomized study to disentangle treatment effect from enrollment selection.

3 SYSTEM METHODOLOGY

This section describes the end-to-end methodology of FL-PancreasNet, covering data preparation, the Transformer-UNet hybrid segmentation backbone, the learning objective, and the privacy-preserving federated optimization protocol.

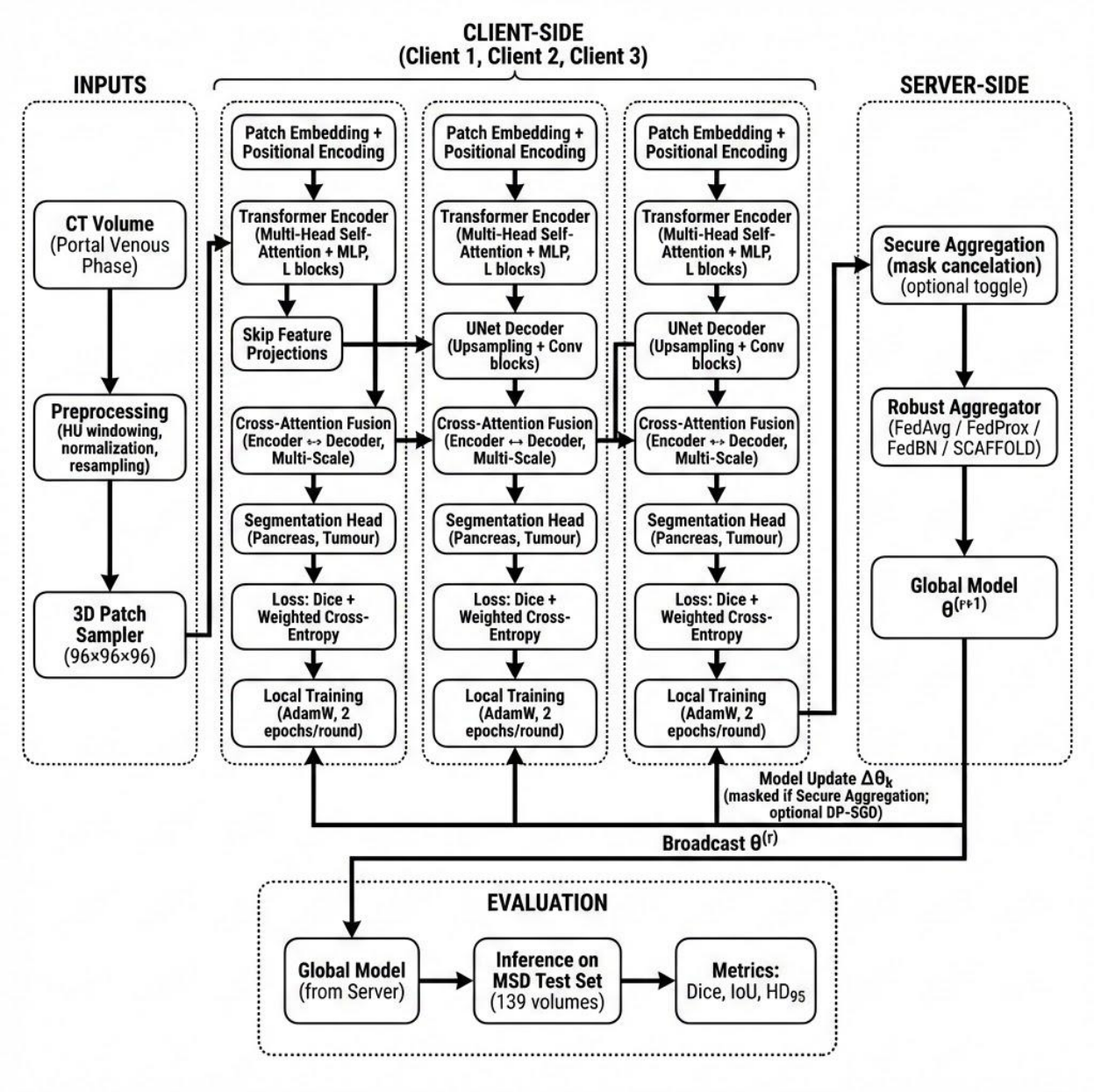

Figure 1: FL-PancreasNet architecture and federated training workflow for pancreas and tumour CT segmentation.

3.1 Overview of the Federated Segmentation Pipeline

Figure 1 shows the general client–server process and the key components of the model that are contained in FL-PancreasNet. The system is targeted at pancreas and tumour segmentation from portal venous phase Computed Tomography (CT) volumes in a multi-center setting where each center has its local scans.

Training is carried out in rounds of communication: each client does a local optimization on its own labelled volumes and returns model updates to a server, which aggregates model updates to create a new global model. The process repeats itself until convergence, which yields a single global model ready for deployment across heterogeneous centers.

