The Visual Computer (2026) 42:363
https://doi.org/10.1007/s00371-026-04558-z

RESEARCH

Explainable AI for accurate diagnosis of papillary thyroid carcinoma via fine-needle aspiration cytopathology

Ahmet Kilicarslan[1] · Canan Tastimur[2] · Gokhan Gundogdu[1] · Erhan Akin[3]

Received: 31 January 2026 / Accepted: 24 May 2026
© The Author(s) 2026

Abstract
Papillary thyroid carcinoma (PTC) represents most thyroid cancer cases worldwide, making early and accurate diagnosis crucial for successful treatment. This study introduces an explainable AI framework for the automated analysis of fine-needle aspiration (FNA) cytopathology images. By integrating handcrafted features with deep learning-derived abstract features, the proposed model achieves high classification performance (~92% accuracy and ~96% AUC). The proposed architecture, incorporating multi-scale dilated and deformable convolutions, effectively captures irregularities in cell morphology. Visualizations demonstrate alignment with pathologists' diagnostic criteria, enhancing the model's clinical applicability and reliability. The code, dataset, and comprehensive documentation are publicly available via Zenodo (https://zenodo.org/records/18428647), ensuring full reproducibility of the experiments and results presented.

Keywords Deep learning in cytopathology · Digital pathology · Explainable AI · Fine-needle aspiration cytology · Papillary thyroid carcinoma

1 Introduction

Papillary thyroid carcinoma (PTC) is the most common malignant cancer in the endocrine system. PTC accounts for 80% of thyroid cancers worldwide, accounting for the vast majority [1, 2]. Furthermore, the incidence of this cancer has recently increased. Early diagnosis of PTC is a major step in improving patient prognosis. However, late diagnosis increases the rate of lymph node metastasis and recurrence, decreasing the success of treatment. Therefore, early diagnosis is critical for patient survival. While it is more common in women than men, it is more common in older adults, and its incidence also increases in younger individuals.

Fine-needle aspiration (FNA) is a frequently used method for examining thyroid nodules. While FNA is a low-cost, reliable, and practical diagnostic tool, it also presents some challenges in accurately assessing the morphological characteristics of PTC. Pathologists require experience in interpreting PTC cytopathology images obtained with FNA, and the interpretation of some cases is ambiguous and may result in different interpretations by pathologists. Pathologists consider criteria such as nuclear features, nuclear indentations, and chromatin arrangement when diagnosing PTC. These criteria are based on subtle structural differences. Differences in interpretation between pathologists in PTC diagnosis and failure to make an accurate diagnosis in ambiguous cases can lead to unnecessary surgical interventions during the patient's treatment process. For these reasons, there has recently been a growing trend toward developing automated and reliable AI-powered tools to support pathologists' diagnostic processes. Studies on deep learning-based AI approaches have increased in the field of digital pathology. These systems accurately capture microscopic and macroscopic structural representations, supporting their success in diagnosis. Numerous studies in this area exist in literature.

✉ Canan Tastimur
ctastimur@erzincan.edu.tr

Ahmet Kilicarslan
ahmetkilicarslan@gmail.com

Gokhan Gundogdu
gokhangundogdu23@gmail.com

Erhan Akin
eakin@firat.edu.tr

[1] Medical Pathology, Fethi Sekin City Hospital, Elazig, Turkey

[2] Computer Engineering, Erzincan Binali Yildirim University, Erzincan, Turkey

[3] Computer Engineering, Firat University, Elazig, Turkey

Published online: 27 June 2026

A deep learning framework inspired by pathologist decision-making processes is proposed for PTC diagnosis in FNA cytopathology images. In the first step, we extracted patches from images of real patients using YOLOv11. This step aims to mimic the diagnostic regions that pathologists pay attention to when examining cytopathological images and making diagnoses, removed completely. After saving the patches as a separate dataset, we extracted clinically relevant nuclear morphology and chromatin distribution features from this dataset and then extracted deep features using CNN architecture. We fed the extracted features from the relevant regions to our model during the classification phase of our full cytopathology dataset. The classification architecture used multi-scale dilated and deformable convolutions and adaptive attention simultaneously represented both tissue architecture and irregular morphological structures. Furthermore, the explainability of the system was demonstrated using Grad-CAM visualizations, demonstrating that the diagnostic criteria used by pathologists in their decision-making processes coincided with the regions focused on by the model. Thus, our proposed method not only achieves high accuracy but also supports improved classification of benign and malignant cases in routine patient visits, may assist pathologists as a decision support system in clinical workflows in clinical workflows, and makes providing a complementary contribution to existing literature. The main contributions of this study can be summarized as follows.

1. We propose a pathology-inspired hybrid representation learning framework that explicitly integrates bio-morphological descriptors (capturing nuclear shape, texture, and structural statistics) with deep CNN-based semantic representations. Unlike conventional feature fusion approaches, this design aligns the feature hierarchy with the diagnostic workflow of cytopathologists, improving interpretability and clinical relevance.
2. We introduce a multi-scale deformable–dilated feature extraction strategy within a unified backbone to explicitly model both global tissue architecture and local nuclear deformation patterns. This is motivated by the dual-scale visual inspection process of pathologists, where tissue-level context and cellular-level irregularities are jointly evaluated.
3. We design an attention-guided fusion mechanism that dynamically learns the interaction between handcrafted morphological descriptors and deep visual embeddings. This is not a simple concatenation scheme; instead, it adaptively weights complementary diagnostic cues based on their relevance to the final classification decision.
4. We further incorporate an explainability-driven validation framework using Grad-CAM and region-level visual alignment, enabling qualitative and quantitative assessment of whether the model focuses on diagnostically meaningful regions, thereby improving clinical trustworthiness.
5. Finally, the proposed system is designed as a second-opinion decision support tool, specifically targeting borderline cytology cases, with an emphasis on reducing diagnostic workload in routine clinical screening pipelines.

2 Related works

There are numerous machine learning and deep learning-based studies in the literature on thyroid carcinoma segmentation and classification. These approaches enable pathologists to assist physicians in the diagnostic process. Wuttisarnwattana et al. [1] compared key cytomorphological criteria for PTC detection using different objective magnifications in a study that automatically recognized intranuclear inclusions (INCI) in FNA smears. An identification accuracy of around 80% was achieved. Automatic PTC segmentation was performed using a mask-piloted transformer [2], which achieved 95.81% accuracy. They presented a comparison of their proposed model with Unet and TransUnet. Intan and Soetikno [3] presented a pattern recognition approach that distinguishes three types of histopathology: papillary, follicular, and anaplastic thyroid carcinoma. It was planned to assist the pathologist in the diagnosis phase of clinical pathology. Two different training scenarios, 400 and 800 iterations, were tested with 30 and 63 training samples, and the highest accuracy was 80.37%.

Yang et al. [4] added self-attention to the VGG16 transfer learning model for machine learning-based PTC detection from ultrasound images. Thus, the accuracy rate increased from 78 to 96.62%. A dataset was prepared with real patient data from Zhuhai People's Hospital. A total of 2047 images were found, 538 female and 709 males. Shabrina et al. [5] classified PTC histopathology images with pre-trained ConvNext Tiny. They also performed image analysis and interpretability with Grad-CAM. The dataset they used consisted of 5 classes: 3 classes belonged to the PTC data class, and 2 classes belonged to non-PTC. They used the Tharun & Thompson dataset, containing 1496 images. They achieved test accuracy of 84.96%.

Stenman et al. [6] trained a CNN network developed for leukocyte segmentation in PTC and continued the training by cropping the relevant areas. They analyzed a total of 238,144 tiles with a size of 500×500 pixels. They obtained a dataset with 1738 pathology images from twelve patients and used it in the training phase. The models they used in their study are ResNet and UNET architecture. Another deep learning approach [7] performed classification as benign, malignant, and normal thyroid. They applied transfer

learning by fine-tuning well-known deep learning architectures such as VGG16, VGG19, ResNet50, AlexNet, and Inception V3. The features extracted by the transfer learning models were classified using well-known machine learning approaches such as K-nearest neighbors (KNN), XG Boost, decision tree, random forest, AdaBoost, logistic regression, and support vector machine (SVM). The highest accuracy rate was 82.12% with ResNet architecture. In another study [8] where thyroid cancer classification was conducted using transfer learning, the architecture used was EfficientNetB0. The features extracted with transfer learning were then classified with SVM. In this study, where three different types of carcinomas were classified, they achieved 99% accuracy. The dataset included 1393 histological images. In [9], which used transfer learning, they used AlexNet and classified six types of thyroid cancer. They achieved 97.07% accuracy. Chen et al. [10] added the Convolutional Block Attention Module (CBAM) to the Xception architecture to predict the risk of follicular thyroid carcinoma. The accuracy rate obtained from a dataset of 323 patients was 93%. In the PTC classification using the ResNet architecture [11], they extracted features and used SVM. They achieved 73% train accuracy and 72% test accuracy.

Payatsuporn et al. [12] applied segmentation to PTC regions in a real histopathology dataset obtained from King Chulalongkorn Memorial Hospital. They applied adaptive convolution and dual decoder to 286 PTC histopathology images and achieved an accuracy of 88.83%. El-Hossiny et al. [13] proposed a cascade CNN model for thyroid carcinoma classification. They achieved an accuracy of 98.74% for three types of classification with the versetinde and 95.90% for four types of classification. The dataset they used consisted of 20× full slide images collected from the International Medical Center in Egypt. Shabrina et al. [14] used a deep residual network for the segmentation of stardist nuclei for PTC identification. After completing the training of the model on the Tharun & Thompson dataset, they achieved an accuracy of 98.55%. Ferreira et al. [15] designed a deep architecture for PTC classification. In this study, they compared the performance of six different autoencoders. The model was completed by fixing the autoencoder weights and fine-tuning the network. They achieved the highest accuracy of 98.77% from six different autoencoder versions. Padrez et al. [16] extracted texture and density features from histopathology images for the classification of papillary and follicular thyroid carcinoma. They applied recursive feature elimination and mutual information-based selection in the feature selection process with a support vector classifier and achieved 84.73% accuracy. Li et al. [17] developed a multi-semantic attention network for binary classification of PTC as benign and malignant. They tested the model with 8475 ultrasound images from Peking Union Medical College Hospital and achieved 98.39% accuracy. Lu et al. [18] obtained data from two different sources for the detection of malignant thyroid nodules using deep learning. They trained these versets with the Cascade Mask R-CNN method, including 1086 and 428 pieces, and achieved 99.70% and 94.30% accuracy. Park et al. [19] proposed NeXtSRGAN, a super-resolution GAN framework that integrates a ConvNeXt-based discriminator to enhance visual realism and perceptual quality. Their approach improved feature discrimination and texture fidelity compared to conventional GAN architectures. The study demonstrated that modern convolutional designs can significantly boost performance in visual reconstruction tasks.

Zhang et al. [20] introduced HADiff, a hierarchy-aggregated diffusion model designed for pathology image segmentation. The model effectively captured multi-scale contextual information and structural details, leading to improved segmentation accuracy across complex medical images. Their work emphasized the value of diffusion-based models for robust and fine-grained analysis in medical imaging.

Noto et al. [21] conducted a meta-analysis of forty-six different studies that improved the success of differentiating between benign and malignant thyroid cancer with diffusion-weighted imaging. Metrics of 91.2% AUC, 84% sensitivity, and 88% specificity were achieved. Furthermore, it was particularly critical in differentiating malignant samples. Chen et al. [22] developed a transformer-based model for predicting cervical lymph node metastasis. It produced results superior to CNN with an 81% AUC value in both internal and external validation cohorts from ultrasound images of 540 patients.

A transformer-based deep architecture was designed to predict lymph node metastasis before surgery in papillary thyroid cancer [23]. The model was trained on 1560 patients, achieving an AUC of 0.83% and reducing neck lymph node dissection by 5% in clinically negative patients. Siu et al. [24] developed a study predicting papillary thyroid microcarcinoma using deep learning architecture. Although multiple deep learning models were tried, the highest performance was obtained with the Swin transformer architecture, and external validation AUC: 0.892 was obtained.

In [25], explainable artificial intelligence and cascade R-CNN method were used in the diagnosis of PTC cytopathology. Graph neural networks were used to model the information between cells, achieving 88.84% accuracy. Thanya and Jeslin [26] combined multimodal deep cascade CNN and ResNet50 architectures to diagnose pancreatic cancer in CT images. They achieved 99.3% accuracy. Kim et al. [27] evaluated 1048 nodule images from 943 patients in comparison with Siamese network and ResNet, DenseNet and EfficientNet architectures in the diagnosis of PTAC. Thyroid ultrasound images were used in this study, which achieved 83% accuracy. Wang et al. [28] used Transformer architecture to differentiate benign and malignant thyroid

nodules of follicular thyroid carcinoma. They achieved 87% accuracy in a dataset with 10,771 images. Pengwei et al. [29] combined XGBoost and random forest with machine learning models in PTAC diagnosis. They achieved an accuracy level of 87.4% with data from 2907 patients. Han et al. [30] designed a capsule network in PTC classification. Semantic features from multiple magnification levels were used, mimicking the diagnostic process of pathologists. They achieved 97.03% accuracy. Model training and testing were performed on a dataset containing 16,667 pathological images collected from 55 patients.

These approaches, which have become popular in recent years, have proven successful in various medical fields. In [31], explainable artificial intelligence was used to diagnose pelvic fractures with 98.5% accuracy on 876 x-ray images. In [32], 61.81% accuracy was obtained on 733 x-ray images with residual CNN model according to the types of knee osteoarthritis. Similarly, Naguib et al. [33] achieved three different disease classifications with knee x-ray images using ResNet50 model. 85.42% accuracy was obtained. In [34], cervical spine injuries, over 99% accuracy was obtained with AlexNet and GoogLeNet for the classification of injury type.

Table 1 summarizes representative studies instead of exhaustively listing all prior work. Recent trends clearly indicate a shift from conventional CNN-based approaches toward transformer-based, hybrid, and explainable AI frameworks.

3 The proposed model

The proposed method is inspired by the cytology evaluation process of pathologists and consists of a hybrid deep learning architecture. This architecture aims not only to achieve a high success rate but also to ensure explainability and clinical reliability.

