NeoNet: An End-to-End 3D MRI-Based Deep Learning Framework for Non-Invasive Prediction of Perineural Invasion via Generation-Driven Classification

Youngung Han[1], Minkyung Cha[1], Kyeonghun Kim[2], Induk Um[3], Myeongbin Sho[4], Joo Young Bae[1], Jaewon Jung[1], Jung Hyeok Park[1], Seojun Lee[1], Nam-Joon Kim[1*], Woo Kyoung Jeong[5*], Won Jae Lee[6], Pa Hong[7], Ken Ying-Kai Liao[8], Hyuk-Jae Lee[1]

[1]Seoul National University, South Korea
[2]OUTTA, South Korea
[3]Chung-Ang University, South Korea
[4]Sookmyung Women's University, South Korea
[5]Department of Radiology and Center for Imaging Science, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea
[6]Department of Radiology, Sungkyunkwan University Samsung Changwon Hospital, Changwon, Republic of Korea
[7]Department of Radiology, Samsung Changwon Hospital, Sungkyunkwan University School of Medicine
[8]NVIDIA AI Technology Center, Taipei, Taiwan
yuhan@snu.ac.kr, cmk4911@snu.ac.kr, kyeonghun.kim@outta.ai, naerun4@cau.ac.kr, bin@sookmyung.ac.kr, domi37@snu.ac.kr, owenjaewon@snu.ac.kr, dylan0122@snu.ac.kr, seojun2658@snu.ac.kr, knj01@snu.ac.kr, jeongwk@skku.edu, wjlee@skku.edu, papa.hong@samsung.com, kenyingkail@nvidia.com, hjlee@capp.snu.ac.kr

Abstract

Minimizing invasive diagnostic procedures is a central goal in medical imaging. Perineural invasion (PNI), a critical prognostic factor where tumors infiltrate nerves, remains difficult to confirm noninvasively, as its features are often imperceptible in conventional MRI. PNI research is severely hampered by data scarcity. Our study utilized a dataset collected over a decade at Samsung Medical Center (SMC), initially comprising 306 patients. After rigorous quality control, the final cohort included 128 T1-weighted hepatobiliary phase MRI scans, exhibiting significant class imbalance (44 PNI-positive/84 PNI-negative). To address these challenges, we present NeoNet, the first integrated end-to-end 3D deep learning framework for PNI prediction in cholangiocarcinoma that avoids reliance on radiomics or handcrafted features. NeoNet integrates three modules: (1) NeoSeg, utilizing a Tumor-Localized ROI Crop (TLCR) algorithm; (2) NeoGen, a 3D Latent Diffusion Model (LDM) with ControlNet, conditioned on anatomical masks to generate synthetic image patches, specifically balancing the dataset to a 1:1 ratio; and (3) NeoCls, the final prediction module. For NeoCls, we developed the PNI-Attention Network (PattenNet), which uses the frozen LDM encoder and specialized 3D Dual Attention Blocks (DAB) designed to detect subtle intensity variations and spatial patterns indicative of PNI. In rigorous 5-fold cross-validation, NeoNet outperformed baseline 3D models. By leveraging synthetic data for balanced training, PattenNet achieved the highest performance with a maximum AUC of 0.7903.

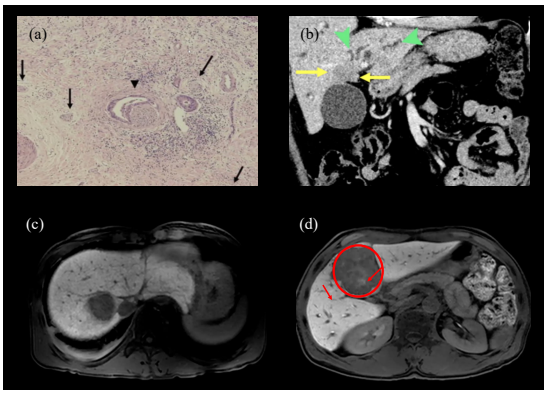

Figure 1: Morphological features of perineural invasion (PNI) demonstrated by histopathology (a), CT (b), and MRI (c-PNI negative, d-PNI positive). Arrows indicate the locations of characteristic PNI features in each modality.

Introduction

Nerves throughout the human body are organized into bundles of fibers surrounded by the perineurium—a protective sheath. When cancer cells infiltrate this structure and spread in or along the nerves, the process is referred to as perineural invasion (PNI) (Liebig et al. 2009).

PNI is recognized as an active process facilitated by biochemical crosstalk between tumor cells and the nerve (Amit, Na'ara, and Gil 2016; Bapat et al. 2011), serving as an independent route of metastasis. It is a critical prognostic factor, particularly in cholangiocarcinoma, where patients with PNI show significantly higher recurrence rates after surgery (Hruban et al. 2022; Qian et al. 2024). Accurate,

*Corresponding Author

pre-surgical diagnosis of PNI is therefore clinically vital.

It enables better risk stratification, informs the extent of surgical resection required, and guides adjuvant therapy decisions, potentially improving patient outcomes significantly.