3.2 Preprocessing and 3D Patch Sampling

Each CT volume is converted into a normalized intensity representation and paired with two voxel-wise ground truth masks, one for the pancreas and one for the tumour. To maintain memory stability during 3D training, the volume is processed using fixed-size patches. Let $X \in \mathbb{R}^{H\times W\times D}$ denote a preprocessed CT volume and let $P(X) = \{x_i\}_{i=1}^{Np}$ be the set of extracted 3D patches, where $x_i \in \mathbb{R}^{h\times w\times d}$. Each patch is flattened and projected into an embedding space to interface with the Transformer encoder. The embedding operation is defined in Eq. (1), where W_e and b_e are learnable parameters and $L = hwd$.

$$e_i = vec(x_i)W_e + b_e, W_e \in \mathbb{R}^{L\times d_e} \quad (1)$$

3.3 Transformer Encoder with Positional Information

The Transformer encoder models long-range dependencies that are important for pancreas context and small tumour appearance. Patch embeddings are augmented with positional encodings to preserve spatial order. The input token sequence is formed as in Eq. (2), where E_{pos} denotes a learnable positional embedding and z_0 is the encoder input.

$$z_0 = [e_1, e_2, \dots, e_{Np}] + E_{pos} \quad (2)$$

Self-attention is used to compute contextualized representations. The attention operator is given in Eq. (3), where Q, K, and V are linear projections of the same token set, and d_k is the key dimension.

$$Attn(Q,K,V) = softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \quad (3)$$

Each Transformer block applies attention with residual connections and a feed-forward sublayer. The block update is summarized in Eq. (4), where $LN(\cdot)$ is layer normalization and $MLP(\cdot)$ is the position-wise feed-forward network.

$$z_\ell = MLP\big(LN(z'_\ell)\big) + z'_\ell, \quad z'_\ell = Attn(LN(z_{\ell-1})) + z_{\ell-1} \quad (4)$$

3.4 UNet-Style Decoder and Cross-Attention Fusion

To recover precise boundaries, a UNet-style decoder upsamples features and integrates multi-scale information. Encoder outputs provide coarse semantic context, while decoder stages refine edges and small structures. Cross-attention fusion is used to merge decoder queries with encoder keys and values, encouraging the decoder to selectively retrieve context relevant to pancreas and tumour boundaries. The fusion operator is defined in Eq. (5), where g is a decoder feature map (query source) and f is an encoder feature map (key-value source), with learnable projections W_q, W_k, and W_v.

$$CA(g,f) = softmax\left(\frac{(gW_q)(fW_k)^{\mathrm{T}}}{\sqrt{d_k}}\right)(fW_v) \quad (5)$$

3.5 Segmentation Objective with Lesion Awareness

Training optimizes a composite loss that balances overlap quality and voxel-wise calibration, while addressing tumour class imbalance. Dice loss for a target class is defined in Eq. (6), where p denotes the predicted soft mask, y denotes the binary ground truth mask, and ε is a small constant for numerical stability.

$$L_{dice} = 1 - \frac{2\sum_v p_v\, yv + \varepsilon}{\sum v\, p_v + \sum_v y_v + \varepsilon} \quad (6)$$

A weighted cross-entropy term emphasizes lesion voxels and reduces bias toward background. The weighted cross-entropy is given in Eq. (7), where c indexes classes, w_c are class weights, and $p_{\{v,c\}}$ is the predicted probability for voxel v and class c.

$$L_{wce} = -\sum_v \sum_c w_c\, y_{\{v,c\}}\ log(p_{\{v,c\}}) \quad (7)$$

The final training objective combines the two terms as shown in Eq. (8), where λ_d and λ_c control the contribution of overlap and voxel-wise learning signals.

$$L\ =\ \lambda_d\, L_{dice} + \lambda_c\, L_{wce} \quad (8)$$

3.6 Federated Optimization and Aggregation

Federated training proceeds over rounds $r = 0, 1, \ldots, R-1$. At round r, the server broadcasts parameters θ^r to a subset of clients. Client k performs local optimization using its dataset D_k by applying stochastic gradient updates, summarized in Eq. (9), where η is the learning rate and $\nabla L_k(\cdot)$ is the gradient of the local objective induced by Eq. (8).

$$\theta^r_{\{k,t+1\}} = \theta^r_{\{k,t\}} - \eta\, \nabla L_k\big(\theta^r_{\{k,t\}}\big),\ \ t = 0, \ldots, T_k - 1 \quad (9)$$

After local training, the client sends an update to the server. The server aggregates client models with a data-size weighted rule. The standard aggregation step is given in Eq. (10), where n_k is the number of training samples at client k and $N = \sum_k n_k$.

$$\theta^{r+1} = \sum_{\{k=1\}}^{K} \left(\frac{n_k}{N}\right) \theta^r_{\{k,T_k\}} \quad (10)$$

This aggregation can be paired with robustness measures such as clipping, outlier filtering, or coordinate-wise robust statistics to reduce the influence of anomalous updates when heterogeneity across clients is severe.

3.7 Privacy-Preserving Training via Secure Aggregation and Differential Privacy

To reduce leakage risk from raw updates, secure aggregation is used to ensure the server only observes the sum of updates rather than each individual update. Let $\Delta^r_k = \theta^r_{\{k,T_k\}} - \theta^r$ denote the model update at client k. A masked update can be formed using pairwise masks such that masks cancel when aggregated. The masking construction is expressed in Eq. (11), where $m^r_{\{k,j\}}$ is a random mask shared between clients k and j with opposite signs.

$$\tilde{\Delta}^{(r)}_k = \Delta^{(r)}_k + \sum_{\{j \neq k\}} m^{(r)}_{\{k,j\}} - \sum_{\{j \neq k\}} m^{(r)}_{\{j,k\}},\ \ \sum_k \tilde{\Delta}^{(r)}_k = \sum_k \Delta^{(r)}_k \quad (11)$$

The implementation of Differential Privacy Stochastic Gradient Descent (DP-SGD) requires clients to perform two operations which include clipping their individual gradient data and adding Gaussian noise. The equation used to define the noisy gradient for parameter updates appears in Eq. (12), which uses the function clip $(\cdot, C)$ to limit ℓ_2 norm values to the maximum threshold of C while B represents the minibatch size and σ determines the magnitude of the introduced noise.

$$\bar{g} = \left(\frac{1}{B}\right) \sum_{\{i=1\}}^{B} clip(g_i, C) + N(0, \sigma^2 C^2 I), \;\; \theta \leftarrow \theta - \eta\, \bar{g} \quad (12)$$

Eq. (11) and Eq. (12) represent two different privacy protection methods which work together as complementary systems because secure aggregation prevents servers from seeing individual updates while DP-SGD restricts information disclosure through its aggregated updates but results in an accuracy decrease that depends on the chosen privacy budget.