Pathologists focus on specific regions of interest when examining cytology. Because this process relies on the pathologist's experience and intuition, it is prone to human error and carries user dependency. In this study, ROI extraction is automated using YOLOv11 applied to FNA cytopathology images, thereby replacing manual ROI selection and reducing inter-observer variability. This eliminates the need for manually selected ROIs. It also minimizes the possibility of process variation due to different expert decisions. We saved these extracted patches as separate versions and then applied a custom feature extraction process to this dataset. The feature extraction process consists of two steps: hand-carved features and deep abstract features. We calculated Hu moments, statistical contrast parameters, and GLCM-based texture features for quantifying nuclear morphology and chromatin patterns. CNN-based extracted deep abstract features, on the other hand, comprise more abstract and higher-dimensional representations in the image. We provided all these extracted features to the model as additional information during the classification phase of our main dataset, the PTC cytopathology patient slide dataset. Our classification architecture consists of multi-scale dilated convolutions and deformable convolutions. Multi-scale dilated convolutions model tissue architecture at the macro-level, while deformable convolutions learn nucleus irregularities at the micro-level. The adaptive attention layer added to our classification architecture enables us to combine information from different scales to produce a more balanced decision-making mechanism. The proposed architecture is supported by Grad-CAM-based visualizations to explain diagnostic decisions. These visualizations reveal that the regions the model pays attention to overlap with those observed by pathologists. Thus, our model demonstrates that it can achieve not only high accuracy but also clinical reliability. Our model stands out as a more reliable, explainable, and clinically applicable system compared to classical black-box CNN approaches. The general flow of the proposed model is shown in Fig. 1.

Table 1 Literature summary of artificial intelligence-based approaches in thyroid cancer diagnosis

Refs	Method	Modality	Dataset Size	AUC/Acc	Category
[21]	Meta-analysis (DWI)	MRI	46 studies	0.912 AUC	Clinical Imaging
[22]	Transformer-based Model	Ultrasound	540 patients	0.81 AUC	Transformer
[23]	Transformer DL Model	Clinical + Imaging	1560	0.83 AUC	Transformer
[24]	Swin Transformer	Ultrasound	–	0.892 AUC	Transformer
[25]	Cascade R-CNN + GNN	Cytology	–	88.84 Acc	Explainable AI
[26]	Multi-modal CNN + ResNet50	CT	–	99.3 Acc	Multi-modal DL
[27]	Siamese + CNNs	Ultrasound	1048 nodules	83 Acc	Metric Learning
[28]	Transformer Model	Ultrasound	10,771	87 Acc	Transformer
[29]	XGBoost + RF	Clinical	2907	87.4 Acc	Ensemble ML
[30]	Capsule Network	Histopathology	16,667	97.03 Acc	Advanced DL

A. Dataset description

Fine-needle aspiration cytology images collected with ethics committee approval were used in this study. Cytology preparations were digitized at 40 × magnification using an Olympus BX53 light microscope in the Pathology Department of Elazığ Fethi Sekin City Hospital. All images were captured under the microscope by an experienced pathologist and labeled according to cytologic diagnoses. The dataset contains whole-slide images from 112 different patients. 1167 of these images are benign nodules and 1167 contain malignant cells from papillary thyroid carcinoma. For this study, images were captured at 2448 × 1920 pixels. Images captured at this resolution provide more data for both capturing cellular details and for the segmentation model. The dimensions of the microscopic images were rescaled to suit the input requirements and computational capacity of the deep learning model. The images, originally captured at 2448×1920 pixels, were downscaled to 640×640 pixels to assess the model's training and real-time prediction performance, as Fig. 2. A total of 14,637 malignant cells were labeled in 1167 malignant images, and these cells were labeled using our proposed ROI extraction with YOLOv11.

To ensure unbiased evaluation, the dataset was split at the slide level prior to any feature extraction, preventing patches from the same slide from appearing in both training and validation sets. Patch-level features were computed separately for each set, and normalization was performed using training set statistics only, as Fig. 3. Slide-level aggregation, including attention-based weighting, was applied exclusively within each set. These measures eliminate any potential data leakage and provide a more reliable estimate of model performance.

B. ROI Extraction with YOLOv11

This first step of the proposed method utilizes automatic pattern extraction from FNA cytology using YOLOv11. The primary purpose of ROI extraction is to eliminate interpretation discrepancies among pathologists while simultaneously reducing human workload. YOLOv11 uses YOLOv11 to extract and crop diagnostically significant regions from cytological images and create another dataset. This simulates the decision-making processes of clinicians, highlighting areas where the model should focus, such as nuclear morphology and chromatin distribution. The necessity and importance of this process is that it ensures that the model executes operations on consistent and clinically meaningful data in subsequent processes. Thus, it bridges the gap between our deep learning model and pathological decision-making mechanisms.

Using YOLOv11 at this stage not only achieves high accuracy but also offers a speed advantage for real-time application of our method. Among the object detection approaches in the literature, YOLOv11 has been applied for the first time in the field of cytology, and our work stands out with its innovative contribution. The newly obtained dataset is directed to the feature extraction process as the second stage.

The ROI extraction stage was implemented using YOLOv11 due to its improved feature representation and enhanced localization performance in small object detection tasks. To validate its effectiveness, YOLOv11 was compared with YOLOv8 under identical training settings.

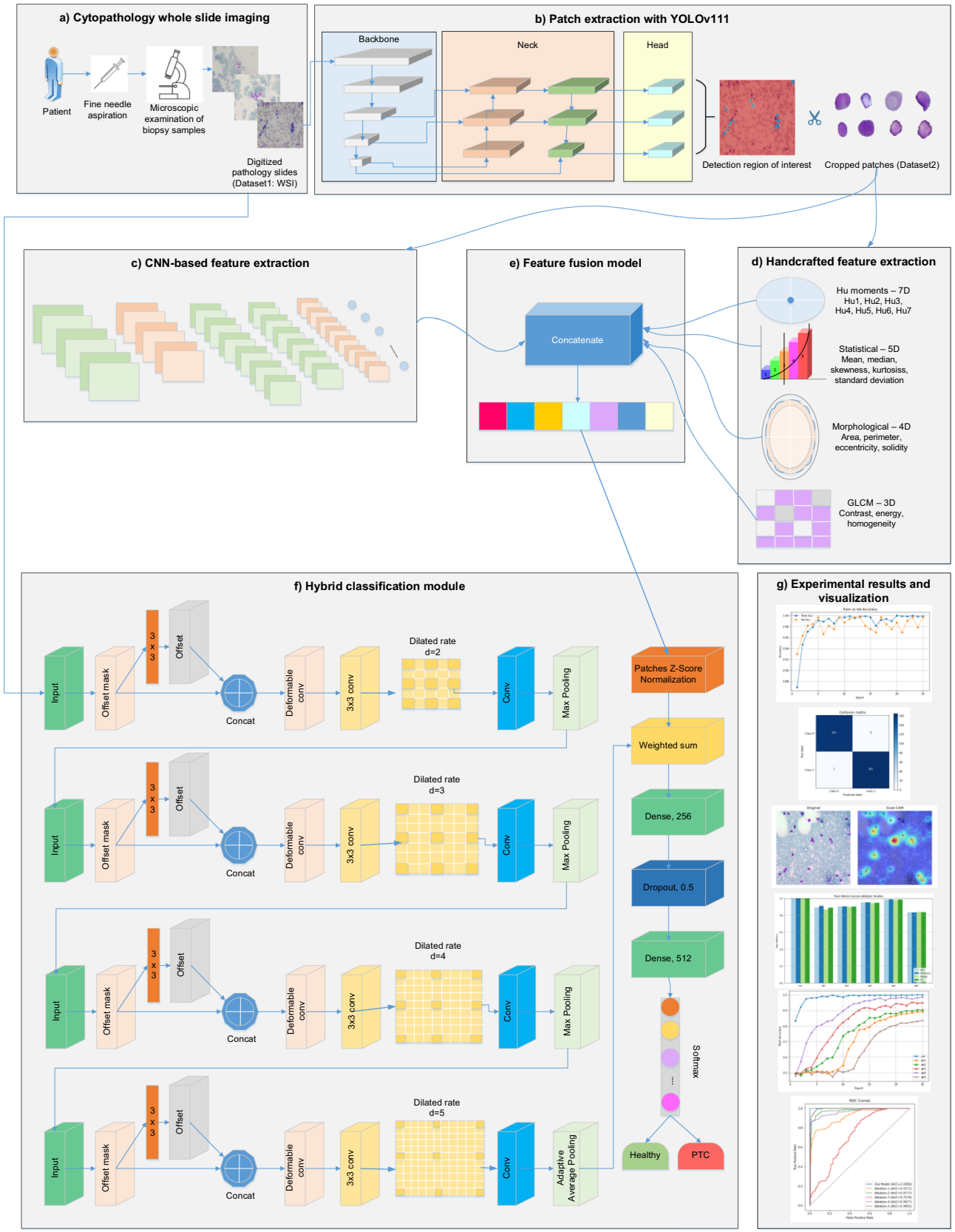
C. Handcrafted Feature Extraction

Handcrafted feature extraction, which we propose as an innovation and distinct from other methods in literature, is the first step in the complex feature extraction process for PTC diagnosis. Unlike other methods in literature, instead of extracting features solely from deep features and conducting the process in a black-box manner, handcrafted feature extraction yields crucial diagnostic information regarding nuclear morphology, tissue heterogeneity, and cellular distribution, all of which pathologists consider when making a PTC diagnosis from cytopathology images. Feature extraction is the numerical representation of meaningful shapes and textures in images. In this study, we extracted 19 handcrafted features: 7 Hu moments, 5 statistical metrics, 4 morphological features, and 3 gray-level co-occurrence matrix (GLCM)-based texture features. Hu moments measure shape changes in nuclear morphology and invariantly quantify nuclear deformities. Statistical features, on the other hand, reflect nuclear morphology. These include information on intracellular density and heterogeneity, including area, perimeter, occupancy, and eccentricity. It allows for the detection of cell density by measuring the central tendency and spread of the gray-level distribution. Morphological features are used to capture structural differences specific to malignancy, yielding basic geometric measurements such as cellular area, perimeter, eccentricity, and occupancy. The GCLM feature quantifies tissue patterns within and around the nucleus. These extracted features allow the model to utilize features specific to and sensitive to PTC classification. To ensure scale consistency and robustness against illumination and staining variations, all handcrafted features were normalized using a percentile-based min–max strategy (2nd–98th percentile clipping), followed by range scaling to [0,1].

1) Hu Moments Invariants

Hu moments consist of seven invariant moments to derive the geometric shape properties of objects in an image [31–33]. Since moment calculations are generally more accurate using ROI regions, we used our patch dataset detected with ROIs. Figure 4 is the visualization of hu moments via the PTC cell.

The image was first converted to grayscale, and Hu invariant moments were computed from normalized central moments defined as (1), (2), (3). The seven Hu invariant


a) Cytopathology whole slide imaging
Patient
Fine needle aspiration
Microscopic examination of biopsy samples
Digitized pathology slides (Dataset1: WSI)
b) Patch extraction with YOLOv111
Backbone
Neck
Head
Detection region of interest
Cropped patches (Dataset2)
c) CNN-based feature extraction
e) Feature fusion model
Concatenate
d) Handcrafted feature extraction
Hu moments – 7D
Hu1, Hu2, Hu3, Hu4, Hu5, Hu6, Hu7
Statistical – 5D
Mean, median, skewness, kurtosiss, standard deviation
Morphological – 4D
Area, perimeter, eccentricity, solidity
GLCM – 3D
Contrast, energy, homogeneity
f) Hybrid classification module
Input
Offset mask
3 x 3
Offset
Concat
Deformable conv
3x3 conv
Dilated rate d=2
Dilated rate d=3
Dilated rate d=4
Dilated rate d=5
Conv
Max Pooling
Adaptive Average Pooling
Patches Z-Score Normalization
Weighted sum
Dense, 256
Dropout, 0.5
Dense, 512
Softmax
Healthy
PTC
g) Experimental results and visualization

◂**Fig. 1** The overview of the proposed method. **a** Obtaining cytopathology whole-slide images from the patient using FNA. **b** Automatically extracting ROI regions using YOLOv11 and saving them as a new version. **c** Extracting features from the patch dataset using CNN. **d** Obtaining handcrafted features from the patch dataset. **e** Combining handcrafted and CNN-based features and transmitting them to the classification process. **f** Classifying images as PTC and healthy using deformable and multi-scale distributed convolution. **g** Experimental performance outputs related to the proposed method

moments were then derived from the normalized central moments.

$$m_{pq} = \sum_{x} \sum_{y} x^p y^q I(x,y) \tag{1}$$

$$\mu_{pq} = \sum_{x} \sum_{y} \left(x - \bar{x}\right)^p \left(y - \bar{y}\right)^q I(x,y) \tag{2}$$

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{1+\frac{p+q}{2}}} \tag{3}$$

2) Statistical Features

Statistical features were extracted from grayscale cell nucleus images to describe the intensity distribution. These features include mean, standard deviation, median, skewness, and kurtosis, capturing first- and higher-order statistics of pixel intensities. Let I_i denote pixel intensities and N the number of pixels in the nucleus.

$$\mu = \frac{1}{N} \sum_{i=1}^{N} I_i \tag{4}$$

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (I_i - \mu)^2} \tag{5}$$

$$median = I_{\left(\frac{N+1}{2}\right)} \tag{6}$$

$$skewness = \frac{\frac{1}{N}\sum_{i=1}^{N}(I_i - \mu)^3}{\sigma^3} \tag{7}$$

$$kurtosis = \frac{\frac{1}{N}\sum_{i=1}^{N}(I_i - \mu)^4}{\sigma^4} \tag{8}$$

Figure 5 is the visualization of statistical features via the PTC cell. These statistical features provided a robust representation for distinguishing between malignant and benign tumors.

3) Morphological Features

In this study, we examined the nuclear morphology of cytopathology images based on four basic attributes: *area*, *perimeter*, *eccentricity*, and *solidity*. Figure 6 is the visualization of morphological features via the PTC cell.

The *area* is the number of pixels covered by the PTC cell nucleus and can be obtained using (9). The p_i in (9) represents the number of pixels in the nucleus.

