Interdisciplinary Sciences: Computational Life Sciences
https://doi.org/10.1007/s12539-026-00824-9

ORIGINAL RESEARCH ARTICLE

Exploring Deep Learning Models for Small Histopathology Datasets: Segmentation and Classification of Glomerular Crescent Lesions with Ablation, Interpretability, and Calibration Analyses

Inayatul Haq[1,2,3] · Haomin Liang[1] · Zheng Gong[1,2] · Zehong Xia[1,2] · Wei Zhang[1] · Rashid Khan[1] · Faizan Ahmad[1] · Yan Kang[4] · Bingding Huang[1,2]

Received: 20 June 2025 / Revised: 29 January 2026 / Accepted: 4 February 2026

Abstract

Glomerular crescent lesions are critical indicators of severe kidney injury and are closely associated with disease progression. However, their automated identification remains challenging due to limited annotated data, class imbalance, and subtle morphological variations. This study proposes a comprehensive deep learning (DL) framework for segmentation and classification of glomerular crescent lesions in histopathology images, with emphasis on robustness under limited data conditions. The ISICDM2024 Challenge dataset is used for evaluation. For segmentation, several baseline models are first evaluated, including DeepLabV3, U-Net, Transformer-based U-Net, and a feature pyramid network (FPN) with a ResNet-34 backbone. Similarly, for classification, multiple baseline models are evaluated, including EfficientNetV2-B0, ResNet-50, DenseNet-121, hybrid CNNs, CTransPath, and RetCCL. Motivated by the strong performance of FPN with ResNet-34 and DenseNet-121, two customized models are developed, namely CrescentSegNet for segmentation and CrescentDenseNet for classification. Comprehensive ablation studies are conducted, and interpretability and reliability are assessed using Grad-CAM, saliency mapping, uncertainty estimation, calibration analysis, and *t*-SNE. Cross-dataset evaluation on SICAPv2 and BreaKHis 400× confirms strong generalization and robustness. The proposed framework achieves competitive performance while maintaining efficiency and interpretability.

✉ Yan Kang
kangyan@sztu.edu.cn

✉ Bingding Huang
huangbingding@sztu.edu.cn

[1] School of Artificial Intelligence, Shenzhen Technology University, Shenzhen 518118, China

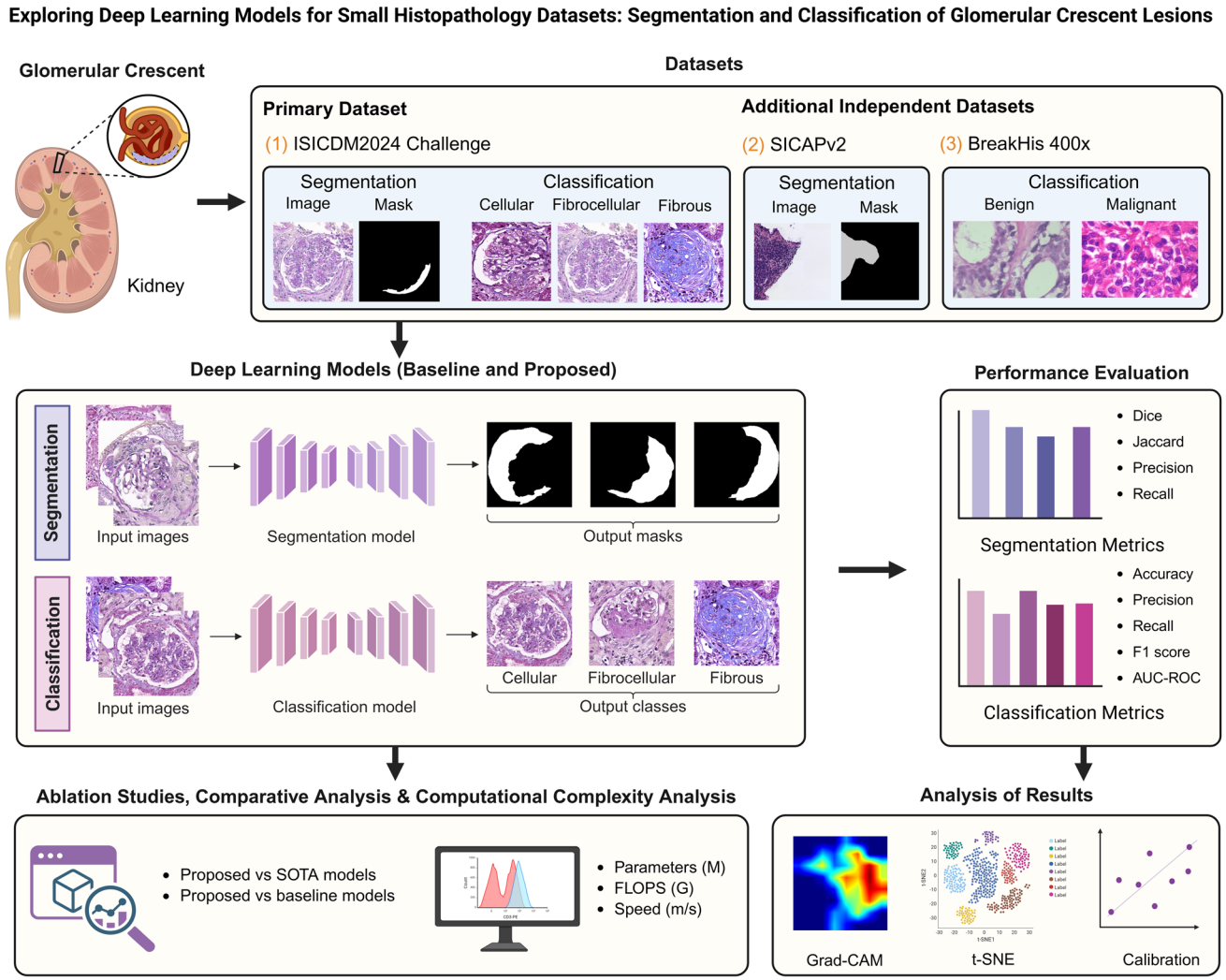
[2] College of Applied Sciences, Shenzhen University, Shenzhen 518060, China

[3] Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, National-Regional Key Technology, Engineering Laboratory for Medical Ultrasound, School of Biomedical Engineering, Shenzhen University Medical School, Shenzhen 518060, China

[4] College of Health Science and Environmental Engineering, Shenzhen Technology University, Shenzhen 518118, China

Published online: 23 March 2026

Graphical Abstract



Highlights

- A lightweight deep learning framework is proposed for crescent lesion segmentation and classification.
- CrescentSegNet and CrescentDenseNet are designed and optimized via systematic ablation studies.
- Interpretability, uncertainty, and calibration analyses ensure reliable and trustworthy predictions.
- Strong cross-dataset generalization is demonstrated on SICAPv2 and BreaKHis 400 × datasets.
- Lightweight models outperform heavier architectures on small, imbalanced histopathology data.

Keywords Glomerular crescent lesion · Histopathology image · Deep learning · Segmentation · Classification

1 Introduction

Glomerular crescent lesions form when epithelial cells and macrophages proliferate in Bowman's capsule and appear as cellular, fibrocellular, or fibrous crescents. While cellular and fibrocellular lesions may regress, fibrous crescents indicate chronic and irreversible damage [1]. Accurate recognition of these lesions is crucial for guiding treatment and prognosis. However, their subtle morphology makes manual detection challenging and prone to observer variability [2–5]. Although histopathology remains the diagnostic gold standard, variations in staining quality, resolution, and background noise complicate automated analysis [6, 7]. Deep learning (DL) has shown potential in medical image segmentation and classification. However, models trained

on small, imbalanced datasets are prone to overfitting, often focusing on dominant background regions and underperforming on rare findings such as crescents. Moreover, inconsistencies in laboratory protocols introduce further variability [6, 8, 9].

Researchers tackle small-data problems with several strategies. Data augmentation increases the size and diversity of the training dataset by generating additional samples through transformations such as rotation, flipping, scaling, and geometric distortion [10]. Transfer learning fine-tunes networks pretrained on large corpora, such as ImageNet, providing a head start in feature extraction [11]. Regularization methods such as dropout and batch normalization prevent overfitting [9]. Compact architectures, including MobileNet and EfficientNet, achieve an effective trade-off between model efficiency and accuracy by significantly reducing parameter counts and computational cost while maintaining high predictive performance [12–14]. Hybrid and attention models enable networks to focus on lesion regions and combine complementary features [15, 16]. Recent studies confirm the impact of these techniques. For example, CycleGAN generates synthetic tissue to expand limited datasets and supports a residual-inception U-Net with a post-processor intensity contour for better nuclei segmentation [17]. In addition, Bend-Net adds a boundary decoder and bending loss to refine overlapped nuclei [18], while a lean U-Net with Lovasz-Softmax loss segments colorectal slides and feeds a random forest biopsy classifier [19]. Other solutions include a weakly supervised CDWS-MIL with area constraints [20], a Bayesian VGG-U-Net that flags uncertain prostate regions [21], a hybrid-attention HanNet for dense nuclei [22], and multi-model ensembles that won the ACDC LungHP challenge despite using data from only 200 annotated slides [23]. Lightweight designs also shine: MobileNetV2 with CBAM and a dual fusion decoder is superior to DeepLabV3 + on the small CAMELYON16 dataset [24], Mu-Net trims U-Net yet keeps accuracy on brain organoids [25], a simple convolutional neural network (CNN) ensemble handles tissue subtyping on tiny whole-slide sets [26], and PATrans uses pixel-adaptive attention for cervical nuclei [27].

Small-sample classification improves via customized learning schedules and feature reuse. Gradual unfreezing with cosine annealing, differential rates, and cycle-length adaptation cuts dermatology training time and lifts accuracy [28]. Segmentation-based feature learning improves downstream classification performance, even with weakly labeled data [29]. Cosine-similarity loss outperforms cross-entropy on few-shot tasks [30]. Selecting data based on effect size and accuracy enables simple models to generalize effectively [31]. Naïve Bayes and AdaBoost tend to remain more stable than decision trees when training data is limited, as they are less prone to overfitting under small-sample conditions [32]. Ensemble models combining MobileNetV2, VGG16, and EfficientNet demonstrate improved performance for breast cancer classification on the Databiox dataset, which comprises histopathological images of invasive ductal carcinoma (IDC) [33]. Vision Transformers (ViTs) are trained on the tiny PH2 skin-lesion set via transfer learning from ISIC 2020 and staged fine-tuning with augmentation [34, 35].

Recent advances in computational pathology have introduced powerful AI frameworks such as CHIEF [36], HiCervix [37], and SAC-Net [38], which use large-scale datasets, weakly supervised learning, or hierarchical feature aggregation to address slide-level diagnostic and prognostic tasks. While these approaches have demonstrated strong performance across diverse pathological applications, they are generally designed under the assumption of abundant training data, large multi-institutional cohorts, and coarse-grained clinical labels. In contrast, glomerular crescent lesion analysis operates in a fundamentally different setting, as the available datasets are small, class-imbalanced, and highly lesion-specific. Moreover, the subtle morphological differences between crescent subtypes require fine-grained spatial and cellular-level reasoning for accurate discrimination.

Consequently, directly adopting large-scale or weakly supervised pathology frameworks may lead to overfitting or suboptimal generalization in this context. To address this challenge, our work adopts a lightweight, task-specific DL strategy explicitly customized to small histopathology datasets. Instead of relying on generic foundation models, we emphasize architectural parsimony, systematic ablation-driven design, and the explicit incorporation of interpretability and calibration analyses. This methodological framework is intended to support stable learning under data-limited conditions and to complement existing pathology AI approaches that primarily focus on large-scale or slide-level inference.

This study is motivated by the need for accurate, interpretable, and clinically reliable DL tools for renal pathology, as emphasized by the ISICDM2024 Challenge [39]. In particular, the challenge underscores the importance of automated segmentation and classification methods that support safer, more consistent assessment of crescent lesions in kidney biopsy images. To address this need, we focus on lightweight, task-specific models that emphasize interpretability and robustness under data-limited conditions. Building on this foundation, the present study aims to automate the segmentation and classification of glomerular crescent lesions in histopathology images. For segmentation, we evaluate representative architectures including DeepLabV3 with a ResNet-101 backbone,

U-Net, a Transformer-based U-Net (MiT-B0 encoder), and a customized ResNet-34 encoder with a feature pyramid network (FPN) based model named CrescentSegNet. To further examine generalization, CrescentSegNet is additionally evaluated on the SICAPv2 prostate cancer histopathology dataset. For classification, we compare baseline models such as CTransPath, RetCCL, EfficientNetV2-B0, ResNet-50, a hybrid CNN model, and a customized DenseNet-121-based model (CrescentDenseNet). To assess robustness beyond renal pathology, the proposed CrescentDenseNet model is further evaluated on the BreaKHis 400 × breast histopathology dataset. Across both tasks, standard data augmentation and regularization strategies are applied, and model performance is evaluated based on accuracy, robustness, and reliability. The complete workflow of the study is illustrated in Fig. 1.

2 Dataset

The primary focus of this research is on the ISICDM2024 Challenge dataset. To assess the robustness and cross-domain generalization of the proposed models, we also used the two public datasets SICAPv2 for segmentation and BreaKHis 400 × for classification.

2.1 ISICDM2024 challenge

This study employed a curated the task 8 dataset designed for the segmentation and classification of renal glomerular crescent lesions in histopathological images. The dataset was released as part of the ISICDM2024 Challenge [39] and contains two primary subsets: a segmentation dataset and a classification dataset. Each subset contains 588 high-resolution images (1360 × 1360 pixels), facilitating both

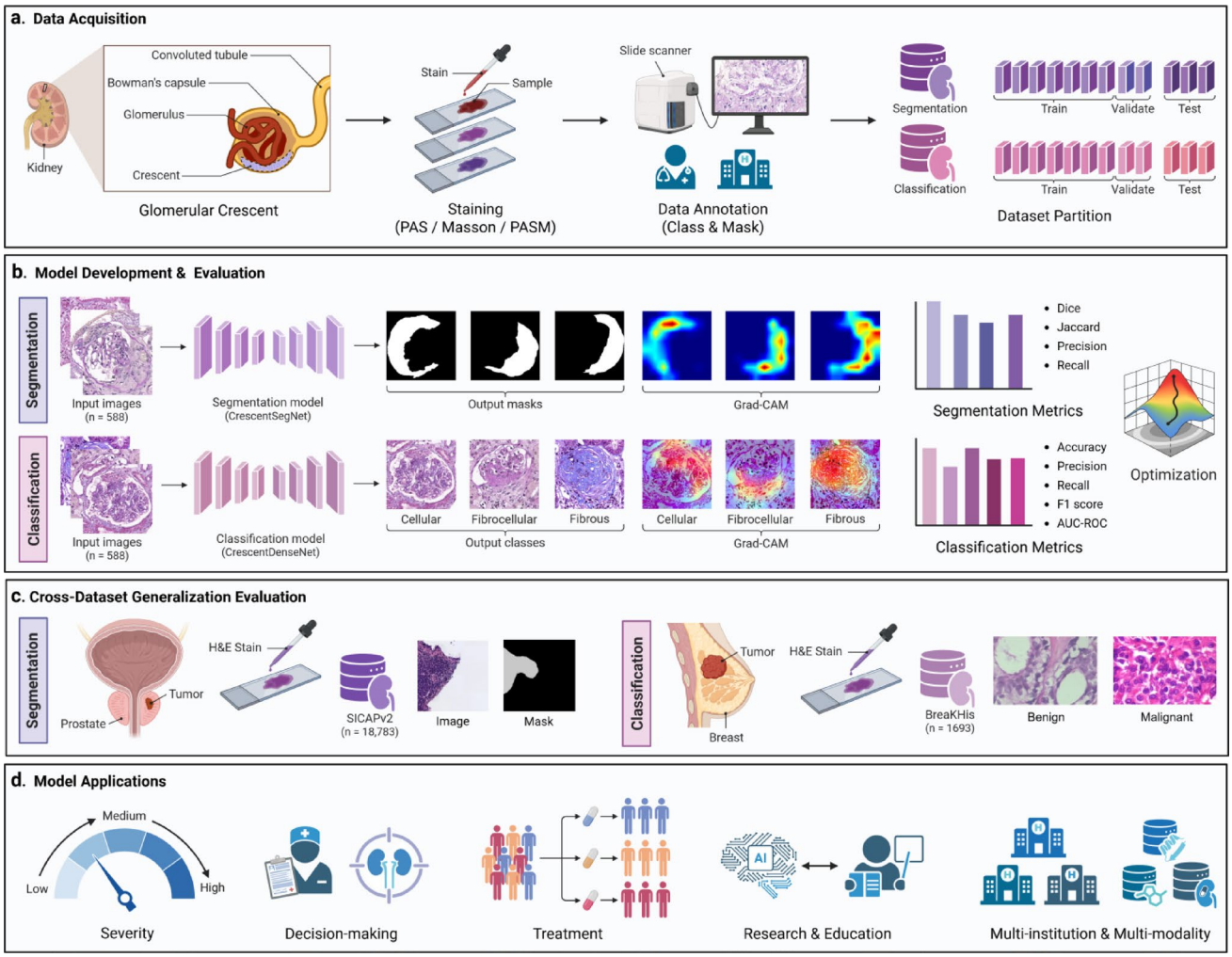


Fig. 1 Overview of the proposed workflow for glomerular crescent analysis. **a** Data acquisition and preparation, including staining, scanning, annotation, and dataset partitioning. **b** Model development and evaluation using CrescentSegNet for segmentation and CrescentDenseNet for classification, with Grad-CAM visualization and quantitative metrics. **c** Domain generalization evaluation on SICAPv2 and BreaKHis using the same training, validation, and testing protocol as ISICDM. **d** Clinical and research applications, including severity assessment and decision support. Created in BioRender. https://BioRender.com/5sjkqzl

Table 1 Overview of ISICDM2024 Histopathology task 8 dataset compositions, split, and annotation types for binary segmentation and multi-class classification tasks

Task	Count of images				Annotation type
	Total	Train	Validate	Test	
Segmentation	588	417	51	120	Crescent region masks
Classification	588	417	51	120	Labels: cellular, fibrocellular, and fibrous

fine-grained structural analysis and automated diagnostic modeling, as shown in Table 1.