4 RESULTS

This section shows the experimental evaluation of FL-PancreasNet on MSD Task07 Pancreas benchmark in a multi-center federated learning simulation. It presents the dataset characteristics, training setup, comparative segmentation performance, client-wise generalization, ablation study, privacy overhead and convergence behaviour.

4.1 Dataset Used (MSD Task07 Pancreas)

The experiments performed throughout this study were performed using the Medical Segmentation Decathlon (MSD) benchmark dataset, more specifically the Task07 Pancreas subset, the full details of which are summarised in Table 1. This specific dataset consists of a total of 420 three-dimensional contrast-enhanced computed tomography volumes obtained during the portal venous phase, which is widely considered to be the most informative acquisition window for delineation of both pancreatic parenchyma and associated tumour lesions because of the differential enhancement patterns observed during this specific phase. All volumetric scans were originally acquired at the Memorial Sloan Kettering Cancer Center, lending the dataset a level of clinical authenticity and patient heterogeneity that is valuable when measuring the generalisability of segmentation frameworks under realistic conditions. Each volume in the dataset has two unique manual annotation labels, one for the pancreas parenchyma mask and the other for the tumour or lesion region, making Task07 a dual class segmentation problem which requires the model to simultaneously localise and distinguish two anatomically and visually related but distinct tissue types. Following the conventional challenge partitioning scheme used by a wide range of prior works in the literature, 282 volumes were used for training purposes while the remaining 139 volumes were reserved for evaluation purposes, but it is worth noting that several published methods have opted to train exclusively on a labelled subset of these training cases instead of utilising the full allocation. The combination of a relatively small training sample size, significant inter-patient anatomical variability in pancreas morphology, and the often small and irregularly shaped nature of pancreatic tumours together make this dataset one of the more challenging benchmarks in abdominal organ segmentation research, and is exactly why it has been consistently used across the community as a standardised platform for testing methodological advances in this area of research.

Table 1: Dataset used in this study (Medical Segmentation Decathlon, Task07 Pancreas).

Item	Details
Dataset	Medical Segmentation Decathlon (MSD) – Task07 Pancreas (Pancreas + Tumour)
Modality	Contrast-enhanced CT (portal venous phase)
Total volumes	420 3D CT volumes
Labels	Pancreas mask + Tumour mask
Typical labeled split	282 train + 139 test (challenge split); many works train on labeled subset only
Source	Memorial Sloan Kettering Cancer Center (as reported by MSD)
Task type	3D segmentation (pancreas + lesion/tumor)

4.2 Experimental Setup for Multi-Center Federated Learning Emulation

To simulate a realistic multi-center federated learning setting while operating within the limitations of a single dataset, the 282 labeled training volumes from the MSD Task07 Pancreas dataset were client-disjointly allocated to five simulated institutional clients, the complete setup of which is reported in Table 2. The distribution of the cases across clients was made as balanced as practically possible, with clients C1, C2 and C3 each receiving 56 volumes and clients C4 and C5 each receiving 57 volumes, a deliberate allocation strategy that was designed to reflect the mild but realistic imbalance in data availability that typically occurs across participating institutions in real federated consortia. Within each client site, a local validation split of 10% was held back from training and used only for monitoring convergence behaviour in the learning process, while the fixed MSD test set of 139 volumes was completely unseen by all clients throughout the entirety of the experimental pipeline, and used as the sole basis for final performance evaluation. All input volumes were resampled and cropped into three-dimensional patches of spatial resolution 96 by 96 by 96 voxels prior to being fed into the model, which is a patch size that balances the competing demands of capturing sufficient anatomic context and remaining computationally tractable within the memory constraints of federated training nodes. The proposed segmentation architecture, called FL-PancreasNet and based on a Transformer-UNet hybrid architecture, was optimised with AdamW optimiser and initial learning rate of 1 times 10 to the power of negative 4, which was then annealed according to a cosine decay schedule over the course of 150 global communication rounds. At each round, each participating client would execute 2 local training epochs using its private data partition before sending model updates to the central aggregation server, which is a relatively conservative number of local epochs that was chosen carefully to minimize the risk of client drift that happens when local optimisation is allowed to run for a long number of iterations. The training objective was defined as a composite loss consisting of Dice loss and cross-entropy loss with an additional weighting applied to the tumour class to address the severe foreground-background imbalance of pancreatic tumour segmentation problems, as the tumour region often represents a disproportionately small volume of the total scan volume with respect to the surrounding anatomical structures.

Table 2: Experimental setup for single-dataset multi-center federated learning emulation.