Despite the critical need, efforts to predict PNI non-invasively have remained challenging. A key reason lies in the intrinsic difficulty of capturing PNI on imaging. While features of PNI are clear in post-surgical histopathology images, corresponding signs on CT and MRI are often subtle and less definitive (Figure 1), appearing as minor nerve thickening or heterogeneous enhancement (Purohit et al. 2019).

Compounding this challenge is the profound scarcity of well-labeled PNI datasets. Definitive PNI identification requires histopathological examination, making data collection slow. This scarcity is a field-wide issue; related studies consistently rely on small cohorts typically fewer than 200 cases (Huang et al. 2021; Qi et al. 2025), leading to significant class imbalance and hindering model generalization.

Methodologically, current approaches predominantly rely on radiomics (Gillies, Kinahan, and Hricak 2016) or handcrafted features and manual segmentation (Huang et al. 2021). Even recent advancements incorporating deep learning often utilize it as part of a hybrid approach rather than a fully end-to-end framework (Qi et al. 2025; Li et al. 2020). These methods do not leverage the full potential of end-to-end deep learning frameworks capable of learning subtle, implicit 3D features directly from the imaging data.

Therefore, we propose NeoNet, an integrated 3D deep learning framework designed to overcome these limitations. Our framework is the first end-to-end 3D deep learning approach for PNI prediction that integrates automated localization, generative data balancing, and specialized classification. By moving beyond traditional radiomics and addressing data scarcity head-on, NeoNet aims to provide a robust, non-invasive tool for PNI assessment.

Our main contributions are summarized as follows:

- We propose NeoNet, the first integrated end-to-end 3D deep learning framework specifically designed for PNI prediction from MRI, overcoming the limitations of radiomics-based approaches by learning features directly from the imaging data.
- We utilize NeoGen, a 3D LDM with ControlNet, to address the critical challenges of data scarcity and severe class imbalance inherent in PNI datasets by generating anatomically constrained synthetic data and balancing the training cohort to a 1:1 ratio.
- We design PattenNet, a lightweight 3D classifier incorporating a frozen LDM encoder and Dual Attention Blocks (DAB) specifically engineered to detect the subtle, localized features of PNI by focusing on critical intensity variations and spatial patterns at the tumor interface.

Related Works

PNI Prediction in Medical Imaging The non-invasive prediction of PNI is challenging due to the subtlety of its features on conventional imaging. Traditional assessment relies on subjective radiological interpretation of MRI features (Purohit et al. 2019).

To achieve quantitative assessment, radiomics (Gillies, Kinahan, and Hricak 2016; Lambin et al. 2017) has emerged as the predominant approach. Methodologically, however, radiomics relies on handcrafted features extracted from predefined regions of interest, often requiring labor-intensive manual segmentation.

Data scarcity is an inherent obstacle, as definitive PNI diagnosis requires post-surgical pathology. Consequently, publicly available datasets are nonexistent, and studies rely on small, private cohorts. For instance, (Huang et al. 2021) utilized a cohort of 101 patients to predict PNI in extrahepatic cholangiocarcinoma using radiomics features and machine learning classifiers. These small datasets invariably suffer from significant class imbalance, typically addressed using statistical oversampling (e.g., SMOTE), which may not be optimal for high-dimensional imaging data.

The adoption of end-to-end deep learning has been severely limited by these constraints. Small datasets hinder the ability of deep networks to learn generalizable features. Recent work by (Qi et al. 2025) utilized a larger cohort of 192 cases but employed a hybrid fusion model combining ResNet101 (He et al. 2016) features, radiomics, and clinical data, rather than a pure end-to-end deep learning approach for feature learning.

Localization in Medical Image Classification For localized pathologies like PNI, which manifest at the tumor-nerve interface, focusing the model's attention on the relevant anatomical region is crucial. While end-to-end frameworks combining segmentation and classification are common in medical imaging, the precise localization strategy is vital for performance.

Generative Models for Medical Data Augmentation Data scarcity and class imbalance are major challenges in medical deep learning. Generative models offer a powerful tool for data augmentation and balancing (Yi, Walia, and Babyn 2019; Kazeminia et al. 2020). Variational AutoEncoders (VAEs) (Kingma and Welling 2013) and Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) have been widely used. More recently, Diffusion models (Sohl-Dickstein et al. 2015) have shown remarkable success in high-fidelity image generation, including DDPM (Ho, Jain, and Abbeel 2020) and Latent Diffusion Models (LDM) (Rombach et al. 2022). LDM operates in a compressed latent space, making it efficient for high-resolution 3D data (Ben Melech Stan et al. 2023), and has shown promise in medical domains such as brain imaging (Pinaya et al. 2022).

To control the generation process, conditional mechanisms are crucial. ControlNet (Zhang, Rao, and Agrawala 2023) allows for integrating external control signals, such as anatomical masks. In this work, we leverage a 3D LDM (NeoGen) with ControlNet to generate realistic synthetic MRI patches conditioned on segmentation masks, which are then used to balance the training dataset.

