Research Article | Vol. 17, No. 5 / 1 May 2026 / Biomedical Optics Express 2428
Biomedical Optics EXPRESS

Label-free diagnosis across the thyroid nodule pathology spectrum using deep learning-enabled optical coherence tomography

WOOJIN LEE,[1] SOONYONG KWON,[1] HYEONG SOO NAM,[1] JAE YEON SEOK,[2,3] AND HONGKI YOO[1,4]

[1]Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, 34141, Republic of Korea
[2]Department of Pathology, Yongin Severance Hospital, Yonsei University College of Medicine, 363 Dongbaekjukjeon-daero, Giheung-gu, Yongin-si, Gyeonggi-do 16995, Republic of Korea
[3]jyseok@yuhs.ac
[4]h.yoo@kaist.ac.kr

Abstract: Thyroid nodules are highly prevalent, yet identifying malignancy remains a persistent challenge due to their significant pathological heterogeneity. Conventional diagnostic workflows rely on invasive and tissue-destructive sampling followed by a time-consuming histopathology tissue process, restricting the ability to obtain pathological insights in real-time or at the bedside. In this context, optical coherence tomography (OCT) has gained attention as a non-invasive, label-free imaging modality; however, its interpretation for detailed pathological assessment has remained challenging. Here, we developed a deep learning (DL)-based framework for diagnostic classification across the spectrum of thyroid nodule pathology using OCT images. OCT datasets were acquired from seven pathological categories, including five thyroid carcinoma subtypes as well as two non-carcinoma tissue types (benign and normal), and were matched with histology for supervised learning. Robust binary differentiation between carcinoma and non-carcinoma was achieved, with an accuracy of 98.37% and an area under the receiver operating characteristic curve of 0.997. Furthermore, multi-class classification across seven pathological categories further demonstrated an overall accuracy of 93.66% on held-out test sets. Diagnostic predictions were visualized as color-coded overlays on e*n face* OCT images, enabling coherent interpretation across tissue volumes. These results demonstrate the feasibility of combining OCT with DL for enhanced thyroid pathology assessment and motivate further optimization toward real-time and point-of-care diagnostic deployment.

1. Introduction

Thyroid nodular disease encompasses a broad pathological spectrum ranging from non-neoplastic and benign lesions to multiple subtypes of thyroid carcinoma, representing a substantial and growing clinical burden worldwide [1]. While malignant nodules necessitate surgical resection, the clinical approach varies drastically by subtype: differentiated thyroid carcinomas generally yield favorable post-surgical outcomes, whereas anaplastic thyroid carcinoma (ATC) is associated with a dismal prognosis, often requiring alternative therapeutic strategies beyond surgery [2]. Accurate preoperative risk stratification across this heterogeneous spectrum remains challenging, particularly in distinguishing malignant from benign nodular lesion or normal thyroid tissue and in subclassifying carcinoma subtypes [2–4]. Current evaluation based on ultrasonography and fine-needle aspiration frequently yields indeterminate results, often leading to additional surgical assessment despite a high prevalence of benign pathology [5]. These limitations highlight the need for a rapid, tissue-sparing, and high resolution modality capable of providing reliable diagnostic

#591801 https://doi.org/10.1364/BOE.591801
Journal © 2026 Received 27 Jan 2026; revised 31 Mar 2026; accepted 4 Apr 2026; published 15 Apr 2026

Biomedical Optics EXPRESS

information across both benign and malignant thyroid conditions, beyond the constraints of existing approaches [6].

Optical coherence tomography (OCT) is a non-invasive, label-free imaging modality that provides micrometer-scale, depth-resolved visualization of tissue microarchitecture through interferometric detection of backscattered light [7,8]. Its real-time, three-dimensional imaging capability with established biosafety has enabled broad clinical adoption across multiple medical fields and has driven growing interest in thyroid and parathyroid applications [9–15]. OCT has been shown to capture thyroid-specific architectural features and correlate with histopathology, supporting its role as a complementary imaging tool. We previously explored OCT-based quantitative features for thyroid tissue assessment through comparisons between thyroid carcinoma subtypes and normal thyroid tissue [16]. However, conventional OCT primarily resolves tissue- and architectural-level features rather than cellular or nuclear detail, which can limit its ability to precisely differentiate among diverse thyroid pathologies based on OCT images alone [9–11].

Meanwhile, deep learning (DL) has been widely adopted across the field of medical image analysis, driven by their ability to learn complex, high-dimensional feature representations directly from image data [17–19]. In microscopic and high-resolution imaging modalities, DL has demonstrated strong performance in image enhancement, and classification tasks that are difficult to address using conventional, explicitly defined image descriptors [18,20,21]. By capturing subtle spatial patterns and contextual relationships that may not be readily discernible by human observers, DLs enable objective and reproducible interpretation of imaging data [22,23]. Such capabilities suggest that DL can provide an effective framework for extracting diagnostically relevant information from OCT images, where meaningful pathological characteristics are embedded in multi-scale architectural patterns rather than explicit cellular detail [20,24].