2.1.1 Composition and Splits

The dataset is partitioned into training, validation, and test sets containing 417, 51, and 120 images, respectively, with no overlap between splits to prevent data leakage and ensure unbiased model evaluation. Expert pathologists precisely annotated the images to provide ground-truth masks and class labels, ensuring the clinical accuracy and reliability of downstream machine-learning tasks. For the classification subset, the training set contained 171 cellular, 191 fibrocellular, and 55 fibrous images; the validation set included 19 cellular, 21 fibrocellular, and 11 fibrous samples; and the test set comprised 58 cellular, 53 fibrocellular, and 9 fibrous images. No additional metadata regarding patient identity, staining batches, or scanning devices were released by the challenge organizers.

2.1.2 Image Characteristics

The dataset includes images stained using three standard histopathological techniques: periodic acid-schiff (PAS), Masson's trichrome (Masson), and periodic acid-silver methenamine (PASM). This staining diversity enhances the dataset's generalizability and supports the development of models robust to color and contrast variations.

2.1.3 Challenges and Relevance

Crescent lesions exhibit significant variability in morphology, size, and texture. Moreover, the subtle inter-class differences and blurred boundaries between lesion types pose significant challenges for automated analysis. Cellular and fibrocellular crescents are typically reversible and indicate treatable disease states, whereas fibrous crescents represent chronic, irreversible damage. Accurate identification of these lesion types is crucial for determining prognosis and planning treatment. The fibrous class has fewer samples than the cellular and fibrocellular classes. The dataset was constructed to address the current paucity of publicly available annotated glomerular crescent pathology images. It serves as a benchmark for the development of DL models designed to support clinical decision-making in nephropathology.

Table 2 Overview of the SICAPv2 prostate cancer histopathology dataset used for binary semantic segmentation

Task	Count of images				Annotation type
	Total	Train	Validate	Test	
Segmentation	18,783	15,026	1878	1879	Pixel-wise binary masks (cancer vs. background)

2.2 SICAPv2

Table 2 provides an overview of the SICAPv2 prostate cancer histopathology dataset used in this study for binary semantic segmentation of cancerous tissue regions. Additional details of the dataset are provided in the Supplementary Information (SI). The dataset comprises H&E-stained prostate biopsy image patches with corresponding pixel-wise binary masks. As the original Kaggle-hosted version provided a highly imbalanced split, the data were reorganized prior to model training by consolidating all image–mask pairs and repartitioning them into training, validation, and test sets using an 80%/10%/10% split, with strict preservation of image–mask correspondence and no overlap between subsets. This restructuring enables robust evaluation, generalization analysis, and independent testing. The dataset was accessed via a Kaggle-hosted mirror [40] for reproducibility and ease of use, while the original form is available from Mendeley data [41]. The VS Code console output of the automated data splitting procedure for the SICAPv2 dataset is shown in Figure S22 of SI.

2.3 BreaKHis 400×

Table 3 summarizes the composition, data splits, and annotation characteristics of the BreaKHis 400× breast cancer histopathology dataset [42] used for binary classification. A comprehensive description of the dataset, including its source, class distribution, and intended use, is provided in SI.

Table 3 Overview of the BreaKHis 400× breast cancer histopathology dataset used for binary classification

Task	Count of images				Annotation type	Magnification
	Total	Train	Validate	Test		
Classification	1693	1,184	339	170	Labels: Benign and Malignant	400×

3 Methods and Techniques

This study utilized a small dataset to enhance DL models for segmenting and classifying renal glomerular crescent lesions in histopathological images. The segmentation and classification approaches are discussed below.

3.1 Proposed Model for Crescent Lesions Segmentation

We explored several models for task segmentation, including DeepLabV3 with a ResNet-101 backbone, a Transformer-based U-Net (MiT-B0 encoder), U-Net, and FPN-ResNet-34. We applied key solutions, including data augmentation, transfer learning, architectural changes, and regularization techniques, to enhance model performance on small datasets. We also focused on lightweight architectures and model customization, including adjustments to output layers and the use of pre-trained weights. Early stopping and learning rate schedulers were employed to enhance convergence and prevent overfitting.

Customizations are applied to FPN with a ResNet-34 backbone to adapt it to the small dataset, and the modified model is named CrescentSegNet. Based on its best performance, CrescentSegNet was selected as the proposed model. This section describes the proposed model, while the Supplementary Information (SI) provides details on the architectures, customizations, fine-tuning strategies, hyperparameters, and settings of the baseline models mentioned above. In SI, Tables S1-S3 summarize the hyperparameters and settings for DeepLabV3, U-Net with MiT-B0, and U-Net, respectively.

3.1.1 Model Architecture, Customization, and Fine-tuning

The proposed segmentation framework, termed CrescentSegNet, is a task-specific architecture developed for the binary segmentation of glomerular crescent lesions in renal histopathology images. While it adopts an FPN with a ResNet-34 encoder as the backbone, CrescentSegNet introduces a custom-made multi-scale feature fusion and optimization strategy designed to address the unique morphological variability, irregular boundaries, and scale heterogeneity of crescentic lesions. By explicitly using pyramid-level feature aggregation and a lesion-aware training objective, the model enhances localization accuracy and boundary delineation compared to a vanilla FPN configuration. Figure 2 presents the complete architectural flow of the proposed CrescentSegNet framework.

3.1.2 Input and Encoder (ResNet-34 Backbone)

The model receives a 3-channel histopathology image of size 256×256 as input. It is first processed by a 7×7 convolutional layer and a max pooling operation, which reduces the spatial dimensions and helps extract low-level visual features. The encoder follows the ResNet-34 architecture, comprising four residual stages (blocks), each responsible for progressively capturing deeper contextual features while reducing spatial resolution. The output of each encoder stage can be mathematically expressed as

$$\boldsymbol{C}_l = f_l(\boldsymbol{I}), \qquad l = 1, 2, 3, 4 \tag{1}$$

where $\boldsymbol{C}_l$ denotes the feature map from the l-th residual stage, and $\boldsymbol{I}$ is the input image. These outputs serve as the basis for the subsequent top-down feature fusion.

3.1.3 Customized FPN Decoder and CrescentSegNet Development

While FPNs were initially introduced for generic object detection tasks, CrescentSegNet re-engineers the FPN decoder for binary medical image segmentation, customized explicitly to glomerular crescent localization. Unlike a vanilla FPN implementation, the proposed framework emphasizes dense spatial reconstruction and lesion-focused feature aggregation, rather than region-level detection. First,

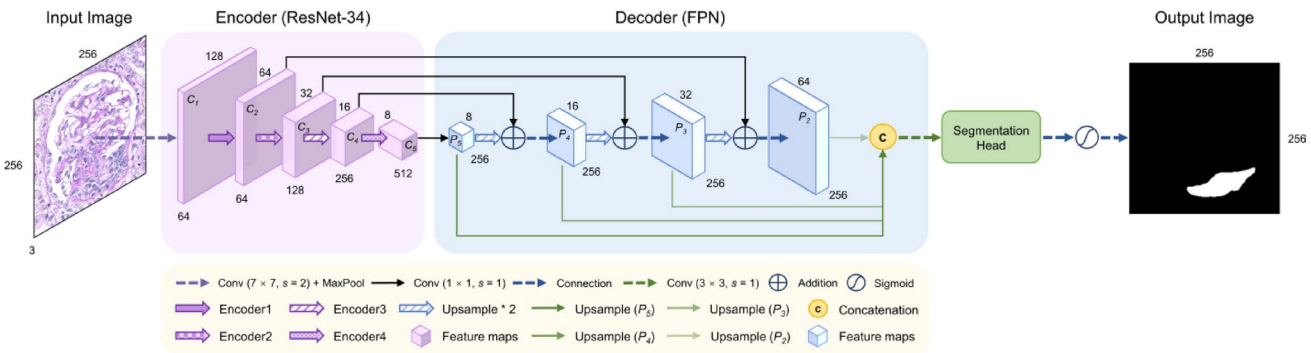


Fig. 2 The architecture of the CrescentSegNet model for glomerular crescent lesion segmentation. The network integrates multi-scale features from the encoder through a top-down FPN decoder and produces a binary segmentation mask via the segmentation head

1×1 lateral convolutions are applied to the encoder outputs to unify channel dimensions. A top-down pathway is then constructed, where higher-level semantic features are progressively upsampled and fused with lower-level spatial features through element-wise addition:

$$\boldsymbol{P}_l = \text{Upsample}(\boldsymbol{P}_{l+1}) + \text{Conv}_{1\times1}(\boldsymbol{C}_l) \tag{2}$$

The top layer of the pyramid is initialized as

$$\boldsymbol{P}_4 = \text{Conv}_{1\times1}(\boldsymbol{C}_4) \tag{3}$$

To further refine the feature maps and smooth the fused outputs, each pyramid level is passed through a 3×3 convolution:

$$\hat{\boldsymbol{P}}_l = \text{Conv}_{3\times3}(\boldsymbol{P}_l) \tag{4}$$

The refined multi-scale features $\widehat{\boldsymbol{P}}_{\boldsymbol{l}}$ are either summed or concatenated before being forwarded to the segmentation head.

3.1.4 Segmentation Head and Output

The fused multi-scale feature map is processed by the segmentation head, which consists of a convolutional layer followed by a sigmoid activation to produce the final binary segmentation mask:

$$\hat{\boldsymbol{Y}} = \sigma\left(\text{Conv}(\hat{\boldsymbol{P}})\right) \tag{5}$$

where $\widehat{\boldsymbol{P}}$ is the aggregated feature map and $\widehat{\boldsymbol{Y}}$ is the predicted binary mask highlighting the crescent lesion regions.

3.1.5 Loss Function

To ensure effective training and handle class imbalance, a hybrid loss function combining binary cross-entropy (BCE) and Dice loss is employed:

$$\mathcal{L}_{\text{total}} = \lambda\mathcal{L}_{\text{BCE}} + (1-\lambda)\mathcal{L}_{\text{Dice}}, \qquad \lambda = 0.5 \tag{6}$$

Here $\mathcal{L}_{\text{BCE}}$ denotes the BCE loss, which is computed as

$$\mathcal{L}_{\text{BCE}} = -[\boldsymbol{Y}\log(\hat{\boldsymbol{Y}}) + (\boldsymbol{1}-\boldsymbol{Y})\log(\boldsymbol{1}-\hat{\boldsymbol{Y}})] \tag{7}$$

Moreover, the Dice loss is computed as

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2|\boldsymbol{Y}\cap\hat{\boldsymbol{Y}}| + \epsilon}{|\boldsymbol{Y}| + |\hat{\boldsymbol{Y}}| + \epsilon} \tag{8}$$

Here $\boldsymbol{Y}$ and $\widehat{\boldsymbol{Y}}$ denote the ground-truth and predicted masks, respectively, and ϵ is a smoothing constant.

3.1.6 Optimization and Regularization

To improve convergence, a cosine annealing scheduler is used to adjust the learning rate dynamically:

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min})\left(1 + \cos\left(\frac{T_{\text{cur}}}{T_i}\pi\right)\right) \tag{9}$$

Here η_t represents the learning rate at iteration t, $\eta_{\max}, \eta_{\min}$ are the maximum and minimum learning rates, T_{cur} is the current iteration, and T_i is the total iterations per restart. A dropout rate of 0.2 is applied in the decoder to mitigate overfitting, which is especially important for small datasets.

3.1.7 Training, Validation, and Testing

The model was trained using the Adam optimizer, with both the learning rate and the weight decay of 1×10^{-4} to control overfitting. The model was trained for 500 epochs with a batch size of 8. Early stopping was applied with a patience of 20 epochs based on validation loss to prevent overfitting. This process halts training if the validation loss does not improve for a specified number of epochs. During training, the model's performance was assessed using four standard metrics, including Dice coefficient, Jaccard index, recall, and precision, which were monitored for both the training and validation datasets to ensure model generalizability. Table 4 presents the hyperparameters and settings for the CrescentSegNet model.

The trained CrescentSegNet was evaluated on unseen glomerular segmentation images. The preprocessed test samples were passed through the model to generate predicted masks, which were then binarized using a threshold of 0.5. Performance was evaluated using Dice coefficient, Jaccard index, recall, and precision metrics, and results were saved for each sample.

For the segmentation task, model performance was assessed using the four standard metrics indicated above. Higher values for these metrics indicate improved segmentation accuracy. The mathematical definition of each metric is shown below:

$$\text{Dice} = \frac{2\times|A\cap B|}{|A| + |B|} \tag{10}$$

$$\text{Jaccard} = \frac{|A\cap B|}{|A\cup B|} \tag{11}$$

Table 4 Hyperparameters and settings of the CrescentSegNet model

Parameter	Setting	Parameter	Setting
Encoder	ResNet-34	Input size	256 × 256
Decoder	FPN	Horizontal flip	0.5
Epochs	500	Vertical flip	0.5
Batch size	8	Rotation	±30°
Learning rate	1×10^{-4}	Normalization (mean, std)	(0.485, 0.456, 0.406), (0.229, 0.224, 0.225)
Optimizer	Adam	Tensor conversion	Yes
Learning rate scheduler	CosineAnnealing	BCE loss weight	0.5
T_{max} (scheduler)	10	Dice loss weight	0.5
Loss function (binary cross-entropy)	BCEWithLogitsLoss	Weight decay	1×10^{-4}
Loss function (Dice)	BCE and Dice	Data augmentation	Flip, rotation, elastic deformation
Input channels	3	Evaluation metrics	Dice, Jaccard, precision, recall
Output classes	1	Patience	20

$$\text{Precision} = \frac{n_{\text{TP}}}{n_{\text{TP}} + n_{\text{FP}}} \tag{12}$$

$$\text{Recall} = \frac{n_{\text{TP}}}{n_{\text{TP}} + n_{\text{FN}}} \tag{13}$$

In Eqs. (10)–(13), let A denote the set of ground-truth crescent pixels and B the set of pixels predicted as crescent by the model. Then $|A \cap B|$ corresponds to the number of true positives n_{TP}, $|A| = n_{\text{TP}} + n_{\text{FN}}$, $|B| = n_{\text{TP}} + n_{\text{FP}}$, and $|A \cup B| = n_{\text{TP}} + n_{\text{FP}} + n_{\text{FN}}$. Here, n_{TP}, n_{FP}, n_{FN}, and n_{TN} denote the numbers of true positives, false positives, false negatives, and true negatives, respectively.