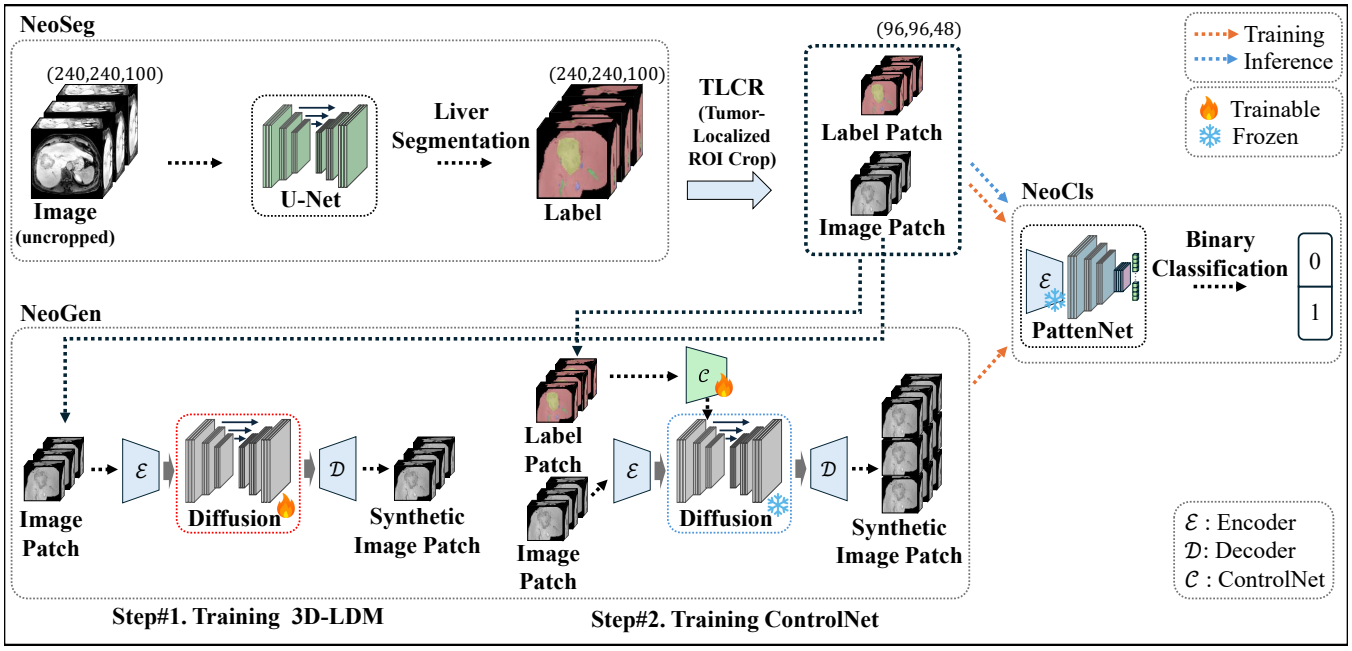

Figure 2: Overview of the NeoNet framework. (1) NeoSeg performs automated liver and tumor segmentation on inputs (e.g., $240 \times 240 \times 100$), producing segmentation labels (e.g., $240 \times 240 \times 100$). The TLCR algorithm utilizes these labels and the original image to extract localized image and label patches ($96 \times 96 \times 48$). (2) NeoGen utilizes a 3D-LDM (Step #1) and trains ControlNet (Step #2) conditioned on label patches to generate synthetic image patches, balancing the dataset. (3) NeoCls utilizes PattenNet, initialized with the frozen LDM encoder ($\mathcal{E}$), to predict PNI status from real and synthetic patches.

Methodology

NeoNet (Figure 2) is an integrated framework composed of three main modules: NeoSeg for segmentation and localization, NeoGen for synthetic data generation, and NeoCls for PNI prediction.

Datasets and Preprocessing

We utilized a retrospective dataset of anonymized T1-weighted, contrast-enhanced MR images from the hepatobiliary phase of cholangiocarcinoma patients collected over a decade at Samsung Medical Center (SMC). An initial pool of 306 patients was manually inspected. 178 patients were excluded due to poor image quality, inaccurate annotations, incomplete sequences, or tumors occupying more than 70% of the liver volume.

The final analysis cohort comprised 128 patients. The presence of PNI was confirmed by post-surgical examination. The dataset exhibits significant class imbalance, with 44 PNI-positive and 84 PNI-negative cases. Images were provided in NIfTI format, along with ground truth annotations.

For NeoSeg training, Intensities were normalized between 0 and 1. Labels were remapped to: background (0), liver (1), and tumor (2).

Experimental Protocol

To ensure robustness and prevent data leakage, we employed a stratified 5-fold cross-validation strategy consistently across the entire framework. In each fold, the data was split into 80% training and 20% validation sets at the patient level. All components, including the LDM pretraining (VAE and U-Net) and ControlNet training, were strictly performed using only the training data of the respective fold.

NeoSeg and Tumor Localization

NeoSeg performs automated segmentation of the liver and tumor. We evaluated four architectures: U-Net (Ronneberger, Fischer, and Brox 2015), SegResNet (Myronenko 2018), DynUNet (Isensee et al. 2021), and SwinUNETR (Tang et al. 2022). SwinUNETR, leveraging advancements in Vision Transformers (Dosovitskiy et al. 2020; Liu et al. 2021), achieved the highest performance and was selected as the backbone.