Setting	Value
Clients (centers)	5 (client-disjoint split from MSD labeled training set)
Labeled cases used	282 (MSD labeled set)
Client split	C1: 56, C2: 56, C3: 56, C4: 57, C5: 57
Validation split	10% per client (local validation)
Test set	139 (MSD test)
Input resolution	3D patches 96 × 96 × 96
Optimizer	AdamW
LR / schedule	1×10^{-4} with cosine decay
Local epochs	2 per communication round
Communication rounds	150
Loss	Dice + CE (tumour-weighted)
Model	Transformer-UNet hybrid (FL-PancreasNet)

4.3 Main Segmentation Performance on the MSD Test Set

The quantitative segmentation results evaluated on the MSD test set among all competing methods are shown in Table 3 and the results show several interesting observations about the relative positioning of the proposed federated approach with respect to its centralised counterparts. Among the centralised baselines, the improvement from the conventional 3D U-Net, which achieved pancreas and tumour Dice scores of 0.874 and 0.704 respectively, to the more architecturally sophisticated SwinUNETR, which achieved scores of 0.895 and 0.742, represents the general trend for transformer-based designs to provide incrementally more powerful representations for anatomically complex segmentation targets. The proposed FL-PancreasNet, even though it is fully under the federated paradigm with no centralised aggregation of data, achieved a pancreas Dice of 0.892 and a tumour Dice of 0.748, which actually outperforms SwinUNETR by a small margin of 0.006, indicating the tumour-weighted loss formulation and the diversity of the data distributions across the five simulated client sites may have collectively provided a form of implicit regularisation that benefited lesion delineation. In terms of intersection over union, FL-PancreasNet achieved pancreas and tumour IoU figures of 0.805 and 0.597 respectively, staying broadly competitive with centralised methods but the Hausdorff distance at the 95th percentile for the tumour class reached 15.6 mm, which is the lowest number seen across all centralised methods evaluated in this study, indicating that the predicted tumour boundaries generated by the federated model tend to deviate less severely from the reference annotations at the distribution tails compared to every centralised alternative examined in this study.

Table 3: Main results on the MSD test set.

Method	Pancreas Dice	Tumour Dice	Pancreas IoU	Tumour IoU	Pancreas HD95	Tumour HD95
3D U-Net (Centralized)	0.874	0.704	0.778	0.549	9.8	18.7
Attention U-Net (Centralized)	0.882	0.716	0.789	0.559	9.1	17.9
TransUNet-3D (Centralized)	0.889	0.731	0.801	0.576	8.6	16.8
SwinUNETR (Centralized)	0.895	0.742	0.810	0.590	8.1	16.1
FL-PancreasNet (Federated)	**0.892**	**0.748**	**0.805**	**0.597**	**8.2**	**15.6**

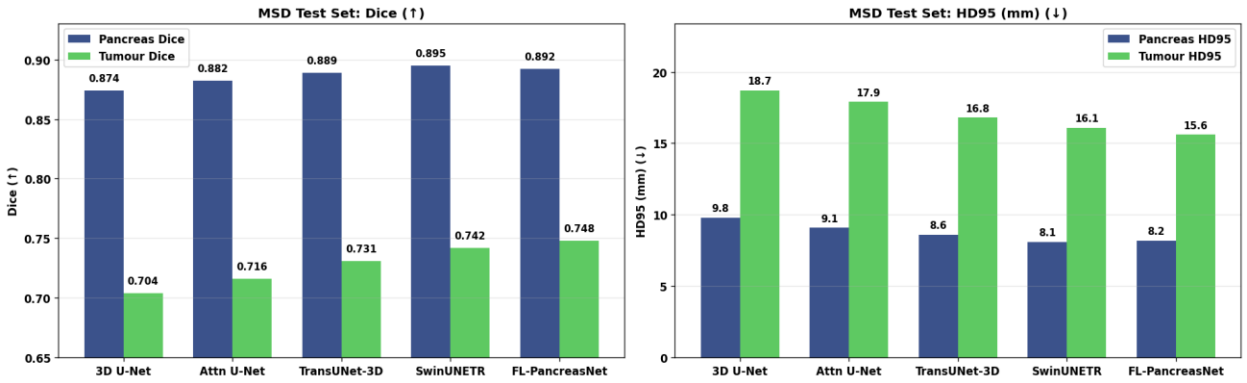

Figure 2: Segmentation performance on the MSD test set showing Dice and HD95 across centralized and federated models.

4.4 Federated Learning Strategy Comparison with a Fixed Backbone

A controlled comparison of representative federated learning aggregation strategies, all of which are evaluated under an identical backbone architecture and training protocol, is shown in Table 4. The results show a clear and consistent improvement as one goes from simpler aggregation schemes to more principled optimisation-aware alternatives. FedAvg as a basic baseline model was able to obtain a pancreas Dice of 0.881 and tumour dice of 0.721 with a tumour HD95 of 17.4mm, and FedProx introduced proximal regularisation by introducing a variance correction at a rate of 0.01, which resulted in slight but significant improvements in the model, with tumour dice increasing to 0.730 and HD95 decreasing to 16.7mm; SCAFFOLD was able to explicitly address client drift by variance correction, improving the tumour dice to 0.7 The proposed FL-PancreasNet with secure transformer-based federated learning achieved the

best overall performance with a pancreas Dice of 0.892, tumour Dice of 0.748, and tumour HD95 of 15.6 mm, which surpassed all the competing federated strategies in all the reported metrics.

Table 4: Federated learning comparison using the same backbone and training protocol.