3.2 Proposed Model for Crescent Lesion Classification

Initially, we employed some baseline models such as EfficientNetV2-B0, ResNet-50, DenseNet-121, a hybrid CNN (EfficientNetV2-B0 + ResNet-50), CTransPath, and RetCCL (ResNet-50). Each model was fine-tuned to optimize performance on the task, with modifications to its architecture and training settings. DenseNet-121 demonstrated the best performance among these models and is therefore considered the proposed model. We further customized DenseNet-121 for our task of enhancing classification accuracy on a small histopathology dataset and named it the CrescentDenseNet model. Furthermore, we conducted five ablation studies on CrescentDenseNet to optimize its performance further. This section focuses on the final (fifth) ablation of CrescentDenseNet; the details of the initial four ablations for CrescentDenseNet and the baseline models are in SI. The hyperparameters and training settings of all baseline models are provided in the SI. Table S4 summarizes the configurations for EfficientNetV2-B0, ResNet-50, and the hybrid CNN model. Tables S5 and S6 present the settings for CTransPath and RetCCL, respectively. Table S7 reports the configuration used for DenseNet-121. In addition, Table S8 details the hyperparameters employed in ablation studies 1–4 for the CrescentDenseNet model.

3.2.1 Model Architecture, Customization, and Fine-tuning

The CrescentDenseNet architecture was customized and optimized for classifying glomerular crescent lesions into three histological categories: cellular, fibrocellular, and fibrous. CrescentDenseNet, a densely connected CNN, promotes efficient feature reuse and mitigates vanishing gradient issues, making it particularly suitable for small-scale medical imaging datasets. The model processes 3-channel histopathology images resized to 224 × 224. These inputs pass through an initial 7 × 7 convolutional layer with a stride of 2, followed by a 3 × 3 max pooling layer to capture low-level features and reduce spatial resolution. The core of the CrescentDenseNet model consists of four dense blocks with 6, 12, 24, and 16 convolutional layers, respectively. Transition layers between these blocks perform 1 × 1 convolutions and 2 × 2 average pooling to reduce feature map dimensions and computational complexity.

A key innovation of the CrescentDenseNet model is its dense connectivity, in which each layer receives input from all preceding layers. This facilitates enhanced feature propagation and learning efficiency, particularly in scenarios with limited data. Mathematically, the input to the l-th layer is defined as

$$\boldsymbol{x}_l = H_l\left([\boldsymbol{x}_0, \boldsymbol{x}_1, ..., \boldsymbol{x}_{l-1}]\right) \tag{14}$$

Here $\boldsymbol{x}_l$ is the output of the l-th layer, $H_l(\cdot)$ is a composite function of Batch Normalization (BN), ReLU activation, and convolution, and $[\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{l-1}]$ represents the concatenation of feature maps from all previous layers.

To adapt the model for crescent lesion classification, several customizations were implemented. The original classifier layer was replaced with a fully connected (FC) layer

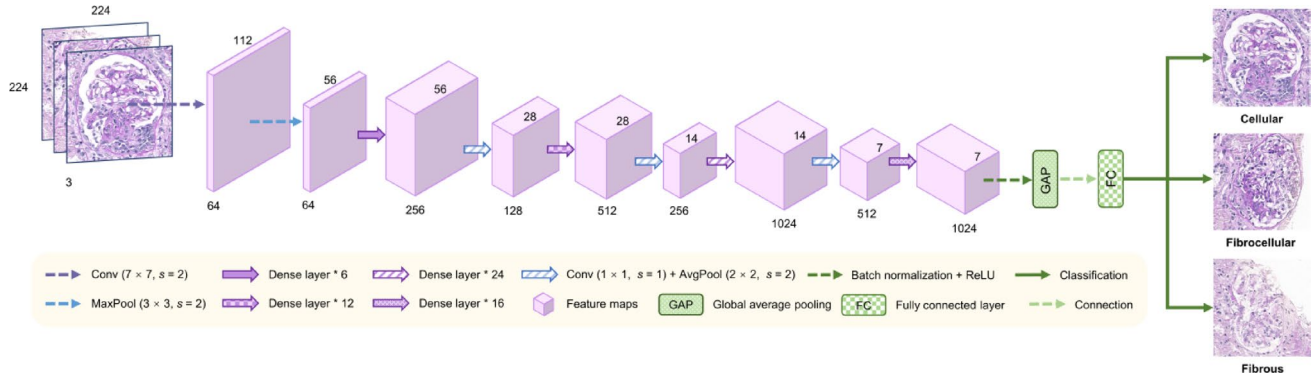


Fig. 3 The architecture of the CrescentDenseNet model was used for classifying glomerular crescent lesions. The network processes 224 × 224 input histopathology images through densely connected convolutional layers, followed by global average pooling and a fully connected layer to classify lesions into cellular, fibrocellular, or fibrous types

with the same number of output classes. A dropout layer with a rate of 0.4 was inserted before the classifier to prevent overfitting. To address the class imbalance, weights were calculated based on the class frequency and incorporated into a weighted cross-entropy (CE) loss function:

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{c=1}^{C} w_c y_c \log(\hat{y}_c) \tag{15}$$

Here C is the number of classes, y_c is the ground truth label, $\widehat{y}_c$ is the predicted probability for class c, and w_c is the weight assigned to class c.

The model was trained using the Adam optimizer with both the learning rate and the weight decay of 1×10^{-4} to enhance convergence and reduce overfitting. In some experimental variants, early layers of the DenseNet-121 backbone were frozen during training to preserve pretrained features and prevent degradation of general representations. A comprehensive data augmentation strategy was also applied, including random horizontal flipping, 20° rotations, resized cropping, and color jittering (brightness, contrast, saturation, hue) to improve robustness and generalization. A global average pooling (GAP) layer compresses the spatial dimensions of the final feature maps before feeding them into the classifier. The final FC layer outputs three class probabilities, and the class with the highest score is selected as the predicted crescent lesion type. To assess the contribution of each modification, an ablation study was conducted to examine the effects of dropout regularization, class weighting, layer freezing, and data augmentation. The results demonstrated that CrescentDenseNet achieved consistently high and balanced classification performance on the small, imbalanced dataset, confirming the effectiveness of the applied enhancements.

Table 5 Hyperparameters and settings of CrescentDenseNet in the final ablation

Parameter	Value
Image size	224 × 224
Batch size	16
Learning rate	1×10^{-4}
Optimizer	Adam
Weight decay (L2 Reg.)	1×10^{-4}
Epochs	200
Dropout rate	0.4
Class weights	Yes
Data augmentation	Random resized crop, random horizontal flip, random rotation, color jitter (brightness, contrast, saturation, hue)
Loss function	CrossEntropyLoss (with class weights)
Evaluation metrics	Accuracy, precision, recall, F1-score

Figure 3 illustrates the architecture of the CrescentDenseNet model, and Table 5 summarizes the key hyperparameters and final training settings.

3.2.2 Training, Validation, and Testing

The CrescentDenseNet model was trained for 200 epochs, with performance monitored on the training and validation sets using standard evaluation metrics. During training, the model achieving the lowest validation loss was saved to ensure optimal generalization. All training statistics, including loss values and performance metrics, were recorded and stored in structured CSV files for further analysis.

For evaluation, the trained model was tested on an independent test set using the same preprocessing and normalization procedures applied during training. Performance was assessed using confusion matrices to analyze class-wise behavior and prediction reliability. Classification performance was quantified using four metrics: accuracy, precision, recall, and F1-score. These metrics were consistently applied across the training, validation, and test phases to

ensure fair and comparable evaluation. Precision and recall are defined in Eqs. (12) and (13), while the formulas for accuracy and F1-score are provided below.

$$\text{Accuracy} = \frac{n_{\text{TP}} + n_{\text{TN}}}{n_{\text{TP}} + n_{\text{TN}} + n_{\text{FP}} + n_{\text{FN}}} \tag{16}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

Here, n_{TP} denotes the number of samples correctly classified as a specific crescent type, n_{FP} denotes the number of samples incorrectly labeled as that crescent type, n_{FN} denotes the number of samples that truly belong to a given crescent class but were misclassified as another type, and n_{TN} denotes the number of samples correctly identified as not belonging to that class.

3.3 Model Interpretability and Evaluation Techniques

To evaluate interpretability, reliability, and generalization, we employed a comprehensive set of post-hoc analysis techniques for both segmentation and classification models. These techniques collectively provided a detailed understanding of each model's decision-making behavior, predictive confidence, and complete robustness in the crescent lesion analysis.

For the segmentation task, the CrescentSegNet model was evaluated using pixel-level explainability methods, including Grad-CAM, Score-CAM, Integrated Gradients, and LIME, which highlighted the spatial regions most strongly influencing lesion localization. To quantify predictive uncertainty, Monte Carlo Dropout was applied to generate entropy- and variance-based uncertainty maps. Model calibration was examined using reliability diagrams and confidence histograms to assess the alignment between predicted probabilities and actual segmentation accuracy. Furthermore, *t*-SNE was used to project high-dimensional encoder features into a 2D space, with Dice-based color coding to visualize class separability. Performance metrics such as Dice coefficient, Jaccard index, recall, and precision were further analyzed through radar plots, distribution histograms, and sample-wise comparisons to evaluate consistency across the dataset.

For the classification task, CrescentDenseNet was evaluated through fivefold cross-validation using class-balanced loss to mitigate data imbalance. Data augmentation techniques, including flipping, rotation, brightness adjustment, and cropping, were employed to enhance model generalization. Interpretability was achieved using Grad-CAM and saliency maps, allowing visualization of class-discriminative regions. Prediction confidence was calibrated through reliability plots, and latent feature distribution was examined using *t*-SNE to visualize inter-class separation in the embedding space.

4 Results and Discussions

This section presents the results for the proposed models with the best performance on segmentation and classification tasks.

4.1 Segmentation Results

Table 6 presents the results of the ablation study conducted on the ISICDM2024 dataset, comparing the performance of different DL architectures and their variants for glomerular crescent segmentation. The table reports average training and validation metrics, including loss, Dice coefficient, Jaccard index, recall, and precision. The results demonstrate progressive performance improvements across ablation stages, with the proposed CrescentSegNet achieving the highest overall accuracy and segmentation quality among all evaluated models. Additional details on these approaches are provided in SI. The performance evaluation plots for the baseline segmentation models are provided in SI, including DeepLabV3 (Figure S1), Transformer-based U-Net (MiT-B0 encoder) (Figures S2–S4), and U-Net (Figures S5–S7).

Table 6 Ablation study comparing the performance of different DL models for glomerular crescent segmentation on the ISICDM2024 dataset. The table reports average training and validation losses and segmentation performance metrics, including Dice coefficient, Jaccard index (*J*), recall (*R*), and precision (*P*), for the baseline, ablation variants, and the proposed CrescentSegNet model

Models and ablations (Ab)	Training metrics (avg)					Validation metrics (avg)				
	Loss	Dice	*J*	*R*	*P*	Loss	Dice	*J*	*R*	*P*
DeepLabV3 [39]	**0.03**	**0.98**	**0.95**	**0.98**	**0.98**	**0.15**	0.81	0.70	0.84	0.82
Transformer-based U-Net [43] Ab1	0.14	0.80	0.69	0.78	0.83	0.16	0.70	0.57	0.71	0.74
Transformer-based U-Net [43] Ab2	0.13	0.81	0.71	0.80	0.85	**0.15**	0.73	0.60	0.74	0.76
Transformer-based U-Net [43] Ab3	0.05	0.93	0.88	0.92	0.94	0.17	0.78	0.66	0.80	0.79
U-Net [44] Ab1	0.40	–	–	–	–	0.79	0.34	0.23	0.49	0.43
U-Net [44] Ab2	0.07	0.89	0.84	0.87	0.94	0.68	0.70	0.58	0.71	0.79
U-Net [44] Ab3	0.09	0.91	0.85	0.91	0.92	0.17	0.80	0.70	0.82	**0.85**
CrescentSegNet Proposed	0.60	0.92	0.86	0.93	0.92	0.20	**0.84**	**0.74**	**0.88**	0.83

Figure 4 illustrates the training and validation performance of the CrescentSegNet model on the ISICDM2024 dataset. The loss decreases steadily, while Dice coefficient, Jaccard index, recall, and precision consistently increase on both the training and validation sets, indicating robust optimization behavior and good generalization performance.

4.1.1 Comprehensive Evaluation of Test Segmentation Results

Figure 5 presents the average segmentation performance of the CrescentSegNet model on the test set of the ISICDM2024 dataset. Panel a shows a balanced radar shape, indicating consistent performance across all four metrics. Panel b confirms this result with high scores: Dice (0.864) and recall (0.913) indicate a strong overlap and sensitivity, while Jaccard (0.775) and precision (0.844) demonstrate solid specificity and reduced false positives. The results show that the model performs robustly, achieving reliable and accurate segmentation of glomerular structures.

Figure 6 summarizes the uncertainty and reliability analysis of the CrescentSegNet model on the ISICDM2024 dataset. The distribution of mean entropy values reflects prediction confidence across samples, while the Dice–entropy relationship shows a clear association between lower uncertainty and higher segmentation accuracy. The boxplot analysis further confirms improved Dice scores with increasing confidence levels. The *t*-SNE visualization illustrates meaningful feature separation aligned with segmentation quality. The reliability curve and probability histogram demonstrate well-calibrated predictions, indicating stable and consistent model behavior across the test set.

Figure 7 shows the distribution of segmentation performance metrics on the test set on the ISICDM2024 data using the CrescentSegNet model. Each subplot shows the dispersion and central tendency of Dice, Jaccard, recall, and precision scores across test samples, with red dashed lines indicating the mean values. These plots highlight strong overall performance, with high recall and Dice scores, and moderate variability in precision and Jaccard scores, driven by occasional outliers.

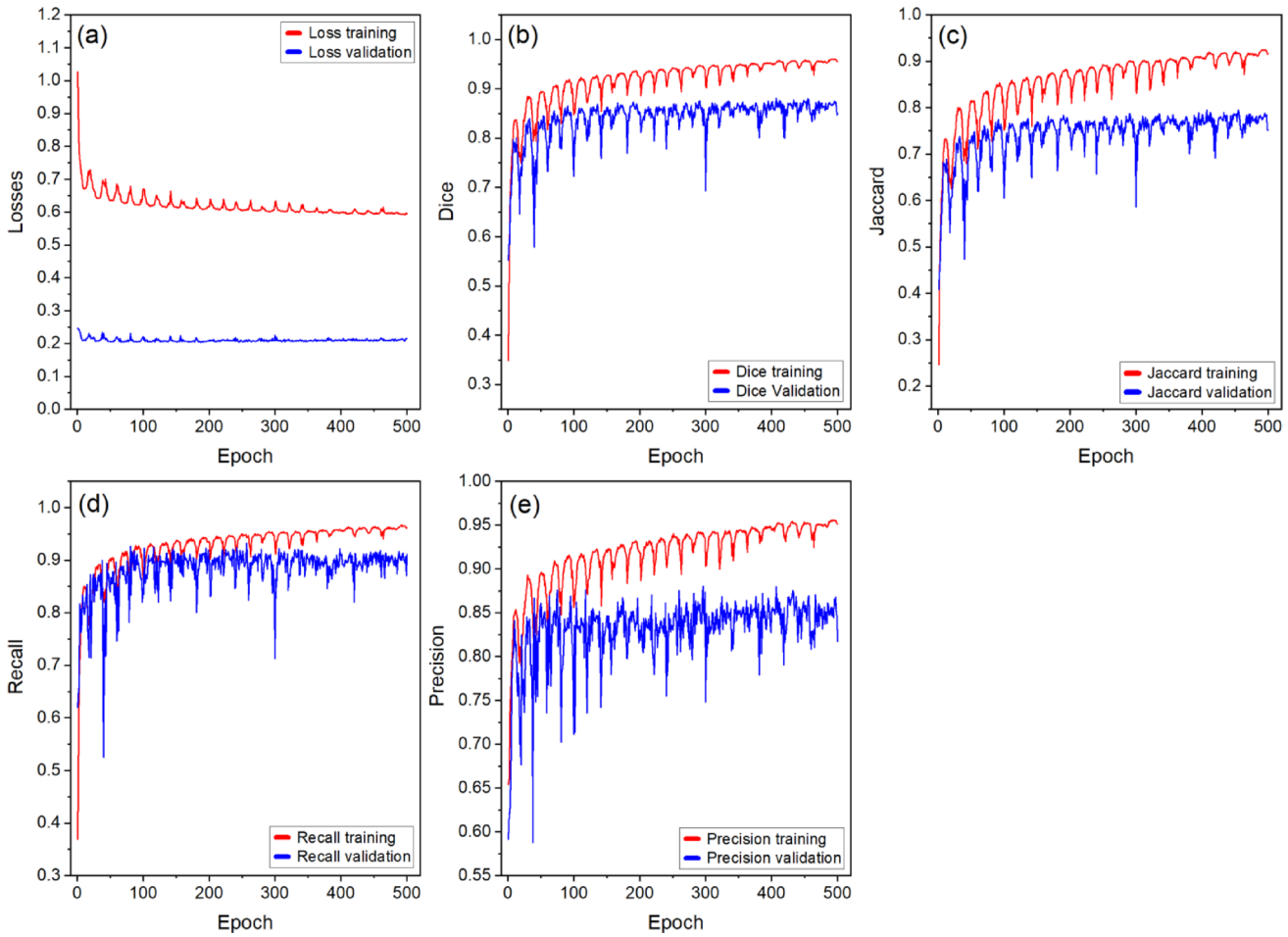


Fig. 4 Segmentation task training and validation results for the CrescentSegNet model on the ISICDM2024 dataset. **a** Loss, **b** Dice coefficient, **c** Jaccard index, **d** recall, and **e** precision