Tumor-Localized ROI Crop (TLCR) Based on the segmentation results, we implement the TLCR algorithm (Algorithm 1) to extract a standardized input volume focused on the tumor region. This approach is based on the assumption that the most salient features for PNI prediction are located in and immediately around the tumor.

TLCR calculates the geometric center of the tumor mask and extracts a fixed-size 3D crop of ($96 \times 96 \times 48$) voxels. The algorithm includes boundary clamping.

The output is a dual-channel volume: Channel 1 contains the peritumoral region (liver tissue and tumor), and Channel 2 contains only the tumor core. Voxels outside these masks are zeroed out. This dual-channel representation guides the classifier to focus on both the tumor characteristics and the tumor-liver interface.

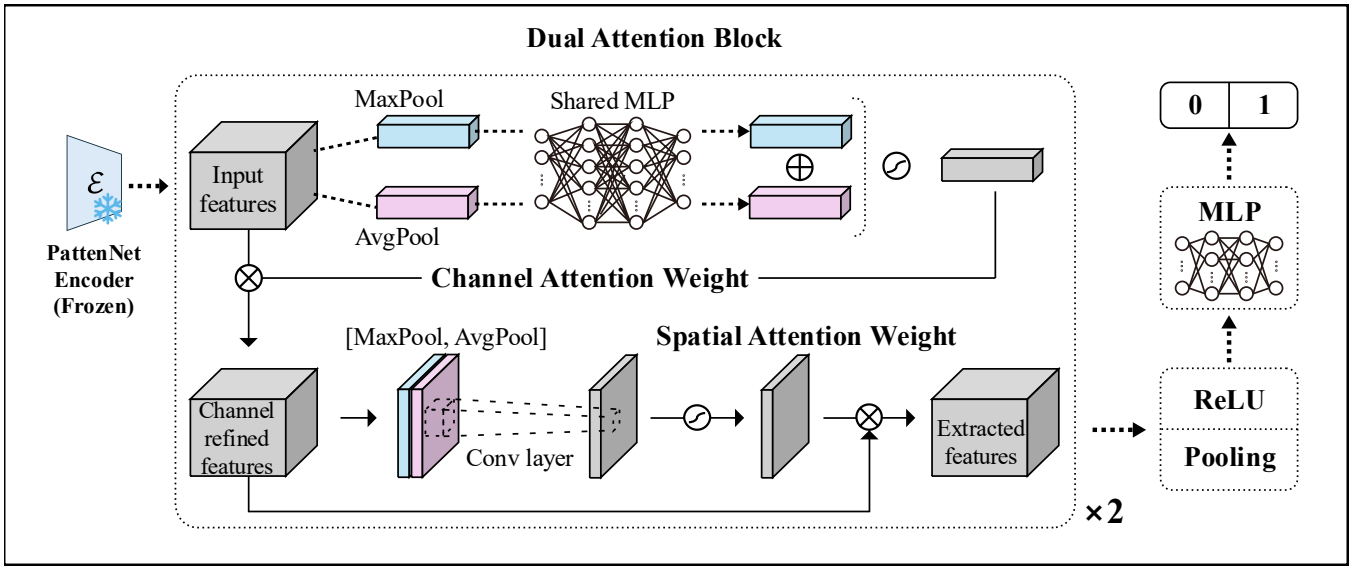

Figure 3: Architecture of PattenNet for PNI prediction. The model utilizes a frozen encoder from the 3D-LDM followed by two Dual Attention Blocks (DAB). Each DAB integrates 3D channel and spatial attention. Spatial attention utilizes a 3D convolution ($7 \times 7 \times 7$) to effectively model 3D spatial relationships. The attended features are processed through global pooling and an MLP for binary classification.

Algorithm 1: Tumor-Localized ROI Crop Algorithm

Input:
3D medical image $\mathbf{I}$, crop size $\mathbf{C} = (c_x, c_y, c_z)$,
label map $\mathbf{L} = \{0 : background,\ 1 : liver,\ 2 : tumor\}$
Output:
Cropped vectors
$\mathbf{T_1} \in \mathbb{R}^{c_x \times c_y \times c_z}$ (peritumoral mask)
$\mathbf{T_2} \in \mathbb{R}^{c_x \times c_y \times c_z}$ (tumor mask)

Find tumor voxel coordinates:
$\mathcal{S} \leftarrow \{(x, y, z) \mid \mathbf{L}[x, y, z] = 2\}$
if $\mathcal{S}$ is empty **then**
  $\mathbf{T_1}, \mathbf{T_2} \leftarrow \mathbf{0}^{c_x \times c_y \times c_z}$
  return $\mathbf{T_1}, \mathbf{T_2}$
end if
$[x_{\min}, y_{\min}, z_{\min}] \leftarrow \min(\mathcal{S})$
$[x_{\max}, y_{\max}, z_{\max}] \leftarrow \max(\mathcal{S}) + 1$
$\mathbf{center} \leftarrow \left(\left\lfloor \frac{x_{\min}+x_{\max}}{2} \right\rfloor, \left\lfloor \frac{y_{\min}+y_{\max}}{2} \right\rfloor, \left\lfloor \frac{z_{\min}+z_{\max}}{2} \right\rfloor\right)$
$\mathbf{start} \leftarrow \max\left(\mathbf{center} - \frac{\mathbf{C}}{2}, 0\right)$
$\mathbf{end} \leftarrow \mathbf{start} + \mathbf{C}$
Extract crop:
$\mathbf{I}_{crop} \leftarrow \mathbf{I}[\mathbf{start} : \mathbf{start} + \mathbf{C}]$
$\mathbf{L}_{crop} \leftarrow \mathbf{L}[\mathbf{start} : \mathbf{start} + \mathbf{C}]$
$\mathbf{T_1} \leftarrow \mathbf{I}_{crop}[\mathbf{L}_{crop} \in \{1, 2\}]$
$\mathbf{T_2} \leftarrow \mathbf{I}_{crop}[\mathbf{L}_{crop} \in \{2\}]$
return $\mathbf{T_1}, \mathbf{T_2}$