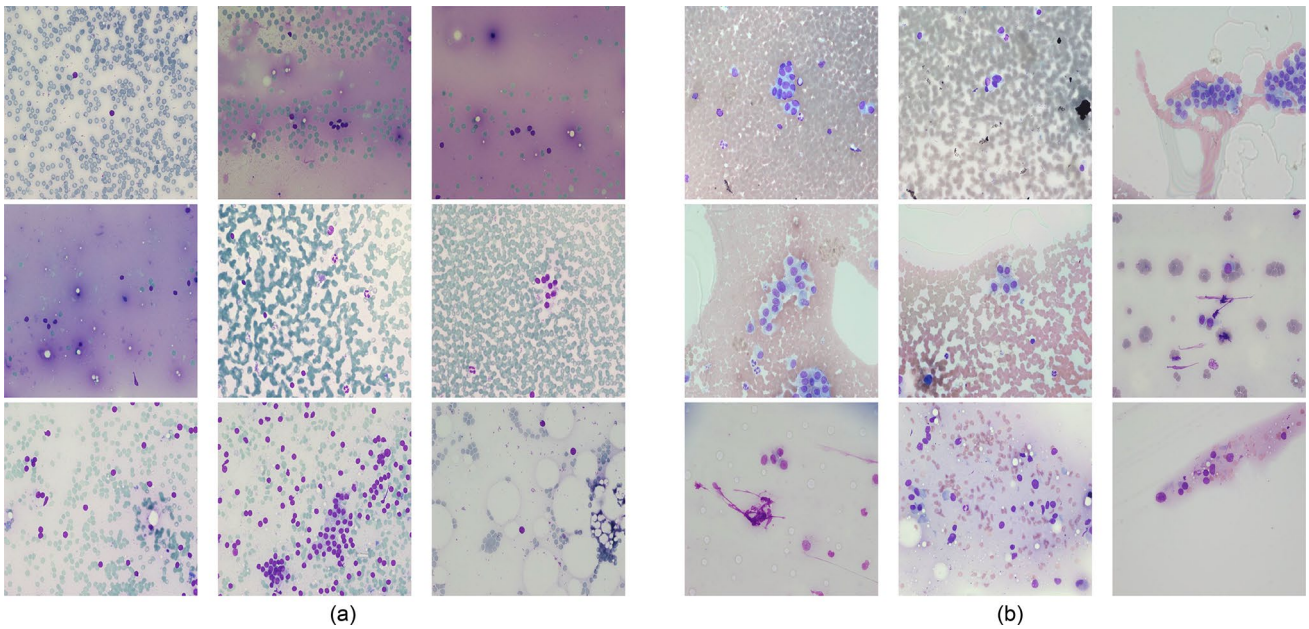


Fig. 2 Representative cytology images. **a** benign samples and **b** malignant samples in whole-slide image dataset



Fig. 3 Representative cropped patches of region of interests in patch extraction dataset

$$A = \sum_{i=1}^{N} p_i \tag{9}$$

The second attribute, *perimeter*, is the pixel length that forms the boundary of the cell nucleus. Irregularities in the contour of the nucleus are quantified and expressed as (10). The d_k in (10) is the distance between consecutive boundary points.

$$P = \sum_{k=1}^{M} d_k \tag{10}$$

The *eccentricity* attributes measure the extent to which the cell is elongated in an elliptical shape, that is, it measures the elliptical deformation in the cell and is calculated using (11). Here, a represents the major semi-axis and b represents the minor semi-axis. The closer the *eccentricity* is to zero, the more circular the shape becomes. The closer it is to 1, the closer the nucleus becomes to an ellipse.

$$e = \sqrt{1 - \frac{b^2}{a^2}} \tag{11}$$

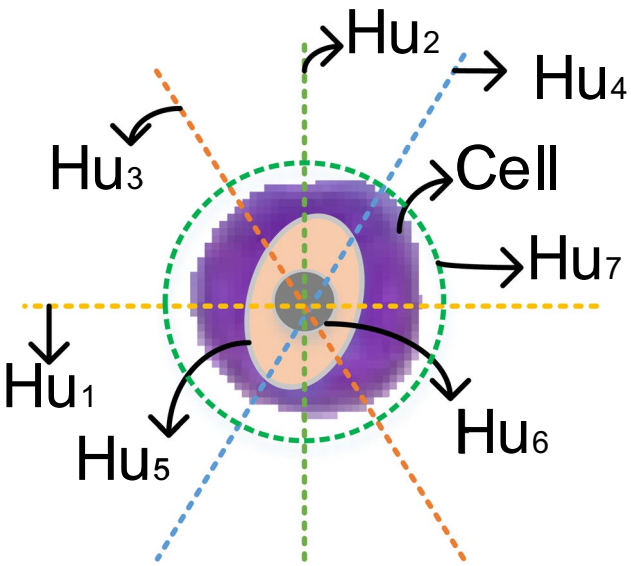


Fig. 4 Visual representation of Hu moments in the PTC cell

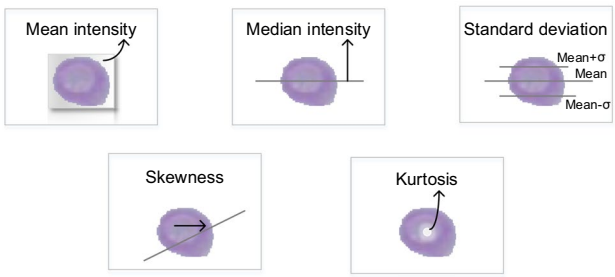


Fig. 5 Visual representation of statistical features in the PTC cell

The fourth attribute, *solidity*, is the ratio of the nuclear area to the convex hull. *Solidity* measures the occupancy of the cell nucleus relative to its convex hull and is calculated using (12). In (12), A represents the nuclear area, while A_c represents the convex hull area of the nucleus.

$$S = \frac{A}{A_c} \tag{12}$$

4) GLCM-Based Texture Features

The relationships between pixel pairs are examined to define tissue structure. In our study, the features extracted in this context are *contrast*, *energy*, and *homogeneity*. Figure 7 is the visualization of GLCM-based texture features via the PTC cell.

$$Contrast = \sum_{i,j} (i-j)^2 P(i,j) \tag{13}$$

The second attribute, energy, measures the regularity of the cell nucleus pattern and is calculated by summing the squared values in the GLCM matrix, as in (14). L in (14) represents the number of input levels.

$$Energy = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} P(i,j)^2 \tag{14}$$

The third metric, *homogeneity*, measures the similarity of gray levels. As the gray tone difference between neighboring pixels decreases, homogeneity increases. Its mathematical equivalent is found in (15).

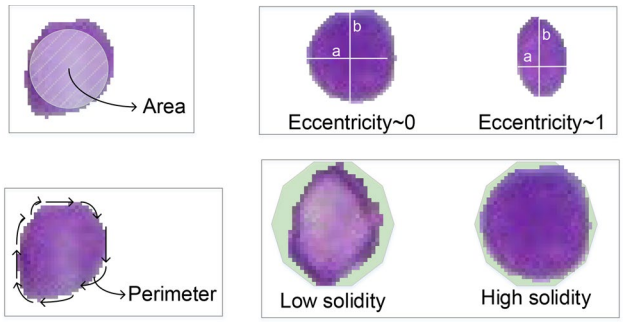


Fig. 6 Visual representation of morphological features in the PTC cell

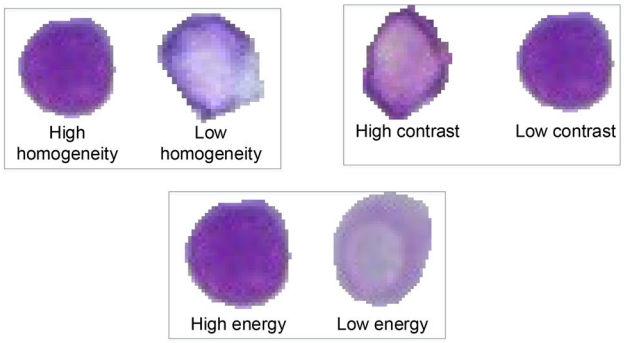


Fig. 7 Visual representation of GLCM-based texture features in the PTC cell

$$Homogeneity = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} \frac{P(i,j)}{1+|i-j|} \tag{15}$$

D. Deep Feature Extraction

Extracting handcrafted features from cytopathological images is intended to approximate the pathologist's decision-making process during pathological diagnosis, while also providing a degree of clinical interpretability. Handcrafted features rely solely on predefined mathematical forms, which can lead to situations where their representational capacity becomes limited. In these cases, deep feature representations are introduced to enhance diagnostic robustness. For example, while Hu moments can be used to calculate deviations from sphericity in the nuclear shape, they are not fully sensitive to fine-scale morphological variations such as small indentations and protrusions in the nuclear membrane. Similarly, GLCM may fail to capture structural irregularities at different scales. To overcome these limitations, deep feature learning is incorporated into the framework. With the proposed CNN-based architecture, deep features are extracted from low-level to high-level abstract features. Structural features are extracted automatically. Figure 8 presents the proposed CNN architecture, which automatically extracts

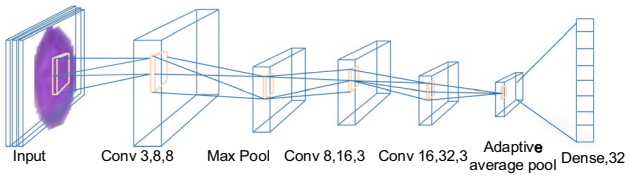


Fig. 8 CNN-based deep feature extraction for cytopathological structures in the proposed model

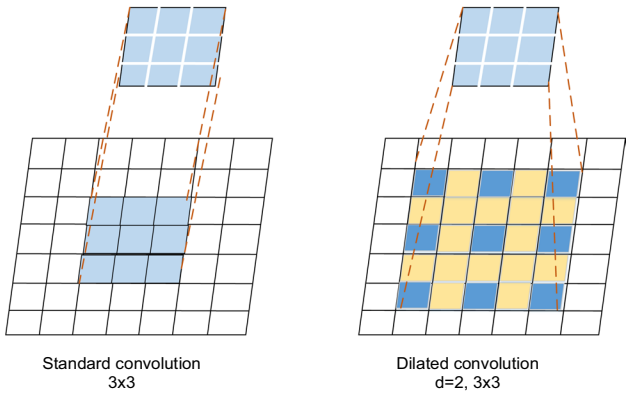


Fig. 9 Sample filtering mechanism of standard 3 × 3 convolution and dilated (d = 2) convolution

multi-layered deep features from cytology images, from low-level structural details to high-level abstract representation.

The first layer in the proposed architecture captures low-level information such as edges and contrast and detects the boundaries of the nuclear membrane. The second layer extracts more complex edge structures and chromatin distribution. The third layer identified higher-order structures by determining irregular nuclear growth and heterogeneous density characteristics of PTC. Our architecture includes a three-layer convolution operation. Mathematically, the convolution layer (16) is calculated.

$$y_{i,j}^{k} = \delta\left(\sum_{m=1}^{M}\sum_{n=1}^{N} x_{i+m,j+n}.w_{m,n}^{k} + b^{k}\right) \tag{16}$$

$$z_k = \frac{1}{H.W}\sum_{i=1}^{H}\sum_{j=1}^{W} y_{i,j}^{k} \tag{17}$$

Here, $x_{i,j}$ are the input pixels, $w_{m,n}^{k}$ is the weights of the k-th filter, b^{k} is the bias, and δ is the relu activation. Following convolution, global pooling is applied, reducing the feature map for each filter to a single value. (17) is the mathematical formulation of the global pooling process, where H and W are the dimensions of the feature map and z_k is the global feature of the k-th channel. Figure 9 represents the architecture we used for deep feature extraction.

We obtained a 32-dimensional deep feature vector by completing the points where handcrafted features were insufficient in PTC diagnosis. This way, our model provided a more comprehensive and robust representation. By combining the explainability of classical features and the ability to distinguish deep features, we obtained a diagnostic kit that is more reliable and less prone to error in PTC diagnosis.

E. Hybrid Classification with Feature Fusion

Feature extraction numerically describes the cell structure in cytopathology images. For this purpose, the features extracted in this study include GLCM-based tissue features, statistical density measures, and Hu moments, in addition to traditional features commonly known in literature. The features we obtained in this section are directly related to the biological data observed by pathologists under the microscope, and therefore, the resulting framework provides clinically interpretable and explainable outputs. However, these handcrafted features alone may be insufficient for capturing heterogeneous malignant patterns in PTC cases. Furthermore, their discriminatory power may be limited in borderline cases. To overcome these limitations, deep feature representations are integrated into the proposed method alongside traditional features. These representations capture subtle variations in nuclear morphology that are difficult to distinguish visually. Using only deep features, however, does not provide explainability due to its black-box nature and poses challenges in integrating them into the pathologist's decision-making process. For these reasons, we designed a hybrid classification pipeline using both classical and deep features. The two feature sets were combined at the vector level, yielding a 51-dimensional hybrid feature space for each patch. We used an innovative attention pooling strategy to transfer the resulting hybrid features to the slide level. We modeled the contribution of each patch to the classification decision using learnable weights. The attention coefficient α_i for each patch feature x_i is calculated using (18). Here, w^T is the learned weight vector. It is the parameter optimized during the model's training process. All patches are weighed by (19). This innovative method allows the model to assign higher importance to malignant cells, thus allocating more attention.

$$\alpha_i = \frac{\exp(w^T x_i)}{\sum_j \exp(w^T x_j)} \tag{18}$$

$$X = \sum_i \alpha_i x_i \tag{19}$$

Structural anomalies such as irregular nuclear contours, and nuclear enlargement seen in PTC require the development of more flexible mechanisms. The fixed grid structure of the classical CNN cannot adequately represent the irregular cells in PTC. Therefore, in our study, we designed special architecture that includes deformable and dilated convolution layers in the classification phase. Dilated convolution adds gaps between the core elements. While the filter size does not increase, the receptive field is expanded. PTC allows for multi-scale examination of structures around the nucleus. Standard convolution is defined by (20) and dilated convolution by (21). Dilated convolution is an extended version of standard convolution. In (20), $y(p_0)$ is the output value, p_0 is the normal pixel, k is the position vector within the filter, $x(.)$ represents the input image, and $w(k)$ represents the weights. The d value in (21) is the dilation rate.

$$y(p_0) = \sum_{k \in \mathbb{R}} w(k).x(p_0 + k) \tag{20}$$

$$y_d(p_0) = \sum_{k \in \mathbb{R}} w(k).x(p_0 + d.k) \tag{21}$$

With dilation convolution, when the nucleus size was kept unchanged from $3x3$ and $d = 2$, the filter's receptive field became $7x7$. This expansion of the receptive field allows for a more comprehensive examination of nuclear clusters, tissue density differences, and cell microarchitecture within the PTC. Dilation convolution allows for the effective capture of nuclear size changes and chromatin density at different scales within PTC cells. This allows for the capture of not only the local shape of the nuclei but also their spatial relationships with surrounding cells.