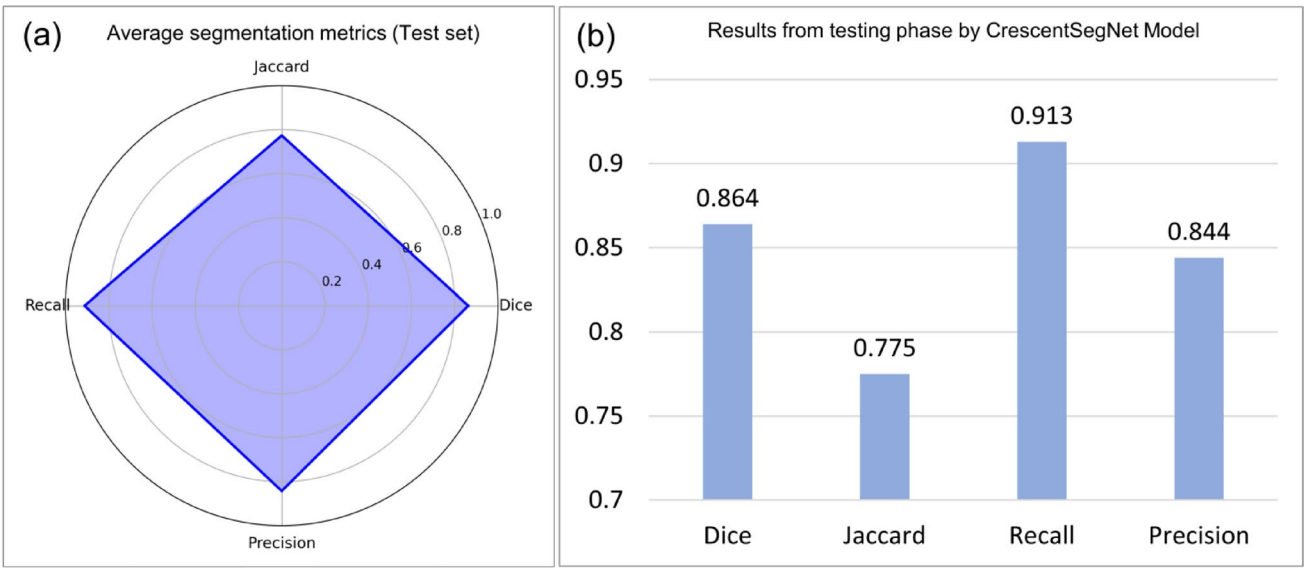


Fig. 5 Summary of CrescentSegNet segmentation performance in the test set of the ISICDM2024 dataset. **a** Radar plot illustrating average Dice, Jaccard, recall, and precision values. **b** Bar chart displaying the same metrics as (**a**) numerically for direct comparison

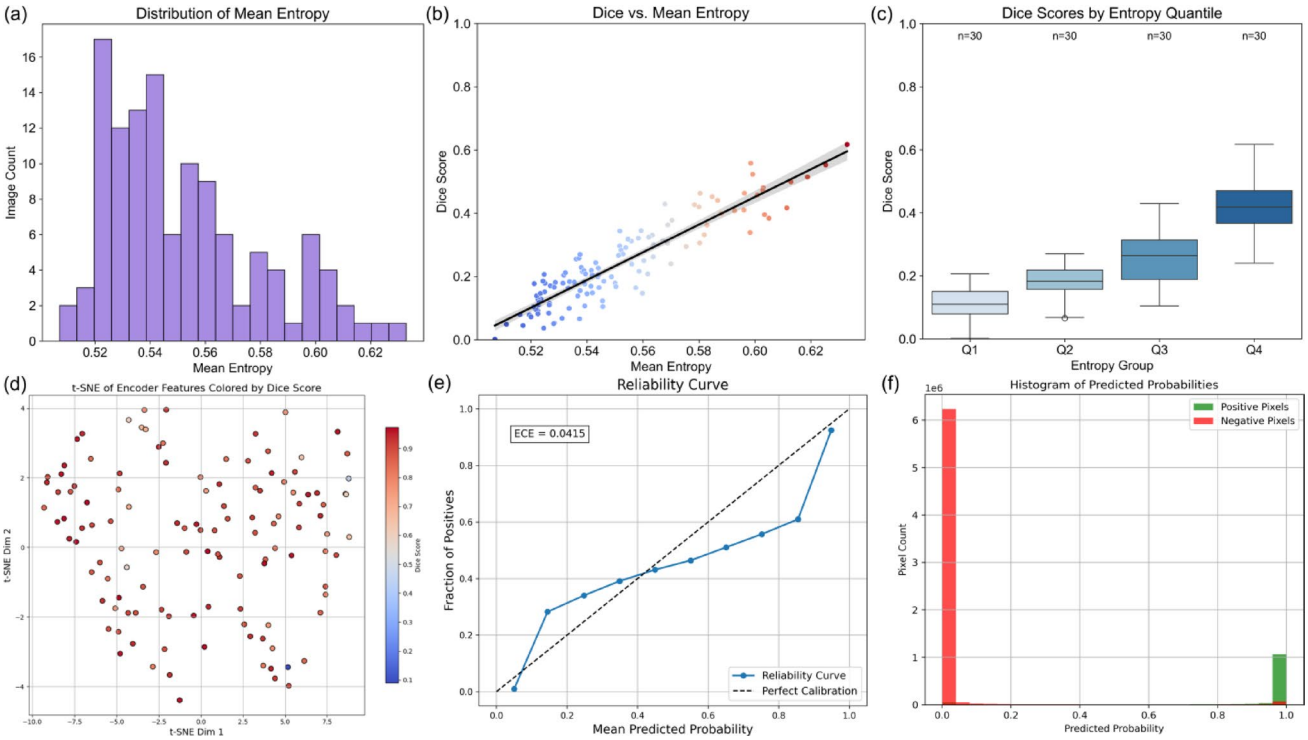


Fig. 6 Comprehensive evaluation of test-phase predictions by the CrescentSegNet model on the ISICDM2024 dataset. **a** Mean entropy distribution. **b** Dice vs. mean entropy. **c** Dice scores by entropy quantile. **d** *t*-SNE plot of encoder features color-coded by Dice score. **e** Reliability curve. **f** Histogram of predicted probabilities

Figure 8 illustrates both the absolute (scatter plots) and relative (ranked line plots) performance of the model across individual test images on the ISICDM2024 dataset. The Dice and recall scores show high consistency with means above 0.86 and 0.91, respectively, indicating strong overall segmentation quality and sensitivity. Jaccard and precision scores are slightly lower and more dispersed, reflecting stricter overlap criteria and occasional over-segmentation. Outliers, shown in red, highlight rare failure cases worth investigating. This comprehensive view confirms robust segmentation across most cases while pinpointing samples that need further analysis or refinement.

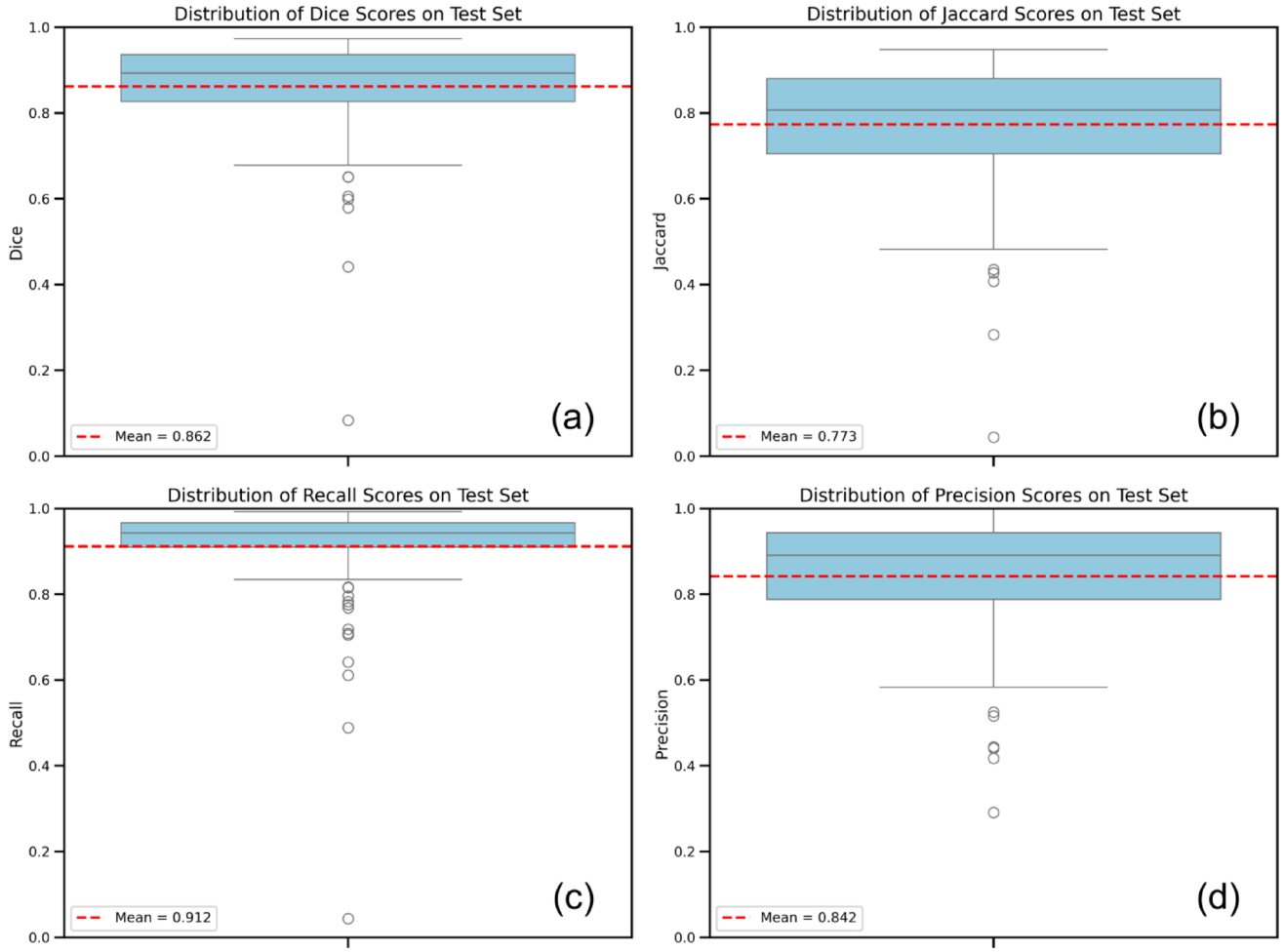


Fig. 7 Distribution of segmentation performance metrics on the ISICDM2024 test set using the CrescentSegNet model. Boxplots summarize the sample-wise distribution of Dice coefficient (**a**), Jaccard index (**b**), recall (**c**), and precision (**d**) across all test images. The red dashed line in each panel indicates the mean value of the corresponding metric, while boxes represent the interquartile range and whiskers denote variability across samples

Figure 9 presents six examples of explainability results from the test set of 120 images of the ISICDM2024 dataset. Column 1 displays the raw histopathological images, and Column 2 shows their corresponding ground truth masks. Column 3 contains predicted segmentation masks generated by the CrescentSegNet model. Column 4 provides an overlay comparison, where red highlights the ground truth and white shows the predicted regions. Columns 5–9 visualize explainability outputs from different XAI methods such as Grad-CAM, LIME, Score-CAM, integrated gradients, and uncertainty estimation, revealing model attention, feature importance, and prediction confidence for each case. Additional qualitative results on the ISICDM2024 test set are provided in SI. These include visual comparisons between predicted segmentation masks and ground-truth annotations in Figure S8. Figure S9 presents the corresponding mean prediction maps and entropy-based uncertainty maps for glomerular crescent segmentation, while Figure S10 illustrates sample-wise performance across test images.

4.2 Classification Results

Table 7 presents the validation performance of multiple baseline DL models for glomerular crescent classification. Additional descriptions of these models are provided in SI. Among them, DenseNet-121 demonstrated the highest overall effectiveness, achieving the best F1-score (0.42) and accuracy (0.57), making it the most suitable baseline for this task. EfficientNetV2-B0 offered moderate recall but underperformed on precision and F1-score, while ResNet-50 achieved higher precision but lacked overall balance. The hybrid CNN model yielded the poorest results across all metrics. Additional qualitative results of the baseline models are provided in SI. Figure S11 presents the

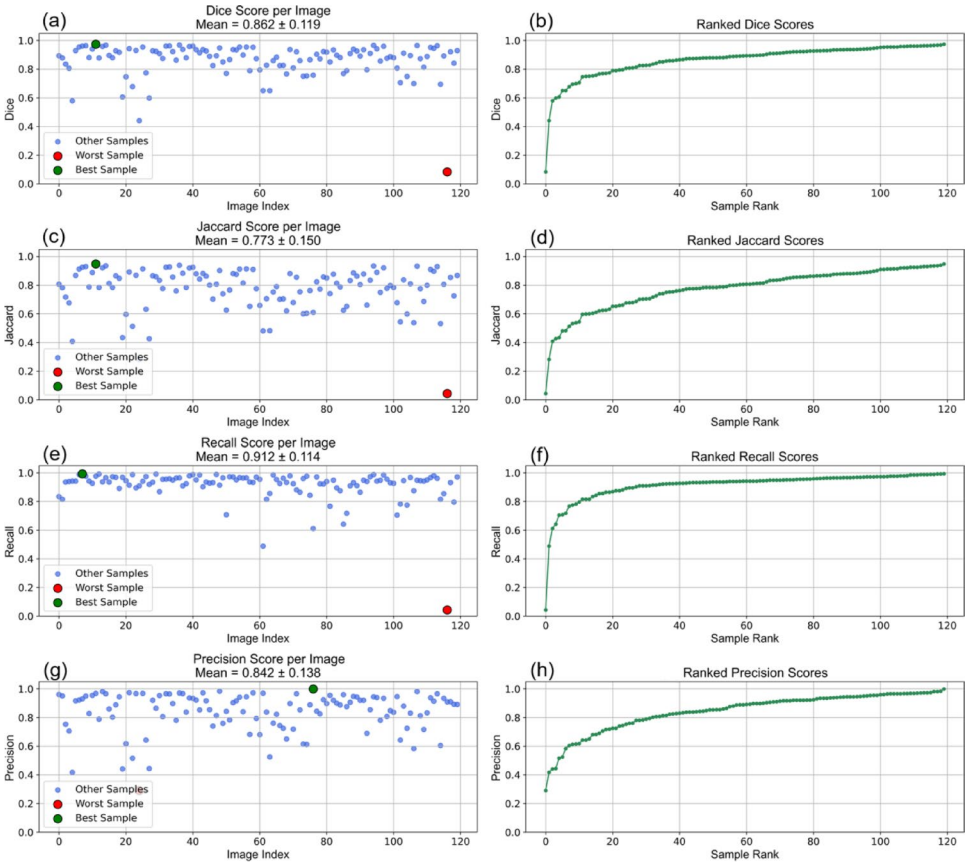


Fig. 8 Per-image and ranked distribution analysis of segmentation performance metrics for CrescentSegNet on the ISICDM2024 test set. Panels **a** and **b** show the per-image Dice scores and the corresponding ranked Dice distribution, respectively. Panels **c** and **d** present the per-image Jaccard scores and ranked Jaccard distribution. Panels **e** and **f** illustrate the per-image recall scores and ranked recall distribution, while panels **g** and **h** depict the per-image precision scores and ranked precision distribution

results of EfficientNetV2-B0, Figure S12 shows the results of ResNet-50, and Figure S13 illustrates the performance of the hybrid CNN model. Figure S14 displays the training and validation curves of CTransPath, Figure S15 shows the corresponding curves for RetCCL, and Figure S16 presents the classification results of the DenseNet-121 model. Table S9 presents class-wise performance metrics for the EfficientNetV2-B0, ResNet-50, hybrid CNN, and DenseNet-121 models.