NeoGen: Conditioned Synthetic Generation

To address data scarcity, NeoGen generates synthetic 3D MRI patches using a 3D Latent Diffusion Model (3D-LDM) (Ben Melech Stan et al. 2023) enhanced with ControlNet (Zhang, Rao, and Agrawala 2023).

Addressing Data Scarcity and Imbalance The limited size of the PNI cohort of 128 cases and the inherent class imbalance pose significant challenges. NeoGen is utilized to mitigate these issues. By generating high-fidelity synthetic patches conditioned on anatomical masks, we augment the training dataset within each cross-validation fold. Crucially, we employ NeoGen to generate sufficient synthetic samples to achieve a balanced 1:1 ratio between PNI-positive and PNI-negative samples in the final augmented training set used for NeoCls.

3D Latent Diffusion Model The 3D-LDM operates in a compressed latent space learned by an autoencoder (VAE). The encoder $\mathcal{E}$ transforms a medical volume x into a latent representation $z = \mathcal{E}(x)$. A diffusion model (3D U-Net) ϵ_θ is trained to denoise the latent features z_t at timestep t. The loss function is:

$$\mathcal{L}_{LDM} = \mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}[||\epsilon - \epsilon_\theta(z_t, t)||_2^2] \quad (1)$$

ControlNet Conditioning ControlNet introduces trainable copies of the LDM blocks to integrate external control signals. Given a network block $F(\mathbf{x}; \theta)$, the controlled output $\mathbf{y}_c$ is:

$$\mathbf{y}_c = F(\mathbf{x}; \theta) + Z_2\left(F\left(\mathbf{x} + Z_1(\mathbf{c}); \theta'\right)\right) \quad (2)$$

where θ' are trainable parameters, and Z_1, Z_2 are zero-initialized $1 \times 1 \times 1$ convolution layers.

We employ the anatomical mask as the conditioning input ($\mathbf{c}$). The dual-channel label patch from TLCR guides the spatial structure of the generated image, ensuring anatomical plausibility.

The overall learning objective incorporates this condition:

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}_0,\mathbf{c},\epsilon,t}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c})\|_2^2\right] \quad (3)$$

NeoCls: PNI-Attention Network (PattenNet)

NeoCls utilizes PattenNet (Figure 3), a lightweight 3D classifier optimized for detecting subtle PNI features. PattenNet leverages the encoder from the trained 3D-LDM as a frozen feature extractor, followed by a stack of two Dual Attention Blocks (DABs).

Dual Attention Block (DAB) To effectively capture the subtle intensity variations and complex spatial patterns characteristic of PNI, we designed the Dual Attention Block (DAB). DAB sequentially applies 3D channel and spatial attention mechanisms, adapted from concepts in CBAM (Woo et al. 2018) and general attention principles (Vaswani et al. 2017), to refine the feature maps extracted by the frozen LDM encoder.

Channel Attention identifies what features are relevant to PNI by adaptively recalibrating channel-wise feature responses. For an input feature map $\mathbf{F} \in \mathbb{R}^{D\times H\times W\times C}$, 3D global max and average pooling are applied, followed by a shared MLP:

$$\mathbf{M}_c(\mathbf{F}) = \sigma(\text{MLP}(\text{AvgPool}(\mathbf{F})) + \text{MLP}(\text{MaxPool}(\mathbf{F}))) \tag{4}$$

3D Spatial Attention identifies where the PNI features are located, focusing attention on critical regions such as the tumor-liver interface. Pooling is applied across the channel dimension, and the resulting map is processed by a 3D convolution layer:

$$\mathbf{M}_s(\mathbf{F}) = \sigma(f^{7\times7\times7}([\text{AvgPool}_c(\mathbf{F}); \text{MaxPool}_c(\mathbf{F})])) \tag{5}$$

where $f^{7\times7\times7}$ denotes a 3D convolution with a $7 \times 7 \times 7$ kernel to capture sufficient 3D spatial context.

The refined feature map is obtained by sequentially applying the attention maps. After the DABs, the features are passed through global pooling and a fully connected layer for binary classification.

Training Protocol and Implementation Details

All modules were trained strictly following the 5-fold CV protocol defined previously. All training was performed on NVIDIA A100 GPUs, utilizing the AdamW optimizer (Loshchilov and Hutter 2017) consistently across all stages.