Figure 9 shows that the standard $3x3$ convolution filter densely samples neighboring pixels, whereas dilated convolution ($d = 2$) adds space between filter elements to collect information from a wider area. This expands the receptive field without increasing the number of parameters, allowing for the capture of more macro-scale structural features such as tissue architecture.

Deformable convolution adds flexible offsets to standard convolution as shown in Fig. 10. It allows filter points to slide freely on the input image. Its mathematical formulation is defined by (22). Here, $\mathcal{R}$ denotes the receptive field of the

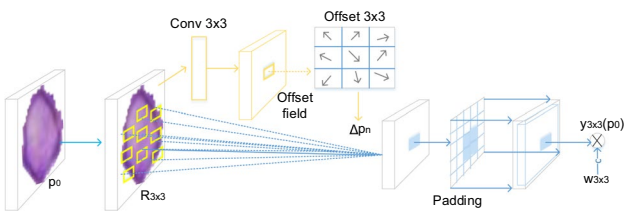


Fig. 10 Visualization of deformable convolution

filter, p_0 indicates the central position, p_n corresponds to the standard sampling points, and Δp_n represents the learned offset values. Thanks to these offsets, the model can learn variations specific to biological structure by going beyond the fixed grid structure. Because filters in deformable convolution operate on a flexible grid rather than a fixed grid, our study allows for more accurate modeling of irregular nuclear boundaries in PTC. While nuclear grove structures in PTC nuclei cannot be captured with standard convolution, deformable convolution allows the filter to be bent to this structure. Thus, irregular cell geometries in PTC are more accurately represented.

$$y_{def}(p_0) = \sum_{p_n \in \mathcal{R}} w(p_n).x(p_0 + p_n + \Delta p_n) \tag{22}$$

Feature maps from both convolutions were obtained, resulting in a hybrid output $\hat{y} = \left[y_d \oplus y_{def}\right]$. The hybrid use of dilated and deformable convolutions provides both geometric adaptation and multi-scale contextual information. In PTC diagnosis, this feature allows for the capture of both nuclear cluster structure and nucleus contour irregularity. While classical convolutional filters struggle to detect nuclear irregularities and chromatin condensation in cytopathological images, the hybrid structure provides increased capturing capacity.

4 Experimental results

A diagnostic tool has been developed using cytology images to diagnose PTC without requiring pathologists to add their own interpretation. The proposed method consists of two different stages. In the first stage, the relevant regions are identified from cytology images using YOLOv11, and these regions are then collected as a separate data source. In the second stage, a classification mechanism is created using both the relevant regional data and the complete cytology data, supported by abstract and concrete features. Grad-CAM visualizations are also presented in detail in this section to demonstrate the performance results and their interpretability.

A. Implementation Details

Our study was conducted on a computer with an NVIDIA GeForce RTX 4070 GPU, 8 GB GDDR6 memory, and 32 GB RAM. We implemented the hyperparameter settings of the model as follows: 25 epochs, Adam optimizer, 64 batch size, 0.001 learning rate. The dataset used was divided into 70/15/15 for training, validation, and testing. To improve model robustness and mitigate overfitting, extensive data augmentation techniques were applied, including random rotations, horizontal and vertical flips, Gaussian blur, sharpness adjustment, grayscale conversion, and color jittering. In addition, a k-fold cross-validation strategy at the slide level was employed to ensure a reliable and unbiased evaluation. This approach prevents data leakage by ensuring that all samples from a given slide are contained within a single fold.

B. Evaluation Metrics

To evaluate model performance, accuracy, precision, recall, specificity, f1-score, and AUC(Area Under the Curve) metrics were calculated. Accuracy, precision, recall, specificity, f1-score, and metrics are all listed in order (23–27). Accuracy indicates the percentage of correct predictions made by the model among all predictions; precision indicates the percentage of malignant samples that it can label as malignant; recall indicates the percentage of truly malignant patients that it can label as malignant; specificity indicates the percentage of benign samples that it can label as benign; and f1-score ensures that malignant cases are not missed and that unnecessary malignant alarms are not generated. In these critical cases where we detect PTC, it is crucial for reducing missed cases. Specificity is the ability to correctly distinguish benign samples. It indicates the percentage of truly benign samples correctly labeled as benign as in (25). This reduces misdiagnosis and unnecessary biopsy. The F1-score is a balanced combination of precision and recall as in (26). If precision is very high but recall is low, it indicates that the model is selective but is missing malignant cases. If precision is low and recall is high, it can be said that the model mostly interprets all cases as cancer, leading to false alarms. A high F1-score indicates that no cases were missed, and no false alarms were generated. AUC indicates that the model can distinguish between malignant and benign samples without needing any threshold value as in (30). A high AUC indicates that the model can distinguish benign and malignant samples well, regardless of the decision threshold. All these metrics not only reveal the classification performance but also demonstrate the clinical reliability of the model.

In Eqs. (23) to (29), TP is (true positive). A malignant sample is predicted as malignant, and a benign sample is predicted as benign, TN is (true negative). A benign sample is falsely predicted as malignant, FP is (false positive), and a malignant sample is falsely predicted as benign, FN is (false negative). In (28), the true positive rate (TPR) of the meta-metric formulation shows the rate at which truly malignant patients are identified as actually malignant, while in (29) the false positive rate (FPR) shows the rate at which healthy patients are falsely labeled as malign. Both rates are quite important for disease diagnosis tools.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{23}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{24}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{25}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{26}$$

$$F1 - \text{Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{27}$$

$$\text{TPR} = \frac{TP}{TP + FN} \tag{28}$$

$$\text{FPR} = \frac{FP}{FP + TN} \tag{29}$$

$$\text{AUC} = \int_0^1 TPR(FPR)dFPR \tag{30}$$

C. Quantitative Results

(1) Model Performance

To present the performance of the proposed model and demonstrate its clinical usability, it was experimentally tested on the dataset in terms of accuracy, precision, recall, specificity, and f1-score metrics. Figure 11 represents the training and validation accuracy and loss curves of the proposed model in the dataset. Examination of the curves shows that the performance is very similar. This indicates that the model has the ability to generalize.

Experimental results demonstrate that YOLOv11 outperforms YOLOv8 in ROI detection performance. As shown in the precision-recall curves, YOLOv11 achieves higher recall while maintaining competitive precision, indicating improved detection of diagnostically relevant regions. Quantitative results further confirm superior mean avearge precision-mAP@0.5 performance, supporting the suitability of YOLOv11 for ROI extraction in cytopathology images.

Figure 12 compares the performance curves of the YOLOv8 and YOLOv11 models during the training process. YOLOv11 achieved higher values compared to YOLOv8, especially in recall (sensitivity) and mAP50-95 metrics, and exhibited more stable convergence throughout training. This indicates that YOLOv11 produces fewer false negatives in the detection of pathological areas and has superior segmentation sensitivity.

Figure 13 shows the precision-recall curves of both models. The area under the curve (AUC) of YOLOv11 is larger, revealing that the model strikes a better balance

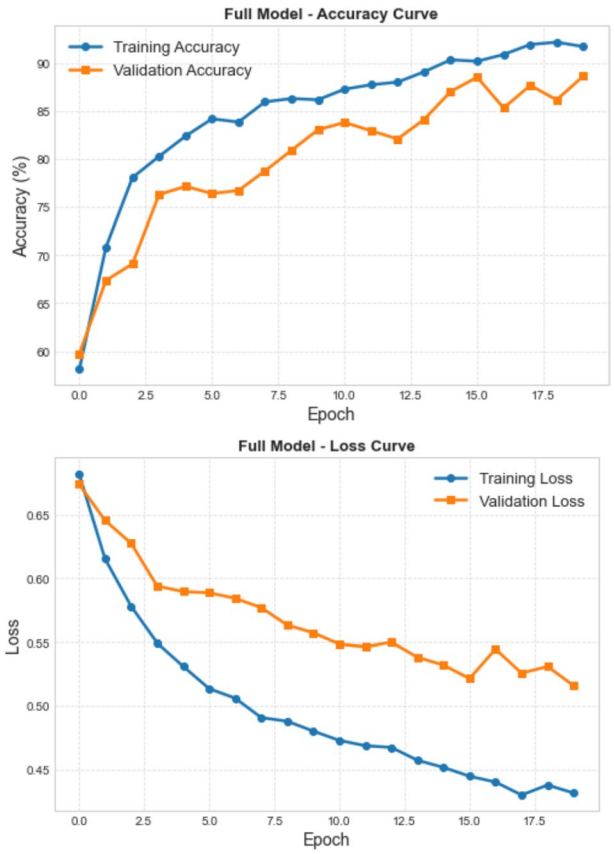


Fig. 11 Train vs validation accuracy and loss ratios of the proposed methodology

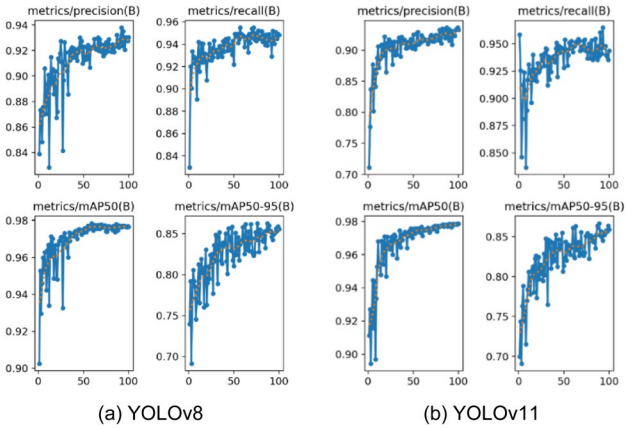


Fig. 12 Precision, recall, and mAP metrics curves of YOLOv8 and YOLOv11 for ROI detection

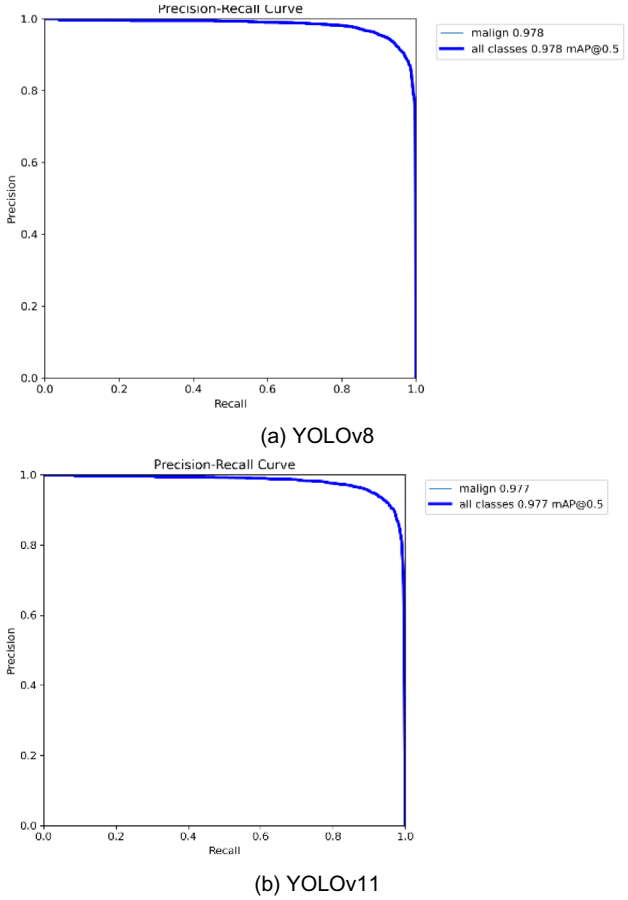


Fig. 13 Precision-recall curves of YOLOv8 and YOLOv11 for ROI detection. YOLOv11 demonstrates improved convergence stability and superior detection performance

Table 2 Literature summary of artificial intelligence-based approaches in thyroid cancer diagnosis

Metrics	YOLOv8	YOLOv11	Improvement
Box Detection			
Precision (B)	0.9381	0.9367	−0.0014
Recall (B)	0.9518	0.9649	+0.0131
mAP50 (B)	0.9784	0.9788	+0.0004
mAP50-95 (B)	0.8624	0.8664	+0.0040
Mask Segmentation			
Precision (M)	0.9381	0.9367	−0.0014
Recall (M)	0.9518	0.9649	+0.0131
mAP50 (M)	0.9784	0.9788	+0.0004
mAP50-95 (M)	0.8624	0.8664	+0.0040
Loss Values			
Box Loss	0.5778	0.5873	+0.0095
Segmentation Loss	0.6321	0.6415	+0.0094
Classification Loss (CLS) Loss	0.3311	0.3453	+0.0142
Distribution Focal Loss (DFL) Loss	0.7953	0.7968	+0.0015

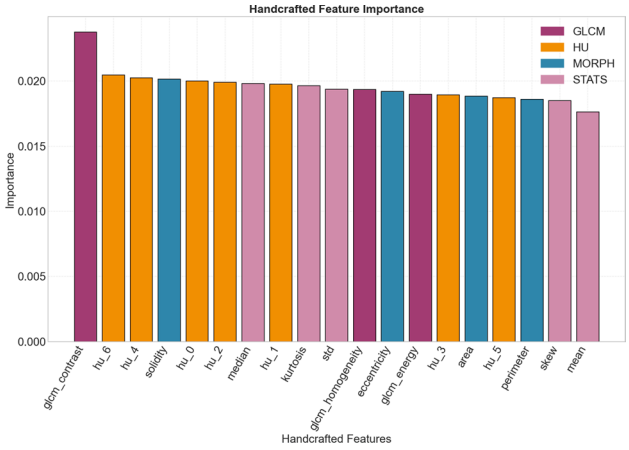


Fig. 14 Order of importance of handcraft features

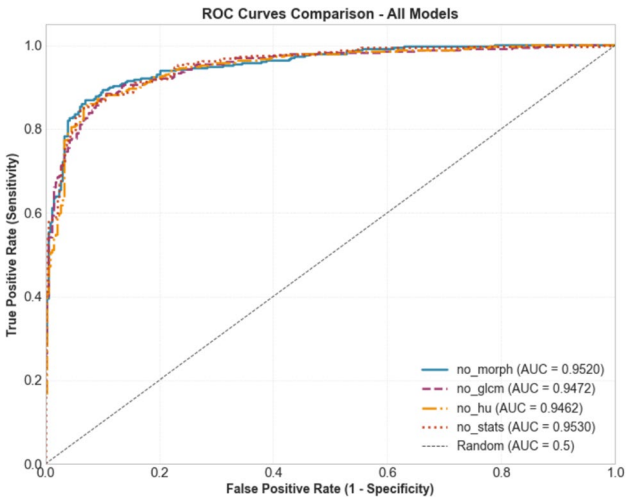


Fig. 15 ROC curve comparison for ablation study results showing the contribution of different handcrafted features

between precision and recall. These findings demonstrate that YOLOv11 can achieve high recall values with low precision loss and is a more reliable option for ROI detection in cytopathology images.