Table 8 summarizes the ablation study results for the CrescentDenseNet model on the ISICDM2024 dataset. The table reports average training and validation performance across different ablation configurations using loss, accuracy, precision, recall, and F1-score. The results show consistent performance improvements across successive ablations, with the final configuration achieving the highest accuracy and F1-score. The findings demonstrate the effectiveness of the proposed design choices in improving model performance.

Complete details of the ablations used for the CrescentDenseNet model are included in the SI and are summarized as follows:

- Ablation 1: Customizations applied to vanilla DenseNet-121, including architectural adjustments and a modified classifier head, with standard input resizing to 224×224 and simple data augmentation (horizontal flip and color jitter). Training was performed using the Adam optimizer with a learning rate of 1×10^{-4}, a weight decay of 1×10^{-4}, batch size of 16, and 200 epochs. Class weighting was enabled to address class imbalance, while no dropout, learning-rate scheduler, or layer freezing was applied. This ablation serves as the baseline fine-tuned configuration.
- Ablation 2: Strong regularization through random resized cropping, horizontal flipping, random rotation, and color jitter, combined with freezing of early DenseNet layers. Input images were resized to 256×256 and trained with batch size of 16 using the Adam optimizer. A reduced learning rate of 5×10^{-5} and a weight decay of 5×10^{-5} were used, with training conducted for 150 epochs. A dropout rate of 0.5 was introduced at the classifier head, and a ReduceLROnPlateau learning rate scheduler was applied. Class weighting was enabled throughout training.
- Ablation 3: Moderate data augmentation, including random resized cropping, horizontal flipping, random rotation, and color jitter, with input images resized to 256×256. The model was trained using the Adam optimizer with a learning rate of 1×10^{-4}, a weight decay

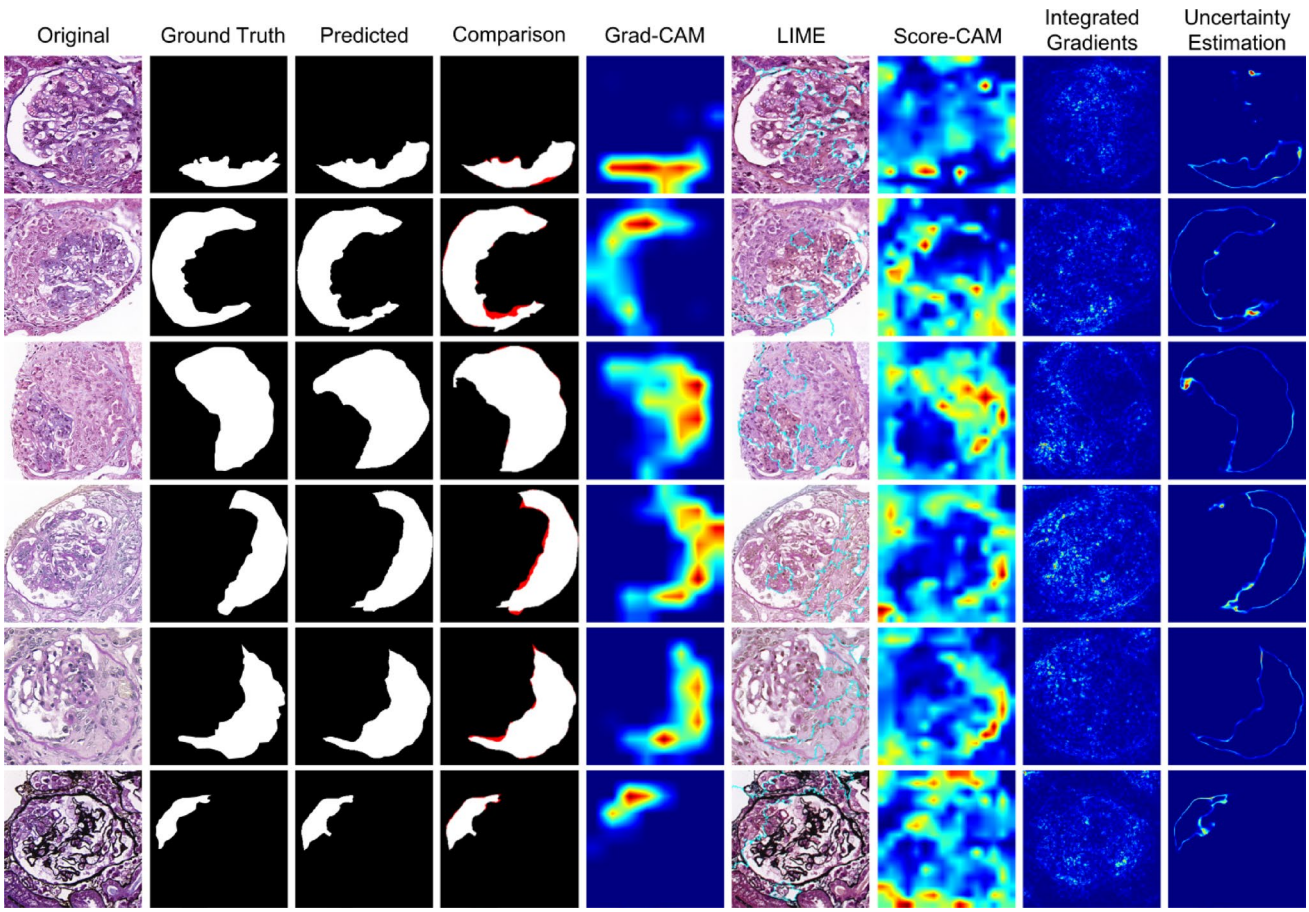


Fig. 9 CrescentSegNet segmentation and explainability results for six representative test images of the ISICDM2024 dataset. The first two columns show the raw images and ground truth masks, respectively; third column shows the predicted masks; and fourth column provides a comparison overlay with ground truth in red and predictions in white. The last five columns display XAI outputs from Grad-CAM, LIME, Score-CAM, integrated gradients, and uncertainty estimation analyses

Table 7 Performance of baseline DL models in terms of average precision, recall, F1-score (F1), and accuracy (Acc) on the validation set for glomerular crescent classification using the ISICDM2024 dataset

Model	Precision	Recall	F1	Acc
EfficientNetV2-B0 [14]	0.28	0.43	0.34	0.45
ResNet-50 [8]	0.47	0.36	0.25	0.43
DenseNet-121 [45]	0.40	0.47	0.42	0.57
Hybrid CNN	0.07	0.33	0.11	0.22
CTransPath [46]	0.46	0.46	0.42	0.46
RetCCL (ResNet-50) [47]	**0.63**	**0.58**	**0.58**	**0.61**

of 1×10^{-4}, batch size of 16, and 150 epochs. A dropout rate of 0.4 was applied in the classifier head, while no learning-rate scheduler or layer freezing was used. Class weighting remained enabled, allowing evaluation of moderate regularization without aggressive optimization constraints.

- Ablation 4: Minimal augmentation consisting of horizontal flipping and color jitter, with standard resizing to 224×224. Training was performed using the Adam optimizer with a learning rate of 1×10^{-4}, a weight decay

Table 8 Training and validation results of the ablation study for the CrescentDenseNet model on the ISICDM2024 dataset. The table reports average performance metrics, including loss, accuracy (Acc), precision (*P*), recall (*R*), and F1-score (F1), for each ablation configuration across training and validation phases

Ablation	Training metrics (avg)					Validation metrics (avg)				
	Loss	Acc	*P*	*R*	F1	Loss	Acc	*P*	*R*	F1
1	**0.07**	**0.99**	**1**	**0.99**	**0.99**	1.85	0.62	0.64	0.62	0.59
2	1.16	0.37	0	0.37	0.38	1.10	0.37	0.36	0.37	0.35
3	0.35	0.85	**1**	0.85	0.86	2.01	0.63	0.65	0.63	0.61
4	**0.07**	**0.99**	**1**	**0.99**	**0.99**	**0.25**	**0.92**	**0.92**	**0.92**	**0.92**
5	0.32	0.87	**1**	0.87	0.87	0.45	0.86	0.87	0.86	0.86

of 1×10^{-4}, batch size of 16, and 200 epochs. A dropout rate of 0.4 was applied to the classifier head, while no learning rate scheduler or layer freezing was used. Class weighting was enabled. This ablation assesses the effect of lighter augmentation and longer training duration.

- Ablation 5: This final ablation incorporates strong data augmentation, including random resized cropping, random rotations, and color jittering (brightness, contrast, saturation, and hue), together with dropout regularization (0.4), class-weighted cross-entropy loss, and L2 weight decay to mitigate overfitting and class imbalance. The model was trained using the Adam optimizer with a learning rate of 1×10^{-4}, batch size of 16, and 200 epochs of training. This configuration represents the fully optimized CrescentDenseNet and is adopted as the final proposed model.

The training and validation performance curves of the CrescentDenseNet model on the ISICDM2024 dataset for initial four ablation studies provided in SI (Figures S17–S21). Figure 10 illustrates the classification performance of the CrescentDenseNet model using Ablation 5 on the ISICDM2024 dataset. Panel a shows a steady decrease in training and validation losses, indicating effective learning. Panels b to e demonstrate strong and consistent performance across accuracy, precision, recall, and F1-score, with training metrics slightly higher than validation metrics, suggesting good generalization with minimal overfitting. However, compared to other ablation studies, overfitting has been significantly reduced in the final ablation. Panel f, a confusion matrix, confirms that most samples across all three crescent classes (cellular, fibrocellular, and fibrous) were correctly classified, with the cellular class showing the highest accuracy.

4.2.1 Comprehensive Evaluation of Test Classification Results

The fifth ablation of the CrescentDenseNet model was selected as the final proposed configuration. In this section, we analyze the test results obtained using this final model to assess its classification performance. Figure 11 presents

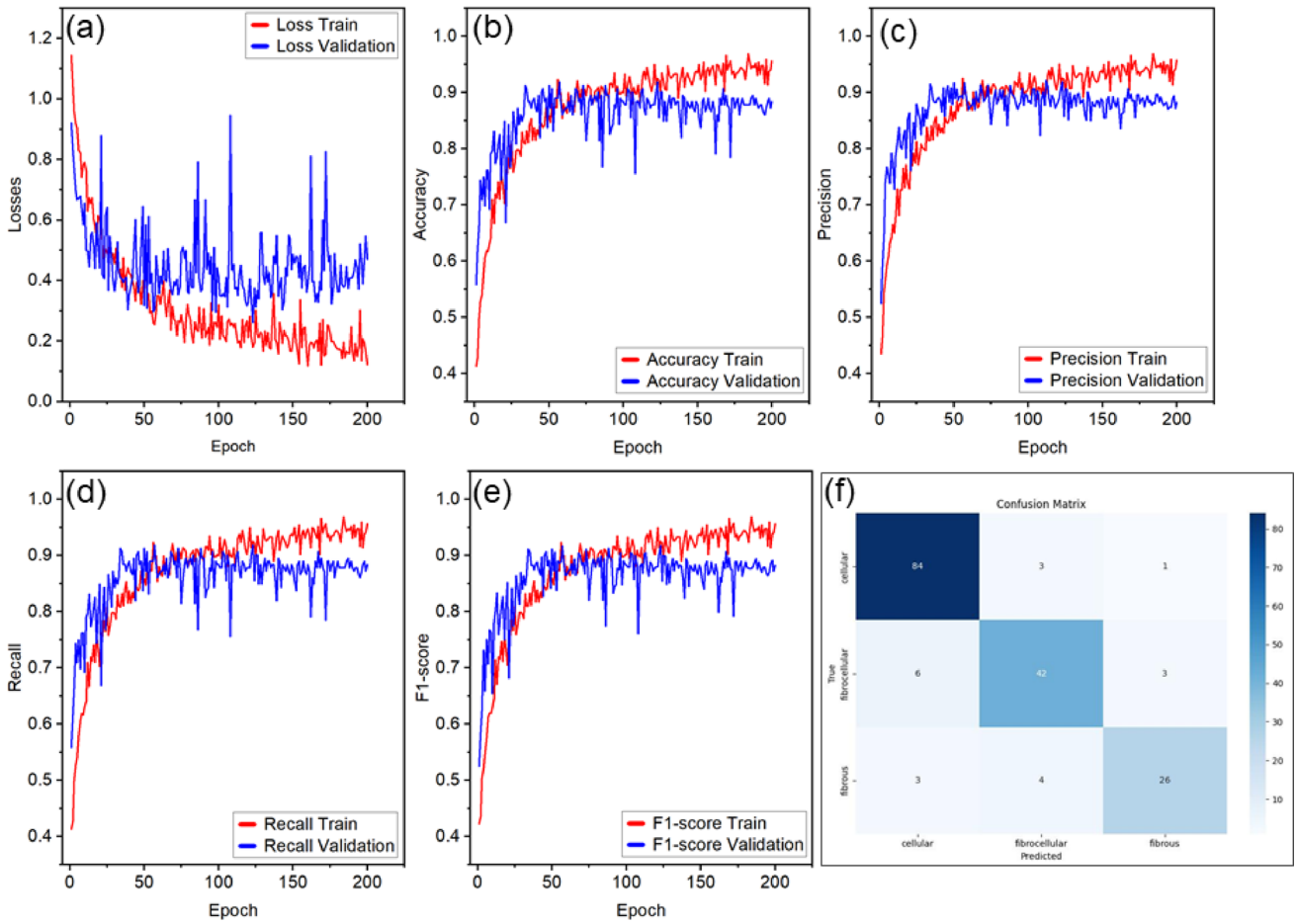


Fig. 10 CrescentDenseNet model Ablation 5 (final) classification results on the ISICDM2024 dataset. The red curves represent the training phase, and the blue curves represent the validation phase. **a** Losses. **b** Accuracy. **c** Precision. **d** Recall. **e** F1-score. **f** Confusion matrix (validation)

the test-set classification results of the CrescentDenseNet model on the ISICDM2024 dataset. Panel a shows that the model achieved the highest accuracy when identifying cellular crescents but was more likely to misclassify fibrocellular and fibrous lesions. Panel b displays the area under the receiver operating characteristic (AUC) curves, with the cellular class showing the best discriminative performance (AUC = 0.86), followed by fibrocellular (AUC = 0.77) and fibrous (AUC = 0.70), indicating that the model more confidently and accurately predicted cellular lesions.

Figure 12 illustrates the classification performance of the CrescentDenseNet model on the ISICDM2024 dataset. Panel a summarizes the overall performance, showing balanced accuracy, precision, recall, and F1-score across the test set. Panel b presents class-wise results, showing strong performance for the cellular class and moderate performance for the fibrocellular class. In contrast, the fibrous class exhibits lower performance, primarily due to the minimal number of available samples (only nine images), which limits the model's ability to learn robust, discriminative features. This data imbalance reduces classification reliability for this class.

Figure 13 presents the feature distribution and calibration behavior of the CrescentDenseNet model on the ISICDM2024 test set. Panel a shows the *t*-SNE visualization of learned feature embeddings, where samples from different classes form partially separable clusters, indicating effective feature discrimination while reflecting some overlap due to class similarity. Panel b illustrates the reliability diagram, demonstrating a close alignment between predicted probabilities and observed accuracies. This indicates good calibration of the model, with predictions closely matching true confidence levels across probability bins.

Figure 14 illustrates how Grad-CAM and saliency maps provide visual insight into the CrescentDenseNet model's decision-making process for crescent classification. The first column presents the original test images with ground-truth cellular, fibrocellular, and fibrous class labels. The second column displays Grad-CAM results, with red-highlighted regions indicating the areas most strongly influencing the model during classification. These heatmaps align well with crescent structures, suggesting that the model focused on relevant pathological features. The third column shows saliency maps that highlight pixel-level sensitivity to the model's predictions. The activation patterns in both XAI methods consistently highlight biologically meaningful areas, supporting the interpretability and reliability of the model's classification behavior.

4.3 Cross-Domain Generalization on Independent Histopathology Datasets

Beyond the ISICDM2024 glomerular crescent dataset, we further evaluated the proposed CrescentSegNet and CrescentDenseNet models on independent public histopathology datasets involving different pathological patterns. This analysis aims to assess the cross-domain generalization and transferability of the proposed architectures under domain shift rather than disease-specific validation.