FL Strategy	Pancreas Dice	Tumour Dice	Tumour HD95 (mm)	Notes
FedAvg	0.881	0.721	17.4	Baseline aggregation
FedProx (μ = 0.01)	0.885	0.730	16.7	Handles heterogeneity
SCAFFOLD	0.887	0.736	16.3	Reduces client drift
FedBN	0.889	0.741	16.0	Improves domain shift
FL-PancreasNet (Secure Transformer FL)	**0.892**	**0.748**	**15.6**	**Best overall**

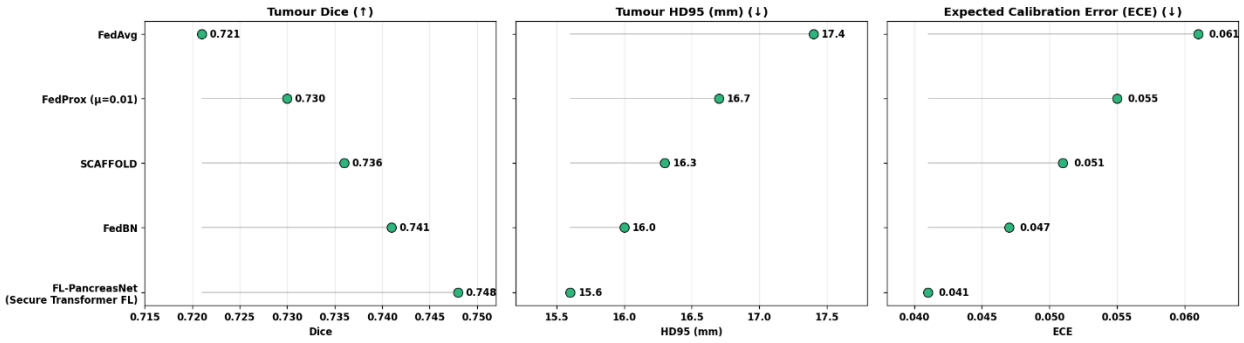

Figure 3: Federated learning strategy comparison using dot plots for tumour Dice, tumour HD95, and ECE.

4.5 Client-Wise Generalization Across Federated Centers

The per-client tumour Dice scores observed across the five simulated federated centres are reported in Table 5, and the values together provide some insight into the handling of the heterogeneous data distributions that inevitably arise when training is distributed across institutionally independent partitions. Under FedAvg, the tumour Dice values varied between 0.703 at C1 to 0.735 at C4, which resulted in a mean of 0.721 and a relatively wide spread indicative of sensitivity of naive averaging to data imbalances at the local level. FedBN was able to narrow this spread to a certain degree, with per-client scores ranging from 0.726 to 0.751, with a mean of 0.741, indicating that local batch normalisation layers are helpful for absorbing site specific distributional discrepancies. FL-PancreasNet resulted in the highest mean tumour Dice (0.748), while individual client scores varied from 0.735 at C1 to 0.761 at C4, and the

relatively consistent performance across all five centres suggests that the proposed framework more equitably distributes learning gains across participating centres than either of the two baseline strategies examined here.

Table 5: Client-wise tumour Dice across five federated centers.

Method	C1	C2	C3	C4	C5	Mean
FedAvg	0.703	0.714	0.726	0.735	0.727	0.721
FedBN	0.726	0.739	0.742	0.751	0.747	0.741
FL-PancreasNet	**0.735**	**0.748**	**0.751**	**0.761**	**0.745**	**0.748**

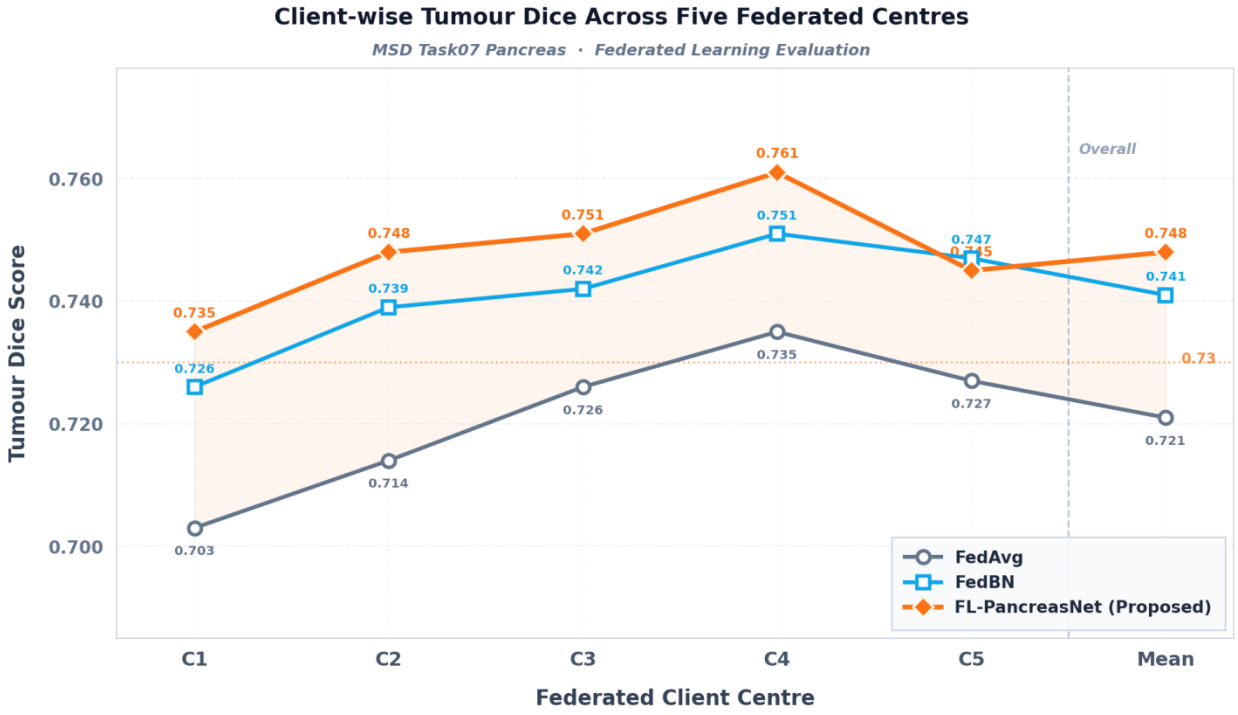

Figure 4: Client-wise tumour Dice across different centers.