YOLOv11 is the most current and advanced version of the YOLO series, incorporating architectural improvements and optimizations compared to previous versions. YOLOv8, on the other hand, is a widely used and reliable reference model in the literature. This study compares the two latest YOLO architectures to demonstrate the effectiveness of the proposed approach. Table 2 show that YOLOv11 exhibits superior performance in ROI detection compared to YOLOv8. Specifically, a 0.0131 improvement in recall (sensitivity) and a 0.0040 increase in the mAP50-95 score were achieved. These findings demonstrate that YOLOv11 is more successful in detecting pathological areas, especially

in challenging cytopathology images, and produces fewer false negatives.

(2) Feature Analysis

To quantify the contribution of each handcrafted feature, a permutation-based feature significance analysis was performed. A random forest classifier was trained on the combined feature space, and feature significance scores were calculated by measuring the decrease in classification performance after random permutation of each feature. This approach allows for an unbiased estimation of feature relevance without relying on model-specific assumptions. The corresponding results are illustrated in Fig. 14, highlighting the most discriminative feature groups.

Figures 14 and 15 show that the statistical properties and Hu moments contribute more to the proposed method in PTC tumor diagnosis. In particular, the contrast feature is prominent in determining whether there is heterogeneity or regularity in the tissue structure. Furthermore, the GLCM properties and the removal of Hu moments from the model make it difficult to discover important semantic structures and are effective in determining intra-tissue variations with GLCM. Statistical and morphological properties support model stability as complementary information for the model. The ROC curve results of the ablation analysis are presented in Fig. 15.

3) Robustness analysis

The proposed model demonstrates stable performance under moderate noise levels, with gradual degradation observed as SNR decreases. This indicates strong robustness and generalization capability in realistic noisy imaging conditions. The results of this analysis are presented in Fig. 16, illustrating model performance across different signal-to-noise ratio (SNR) levels. The overall ROC performance is presented in Fig. 17. The performance of the proposed method was evaluated by progressively adding Gaussian noise to the noise-free data up to an SNR level of 10 dB (i.e., 30 dB, 25 dB, 20 dB, 15 dB, and 10 dB). Table 3 and Fig. 16 show the evaluation of the model in terms of accuracy, AUC, and f1-scoremetrics for each SNR level. As can be seen from the ROC curve in Fig. 17, the proposed model maintains its class discrimination power despite moderate noise levels, and its performance gradually decreases as the noise level increases. In scenarios where the integrity of the main data is compromised by large amounts of noise such as 15 dB and 10 dB, performance loss occurred due to the distortion of fine nuclear structures and the loss of morphological details in the PTC visualizations. The decrease in f1-score as noise level increased indicated that this hindered the model's ability to accurately classify PTC malignant cases. However, the fact that the AUC remained high suggests that interclass discrimination was still maintained.

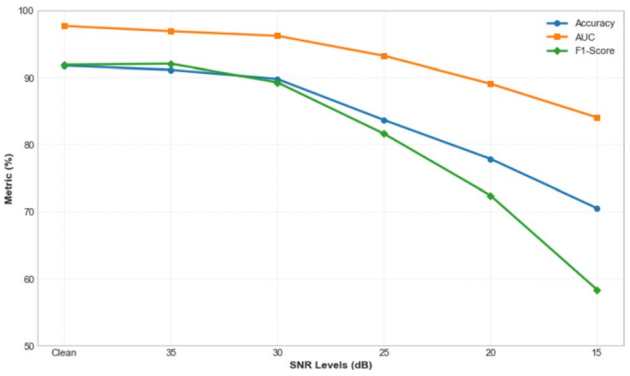


Fig. 16 Model performance under different SNR levels

D. Ablation study

Because our model consists of different components, an ablation study was conducted to examine the contribution

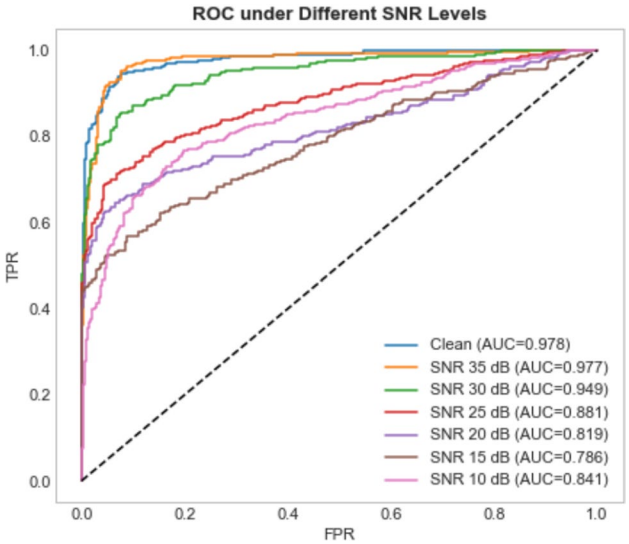


Fig. 17 ROC curves under different SNR levels

Table 3 Performance comparison of the proposed model under progressive SNR degradation Levels

SNR	Accuracy	AUC	F1-Score
Clean	0.9173	0.9764	0.9186
35 dB	0.9110	0.9685	0.9201
30 dB	0.8975	0.9616	0.8924
25 dB	0.8363	0.9318	0.8158
20 dB	0.7782	0.8904	0.7238
15 dB	0.7048	0.8401	0.5831
10 dB	0.6391	0.8311	0.4433

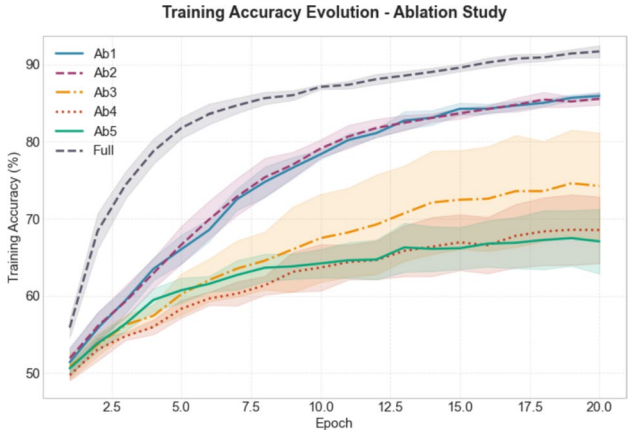


Fig. 18 Comparison of the training accuracy rate of our model with ablation studies

of each component to the model's performance. In the first stage, we removed the handcrafted extracted features from the proposed model and referred to them as *Ab1*. This allows us to observe the main contribution of the handcrafted features in the model. We named the model obtained by extracting the abstract features extracted using CNN as *Ab2*. We labeled the model obtained by removing both CNN and handcrafted features as *Ab3*. This way, the patch dataset will be removed from the model, and the hybrid classification process will be conducted using only the whole-slide image. Our hybrid classification model consists of multi-scale dilated and deformable convolution layers. By removing the dilated convolution layers and adding normal convolution (dilated rate = 1), we obtained the ablation study, *Ab4*. Finally, we removed the deformable convolution from our model and replaced it with normal convolution, completing our ablation study with the *Ab5* model. Figure 18 presents a visual representation of the proposed model's trend in train accuracy during fivefold cross-validation and the ablation models' trend in train accuracy. Figure 19 shows comparative graphs of the proposed model and ablation models in terms of precision, recall, F1-score, and specificity metrics, again during fivefold cross-validation. As can be seen from these graphs, the proposed model has demonstrated superiority over ablation models in all metrics (Fig. 20).

All evaluation metrics presented in Table 4, including accuracy, precision, recall, F1-score, specificity, AUC, and loss, are compared within the ablation study framework. Consistent with the accuracy comparisons illustrated in Figs. 18, 19, 21 and 22 further confirm that the Ab3 configuration exhibits the most significant performance degradation across all metrics. Specifically, validation accuracy drops from 0.8797 (Our model) to 0.6334 (Ab3), along with substantial decreases in precision (0.8652 → 0.6053), specificity (0.8574 → 0.4866), and AUC (0.9502 → 0.6962).

Following Ab3, the performance ranking based on validation accuracy is observed as Ab1 (0.7993), Ab2 (0.8011), Ab4 (0.6221), and Ab5 (0.5912). Ab5 exhibits a larger performance drop than Ab4, contrary to the previous ordering. The highest performance across all metrics is achieved by the proposed model (Accuracy: 0.8797, Precision: 0.8652, Recall: 0.9021, F1-score: 0.8823, Specificity: 0.8574, AUC: 0.9502, Loss: 0.5321), confirming the effectiveness of the full architecture.

The severe degradation observed in Ab3 highlights the critical importance of hybrid feature extraction from patches in the proposed framework. The performance reductions in Ab1 and Ab2 (approximately 9.1% and 8.9% accuracy drop

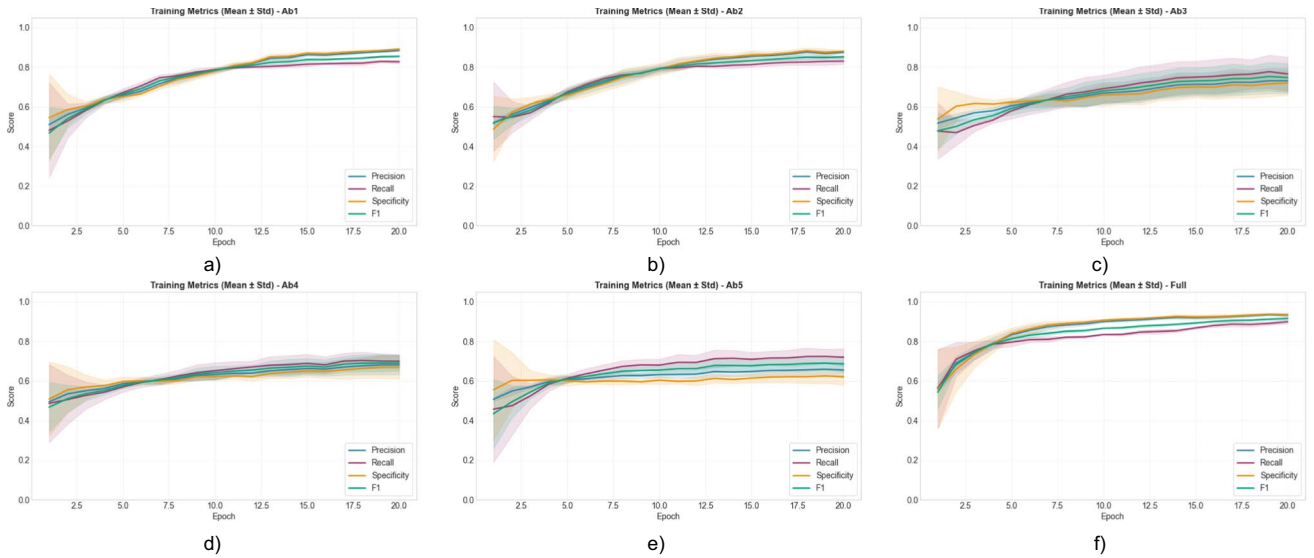


Fig. 19 Comparison of all evaluation metrics across the ablation studies **a** Ab1, **b** Ab2, **c** Ab3, **d** Ab4, **e** Ab5, and **f** our model

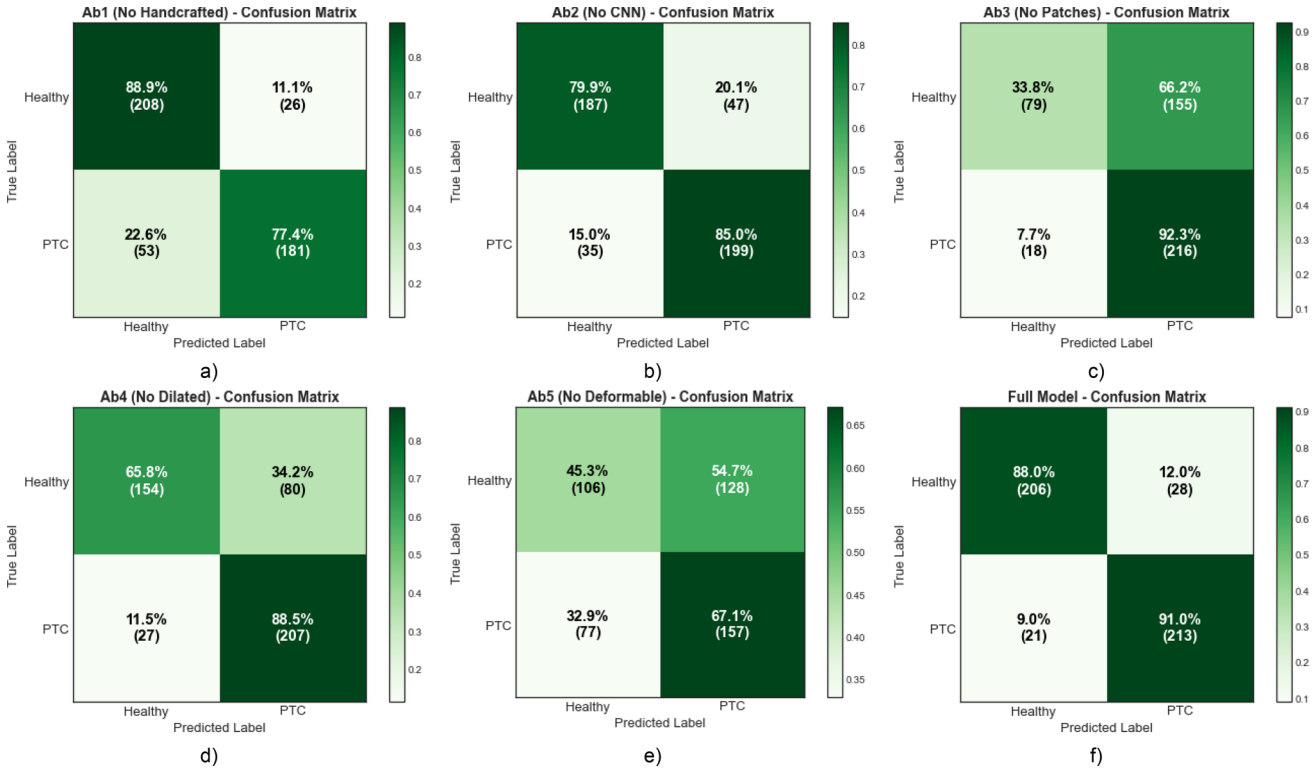


Fig. 20 Comparison of confusion matrix of f) our model with **a** Ab1, **b** Ab2, **c** Ab3, **d** Ab4 and **e** Ab5 ablation studies