4.3.1 CrescentSegNet on SICAPv2 Segmentation Dataset

To assess the robustness and cross-domain generalization of CrescentSegNet, the model was further evaluated on the SICAPv2 prostate cancer histopathology segmentation dataset. Table 9 shows that CrescentSegNet achieves consistently high segmentation performance on the SICAPv2 dataset, with closely aligned Dice coefficient, Jaccard index, precision, and recall scores across all phases. The lower validation loss compared to training loss indicates

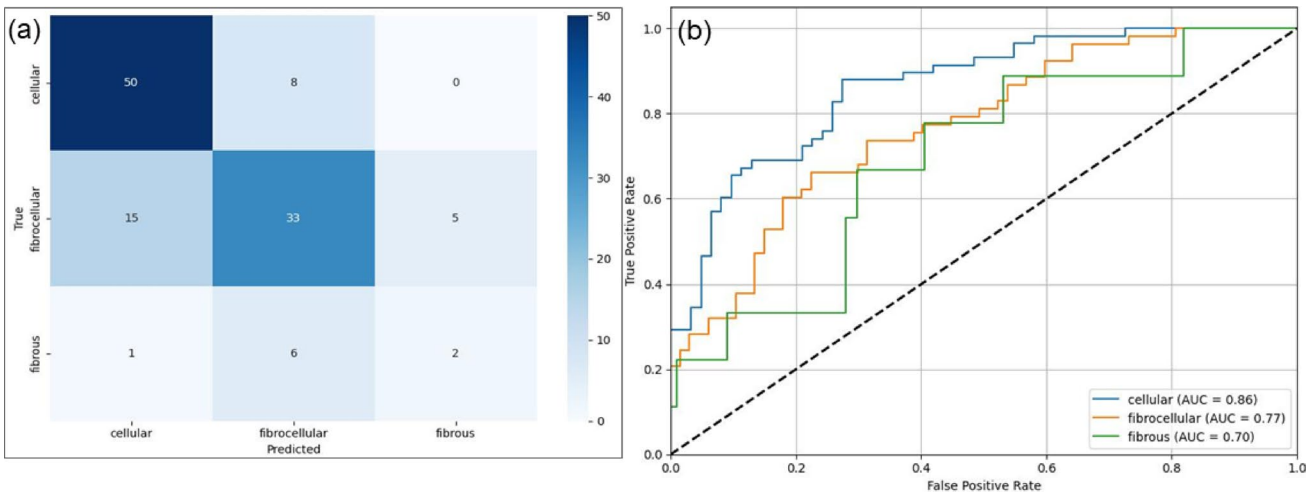


Fig. 11 Classification performance of the proposed CrescentDenseNet model on the test set on the ISICDM2024 dataset. **a** Confusion matrix. **b** AUC curves for each class

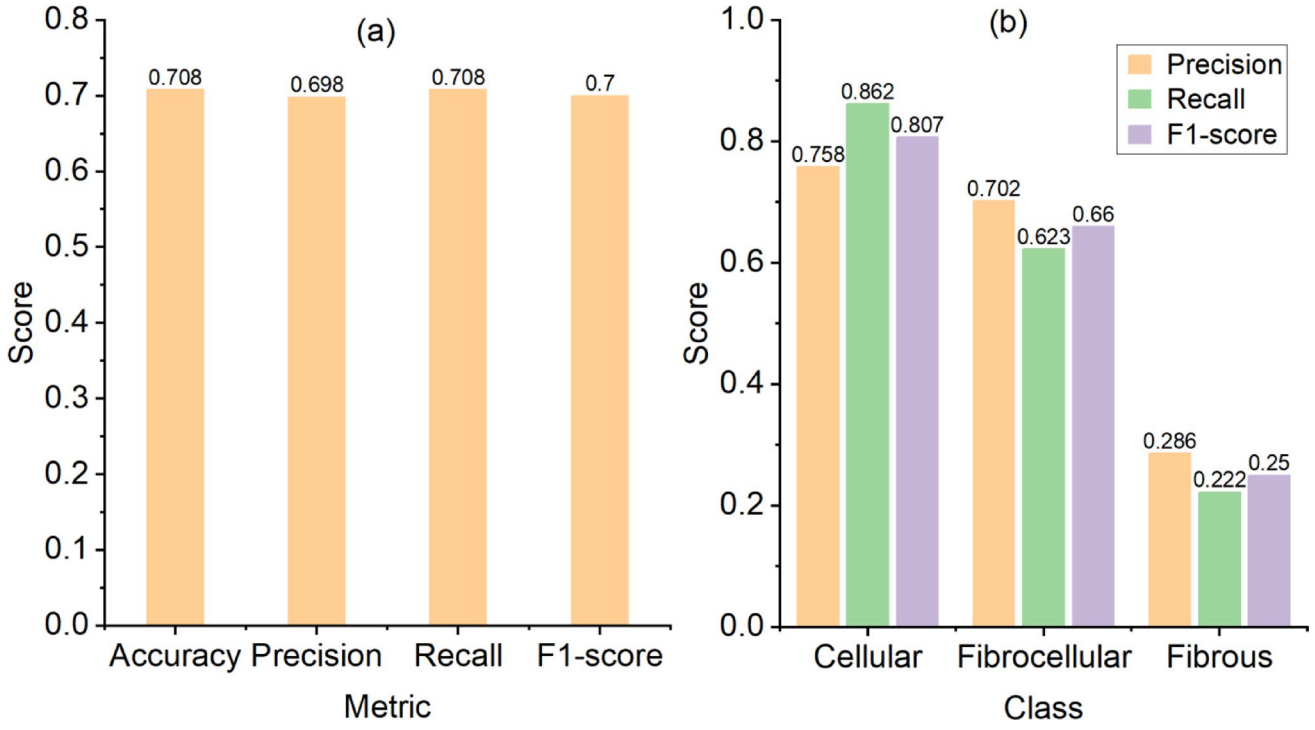


Fig. 12 Performance metrics for CrescentDenseNet classification during the testing phase on the ISICDM2024 dataset. **a** Overall model performance. **b** Class-specific performance metrics

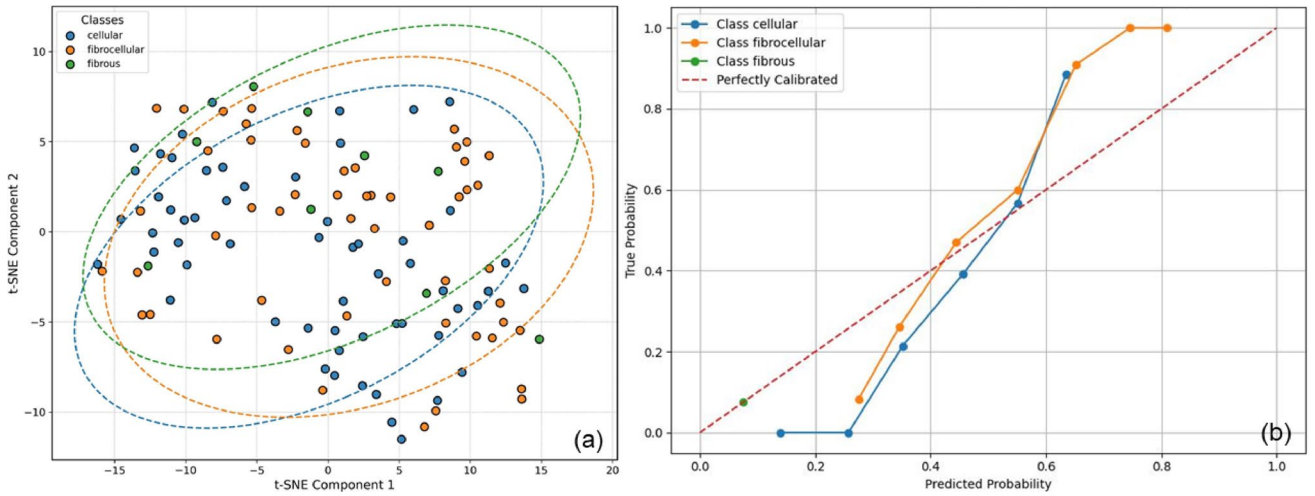


Fig. 13 Feature representation and confidence calibration of CrescentDenseNet on the ISICDM2024 test set. **a** *t*-SNE visualization illustrating the learned feature distributions across classes. **b** Reliability diagram assessing the calibration of predicted class probabilities during the testing phase

stable optimization and good generalization, supporting the robustness of the proposed model for prostate cancer histopathology segmentation. The training and validation curves of CrescentSegNet on the SICAPv2 segmentation dataset are presented in Figure S23 of SI.

Figure 15 provides complementary qualitative evidence of the robustness of CrescentSegNet on the SICAPv2 test set. Panel a presents a *t*-SNE projection of the learned deep feature embeddings, revealing a clear global separation between cancerous and background tissue samples, with limited overlap at class boundaries. This distribution indicates that the model captures discriminative representations while preserving intrinsic histopathological variability. Panel b shows the reliability (calibration) diagram, where the predicted confidence closely aligns with the diagonal reference line, demonstrating strong agreement between confidence estimates and observed accuracy. The low expected calibration error (ECE = 0.0065) further confirms

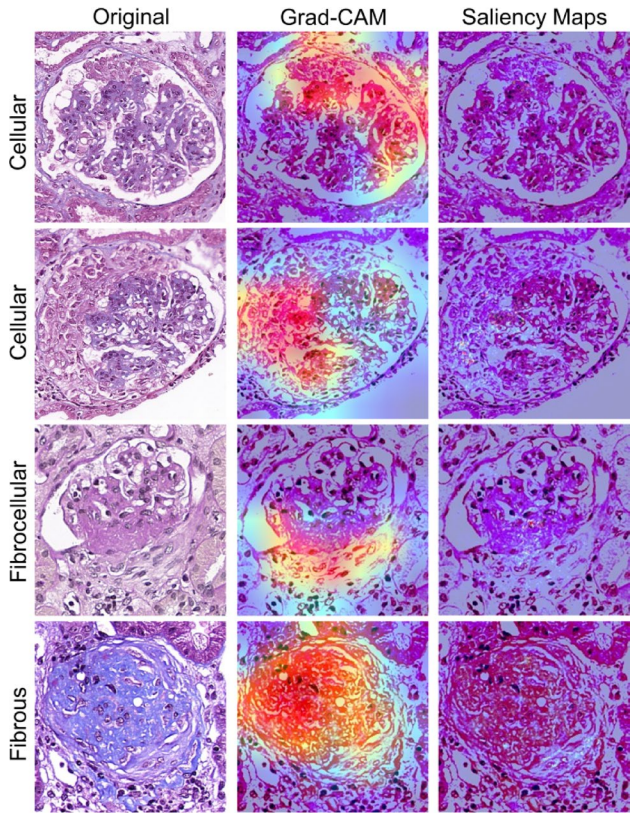


Fig. 14 Visualization of Grad-CAM and saliency maps showing key regions influencing CrescentDenseNet predictions for different crescent classes

Table 9 Quantitative performance of CrescentSegNet on the SICAPv2 prostate cancer histopathology segmentation dataset across training, validation, and testing phases, reported in terms of Dice coefficient, Jaccard index, recall, and precision

Phase	Loss	Dice	Jacard	Recall	Precision
Training	0.38	0.90	0.87	0.93	0.93
Validation	**0.32**	0.89	0.86	0.92	0.92
Testing	–	**0.92**	**0.90**	**0.95**	**0.94**

that CrescentSegNet produces well-calibrated predictions with minimal overconfidence.

Figure 16 presents Grad-CAM attention maps obtained from the final convolutional layer of CrescentSegNet on SICAPv2 test images. The model predominantly attends to histologically relevant regions, such as areas with dense cellular structures and glandular abnormalities, while suppressing background tissue. The spatially localized responses indicate that segmentation decisions are driven by meaningful morphological cues rather than spurious artifacts. Minor activations near patch boundaries reflect contextual dependencies of patch-based analysis and do not indicate systematic bias, supporting the interpretability and reliability of the proposed model.

4.3.2 CrescentDenseNet on BreaKHis 400 × Classification Dataset

To assess the robustness and cross-domain generalization of CrescentDenseNet, the model was further evaluated on the BreaKHis 400 × histopathology dataset, which is substantially larger than the ISICDM2024 kidney glomerular crescent dataset. In contrast to the small-scale ISICDM2024 dataset, where several heavy baseline models suffered from overfitting, CrescentDenseNet demonstrated stable, consistent performance on this larger, more diverse dataset. As shown in Table 10, the model achieved 85% accuracy on the independent test set, with balanced precision, recall, and F1-score, indicating reliable class discrimination. Similarly, the lower validation loss relative to the training loss indicates stable learning and confirms the absence of overfitting, highlighting the strong generalization capability of CrescentDenseNet beyond the original disease domain. Training and validation curves of CrescentDenseNet on the BreaKHis 400 × are shown in Figure S24 of SI.

Figure 17 summarizes the test-set performance of CrescentDenseNet on the BreaKHis 400 × dataset. The class-wise metrics shown in panel a indicate balanced precision, recall, and F1-score for both benign and malignant classes, demonstrating stable discrimination across classes despite class imbalance. The confusion matrix in panel b further confirms this behavior, with a high number of correctly classified samples and limited misclassification between classes.

Figure 18 illustrates feature-level representation and probabilistic calibration of CrescentDenseNet on the BreaKHis 400 × test set. The *t*-SNE plot in panel a shows a clear tendency toward separation between benign and malignant samples, with partial overlap that reflects realistic histopathological similarity rather than forced clustering. This indicates effective yet generalizable feature learning. The reliability diagram in panel b shows that predicted probabilities closely follow the ideal calibration line, especially at higher confidence levels, with moderate ECE and maximum calibration error (MCE) values. These results confirm that CrescentDenseNet achieves both discriminative feature representation and reliable confidence estimation on an external histopathology dataset.

Figure 19 illustrates Grad-CAM-based visual explanations of CrescentDenseNet predictions on the BreaKHis 400 × test set. For benign samples, the model primarily attends to well-organized glandular and stromal structures, while for malignant samples, it focuses on regions with dense cellularity, nuclear atypia, and disorganized tissue patterns. These attention maps align with known histopathological characteristics of benign and malignant breast tumors, indicating that CrescentDenseNet bases its

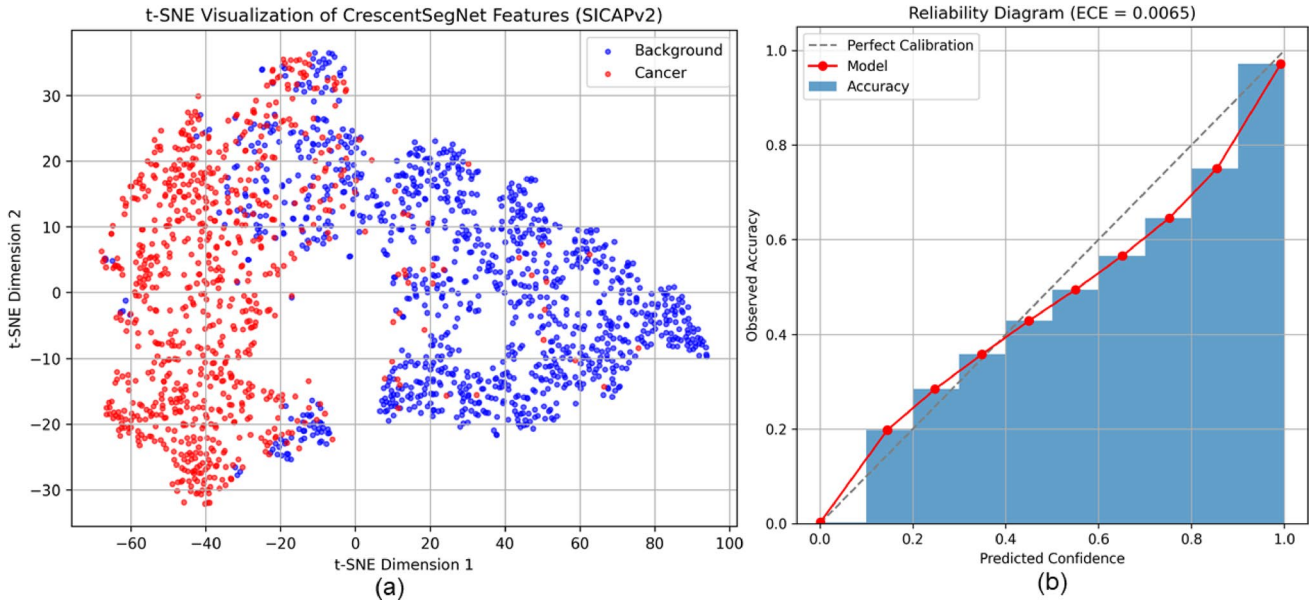


Fig. 15 Feature representation and confidence calibration of CrescentSegNet on the SICAPv2 data test set. **a** *t*-SNE visualization of learned feature embeddings. **b** Reliability (calibration) diagram showing the agreement between predicted confidence and observed accuracy

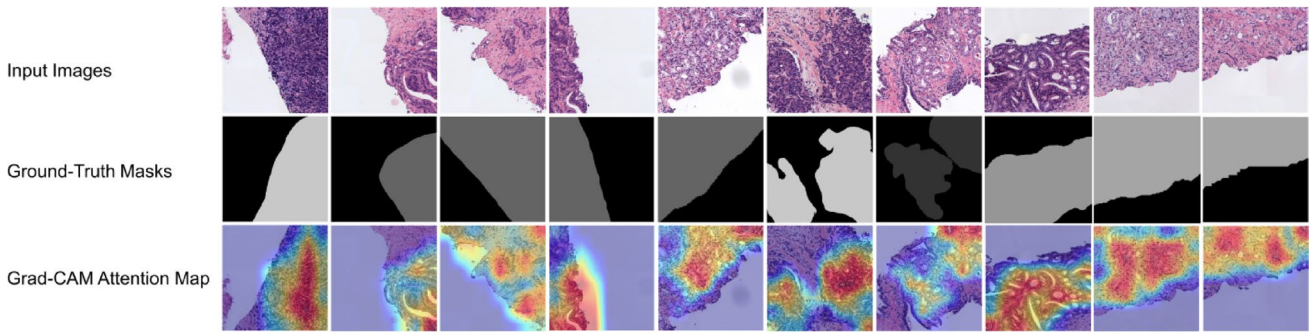


Fig. 16 Qualitative visualization of CrescentSegNet explanations on the SICAPv2 test set. The first row shows the input histopathology images, the second row shows the corresponding ground-truth segmentation masks, and the third row shows the Grad-CAM attention maps

Table 10 Performance of CrescentDenseNet on the BreaKHis 400 × histopathology dataset across training, validation, and testing phases, evaluated using accuracy, precision, recall, and F1-score (F1)

Phase	Loss	Accuracy	Precision	Recall	F1
Training	0.60	0.68	0.71	0.68	0.69
Validation	**0.54**	0.76	0.79	0.76	0.77
Testing	–	**0.85**	**0.86**	**0.85**	**0.85**

decisions on clinically meaningful regions rather than background artifacts.