NeoSeg NeoSeg (SwinUNETR) training utilized a combined Dice and Cross-Entropy loss, with a learning rate (LR) of 10^{-4} for 400 epochs with a batch size of 2.

NeoGen The LDM training involved two stages within each fold. **Stage 1 (LDM Training):** The VAE and 3D U-Net were trained from scratch using the fold's training split. The VAE was trained for 6,000 epochs with L1 loss, KL regularization of 10^{-7}, and a learning rate of 10^{-6}. The 3D U-Net was trained for 6,000 epochs with L2 loss and a learning rate of 10^{-5}. **Stage 2 (ControlNet Training):** ControlNet was trained for 3,000 epochs, keeping the LDM weights frozen. It was optimized using MSE loss with a learning rate of 10^{-5} and batch size of 8, conditioned on the anatomical masks. Synthetic image patches were generated post-training to balance the dataset.

NeoCls PattenNet and baseline 3D models, which include ResNet (He et al. 2016), DenseNet (Huang et al. 2017), EfficientNet (Tan and Le 2019), and SwinTransformer (Liu et al. 2021), were trained using the dual-channel TLCR outputs. We evaluated models trained on the original imbalanced data (R) and models trained on the augmented dataset (R+S). For R+S training, NeoGen was utilized to generate synthetic data until a balanced 1:1 ratio was achieved.

Training utilized a learning rate of 10^{-4} with batch size 4 for up to 300 epochs, with early stopping that had patience of 20 epochs based on validation AUC. The LDM encoder weights in PattenNet, which were trained in Stage 1 of the respective fold, were kept frozen.

Evaluation Metrics

Segmentation performance was evaluated using the Dice Similarity Coefficient (Dice). Generation quality was evaluated using Fréchet Inception Distance (FID) (Heusel et al. 2017), PSNR, SSIM, and LPIPS (Zhang et al. 2018). Classification performance was evaluated using the Area Under the Receiver Operating Characteristic Curve (AUC), reporting the mean AUC across the 5-fold cross-validation.

Model	Mean Dice
U-Net	0.9453
SegResNet	0.9416
DynUNet	0.9482
SwinUNETR	**0.9516**

Table 1: Performance comparison of different segmentation models for liver and tumor segmentation.

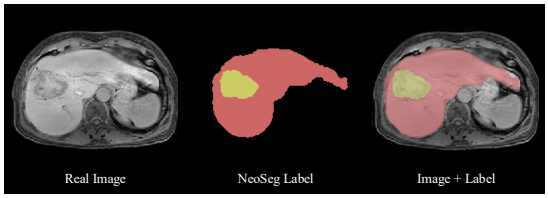

Figure 4: Labels generated by NeoSeg. Real Image (left), Segmentation masks (middle: peritumoral mask in red, tumor in yellow), and Overlay (right).

Experiments

Liver and Tumor Segmentation Performance

Table 1 presents the segmentation performance. SwinUNETR achieved the highest mean Dice score (0.9516) and was selected as the NeoSeg backbone. Figure 4 shows representative segmentation results, illustrating the accuracy of the automated localization.

Synthetic MRI Generation Quality

We evaluated the quality of synthetic images generated by NeoGen. Figure 5 compares a real patch with a synthetic

Model (3D)	R (Imbalanced)	R + S (25%)	R + S (50%)	R + S (75%)	R + S (100% Balanced)
ResNet-50	0.6938	0.7388	0.7599	0.7551	0.7618
ResNet-101	0.6284	0.6989	0.7024	0.7117	0.7203
ResNet-152	**0.7078**	0.7233	0.7421	0.7382	0.7526
ResNet-200	0.6687	0.7250	0.7307	0.7397	0.7412
DenseNet-121	0.6788	0.7443	0.7380	0.7490	0.7551
DenseNet-169	0.6725	0.7021	0.7387	0.7434	0.7584
DenseNet-201	0.6911	0.7133	0.7226	0.7278	0.7311
DenseNet-264	0.6712	0.6899	0.7012	0.7037	0.7155
EfficientNet-B0	0.6798	0.6897	0.6995	0.6910	0.6998
EfficientNet-B1	0.6725	0.7127	0.7098	0.7101	0.7186
EfficientNet-B2	0.6623	0.6899	0.7085	0.7024	0.7110
EfficientNet-B3	0.6533	0.7215	0.7462	0.7671	0.7755
SwinTransformer	0.6829	0.7423	0.7454	0.7503	0.7522
PattenNet (Ours)	0.7001	**0.7577**	**0.7670**	**0.7812**	**0.7903**
Mean AUC	0.6760	0.7178	0.7294	0.7336	0.7417

Table 2: AUC comparison of PNI classifiers trained on different R (Real) to S (Synthetic) ratios using 5-fold cross-validation. R+S columns represent progressive augmentation towards a fully balanced (1:1) dataset (R+S 100% Balanced). Results are presented as Mean AUC. All baseline models are 3D implementations.