Table 4 Quantitative performance metrics of proposed model and ablation variants under fivefold cross-validation

Model	Accuracy	Precision	Recall	F1-Score	Specificity	AUC	Loss	Compt Time (sec/epoch)	Inf Time (msec)
Ab1 Train	0.8537 ± 0.0026	0.8766 ± 0.0064	0.8233 ± 0.0044	0.8491 ± 0.0023	0.8840 ± 0.0071	0.9276 ± 0.0037	0.5001 ± 0.0038	34.2990 ± 0.2658	0.5602 ± 0.0055
Ab1 Val	0.7993 ± 0.0151	0.8034 ± 0.0166	0.7938 ± 0.0435	0.7976 ± 0.0200	0.8048 ± 0.0264	0.8852 ± 0.0121	0.5752 ± 0.0091		
Ab2 Train	0.8499 ± 0.0090	0.8673 ± 0.0119	0.8265 ± 0.0173	0.8463 ± 0.0098	0.8733 ± 0.0142	0.9260 ± 0.0095	0.5031 ± 0.0089	34.1221 ± 0.0965	0.5697 ± 0.0035
Ab2 Val	0.8011 ± 0.0155	0.7991 ± 0.0287	0.8072 ± 0.0338	0.8022 ± 0.0157	0.7950 ± 0.0411	0.8796 ± 0.0160	0.5810 ± 0.0098		
Ab3 Train	0.7336 ± 0.0766	0.7210 ± 0.0716	0.7602 ± 0.0893	0.7397 ± 0.0785	0.7070 ± 0.0724	0.8064 ± 0.0853	0.5960 ± 0.0481	34.2548 ± 0.0777	0.5975 ± 0.0049
Ab3 Val	0.6334 ± 0.0506	0.6053 ± 0.0443	0.7804 ± 0.1052	0.6782 ± 0.0549	0.4866 ± 0.1093	0.6962 ± 0.0643	0.6658 ± 0.0223		
Ab4 Train	0.6866 ± 0.0452	0.6816 ± 0.0497	0.7038 ± 0.0399	0.6922 ± 0.0426	0.6694 ± 0.0584	0.7504 ± 0.0514	0.6280 ± 0.0268	33.7460 ± 0.0973	0.5565 ± 0.0006
Ab4 Val	0.6221 ± 0.0468	0.5931 ± 0.0454	0.8005 ± 0.0259	0.6802 ± 0.0299	0.4437 ± 0.0955	0.6813 ± 0.0713	0.6711 ± 0.0273		
Ab5 Train	0.6593 ± 0.0322	0.6446 ± 0.0287	0.7095 ± 0.0385	0.6754 ± 0.0327	0.6092 ± 0.0293	0.7219 ± 0.0409	0.6410 ± 0.0171	32.2279 ± 0.1168	0.5465 ± 0.0008
Ab5 Val	0.5912 ± 0.0225	0.5673 ± 0.0174	0.7717 ± 0.0295	0.6537 ± 0.0188	0.4107 ± 0.0384	0.6234 ± 0.0263	0.6887 ± 0.0075		
Our model Train	0.9123 ± 0.0059	0.9330 ± 0.0054	0.8885 ± 0.0094	0.9102 ± 0.0063	0.9362 ± 0.0052	0.9748 ± 0.0028	0.4364 ± 0.0051	35.2492 ± 1.6166	0.5833 ± 0.0022
Our model Val	0.8797 ± 0.0186	0.8652 ± 0.0303	0.9021 ± 0.0375	0.8823 ± 0.0184	0.8574 ± 0.0413	0.9502 ± 0.0132	0.5321 ± 0.0069		

relative to the proposed model, respectively) indicate that handcrafted and CNN-based features provide complementary information for classification. Notably, the removal of convolutional modules (Ab4 and Ab5) leads to more substantial performance degradation than initially suggested, with accuracy decreases of approximately 29.1 and 32.7%, respectively, indicating their critical role in multi-scale feature representation rather than a minor contribution.



Fig. 21 Comparison of the training evaluation metrics of our model with ablation studies

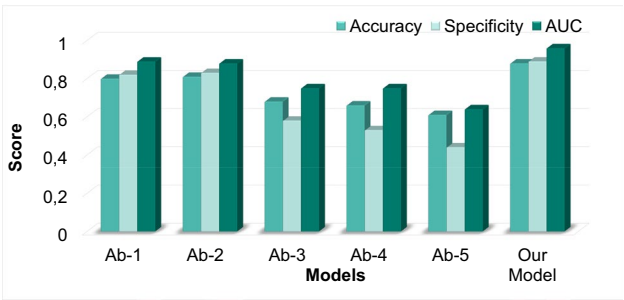


Fig. 22 Comparison of the validation evaluation metrics of our model with ablation studies

Furthermore, analysis of the loss values shows that the proposed model achieves the lowest validation loss (0.5321), while all ablation scenarios exhibit higher loss values, with the most significant increase observed in Ab5 (0.6887). This further supports the conclusion that each component contributes meaningfully to the overall optimization and generalization capability of the model.

Despite incorporating attention-based aggregation, the proposed model maintains a highly efficient inference time of approximately 0.55 ms per sample, ensuring real-time applicability in cytology workflows. Although a marginal increase in computational cost is observed during training, this overhead remains within an acceptable range and does not affect deployment efficiency, where inference speed is the primary constraint.

The confusion matrix in Fig. 20 shows a false positive count of 26, indicating that the model is prone to misclassifying negative examples. The false negative count is 53, which is relatively higher and indicates difficulty in correctly identifying positive samples. This suggests that while the model achieves reasonable overall performance, class separation is still not fully optimal. In Fig. 20b, when only handcrafted features are used, the number of misclassifications changes, with false positives increasing to 47 and false negatives decreasing to 35. This demonstrates that handcrafted features provide a more balanced but still imperfect classification performance. In Fig. 20c, the model performs relatively better in terms of false positives (155) but shows a significant reduction in false negatives (18), indicating a bias toward predicting the positive class more accurately, though at the cost of increased misclassification of negatives. In Fig. 20d, the false positive count is 80, while the false negative count is 27. Although the model improves recall compared to earlier configurations, the relatively high false positive rate suggests that class boundary learning remains unstable. In Fig. 20e, when dilated convolutions are used, the false positives decrease to 128, while false negatives increase significantly to 77. This indicates that dilated convolutions improve sensitivity to negative samples but negatively affect positive class recognition, leading to an unbalanced trade-off. Finally, Fig. 20f shows that the proposed model achieves a strong balance between classes, with only 28 false positives and 21 false negatives. This demonstrates a clear improvement over all ablation scenarios, highlighting the effectiveness of the full architecture in achieving robust and balanced classification performance.

Figures 18, 19, 21, 22 and 23 represent visualizations of the data in Table 4. Figure 24 shows the ROC curve comparison of our model with all ablation studies. Examining this curve, which visualizes the model's discrimination between classes, the lowest value is observed in Ab3 with an AUC of 0.675. The model still maintains a reasonable discriminatory capacity in Ab1 with an AUC of 0.880, and in Ab2 with an AUC of 0.871, indicating that both handcrafted and CNN-based feature representations individually contribute meaningful discriminative information.

Ab4 and Ab5 exhibit lower performance levels with AUC values of 0.662 and 0.604, respectively, suggesting that the removal or modification of these components leads to a notable reduction in class separability and weakens overall feature representation capability. These results indicate that each module contributes in a complementary manner to the final performance, even when individual components alone are insufficient to achieve high discrimination.

Our model, however, achieves the highest performance with an AUC of 0.954, demonstrating strong interclass separability and robust generalization capability. This confirms that the integration of all components significantly enhances the model's discriminative power compared to all ablation configurations.

E. Comparison with State-of-the-Art Methods

The proposed model was experimentally evaluated by comparing it with deep learning-based state-of-the-art (SOTA) methods in the literature. The aim of this analysis was not only to compare performance metrics but also to

Table 5 Quantitative comparison with SOTA

Metric	Swin-T	ConvNeXt	ResNet50	EffNet-B0	Our Model
Acc	0.8899	0.8700	0.6147	0.6300	0.8991
Prec	0.9352	0.8712	0.5720	0.5750	0.8805
F1	0.8839	0.8698	0.7028	0.7293	0.9015
Spe	0.9419	0.8716	0.3180	0.2630	0.8746
AUC	0.9228	0.9235	0.6882	0.8361	0.9541
Param(M)	28.29	28.59	25.56	5.29	1.85
Inf. Time (ms/img)	1.6	1.7	1.4	1.3	0.55

examine the overall behavior and generalizability capabilities of the models. To create a fair experimental environment, the same data splitting strategy, the same data preprocessing steps, and the same training hyperparameters were applied to all models.

As shown in Table 5, the proposed model achieved higher overall performance compared to other models. Figure 25 shows the accuracy rates, with the proposed model reaching 89.91%, followed by the Swin Transformer model at 88.99%, the ConvNeXt model at 87.00%, and the lowest performance being shown by the ResNet50 and EfficientNetB0 models at 61.47% and 63.00%, respectively.

Figure 26 presents the ROC curve comparatively. The proposed model exhibits a more consistent and robust classification behavior against different threshold values with an AUC value of 0.9541. Although the Swin Transformer (0.9228) and ConvNeXt (0.9235) models also showed competitive performance, they lagged behind the proposed model. EfficientNetB0 (0.8361) demonstrated moderate discrimination capability, while ResNet50 (0.6882) had the lowest AUC value, failing to provide sufficient discrimination between classes.

A similar trend is observed in terms of precision and F1-score. While the proposed model offered balanced performance with 0.8805 precision and 0.9015 F1-score, the Swin Transformer and ConvNeXt models produced results close to these values. In contrast, the ResNet50 and EfficientNetB0 models showed a significant decrease, particularly in specificity (ResNet50: 0.3180, EfficientNetB0: 0.2630), indicating their weakness in negative class discrimination (Fig. 25).

When evaluated in terms of model complexity and computational cost, the proposed model, despite being the lightest structure with only 1.85 M parameters, achieved the highest performance. In contrast, Swin Transformer (28.29 M) and ConvNeXt (28.59 M) have significantly higher parameter numbers. EfficientNetB0 (5.29 M), although more compact, lagged behind the proposed model in terms of performance. When examining the inference time, the proposed model was the fastest at 0.55 ms/image, while Swin Transformer and ConvNeXt were slower at 1.6 ms and 1.7 ms, respectively.

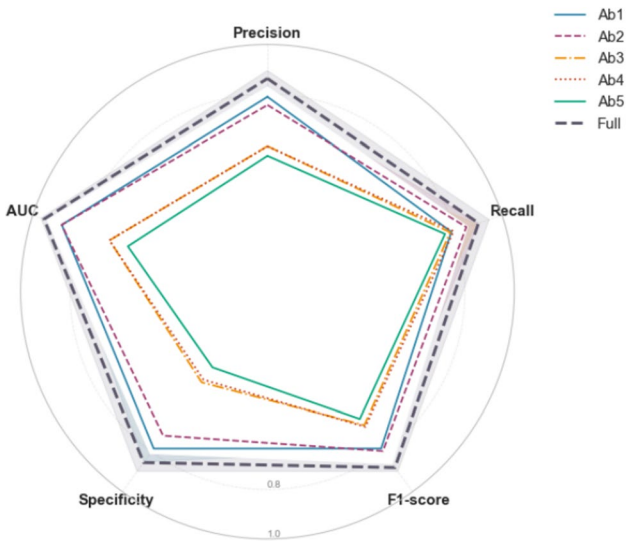


Fig. 23 Comparison of all metrics of our model with ablation studies

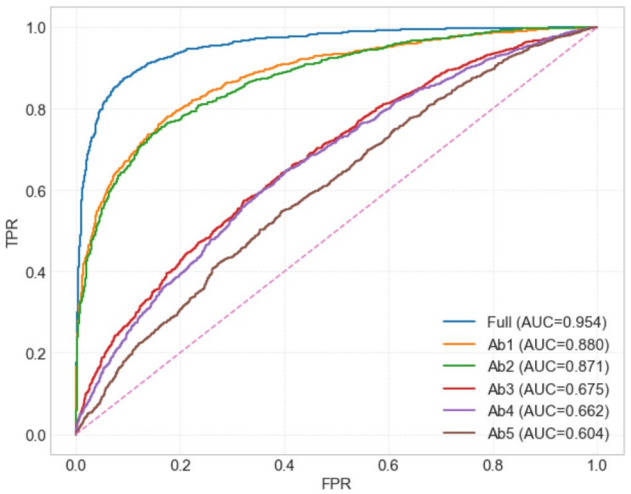


Fig. 24 Comparison of the ROC curves of our model with ablation studies

Considering all these findings together, it is clear that the proposed approach provides a significant advantage not only in terms of performance but also in terms of model lightness

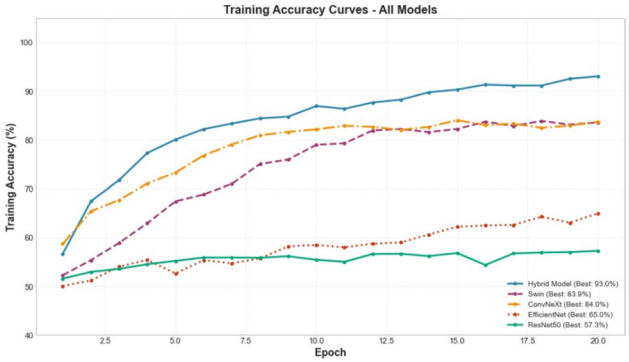


Fig. 25 Training accuracy curves: comparison of proposed model and SOTA methods

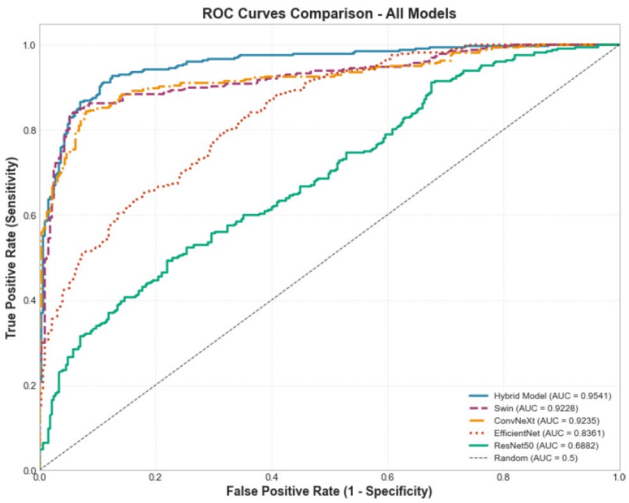


Fig. 26 ROC curves: proposed model vs. SOTA methods

and suitability for real-time applications (Figs. 23, 24 and Table 4).