4.4 Computational Complexity and Efficiency Analysis

Table 11 compares the computational complexity and inference efficiency of different segmentation models used for glomerular crescent lesion analysis. The evaluated architectures range from lightweight to high-capacity designs. DeepLabV3 (with ResNet-101) has the largest number of parameters and FLOPs, which results in slower inference. U-Net with MiT-B0 is computationally lighter but shows weaker segmentation performance on the small dataset. Some baseline models achieve lower inference time using pretrained encoders, which provide optimized feature representations and faster convergence.

In contrast, CrescentSegNet is trained entirely from scratch, which slightly increases its inference time. Despite this, CrescentSegNet maintains a favorable balance between efficiency and accuracy, achieving competitive performance with moderate computational cost. These results highlight that effective segmentation requires balancing model complexity, training strategy, and inference efficiency rather than relying solely on lightweight architectures. Representative terminal outputs from computational profiling of classification models are presented in Figure S25 of SI.

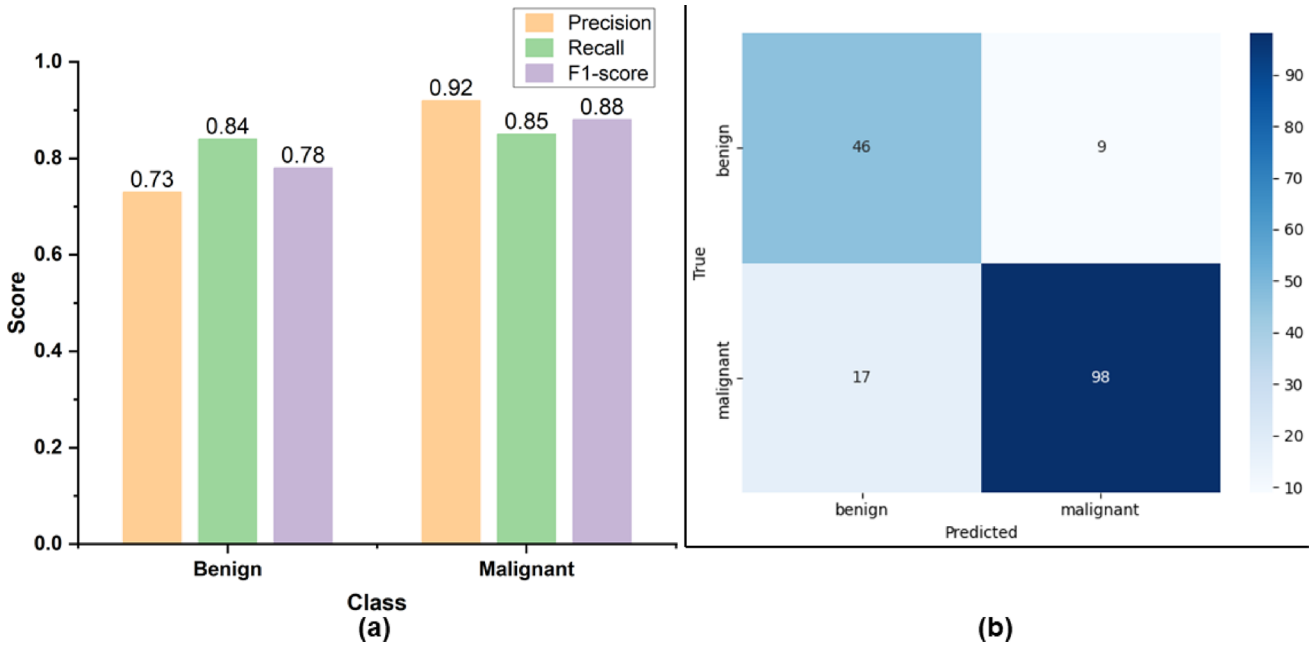


Fig. 17 Test-set performance of CrescentDenseNet on the BreaKHis 400× dataset. **a** Class-wise precision, recall, and F1-score for benign and malignant classes. **b** The corresponding confusion matrix illustrating correct and misclassified samples

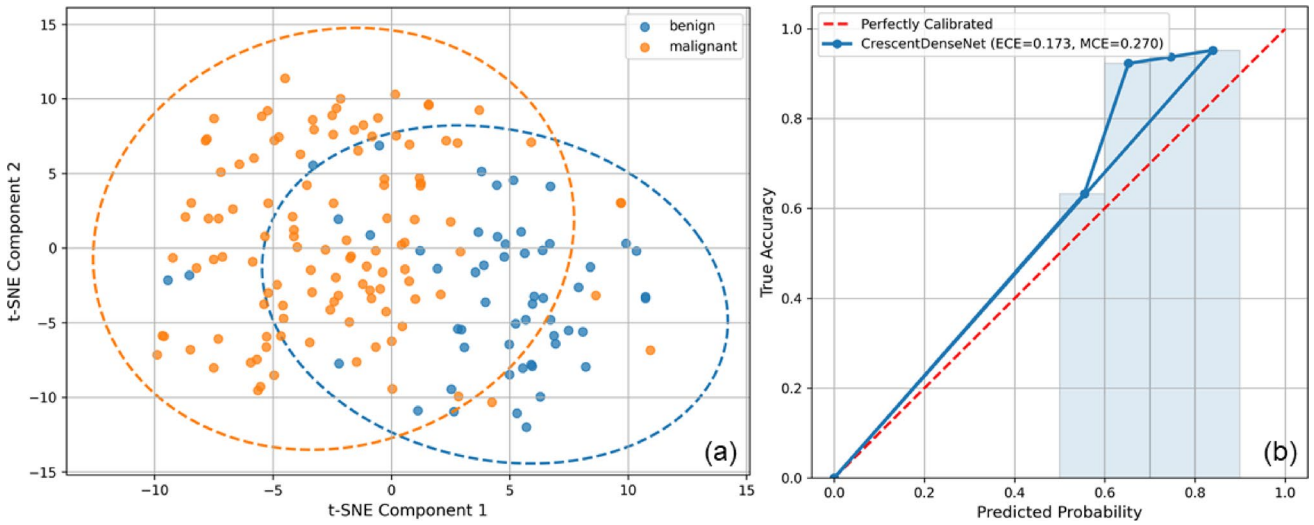


Fig. 18 Feature representation and confidence calibration of CrescentDenseNet on the BreaKHis 400× test set. **a** *t*-SNE visualization highlighting discriminative yet overlapping feature distributions for benign and malignant samples. **b** Reliability diagram demonstrating well-calibrated prediction confidence

The computational complexity and inference efficiency of the proposed CrescentDenseNet were compared with widely used baseline architectures (DenseNet-121, ResNet-50, EfficientNetV2-B0) and pathology-specific encoders (CTransPath and RetCCL) as summarized in Table 12. All models were evaluated under identical hardware conditions using an NVIDIA RTX 3090 GPU to ensure a fair comparison. CrescentDenseNet demonstrates a lightweight architectural design with only 6.957 million parameters and 3.78 GFLOPs, substantially lower than those of ResNet-50, EfficientNetV2-B0, CTransPath, and RetCCL. This reduced computational footprint highlights the proposed model's efficiency and suitability for resource-constrained clinical environments. Although CrescentDenseNet exhibits slightly higher inference time compared with some pretrained models (e.g., ResNet-50 and CTransPath), this behavior is architecturally and methodologically justified. Unlike the competing approaches, CrescentDenseNet is trained from scratch and incorporates task-specific architectural modifications and ablation-driven refinements customized to crescentic glomerular classification.

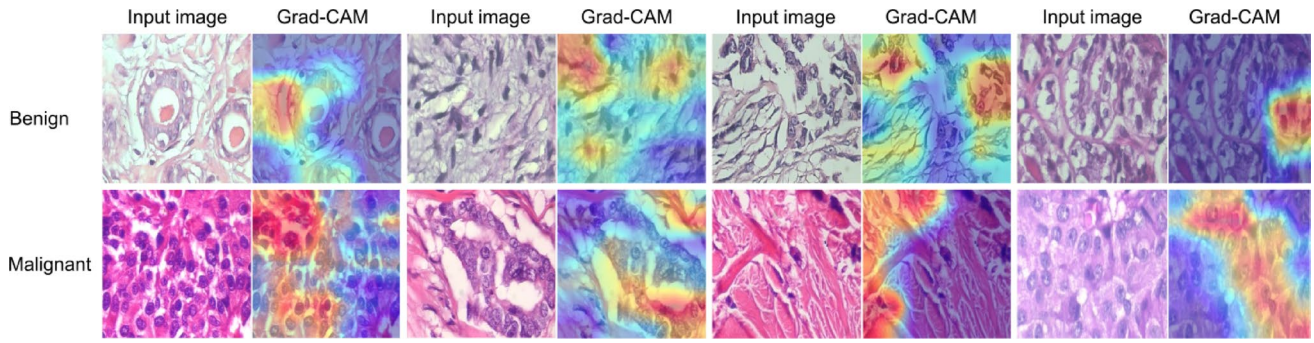


Fig. 19 Grad-CAM-based visual explanations of CrescentDenseNet predictions on the BreaKHis 400 × test set. The first row shows benign samples, and the second row shows malignant samples. For each case, the left image is the original histopathology image and the right image is the corresponding Grad-CAM attention map, highlighting regions that contribute most to the model's classification decision

Table 11 Comparison of computational complexity and inference efficiency of different segmentation architectures for glomerular crescent lesion segmentation, including input resolution, number of parameters, FLOPs, and average inference time measured on an NVIDIA RTX 3090 GPU

Model	Image input size	Total parameters ($\times 10^6$)	FLOPs ($\times 10^9$)	Inference time (ms/image)	Device
U-Net	256 × 256	31.044	54.738	5.918	CUDA
U-Net with MiT-B0	256 × 256	5.549	2.926	19.583	CUDA
DeepLabV3	256 × 256	61.000	62.960	27.510	CUDA
CrescentSeg-Net (Ours)	256 × 256	23.156	6.878	9.423	CUDA

Table 12 Comparison of computational complexity and inference efficiency of different DL architectures for glomerular crescent lesion classification measured on an NVIDIA RTX 3090 GPU

Model	Image input size	Total parameters ($\times 10^6$)	FLOPs ($\times 10^9$)	Inference time (ms/image)	Device
DenseNet-121	256 × 256	6.957	3.783	23.144	CUDA
ResNet-50	256 × 256	23.514	5.397	8.747	CUDA
Efficient-NetV2-B0	256 × 256	20.181	3.787	23.741	CUDA
CTransPath	224 × 224	27.519	4.371	16.113	CUDA
RetCCL (ResNet-50)	256 × 256	23.508	5.396	13.538	CUDA
Crescent-DenseNet (Ours)	256 × 256	6.957	3.783	22.426	CUDA

In contrast, models such as ResNet-50, RetCCL, and CTransPath benefit from large-scale pretraining and highly optimized weight initialization, which typically leads to faster convergence and more efficient inference pipelines. Furthermore, the inference speed of pretrained models is often influenced by highly optimized backbone implementations and feature reuse learned during large-scale pretraining, rather than purely by parameter count. CrescentDenseNet prioritizes feature specialization and discriminative learning for pathology, introducing additional computational overhead but resulting in improved task-specific representation learning. The representative terminal outputs from computational profiling of segmentation models are shown in Fig. S26 of SI.

5 Comparison with State-of-the-art (SOTA) Methods

Tables 13 and 14 compare our proposed models with representative SOTA methods on small medical imaging datasets. Table 13 reports segmentation performance, showing that CrescentSegNet achieves competitive results on the ISICDM2024 dataset for Dice coefficient, Jaccard index, precision, and recall. Table 14 reports classification performance, where CrescentDenseNet attains competitive accuracy and robust precision, recall, and F1-score on the ISICDM2024 dataset compared with some prior small-dataset classification approaches. Both comparisons support the effectiveness of the proposed lightweight models under limited-data conditions.

6 Discussion of Experimental Findings

From the experiments and ablation studies conducted on baseline and proposed models across different datasets and data scales, we derive the following key findings:

(1) Effect of Data Scarcity, Class Imbalance, and Visual Ambiguity: Model performance is strongly affected by limited sample availability and poor visual quality in the ISICDM2024 dataset. The fibrous crescent class contains only a few test images and exhibits low cellularity and weak visual contrast, making it difficult to distinguish from other classes. This scarcity and

Table 13 Comparison of our proposed CrescentSegNet model with SOTA techniques for the segmentation task on a small dataset, evaluated using Dice coefficient, Jaccard index, precision, and recall

Model and reference	Dataset	Dice	Jaccard	Precision	Recall
Bend–Net [18]	MoNuSegv1	0.83	0.63	–	–
U–Net (Lovasz loss) [19]	Dtest_seg	0.72	–	–	–
U–Net (Lovasz loss) [19]	CRAG	0.76	–	–	–
U–Net (Lovasz loss) [19]	GLAS	0.80	–	–	–
Han–Net [22]	MoNuSeg (Test 1)	0.80	0.61	–	–
Han–Net [22]	MoNuSeg (Test 2)	0.81	0.64	–	–
Ensemble (DenseNet + U–Net + Dilation Block) [23]	ACDC_LungHP	0.83	–	–	–
Proposed lightweight CNN + CBAM [24]	CAM-ELYON16 (1200 slices)	0.74	0.77	–	–
Mu–Net (proposed) [25]	Bright–field BOs	0.88	–	–	0.90
Our CrescentSegNet (Training, validation)	ISICDM2024 Challenge	**0.92**, 0.84	**0.86**, 0.74	**0.91**, 0.82	**0.92**, 0.85

morphological overlap introduce label ambiguity at the patch level, leading to unstable learning and unreliable performance metrics such as F1-score. Similar degradation was observed across all baseline and proposed models, indicating that the observed limitations arise primarily from dataset characteristics and pathological complexity rather than from model design or architecture.

(2) Generalizability of the Evaluation Framework: Although the proposed models are designed for glomerular crescent analysis, the overall evaluation framework is broadly applicable to other histopathology tasks. As demonstrated in Sect. 4.3, the proposed model achieves strong performance on two additional independent datasets. Systematic ablation studies, interpretability analysis, calibration, and statistical evaluation provide a general assessment strategy independent of disease type. These evaluation components can be reused in future studies, provided suitable data and annotations are available.

(3) Effect of Dataset Size on Model Complexity: Small datasets severely limit the effectiveness of high-parameter models. In both classification and segmentation tasks, complex models tended to overfit the training data and showed poor generalization. In contrast, lightweight models learned more stable and transferable features under limited data conditions. These results indicate that model complexity should be carefully matched to dataset size to avoid overfitting in small medical imaging datasets.