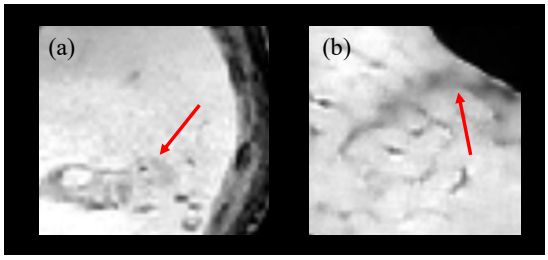

Figure 5: Real Patch (a) and Synthetic Patch (b) generated by NeoGen using the corresponding anatomical mask. Morphological features are highlighted by red arrows.

Model	LPIPS ↓	SSIM ↑	PSNR ↑
QNeoGen VQVAE	0.3003	0.4958	15.2210
NeoGen VAE	**0.2010**	**0.7408**	**22.7427**

Table 3: Reconstruction quality comparison between NeoGen VAE and QNeoGen VQVAE.

patch generated using the corresponding anatomical mask, demonstrating the fidelity of the generated morphology.

Reconstruction Quality We compared the standard VAE used in NeoGen against a Vector Quantized VAE (VQ-VAE) variant (QNeoGen). As shown in Table 3, NeoGen VAE demonstrated superior performance across all metrics (LPIPS, SSIM, PSNR).

Generation Fidelity Table 4 compares the FID scores. NeoGen (3D LDM + ControlNet with VAE) achieved the best average FID (5.3314), significantly outperforming the baseline 3D LDM (NeoGen DM, 12.1008) and the QNeoGen variants.

Method	FID ↓ (A)	FID ↓ (S)	FID ↓ (C)	FID ↓ (Avg.)
QNeoGen DM	18.9751	22.1884	39.2260	26.7965
QNeoGen	14.5820	18.4123	28.6891	20.5611
NeoGen DM	7.6284	10.8560	17.8180	12.1008
NeoGen	**2.9440**	**4.8279**	**8.2223**	**5.3314**

Table 4: Fréchet Inception Distance (FID) comparison across Axial (A), Sagittal (S), and Coronal (C) views.

PNI Prediction Performance

Table 2 summarizes the PNI prediction performance of various 3D classification models. We evaluated performance using the original imbalanced real data (R) and augmented datasets where synthetic data (S) was progressively added to achieve a fully balanced (1:1) dataset.

Impact of Synthetic Data Augmentation and Balancing The inclusion of synthetic data generated by NeoGen consistently improved performance across all models. When trained on the original imbalanced real data only (R), the mean AUC across all models was 0.6760. When the dataset was fully balanced using synthetic data (R+S 100% Balanced), the mean AUC boosted to 0.7417.

Model Comparison Our proposed PattenNet consistently outperformed all baseline 3D architectures, including established models such as ResNet (He et al. 2016), DenseNet (Huang et al. 2017), EfficientNet (Tan and Le 2019), and SwinTransformer (Liu et al. 2021). When trained on real data only, PattenNet achieved an AUC of 0.7001.

With the fully balanced dataset, PattenNet achieved the highest AUC of 0.7903. This highlights the effectiveness of the specialized dual attention mechanism and the use of the frozen LDM encoder for PNI detection.

Configuration	AUC (R+S)
PattenNet (Full Model, 2 DABs)	**0.7903**
DAB Component Ablation (2 Blocks)	
w/ Channel Attention only	0.7276
w/ Spatial Attention only	0.7410
Number of DABs	
0 Blocks (Base LDM Encoder)	0.7001
1 Block	0.7589
3 Blocks	0.7713

Table 5: Ablation study evaluating the impact of attention components and the number of Dual Attention Blocks (DABs) in PattenNet.

Ablation Study

We conducted ablation studies to evaluate the contribution of key components in the PattenNet architecture (Table 5). All ablation experiments were conducted using the R+S (100% Balanced) data regime.

Impact of Attention Mechanisms We first evaluated the importance of the specific components within the DABs in the proposed 2-block architecture. Removing both channel and spatial attention (equivalent to 0 DABs, AUC 0.7001) significantly degraded performance. Using only channel attention (AUC 0.7276) or only spatial attention (AUC 0.7410) provided substantially inferior performance compared to the full configuration (AUC 0.7903). This confirms that the synergistic integration of both channel and spatial attention mechanisms is crucial for capturing the complex features of PNI.

Impact of the Number of DABs We investigated the optimal depth of the attention mechanism by varying the number of stacked DABs from 0 to 3. The configuration with 0 blocks (Base LDM Encoder) yielded the lowest AUC of 0.7001. Adding a single DAB significantly improved performance to 0.7589. The performance peaked with two DABs (AUC 0.7903). Increasing the depth further to three DABs resulted in a decrease in performance (AUC 0.7713), suggesting that two blocks provide the optimal balance of feature refinement for this task and dataset size.

Limitations Despite the promising results, this study has limitations. The PNI prediction cohort is relatively small and sourced from a single institution, which may limit generalizability. External validation using multi-center datasets is required. Furthermore, the current framework relies on single-phase (hepatobiliary) MRI. Integrating multi-phase MRI data could provide complementary information.