F. Qualitative results

The qualitative perfoance of the proposed method, Grad-CAM visualizations are provided to illustrate the regions of interest attended by the model in whole-slide images. It is evident that the regions focused on by the model during the decision-making process are consistent with those observed by pathologists in clinical practice. The model's attention intensity is particularly high in regions where nuclear deformities and cytoplasmic features are prominent, which not only increases the model's classification accuracy but also enhances its explainability. Grad-CAM maps also demonstrate that the model focuses only on pathologically significant regions and ignores irrelevant background. Quantitative explainability evaluation revealed that our model achieves exceptional coverage of 91.8% (95% CI [77.3%, 106.2%]) of pathologist-annotated regions, demonstrating its ability to identify nearly all diagnostically relevant areas. The Dice coefficient of 0.652 ± 0.151 (95% CI [0.356, 0.948]) indicates substantial spatial agreement exceeding conventional 'good' agreement thresholds (0.6), while the moderate IoU of 0.503 ± 0.168 (95% CI [0.174, 0.831]) reflects expected discrepancies between coarse YOLO bounding boxes and fine-grained Grad-CAM activations. Notably, the Center-aligned Dice (0.595 ± 0.134, 95% CI [0.333, 0.857]) substantially exceeds Weighted IoU (0.371 ± 0.116, 95% CI [0.143, 0.599]), confirming the model's clinically desirable focus on lesion centers rather than boundaries. The narrow confidence intervals across all metrics indicate stable and reliable explainability performance without outlier-dominated results. Optimal threshold analysis suggests further improvement potential (Max IoU: 0.428 ± 0.137, 95% CI [0.159, 0.696]), indicating that adaptive post-processing could enhance clinical utility. Collectively, these quantitative metrics provide rigorous validation of model explainability, transforming previous qualitative Grad-CAM visualizations into statistically grounded evidence supporting clinical reliability. In addition to qualitative assessment, a quantitative localization evaluation was performed by comparing Grad-CAM activation maps with expert pathologist-annotated bounding boxes (generated using a YOLO-based detector trained on pathologist labels) in Fig. 28. Representative examples corresponding to high, moderate, and low IoU values are provided in Fig. 27 to illustrate the variability in localization performance across different cases. These qualitative observations are consistent with the quantitative explainability results reported in Table 6. Grad-CAM and whole-slide image (WSI) overlay results further confirm that the proposed model achieves clinically interpretable attention maps, which can effectively support pathologists in decision-making processes.

5 Discussion

This study presents a clinically inspired, explainable deep learning framework for the accurate diagnosis of papillary thyroid carcinoma (PTC) from fine-needle aspiration (FNA) cytopathology images. Unlike conventional end-to-end black-box models, our approach mimics the diagnostic workflow of pathologists by integrating region-based analysis, handcrafted bio-morphological features, and deep abstract representations. This design not only improves classification performance but also enhances clinical trustworthiness and interpretability.

Compared to prior studies that rely solely on transfer learning or end-to-end CNN architectures [4–11], our method demonstrates superior performance by combining domain knowledge with data-driven learning. While several

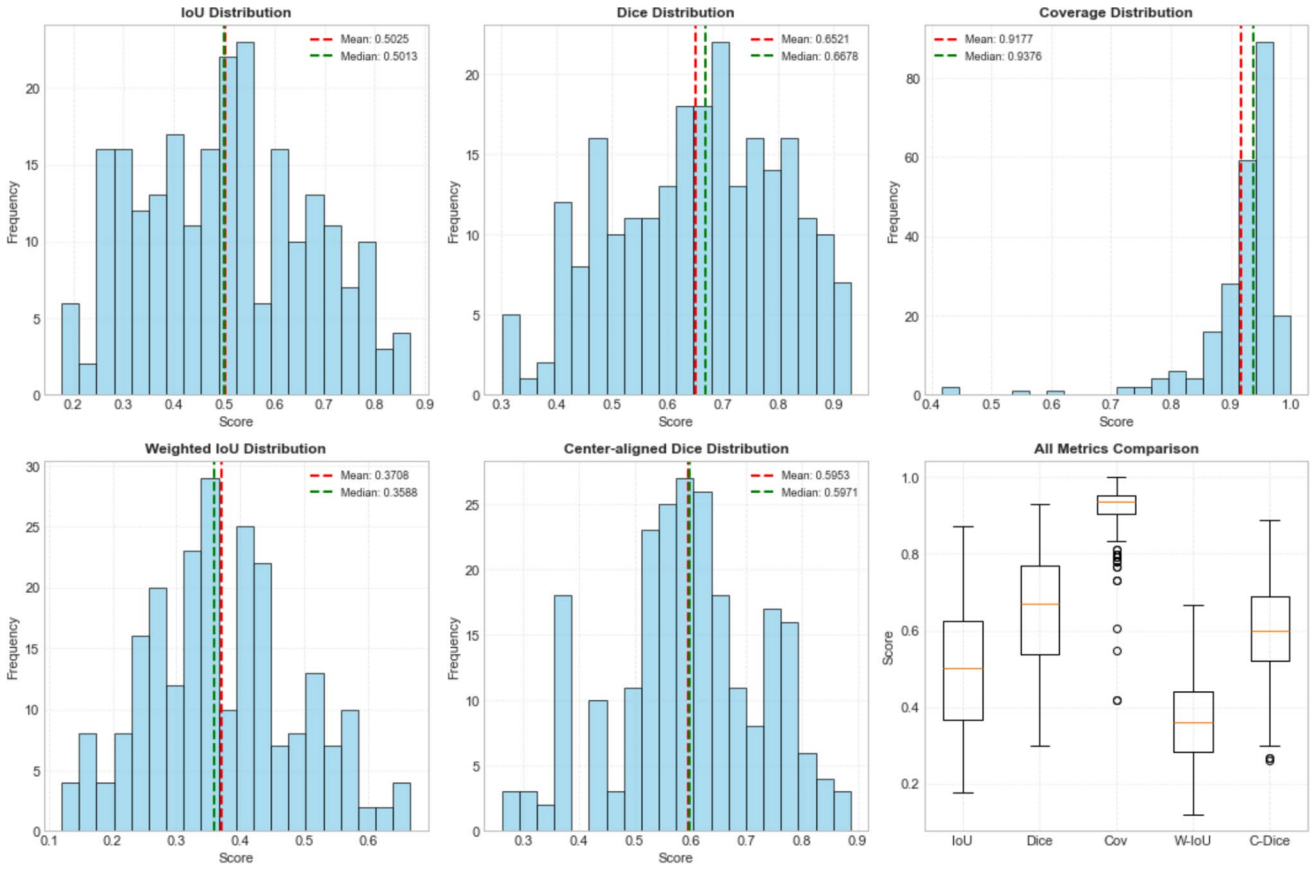


Fig. 27 Quantitative explainability metrics distribution across test samples

works achieved high accuracy using deep learning, such as EfficientNetB0 with SVM [8] and deep autoencoder-based architectures [15], these methods primarily operate on global image representations without explicitly modeling diagnostic regions or nuclear morphology. In contrast, our use of YOLO-based patch extraction enables the model to focus on diagnostically relevant cellular regions, closely resembling the visual attention patterns of expert pathologists. Recent advances in medical image analysis have emphasized the importance of architectural innovation for capturing multi-scale and irregular patterns. In this context, Park et al. [19] demonstrated that modern convolutional designs such as ConvNeXt significantly improve feature discrimination in visual reconstruction tasks. Similarly, Zhang et al. [20] introduced a hierarchy-aggregated diffusion framework that effectively captures multi-scale contextual information in pathology image segmentation. Inspired by these findings, our architecture integrates multi-scale dilated convolutions to represent tissue architecture and deformable convolutions to model irregular nuclear morphology. This dual modeling strategy allows our network to simultaneously capture macro-level tissue organization and micro-level cellular irregularities, which are critical for accurate cytopathological diagnosis.

Moreover, explainability has emerged as a key requirement for clinical AI systems. Several recent studies have incorporated attention mechanisms or visualization techniques to enhance interpretability [5, 10, 14]. Our use of Grad-CAM visualizations demonstrates that the model's attention aligns with established cytopathological diagnostic criteria, such as nuclear enlargement, chromatin irregularity, and intranuclear inclusions. This alignment strengthens the clinical applicability of our system by providing visual evidence that supports the model's predictions, thereby facilitating trust and adoption in real-world clinical workflows.

From a practical perspective, our model offers significant benefits in routine clinical practice. By enabling rapid pre-elimination of benign samples, the system can reduce the diagnostic workload of pathologists and allow them to focus on borderline or complex cases. Furthermore, the system provides reliable second-opinion support, which is particularly valuable in settings with limited access to expert cytopathologists. These clinical advantages distinguish our approach from purely computational models

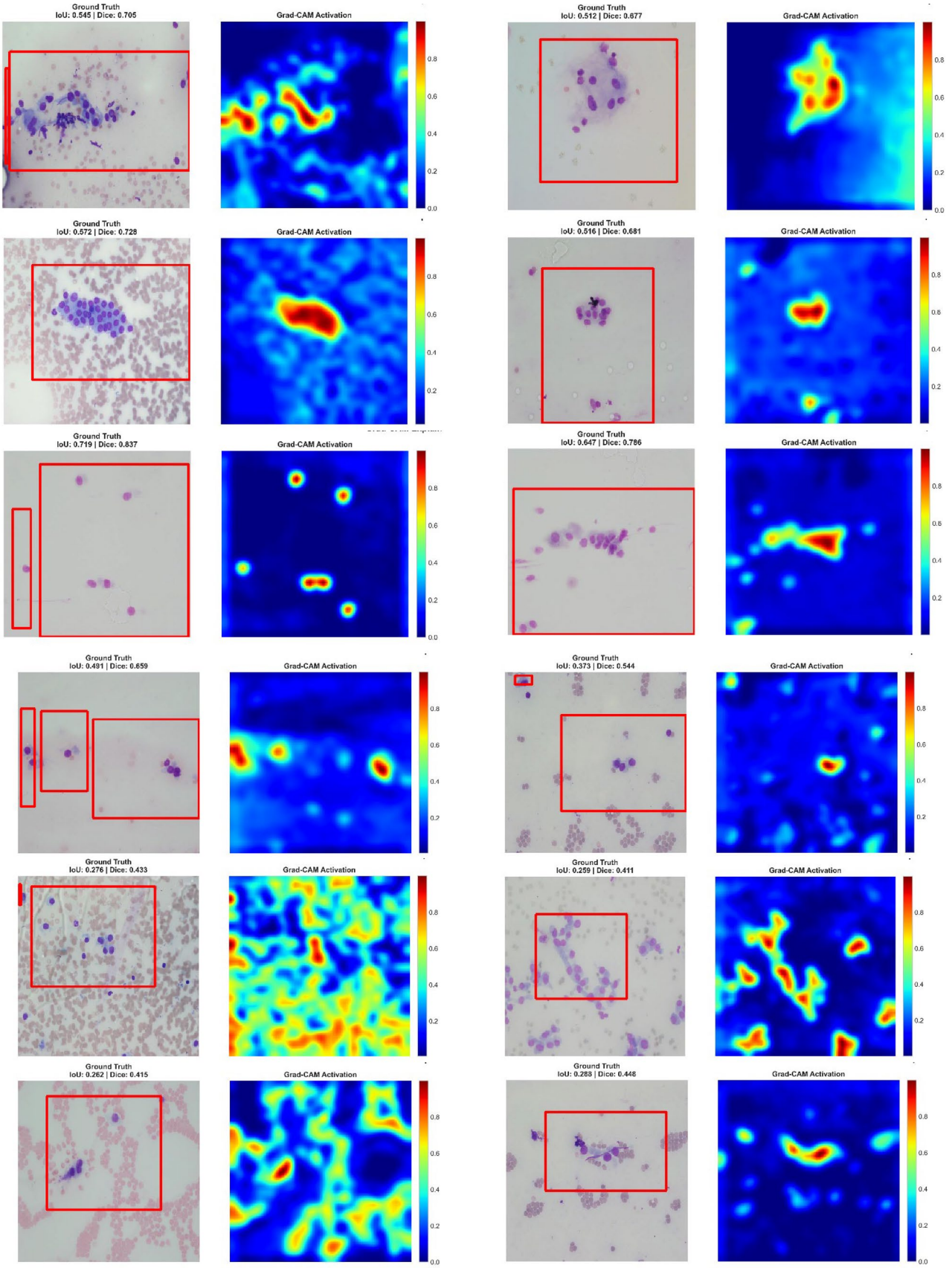


Fig. 28 Comparison of model attention (Grad-CAM) with pathologist-annotated ground-truth bounding boxes

Table 6 Quantitative explainability metrics for proposed model

Metric	Mean	Std	%95 CI
IoU	0.5025	±0.1678	[0.1735, 0.8314]
Dice	0.6521	±0.1512	[0.3558, 0.9484]
Coverage	0.9177	±0.0737	[0.7733, 1.0622]
Weighted IoU	0.3708	±0.1163	[0.1427, 0.5988]
Center-aligned Dice	0.5953	±0.1337	[0.3332, 0.8573]
Max IoU	0.4275	±0.1368	[0.1593, 0.6957]

and highlight its potential translational impact. In addition, variability in cell preparation—such as staining artifacts, blood contamination, cell overlap, and debris—may affect image quality and potentially influence both feature extraction and model predictions.