(4) Sensitivity to Hyperparameter Selection: Ablation studies show that model performance is sensitive to training hyperparameters, including batch size, learning rate, loss function, and input image size. Inappropriate hyperparameter choices led to unstable training

Table 14 Comparison of our proposed CrescentDenseNet model with the SOTA technique for the classification task on a small dataset, evaluated using accuracy, precision, recall, and F1-score (F1)

Method	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
ResNet–34 + DLR + CLA [28]	Derm images	83.40	–	–	–
ResNet–50 + DLR + CLA [28]	Derm images	83.26	–	–	–
ResNet–101 + DLR + CLA [28]	Derm images	83.19	–	–	–
ResNet–152 + DLR + CLA [28]	Derm images	84.90	–	–	–
DenseNet–161 + DLR + CLA [28]	Derm images	84.61	–	–	–
Cosine loss + cross–entropy [30]	CUB	68.00	–	–	–
Cosine loss + cross–entropy [30]	NAB	71.90	–	–	–
Cosine loss + cross–entropy [30]	Cars	85.00	–	–	–
Cosine loss + cross–entropy [30]	Flowers–102	70.60	–	–	–
Cosine loss + cross–entropy [30]	MIT Indoor	52.70	–	–	–
Cosine loss + cross–entropy [30]	CIFAR–100	76.40	–	–	–
Ensemble (EffNetB0 + VGG16 + MobileNetV2) [33]	Databiox	81.00	–	–	–
ViT (fine–tuned all layers) [34]	PH2	85.00	–	–	–
RegNetY–3.2G + MWNL–CLS [35]	ISIC 2018	86.40	–	87.10	–
RegNetY–8.0G + MWNL–CLS [35]	ISIC 2019	85.90	–	85.60	–
RegNetY–1.6G + MWNL [35]	ISIC 2017	87.50	–	–	–
RegNetY–800M + MWNL [35]	7–PT	86.60	–	–	–
Our CrescentDenseNet (Training and validation)	ISICDM2024 Challenge	**87.00, 86.00**	**100,87**	**87, 86**	**87, 86**

or reduced generalization. Careful tuning is therefore essential to achieve balanced and reliable performance on small datasets.

(5) Computational Limits in High-Resolution Training: High-resolution inputs, strong augmentation, and complex models significantly increase computational demand, requiring a careful balance between model design and practical training feasibility.

(6) Effect of Model Size and Pretraining: Models with a large number of parameters often exhibit slower inference speeds due to higher computational cost during the forward pass. In contrast, models initialized with pretrained weights typically converge faster during training than models trained from scratch.

(7) Importance of Interpretability and Prediction Confidence: High-performance metrics, such as accuracy alone, are insufficient for medical image analysis. It is essential to understand whether a model detects clinically relevant lesions correctly or produces false positives and false negatives. Interpretability and confidence calibration help verify that model decisions are reliable, explainable, and aligned with accurate pathological findings.

7 Conclusion

In this study, we presented a comprehensive investigation of deep learning-based approaches for the segmentation and classification of glomerular crescent lesions in histopathology images, with particular emphasis on performance under limited-data conditions. By systematically evaluating multiple baseline architectures and developing task-specific models, namely CrescentSegNet for segmentation and CrescentDenseNet for classification, we demonstrated that carefully designed lightweight architectures can achieve robust and reliable performance despite data scarcity and class imbalance. Extensive experiments on the ISICDM2024 dataset, complemented by cross-domain validation on the SICAPv2 and BreaKHis 400 × datasets, highlight the generalizability of the proposed framework. The results show that appropriate architectural design, combined with targeted regularization, data augmentation, and calibration strategies, can effectively moderate overfitting and enhance model stability. Furthermore, integrating interpretability and uncertainty analysis provides valuable insights into model behavior, supporting transparent and trustworthy decision-making in histopathological analysis.

Importantly, our findings indicate that high model complexity is not required for strong performance in small-scale medical imaging tasks. Instead, balanced model design and careful optimization play a more critical role in achieving reliable outcomes. The proposed framework, therefore, offers a practical, extensible solution for automated glomerular crescent analysis and has the potential to support future research in computational pathology.

Future work will focus on extending the proposed approach to multi-center and multi-stain datasets, incorporating weakly supervised and semi-supervised learning strategies, and further exploring multimodal integration to enhance robustness and clinical applicability.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12539-026-00824-9.

Acknowledgements The authors thank the organizers of ISICDM2024 for providing the histopathology task 8 dataset used in this study.

Author Contributions Inayatul Haq: Methodology, Implementation, Software, Validation, Formal Analysis, Data Curation, Preparing Original Draft, and Visualization. Haomin Liang: Data Curation, Formal Analysis. Zheng Gong: Validation, Data Curation. Zehong Xia: Data Curation. Wei Zhang: Investigation, Validation. Rashid Khan: Validation, Proofreading. Faizan Ahmad: Formal Analysis, Proofreading. Yan Kang: Investigation, Supervision. Bingding Huang: Investigation, Resources, Review, Supervision, Data Curation, and Funding Acquisition.

Funding This study was supported by the Shenzhen Science and Technology Program (Nos. KJZD20240903095605007, JCYJ20250604145033042) and the Shenzhen Medical Research Fund (No. D250402003).

Data and Code Availability The datasets used in this study include the ISICDM2024 Challenge dataset and two publicly available histopathology datasets. The ISICDM2024 dataset was provided by the challenge organizers. Description is available at https://www.imagecomputing.org/isicdm2024/index.html#/Challenge/Eight. The SICAPv2 and BreaKHis 400 × datasets are publicly available on Kaggle. The sources of all datasets are appropriately cited in the aforementioned dataset section.

Declarations

Conflict of interest The authors declare no known financial or personal conflicts of interest that could have influenced the work presented in this paper.

Declaration of Using Generative AI Tools During the preparation of this manuscript, the authors used Grammarly to assist with English language editing and BioRender to create schematic diagrams.

Ethical Approval The data used in this study were obtained from the ISICDM2024 Challenge and two publicly available datasets, SICAPv2 and BreaKHis 400 ×. All datasets consist of fully anonymized medical images, and no new patient data were collected. Therefore, institutional ethical approval and informed consent were not required for this study.

References

1. Anguiano L, Kain R, Anders H-J (2020) The glomerular crescent: triggers, evolution, resolution, and implications for therapy. Curr Opin Nephrol Hypertens 29(3):302–309. https://doi.org/10.1097/MNH.0000000000000596
2. Gurcan MN, Boucheron LE, Can A et al (2009) Histopathological image analysis: a review. IEEE Rev Biomed Eng 2:147–171. https://doi.org/10.1109/RBME.2009.2034865
3. Fuchs TJ, Buhmann JM (2011) Computational pathology: challenges and promises for tissue analysis. Comput Med Imaging Graph 35(7–8):515–530. https://doi.org/10.1016/j.compmedimag.2011.02.006
4. Ramachandran R, Sulaiman S, Chauhan P et al (2022) Challenges in diagnosis and management of glomerular disease in resource-limited settings. Kidney Int Rep 7(10):2141–2149. https://doi.org/10.1016/j.ekir.2022.07.002
5. Kong X-Y, Zhao X-S, Sun X-H et al (2023) Classification of glomerular pathology images in children using convolutional neural networks with improved SE-ResNet module. Interdiscip Sci Comput Life Sci 15(4):602–615. https://doi.org/10.1007/s12539-023-00579-7
6. Abdel-Nabi H, Ali M, Awajan A et al (2023) A comprehensive review of the deep learning-based tumor analysis approaches in histopathological images: segmentation, classification and multi-learning tasks. Clust Comput 26(5):3145–3185. https://doi.org/10.1007/s10586-022-03951-2
7. Haq I, Gong Z, Liang H et al (2025) A review of breast cancer histopathology image analysis with deep learning: challenges, innovations, and clinical integration. Image Vis Comput. https://doi.org/10.1016/j.imavis.2025.105708
8. He K, Zhang X, Ren S et al (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778. https://doi.org/10.1109/CVPR.2016.90
9. Hoque MZ, Keskinarkaus A, Nyberg P et al (2024) Stain normalization methods for histopathology image analysis: a comprehensive review and experimental comparison. Inf Fusion 102:101997. https://doi.org/10.1016/j.inffus.2023.101997
10. Kebaili A, Lapuyade-Lahorgue J, Ruan S (2023) Deep learning approaches for data augmentation in medical imaging: a review. J Imaging 9(4):81. https://doi.org/10.3390/jimaging9040081
11. Sharmay Y, Ehsany L, Syed S et al (2021) HistoTransfer: understanding transfer learning for histopathology. In: 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), pp 1–4. https://doi.org/10.1109/BHI50953.2021.9508542
12. Tan M, Le Q (2019) EfficientNet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp 6105–6114. https://doi.org/10.48550/arXiv.1905.11946
13. Howard AG, Zhu M, Chen B et al (2017) MobileNets: efficient convolutional neural networks for mobile vision applications. arXiv. https://doi.org/10.48550/arXiv.1704.04861
14. Tan M, Le Q (2021) EfficientNetV2: smaller models and faster training. In: International Conference on Machine Learning, pp 10096–10106. https://doi.org/10.48550/arXiv.2104.00298
15. Alruwaili M, Mohamed M (2025) An integrated deep learning model with EfficientNet and ResNet for accurate multi-class skin disease classification. Diagnostics 15(5):551. https://doi.org/10.3390/diagnostics15050551
16. Xiao H, Li L, Liu Q et al (2023) Transformers in medical image segmentation: a review. Biomed Signal Process Control 84:104791. https://doi.org/10.1016/j.bspc.2023.104791
17. Hossain MS, Armstrong LJ, Cook DM et al (2024) Application of histopathology image analysis using deep learning networks. Hum Centric Intell Syst 4(3):417–436. https://doi.org/10.1007/s44230-024-00077-z
18. Wang H, Vakanski A, Shi C et al (2024) Bend-Net: bending loss regularized multitask learning network for nuclei segmentation in histopathology images. Information 15(7):417. https://doi.org/10.3390/info15070417
19. Bokhorst J-M, Nagtegaal ID, Fraggetta F et al (2023) Deep learning for multi-class semantic segmentation enables colorectal cancer detection and classification in digital pathology images. Sci Rep 13(1):8398. https://doi.org/10.1038/s41598-023-35491-z
20. Jia Z, Huang X, Eric I et al (2017) Constrained deep weak supervision for histopathology image segmentation. IEEE Trans Med Imaging 36(11):2376–2388. https://doi.org/10.1109/TMI.2017.2724070
21. Xie A, Elfatimi E, Ghosal S et al (2024) Deep learning with uncertainty quantification for predicting the segmentation dice coefficient of prostate cancer biopsy images. In: 2024 International Conference on Machine Learning and Applications (ICMLA), pp 1158–1163. https://doi.org/10.1109/ICMLA61862.2024.00178
22. He H, Zhang C, Chen J et al (2021) A hybrid-attention nested UNet for nuclear segmentation in histopathological images. Front Mol Biosci 8:614174. https://doi.org/10.3389/fmolb.2021.614174
23. Li Z, Zhang J, Tan T et al (2020) Deep learning methods for lung cancer segmentation in whole-slide histopathology images—the ACDC@LungHP Challenge 2019. IEEE J Biomed Health Inform 25(2):429–440. https://doi.org/10.1109/JBHI.2020.3039741
24. Luo X, Ma T, Fan Z et al (2024) A lightweight deep learning model for breast cancer segmentation on small datasets. In: 2024 IEEE International Conference on Medical Artificial Intelligence (MedAI), pp 313–318. https://doi.org/10.1109/MedAI62885.2024.00047
25. Brémond Martin C, Simon Chane C, Clouchoux C et al (2023) Mu-Net a light architecture for small dataset segmentation of brain organoid bright-field images. Biomedicines 11(10):2687. https://doi.org/10.3390/biomedicines11102687
26. Frank SJ (2023) Accurate diagnostic tissue segmentation and concurrent disease subtyping with small datasets. J Pathol Inform 14:100174. https://doi.org/10.1016/j.jpi.2022.100174
27. Hu H, Zhang J, Yang T et al (2024) PATrans: pixel-adaptive transformer for edge segmentation of cervical nuclei on small-scale datasets. Comput Biol Med 168:107823. https://doi.org/10.1016/j.compbiomed.2023.107823
28. Mishra S, Yamasaki T, Imaizumi H (2019) Improving image classifiers for small datasets by learning rate adaptations. In: 2019 16th International Conference on Machine Vision Applications (MVA), pp 1–6. https://doi.org/10.23919/MVA.2019.8757890
29. Wong KC, Syeda-Mahmood T, Moradi M (2018) Building medical image classifiers with very limited data using segmentation networks. Med Image Anal 49:105–116. https://doi.org/10.1016/j.media.2018.07.010
30. Barz B, Denzler J (2020) Deep learning on small datasets without pre-training using cosine loss. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 1371–1380. https://doi.org/10.1109/WACV45572.2020.9093286
31. Rajput D, Wang W-J, Chen C-C (2023) Evaluation of a decided sample size in machine learning applications. BMC Bioinform 24(1):48. https://doi.org/10.1186/s12859-023-05156-9
32. Althnian A, AlSaeed D, Al-Baity H et al (2021) Impact of dataset size on classification performance: an empirical evaluation in the medical domain. Appl Sci 11(2):796. https://doi.org/10.3390/app11020796

33. Jaryani F, Amiri M (2023) A pre-trained ensemble model for breast cancer grade detection based on small datasets. Iran J Health Sci 11(1):47–58. https://doi.org/10.32598/ijhs.11.1.883.1
34. Abla R, Abdelouahed SM, Abdellah A (2024) Fine-tuning vision transformers for enhanced skin lesion classification: navigating the challenges of small datasets. In: 2024 International Conference on Intelligent Systems and Computer Vision (ISCV), pp 1–5. https://doi.org/10.1109/ISCV60512.2024.10620127
35. Yao P, Shen S, Xu M et al (2021) Single model deep learning on imbalanced small datasets for skin lesion classification. IEEE Trans Med Imaging 41(5):1242–1254. https://doi.org/10.1109/TMI.2021.3136682
36. Wang X, Zhao J, Marostica E et al (2024) A pathology foundation model for cancer diagnosis and prognosis prediction. Nature 634(8035):970–978. https://doi.org/10.1038/s41586-024-07894-z
37. Cai D, Chen J, Zhao J et al (2024) HiCervix: an extensive hierarchical dataset and benchmark for cervical cytology classification. IEEE Trans Med Imaging 43(12):4344–4355. https://doi.org/10.1109/TMI.2024.3419697
38. Guo R, Xie K, Pagnucco M et al (2023) SAC-Net: learning with weak and noisy labels in histopathology image segmentation. Med Image Anal 86:102790. https://doi.org/10.1016/j.media.2023.102790
39. ISICDM2024 (2024) Segmentation and classification of glomerular crescent lesions in renal biopsy histopathological images. In: The 7th International Symposium on Image Computing and Digital Medicine. Accessed 1 April 2025. http://www.imagecomputing.org/isicdm2024/index.html#/Challenge/Eight
40. Pol S (2023) SICAPv2: prostate cancer histopathology segmentation dataset (Kaggle Mirror). Kaggle. Accessed 20 Dec 2026. https://www.kaggle.com/datasets/shridharspol/sicapv2
41. Silva-Rodríguez J (2020) SICAPv2—prostate whole slide images with Gleason grades annotations. Mendeley Data. Accessed 20 Dec 2026. https://doi.org/10.17632/9xxm58dvs3.1
42. Spanhol FA, Oliveira LS, Petitjean C et al (2018) BreaKHis 400×: breast cancer histopathological images. Kaggle. Accessed 20 Dec 2026. https://doi.org/10.34740/kaggle/dsv/11351640
43. Xie E, Wang W, Yu Z et al (2021) SegFormer: simple and efficient design for semantic segmentation with transformers. Adv Neural Inf Process Syst 34:12077–12090. https://doi.org/10.48550/arXiv.2105.15203
44. Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
45. Huang G, Liu Z, Van Der Maaten L et al (2017) Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4700–4708. https://doi.org/10.1109/CVPR.2017.243
46. Wang X, Yang S, Zhang J et al (2022) Transformer-based unsupervised contrastive learning for histopathological image classification. Med Image Anal 81:102559. https://doi.org/10.1016/j.media.2022.102559
47. Wang X, Du Y, Yang S et al (2023) RetCCL: clustering-guided contrastive learning for whole-slide image retrieval. Med Image Anal 83:102645. https://doi.org/10.1016/j.media.2022.102645