Conclusion

We proposed NeoNet, an end-to-end 3D MRI-based deep learning framework for non-invasive prediction of perineural invasion (PNI). Our framework combines three modules: automated tumor segmentation (NeoSeg), conditioned synthetic generation using 3D latent diffusion models (NeoGen), and an optimized PNI prediction Network (PattenNet). This integration addresses the fundamental challenges of subtle PNI features and data scarcity inherent to PNI research. Our results demonstrate that anatomically constrained synthetic data augmentation enhances performance, and PattenNet achieves a superior AUC of 0.7903 for PNI prediction in cholangiocarcinoma. This work advances non-invasive diagnostic capabilities, enabling improved preoperative risk assessment and individualized treatment planning.

Acknowledgement

This work was supported by the Next Generation Semiconductor Convergence and Open Sharing System, by the Institute of Information & Communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Semiconductor Support Program to Nurture the Best Talents (IITP-2023-RS-2023-00256081), and by the High-Performance Computing Support Project, all funded by the Korea government (Ministry of Science and ICT).

References

Amit, M.; Na'ara, S.; and Gil, Z. 2016. The role of the nervous system in cancer: a review. *Nature Reviews Cancer*, 16(6): 393–403.

Bapat, A. A.; Hostetter, G.; Von Hoff, D. D.; and Han, H. 2011. Perineural invasion and an inflammatory microenvironment in pancreatic cancer. *Pancreas*, 40(5): 724–731.

Ben Melech Stan, S.; et al. 2023. 3D-LDM: A latent diffusion model for 3D medical image generation. *arXiv preprint arXiv:2305.00323*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

Gillies, R. J.; Kinahan, P. E.; and Hricak, H. 2016. Radiomics: images are more than pictures, they are data. *Radiology*, 278(2): 563–577.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 770–778.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances*

in Neural Information Processing Systems (NeurIPS), volume 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 6840–6851.

Hruban, R. H.; et al. 2022. *WHO classification of tumours: digestive system tumours*. IARC.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 4700–4708.

Huang, X.; Cheng, J.; Zhang, L.; Li, X.; Zheng, C.; Zhang, Y.; Xu, X.; and Li, M. 2021. Feasibility of magnetic resonance imaging-based radiomics features for preoperative prediction of extrahepatic cholangiocarcinoma stage. *European Journal of Cancer*, 155: 227–235.

Isensee, F.; Jaeger, P. F.; Kohl, S. A.; Petersen, J.; and Maier-Hein, K. H. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2): 203–211.

Kazeminia, S.; Baur, C.; Kuijper, A.; van Ginneken, B.; Navab, N.; Albarqouni, S.; and Mukhopadhyay, A. 2020. GANs for medical image analysis. *Artificial intelligence in medicine*, 109: 101938.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*.

Lambin, P.; Leijenaar, R. T.; Starmans, M. W.; Rios-Velazquez, E.; Nalbantov, G.; Carvalho, S.; et al. 2017. Radiomics: the bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology*, 14(12): 749–762.

Li, M.; Zhang, J.; Wu, N.; Huang, W.; Li, Z.; Gao, Y.; Wang, X.; and Li, X. 2020. MRI-based radiomics signature for the preoperative prediction of perineural invasion status in patients with rectal cancer. *Abdominal Radiology*, 45(9): 2669–2677.

Liebig, C.; Ayala, G.; Wilks, J. A.; Berger, D. H.; and Albo, D. 2009. Perineural invasion in cancer: a review of the literature. *Cancer*, 115(15): 3379–3391.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 10012–10022.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Myronenko, A. 2018. 3D MRI brain tumor segmentation using autoencoder regularization. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, 311–320.

Pinaya, W. H. L.; Tudosiu, P.-D.; Gray, R.; Rees, G.; Cardoso, M. J.; Ourselin, S.; and Barkhof, F. 2022. Brain imaging generation with latent diffusion models. In *International Workshop on Simulation and Synthesis in Medical Imaging (SASHIMI)*, 117–126. Springer.

Purohit, B. S.; D'Oria, M.; Zarea, A.; D'Mello, A.; Keraliya, A. R.; O'Sullivan, B.; and Mukherji, S. K. 2019. Imaging of perineural tumor spread in head and neck cancer. *Radiographics*, 39(1): 205–225.

Qi, Z.; Yuan, H.; Li, D.; et al. 2025. An MRI-based fusion model for preoperative prediction of perineural invasion status in patients with intrahepatic cholangiocarcinoma. *World Journal of Surgical Oncology*, 23(1): 1–12.

Qian, Y.; et al. 2024. Neural regulation of cholangiocarcinoma progression. *Nature Communications*, 15: 1234.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 234–241. Springer.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. *Proceedings of the International Conference on Machine Learning (ICML)*, 2256–2265.

Tan, M.; and Le, Q. V. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning (ICML)*, 6105–6114. PMLR.

Tang, Y.; Yang, D.; Li, W.; Roth, H. R.; Landman, B. A.; Xu, D.; Nath, V.; and Hatamizadeh, A. 2022. Self-supervised pre-training of Swin Transformers for 3D medical image analysis. 20730–20740.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems (NeurIPS)*, volume 30.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Yi, X.; Walia, E.; and Babyn, P. 2019. Generative adversarial network in medical imaging: A review. *Medical image analysis*, 58: 101552.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3836–3847.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 586–595.