Nevertheless, certain limitations should be acknowledged. Although our model demonstrates high accuracy and robustness, its performance has been evaluated on a specific dataset. Future studies should assess generalizability across multi-center datasets, different staining protocols, and diverse imaging devices. Additionally, further investigation into lightweight model variants could facilitate deployment in resource-constrained clinical environments.

In summary, this work contributes a clinically inspired, explainable, and high-performance AI framework for PTC diagnosis from FNA cytopathology. By integrating handcrafted morphological features, advanced convolutional architectures, and interpretable visualizations, the proposed method bridges the gap between computational accuracy and clinical usability, offering a meaningful advancement in the field of medical image analysis and computer-aided diagnosis.

6 Conclusions

In this study, we propose a novel hybrid deep learning framework for the automated classification of cytopathology images. Our model is inspired by the decision-making processes of pathologists, capturing the regions of interest (ROIs) that clinicians focus on when examining whole-slide images. To emulate the morphological and statistical features emphasized by pathologists, we extracted both handcrafted features and CNN-based abstract features from these automatically detected ROIs. These multi-scale features were then integrated into a classification network comprising dilated and deformable convolutional layers to enhance discriminative power. The cytopathology dataset employed in this study consists of in-house slides obtained from real patients, ensuring clinical relevance. The proposed model is composed of multiple components, and their individual contributions were systematically analyzed through ablation studies. The results demonstrate that our approach achieves high classification performance across all evaluation metrics. Additionally, Grad-CAM visualizations were utilized to reveal the regions deemed important by the model, illustrating that its decision-making aligns closely with pathologist reasoning. Overall, the proposed method provides a reliable and explainable AI-based decision support system for clinical diagnostic workflows. Beyond its academic contribution, the framework exhibits strong potential for clinical application. Future work will focus on testing the model on larger, multi-center datasets to further validate its generalizability, as well as developing optimized and lightweight versions for real-time clinical integration.

Author contribution A.K. and C.T. conceptualized the study and designed the methodology. A.K. and G.G. conducted the experiments and performed data analysis. E.A. contributed to data preprocessing and validation. C.T. wrote the main manuscript text. G.G. and E.A. prepared the figures and tables. All authors reviewed and approved the final manuscript.

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK).

Data availability The dataset generated and analyzed during the current study is publicly available at Zenodo (https://zenodo.org/records/18428647).

Declarations

Conflict of interest The authors declare no competing interests.

Ethical approval The fine-needle aspiration cytology images employed in this study were obtained under the approval of the Fırat University Ethics Committee (Approval Number: BS48867PMP).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Wuttisarnwattana, P., Supanuch Duangsanit, and Napatsawan Nugent: Automated identification of intranuclear inclusions in fine needle aspiration smears of papillary thyroid carcinoma.

In: 2025 17th Biomedical Engineering International Conference (BMEiCON). IEEE, (2025)
2. Zhou, C., et al.: Thyroid papillary carcinoma segmentation in ultrasound image using a MP-former based model. In: 2024 6th International Conference on Frontier Technologies of Information and Computer (ICFTIC). IEEE, (2024)
3. Intan, I., and Y. Johny W. Soetikno: classification of thyroid carcinoma using sobel edge detection and adaptive neuro fuzzy inference system approach. In: 2019 7th International Conference on Cyber and IT Service Management (CITSM). Vol. 7. IEEE, (2019)
4. Yang, J., et al.: Skip-attention optimization methods for machine learning models for papillary thyroid carcinoma detection based on ultrasound images. In: 2023 IEEE International Conference on Electrical, Automation and Computer Engineering (ICEACE). IEEE, 2023.
5. Shabrina, N.H., et al.: Papillary thyroid cancer histopathological image classification using pretrained ConvNeXt tiny and grad-CAM interpretation. In: 2023 IEEE 11th joint international information technology and artificial intelligence conference (ITAIC). Vol. 11. IEEE, (2023)
6. Stenman, S., et al.: Antibody supervised training of a deep learning based algorithm for leukocyte segmentation in papillary thyroid carcinoma. IEEE J. Biomed. Health Inf. **25**(2), 422–428 (2020)
7. Deepana, D., et al.: Thyroid cancer prediction using deep learning techniques. In: 2024 International Conference on Computing and Intelligent Reality Technologies (ICCIRT). IEEE, (2024)
8. Linda, B., Bendjillali R.I., and Bendelhoum M.S.: Hybrid classification approach of thyroid cancer histological images using deep feature extraction and classical classifiers. In 2025 3rd International Conference on Electronics, Energy and Measurement (IC2EM). IEEE, (2025)
9. El-Hossiny, Ahmed S., et al.: A robust CNN classification of whole slide thyroid carcinoma images. In: 2021 9th International Japan-Africa Conference on Electronics, Communications, and Computations (JAC-ECC). IEEE, (2021)
10. Chen, L., et al.: A multimodal deep learning model for preoperative risk prediction of follicular thyroid carcinoma. In: 2023 IEEE International Conference on E-health Networking, Application & Services (Healthcom). IEEE, (2023)
11. Jiang, W., et al.: The application of data imputation and deep learning network in the papillary thyroid carcinoma classification. In: 2021 IEEE International Ultrasonics Symposium (IUS). IEEE, (2021)
12. Payatsuporn, T., et al.: Papillary thyroid carcinoma semantic segmentation using multi-scale adaptive convolutional network with dual decoders. IEEE Access. (2025). https://doi.org/10.1109/access.2025.3532505
13. El-Hossiny, A.S., et al.: Classification of thyroid carcinoma in whole slide images using cascaded CNN. IEEE Access. **9**, 88429–88438 (2021)
14. Shabrina, N.H., et al.: Deep residual network with integrated StarDist nuclei segmentation for papillary thyroid cancer identification: a pathologist-inspired approach. IEEE Access. (2025). https://doi.org/10.1109/access.2025.3584941
15. Ferreira, M.F., Rui C., and Luís F.T.: Autoencoders as weight initialization of deep classification networks applied to papillary thyroid carcinoma. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, (2018)
16. Padrez, Y., et al.: Supervised machine learning for thyroid carcinoma diagnosis using wide-field SHG microscopy. IEEE Access (2025)
17. Li, S., et al. Fine-grained thyroid nodule classification via multi-semantic attention network. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, (2019)
18. Lu, Y., Yang, Y., Chen, W.: Application of deep learning in the prediction of benign and malignant thyroid nodules on ultrasound images. IEEE Access **8**, 221468–221480 (2020)
19. Leem, S., Jeong, H., Lee, Y., and Kim S.: A robust hand gesture-erecognition using combined moment invariants in hand shape. In: 4th International Conference on Interdisciplinary Research Theory andTechnology, pp 89–94 (2016)
20. Wibowo, E.P., Harseno, S.A., and Harahap, R.K.: Feature extraction using histogram of oriented gradient and hu invariant moment for face recognition. In: 2018 Third International Conference on Informatics and Computing (ICIC). IEEE, (2018)
21. Noto, B., Bobe, C., Brandt, J., Raum, H.N., Nacul, N.G., Riemann, B., Helfen, A.: Diagnostic accuracy of diffusion-weighted MRI for differentiating benign and malignant thyroid nodules: systematic review and meta-analysis. Cancers **17**(16), 2677 (2025). https://doi.org/10.3390/cancers17162677
22. Chen, H., Ruan, F., Zhu, L., Zhuang, Y., Ye, X., Liu, X., et al.: Utilizing the transformer mechanism to predict cervical lymph node metastasis in patients with papillary thyroid carcinoma. PLoS ONE **21**(4), e0345937 (2026). https://doi.org/10.1371/journal.pone.0345937
23. Xu, S., Yan, X., Su, Y., Qi, J., Chen, X., Li, Y., ... & Du, Y.: A transformer-based 2.5 D deep learning model for preoperative prediction of lymph node metastasis in papillary thyroid carcinoma. *medRxiv*, 2026–04 (2026)
24. Sui, L., Feng, B., Chen, X., Jin, Z., Zhu, X., Jiang, T., Yan, Y., Zhou, Y., Chen, C., Yao, J., Lai, M., Lv, L., Wang, Y., Wang, L., Li, C., Feng, L., Yue, W., Yu, D., Shi, K., Wang, V.Y., Zhang, Y., Xu, D.: Early prediction of thyroid capsule invasion in papillary microcarcinoma using ultrasound-based deep learning models: a retrospective multicenter study. Insights Imaging **16**(1), 265 (2025)
25. Wu, L., Jiang, Y., Luo, T., et al.: Interpretable AI-assisted diagnosis of papillary thyroid cancer cytopathology using graph neural networks and knowledge graphs. Sci. Rep. **15**, 32165 (2025). https://doi.org/10.1038/s41598-025-18235-z
26. Thanya, T., Jeslin, T.: DeepOptimalNet: optimized deep learning model for early diagnosis of pancreatic tumor classification in CT imaging. Abdom. Radiol. **50**(9), 4181–4211 (2025)
27. Kim, J., Kim, M.H., Lim, D.J., Lee, H., Lee, J.J., Kwon, H.S., Baek, K.H.: Deep learning technology for classification of thyroid nodules using multi-view ultrasound images: potential benefits and challenges in clinical application. Endocrinol. Metabolism **40**(2), 216–224 (2025)
28. Wang, Y., Lu, W., Xu, L., Xu, H., Kong, D.: Deep learning-based ultrasound diagnostic model for follicular thyroid carcinoma. Eur. Radiol. **36**(1), 357–366 (2026)
29. Lou, P., Huang, Y., Zeng, H., Xu, J.: Predictive diagnosis of papillary thyroid carcinoma using interpretable machine learning. Chin. J. Eng. **47**(9), 1896–1904 (2025)
30. Han, B., Han, Y., Li, H., Gao, X.: PTC-CapsNet: capsule network for papillary thyroid carcinoma pathological images classification. Multim. Tools. Appl. **84**(4), 1907–1925 (2025)
31. Kassem, M.A., Naguib, S.M., Hamza, H.M., Fouda, M.M., Saleh, M.K., Hosny, K.M., Hošovský, A.: Explainable transfer learning-based deep learning model for pelvis fracture detection. Int. J. Intell. Syst. (2023). https://doi.org/10.1155/2023/3281998
32. Soaad, M., Mohamed, A.K., Hanaa, M.H., Mostafa, M.F., Khalid, S.M., Khalid, M.H.: Automated system for classifying uni-bicompartmental knee osteoarthritis by using redefined residual learning with convolutional neural network. Heliyon **10**(10), e31017 (2024)
33. Naguib, S.M., Saleh, M.K., Hamza, H.M., Hosny, K.M., Kassem, M.A.: A new superfluity deep learning model for detecting knee osteoporosis and osteopenia in X-ray images. Sci. Rep. **14**(1), 25434 (2024)

34. Naguib, S.M., Hamza, H.M., Hosny, K.M., Saleh, M.K., Kassem, M.A.: Classification of cervical spine fracture and dislocation using refined pre-trained deep model and saliency map. Diagnostics **13**(7), 1273 (2023). https://doi.org/10.3390/diagnostics13071273
35. Urooj, S., and Satya Singh, P.: Geometric invariant feature extraction of medical images using Hu's invariants. In: 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom). IEEE, (2016)
36. Park, S.-W., Jung, S.-H., Sim, C.-B.: NeXtSRGAN: enhancing super-resolution GAN with ConvNeXt discriminator for superior realism. Visual Comput. **41**, 1–27 (2025)
37. Zhang, X., et al.: HADiff: hierarchy aggregated diffusion model for pathology image segmentation. Vis. Comput. **41**(8), 5689–5700 (2025)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Ahmet Kilicarslan was born in Elazig, Turkey, in 1979. He graduated from the Faculty of Medicine at Firat University in 2003 and obtained his specialty in medical pathology in 2014. He is married and father of three children. He is currently serving as a pathologist at Elazig Fethi Sekin City Hospital. His research interest focuses on histotechnology, digital pathology, and endocrine pathology.



Canan Tastimur was born in Elazig, Turkey, in 1991. She received the B.S., M.S., and Ph.D. degrees in computer engineering from Firat University, in 2013, 2017, and 2022, respectively. She is currently an Assistant Professor with the Computer Engineering Department, Erzincan Binali Yildirim University. Her research interests include fault diagnosis, deep learning, and image processing.



Gokhan Gundogdu was born in Elazig in 1989. He received his bachelor's degree in Computer Engineering from Karabuk University in 2016 and his master's degree from Firat University in 2025. He currently works as a Hospital Information Management Systems Software Support Specialist at Fethi Sekin City Hospital. His research areas include data analysis, deep learning, and image processing.



Erhan Akin was born in Erzincan, Turkey, in 1963. He received the B.S. and M.S. degrees in electrical engineering and the Ph.D. degree in ac drives from Firat University, Elazig, Turkey, in 1984, 1987, and 1994, respectively. He was an Assistant Professor with the Department of Electrical Engineering, Firat University, where he was an Associate Professor of electrical machines. He is currently a Full Professor of computer engineering with Firat University. His main research interests include power electronics, digital control of variable speed ac drives, fuzzy control and soft computing techniques, railway systems, and image processing.