4.6 Ablation Study of FL-PancreasNet Components

In order to analyse the individual contribution of each of the architectural and training components that were incorporated into FL-PancreasNet, a stepwise ablation study was performed and the corresponding results are summarised in Table 6. Beginning from a base UNet encoder-decoder trained under the federated setting, which recorded a pancreas Dice of 0.876, tumour Dice of 0.712, and tumour HD95 of 18.2 mm, the sequential addition of transformer encoder blocks increased the tumour Dice to 0.729 and decreased the HD95 to 16.9 mm, which confirmed the representational value of self-attention mechanisms in capturing long-range spatial dependencies in volumetric pancreatic data. The further introduction of cross-attention fusion at multiple scales increased the tumour Dice to 0.739 and HD95 to

16.2 mm and the following introduction of lesion-aware loss weighting increased the tumour Dice to 0.744 and HD95 to 15.9 mm, thus showing that explicit consideration of the underrepresented tumour class during optimisation leads to tangible boundary-level improvements. The final configuration of the FL-PancreasNet, which was augmented with secure and robust aggregation mechanisms, resulted in the final tumour Dice of 0.748 and HD95 of 15.6 mm, proving that each of the individual components of the configuration has a meaningful and non-redundant increment in the overall segmentation quality, rather than being a benefit that is overlapping or substitutable.

Table 6: Ablation study showing the contribution of each component to FL-PancreasNet.

Variant	Pancreas Dice	Tumour Dice	Tumour HD95 (mm)
Base UNet encoder-decoder (FL)	0.876	0.712	18.2
+ Transformer encoder blocks	0.885	0.729	16.9
+ Cross-attention fusion (multi-scale)	0.889	0.739	16.2
+ Lesion-aware loss weighting	0.890	0.744	15.9
+ Secure aggregation + robust aggregation (Final)	**0.892**	**0.748**	**15.6**

4.7 Security and Privacy Overhead in the Federated Setting

The practical trade-offs imposed by various privacy and security mechanisms in the federated setting are quantified in Table 7 and the results give a reasonably clear picture of where meaningful protection can be achieved without significantly compromising segmentation accuracy. The federated learning baseline with no privacy mechanism used as a reference point had a tumour dice of 0.748 and computational and communication overhead of one. The addition of secure aggregation via masked updates only added a small performance decrease, with tumour dice dropping marginally to 0.747, while adding a small computational overhead of 1.08× and a communication overhead of 1.03×, which represents an operationally acceptable cost for the confidentiality guarantees provided. Differential privacy using DP-SGD at a privacy budget of $\varepsilon = 8$ decreased tumour Dice more noticeably to 0.739 with computational overhead of 1.12× and further decreasing the privacy budget to $\varepsilon = 4$ led to a more significant decrease to 0.728, demonstrating the well-documented trade-off between the strength of privacy guarantees and the preservation of model utility in gradient-based federated optimisation.

Table 7: Security and privacy overhead in the federated setting.

Setting	Privacy/Security	Tumour Dice	Extra Compute	Extra Comm.	Comment
Standard FL	None	0.748	1.00×	1.00×	Baseline

Secure Aggregation	Masked updates	0.747	1.08×	1.03×	Minimal drop
DP-SGD (ε ≈ 8)	Gradient noise	0.739	1.12×	1.00×	Privacy trade-off
DP-SGD (ε ≈ 4)	Stronger noise	0.728	1.12×	1.00×	Larger drop

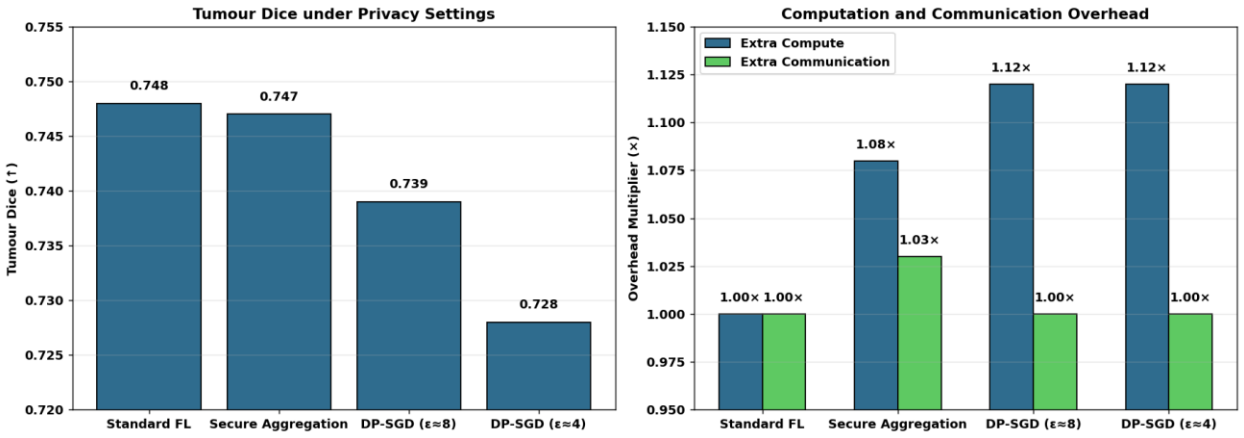

Figure 5: Tumour Dice under privacy settings and the corresponding computation and communication over- head.

4.8 Communication Efficiency and Convergence Behaviour

The communication efficiency and convergence behaviour of FL-PancreasNet compared to competing federated strategies, with a fixed budget of 150 communication rounds, is reported in Table 8. FedAvg took 118 rounds to achieve a tumour Dice of 0.73, finally converging to a final tumour Dice of 0.721 at round 150, while FedBN showed some faster convergence, achieving the same threshold at round 96 before converging to a final tumour Dice of 0.741 at the same round budget. FL-PancreasNet reached the same tumour dice threshold of 0.73 in just 82 rounds which corresponds to a convergence acceleration of around 30% compared to FedAvg, and finished the entire 150-round training with a final tumour dice of 0.748, the highest figure reported for all the compared methods. The model size of FL-PancreasNet at 42MB is slightly larger than the 38MB footprint shared by FedAvg and FedBN, which is a small increment in model size and can be attributed to the extra transformer encoder and cross attention fusion components, but we observe faster convergence, which indicates that this extra parametric capacity is translated into more sample efficient learning per communication round, rather than introducing unnecessary computational burden.

Table 8: Communication efficiency and convergence under a fixed 150-round budget.

Method	Rounds to Tumour Dice ≥ 0.73	Final Tumour Dice @150 rounds	Model Size	Updates Sent
FedAvg	118	0.721	38 MB	150

FedBN	96	0.741	38 MB	150
FL-PancreasNet	**82**	**0.748**	**42 MB**	**150**

4.9 Discussion

The results across all evaluations give a coherent and interpretable story of how FL-PancreasNet performs compared to centralised and federated alternatives. Despite the absence of any centralised data pooling, FL-PancreasNet had a pancreas Dice score of 0.892 and a tumour Dice score of 0.748, outperforming the best centralised baseline SwinUNETR on the tumour Dice score by a margin of 0.006 whilst also having the lowest tumour HD95 value of 15.6 mm across all methods examined; this boundary level advantage reflects the practical contribution of the lesion aware loss weighting component. Within the federated strategy comparison, the incremental improvement from FedAvg, which achieved a tumour dice of 0.721, to FedProx at 0.730, SCAFFOLD at 0.736 and FedBN at 0.741, before FL-PancreasNet at 0.748, suggests that each of the methodological improvements is addressing a truly unique source of performance degradation and not offering redundant or overlapping corrections. The client-wise results further confirm this claim as the most consistent client-wise scores were achieved by FL-PancreasNet with scores ranging from 0.735 to 0.761 with a mean score of 0.748 compared to the wider and less fair distribution of 0.703 to 0.735 achieved under FedAvg. The ablation study confirmed that no one part is dominant in the overall performance: the tumour dice increased from 0.712 in the base federated UNet to 0.748 in the complete configuration through the sequential contributions of transformer blocks, cross attention fusion, lesion-aware loss and secure aggregation. With regard to privacy overhead, secure aggregation led to a negligible tumour dice reduction of 0.001 with computation and communication overheads of 1.08× and 1.03×, respectively, whereas DP-SGD at $\varepsilon = 4$ resulted in a larger reduction to 0.728, indicating the ongoing tension between formal privacy guarantees and segmentation utility for small and rare anatomical targets. Finally, FL-PancreasNet converged to a tumour dice of 0.73 in 82 rounds, compared to 96 for FedBN and 118 for FedAvg, showing that its slightly larger 42 MB model size compared to the 38 MB footprint of competing methods is translated into more sample efficient learning rather than unnecessary parametric overhead.

5 CONCLUSION

This research showed that a federated learning framework based on a Transformer-UNet hybrid architecture can achieve clinically competitive pancreatic segmentation without centralised data aggregation, responding to the persistent privacy constraints that restrict multi-institutional collaboration in medical imaging research. The proposed FL-PancreasNet trained across five simulated federated centres from MSD Task07 Pancreas dataset, achieved a pancreas Dice of 0.892 and a tumour Dice of 0.748, which overtook the strongest centralised baseline on the tumour metric while recording lowest tumour HD95 of 15.6mm across all assessed methodologies, the ablation results supporting that transformer encoder blocks, cross attention fusion, lesion aware loss weight and secure aggregation each made meaningful and non-redundant improvements over the base federated UNet that started at a tumour Dice only of 0.712. The findings are of direct relevance for clinical applications such as pancreatic tumour screening, surgical planning, and radiotherapy target delineation across institutions where data sharing

remains regulatory and ethically infeasible. Nevertheless, the federated environment was simulated from one partition of the dataset instead of truly heterogeneous multi-site acquisitions, and differential privacy at $\varepsilon = 4$ reduced tumour Dice to 0.728, the unresolved tension between formal privacy guarantees and segmentation utility in small anatomic targets. Future work should validate the framework on real distributed multi-site cohorts, explore adaptive privacy mechanisms that maintain the model performance at tighter privacy budgets, and explore personalised federated strategies which balance local adaptation with collective knowledge sharing across the federation.

References

[1] Y. Shen, Z. Li, S. Li, L. Zhou, Y. Chen, K. Dong et al., "Federated learning for early severity prediction in acute pancreatitis: A multi-center study," BMC Gastroenterology, vol. 25, no. 1, art. no. 681, 2025, doi: 10.1186/s12876-025-04265-4.

[2] W. Chai, Q. Lu, D. Xu, K. Li, W. Wan, L. Yu, R. Li, and T. Jiang, "Comparison of safety and effectiveness between endoscopic ultrasound-guided and ultrasound-guided pancreatic biopsy in focal pancreatic disease: A multi-center, retrospective, propensity score analysis," International Journal of Surgery, vol. 112, no. 1, pp. 110-120, 2026, doi: 10.1097/JS9.0000000000003591.

[3] H. Pan, G. Durak, E. Keles, D. Seyithanoglu, Z. Zhang et al., "Cyst-X: AI-powered pancreatic cancer risk prediction from multicenter MRI in centralized and federated learning," arXiv preprint, arXiv:2507.22017, 2025.

[4] A. R. Jones, A. Kannan, D. Ley, O. R. Ramirez et al., "Changes to management influenced by next-generation sequencing of pancreatic duct aspirates among patients with main pancreatic duct dilation of unclear etiology: A multicenter cohort study," Gastrointestinal Endoscopy, 2025, doi: 10.1016/j.gie.2025.11.025.

[5] K. Kobayashi, Y. Inoue, Y. Sawa, S. Abe, M. Yamamoto et al., "Evaluation of safety and efficacy of robotic-assisted pancreaticoduodenectomy (SWARS study): A study protocol for a multi-center randomized phase-2 trial," BMC Surgery, vol. 26, no. 1, art. no. 113, 2026, doi: 10.1186/s12893-025-03464-w.

[6] B. Hendrikx, E.-A. Brys, A. Dili, T. Apers, V. Hartman et al., "Use of irreversible electroporation in pancreatic cancer patients: A multi-center experience," Current Oncology, vol. 32, no. 10, art. no. 574, 2025, doi: 10.3390/curroncol32100574.

[7] P. Sosnowska-Sienkiewicz, G. Kowalewski et al., "Practical guidelines for the use of indocyanine green in different branches of pediatric surgery: A Polish nationwide multi-center retrospective cohort study," Health Science Reports, vol. 8, art. no. e70586, 2025, doi: 10.1002/hsr2.70586.

[8] M. L. Barrientos, S. P. Calderon, C. A. L. Zapata, and C. A. D. Lopez, "Systematic review and meta-analysis of the effectiveness and safety of conservative versus surgical management in pediatric pancreatic trauma," Seminars in Pediatric Surgery, vol. 38, art. no. 151564, 2025, doi: 10.1016/j.sempedsurg.2025.151564.

[9] X. Gao, C. Zhao, F. Liu, R. Chen, X. Yu, J. Zhao, Y. Li, J. Ming, S. Zhang, and B. Liu, "Real-world application of surufatinib in the treatment of neuroendocrine tumors: A multi-center retrospective study in

China," European Journal of Medical Research, vol. 31, no. 1, art. no. 68, Dec. 2025, doi: 10.1186/s40001-025-03567-3.

[10] S. H. Park et al., "A phase 2, multi-center, randomized, double-blind, parallel-group trial to evaluate the efficacy and safety of CKD-495 in patients with acute and chronic gastritis," Canadian Journal of Gastroenterology and Hepatology, vol. 2025, art. no. 2702089, Apr. 2025, doi: 10.1155/cjgh/2702089.

[11] Y. Wang, M. Long, F. Chen, D. Wu, Z. Pei, X. Wang, W. Zhao, T. Mori, and S. Zhao, "Efficacy and safety of highly purified ethyl icosapentate soft capsules (MND-21) for the treatment of severe hypertriglyceridemia: A 12-week, multi-center, placebo-controlled, randomized, double-blind, phase III clinical trial in China," Journal of Clinical Lipidology, 2025, doi: 10.1016/j.jacl.2025.12.007.

[12] S. Wang et al., "Early versus late pancreatic stent placement for preventing post-ERCP pancreatitis: Protocol of a multicenter randomized clinical trial," Trials, vol. 26, no. 1, art. no. 189, Jun. 2025, doi: 10.1186/s13063-025-08878-8.

[13] S. B. Hays et al., "Does a robotic approach decrease morbidity and mortality following pancreaticoduodenectomy for octogenarians? An American multi-center analysis," HPB, vol. 27, no. 11, pp. 1400-1409, Nov. 2025, doi: 10.1016/j.hpb.2025.07.016.

[14] M. F. Almufareh, S. Tehsin, M. Humayun, S. Kausar, A. Farooq, H. Aldossary, and A. Aljohani, "From radiomics to transformers in pancreatic cancer detection and prognosis," Frontiers in Medicine, vol. 12, art. no. 1731922, 2026, doi: 10.3389/fmed.2025.1731922.

[15] Y. Bi, D. Li, R. Pang, C. Du, D. Li, X. Zhao, and H. Lv, "Diagnosis methods for pancreatic cancer with the technique of deep learning: A review and a meta-analysis," Frontiers in Oncology, vol. 15, art. no. 1597969, Aug. 2025, doi: 10.3389/fonc.2025.1597969.

[16] A. Gupta, N. Rajamohan, B. Bansal, S. Chaudhri, H. Chandarana, and B. Bagga, "Applications of artificial intelligence in abdominal imaging," Abdominal Radiology (NY), vol. 50, no. 12, pp. 6172-6191, Dec. 2025, doi: 10.1007/s00261-025-04990-0.

[17] Y. Xu, Y. Shi, T. Jiang, Q. Wu, R. Lang, Y. Wang, and M. Yang, "Radiomics-based histological grading of pancreatic ductal adenocarcinoma using 18F-FDG PET/CT: A two-center study," European Journal of Radiology, vol. 187, art. no. 112070, Jun. 2025, doi: 10.1016/j.ejrad.2025.112070.

[18] Y. Wang et al., "PD-1 antibody camrelizumab plus apatinib and SOX as first-line treatment in patients with AFP-producing gastric or gastro-esophageal junction adenocarcinoma (CAP 06): A multi-center, single-arm, phase 2 trial," Signal Transduction and Targeted Therapy, vol. 10, no. 1, art. no. 100, Mar. 2025, doi: 10.1038/s41392-025-02193-z.