Building on these advances, various studies have demonstrated the feasibility of DL-based interpretation of OCT images for pathological tissue assessment [25–29]. Foo, et al. [25] demonstrated that DL models can classify diverse histopathologic patterns in human breast tissue using OCT imaging, highlighting the diagnostic relevance of OCT-derived features. Butola, et al. [26] showed that DL enables reliable classification of OCT data from a wide range of biological samples, including structurally complex and heterogeneous tissues. Wang, et al. [27] applied DL to OCT images of human tissues, including the thyroid, to achieve tissue-specific identification. In addition, Tampu, et al. [29] demonstrated distinction between normal and diseased thyroid tissue using OCT images and DL. Collectively, these studies establish the technical feasibility of DL-assisted OCT interpretation. However, despite the inclusion of thyroid data in some prior work [27,29], no study has systematically evaluated whether DL-based OCT can achieve subtype-level diagnostic differentiation across the full pathological spectrum of thyroid nodules. Such capability is essential for laying the groundwork to determine whether OCT can contribute to pathological diagnosis in routine clinical practice.

In this study, a DL-based framework is introduced for histopathologic differentiation across the spectrum of thyroid nodule pathology using OCT images. To evaluate the practical feasibility of OCT in thyroid pathology, this approach was designed to achieve detailed pathological classification encompassing five major thyroid carcinoma subtypes as well as benign and normal thyroid tissue. While such distinctions are challenging to achieve through direct interpretation of conventional OCT images alone, the integration of DL enabled reliable differentiation by capturing complex, multi-scale architectural patterns embedded in the OCT signal. By demonstrating subtype-level differentiation across a broad pathological spectrum, these findings support the potential of OCT as a label-free imaging modality that, when enabled by DL, can facilitate clinically relevant tissue characterization and inform future diagnostic and real-time point-of-care applications in thyroid disease.

Biomedical Optics EXPRESS

2. Methods

2.1. OCT dataset acquisition

OCT data were acquired using a custom-built benchtop swept-source OCT system with axial and lateral resolutions of 10 µm (in air) and 13 µm, respectively (Fig. 1(a)) [30]. The system operated at a central wavelength of 1,290 nm with a bandwidth of 110 nm and an average output power of 40 mW, with an imaging speed of 117 frames per second. Volumetric imaging was performed using a two-axis galvanometer scanner. During slow-axis scanning, B-scans were acquired at 6 µm intervals, satisfying Nyquist sampling requirements relative to the lateral resolution. Each A-scan represents a depth-resolved axial reflectivity profile acquired at a single lateral position, and a B-scan corresponds to a single cross-sectional frame in the XZ plane formed by sequentially assembled A-scans. The resulting three-dimensional datasets covered an effective field of view of $5.03 \times 5.79 \times 1.85\ \text{mm}^3$ along the X, Y, and Z axes, with the axial extent defined to encompass the depth range used for patch-based analysis.

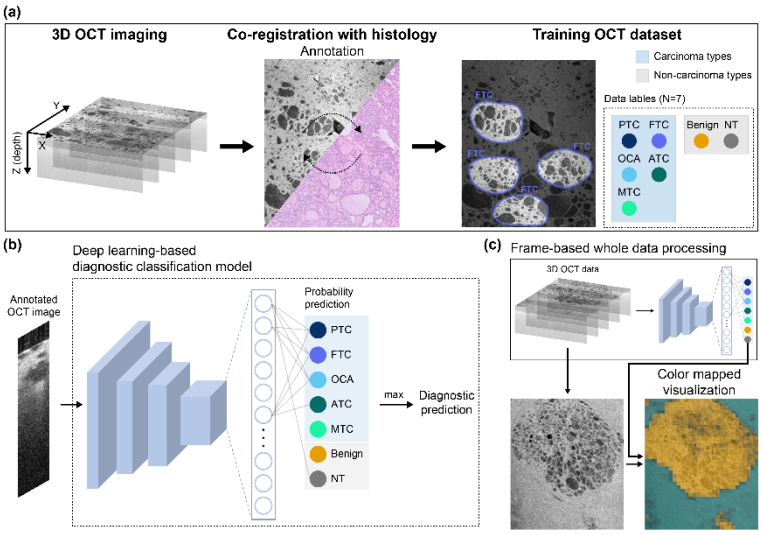

Fig. 1. Overview of the proposed deep learning–based OCT diagnostic framework. (a) Workflow for acquiring and preparing OCT training data with histology-matched annotation. (b) Schematic illustration of the OCT image–based diagnostic classification model. (c) Visualization of diagnostic predictions overlaid on OCT images.

2.2. Histopathologic annotation and training data preparation

Three-dimensional OCT datasets were acquired from formalin-fixed human thyroid tissue specimens (n = 41), with each specimen obtained from a different patient. Corresponding *en face* OCT images were subsequently reconstructed from the volumetric data. These en face images were generated by extracting depth-resolved slices near the tissue surface, corresponding to approximately 0-10 µm beneath the sample surface according to the system resolution. All *en face* images presented in this study were reconstructed under the same condition. For histopathologic ground truth, sister tissue sections (~ 4 µm thickness) immediately adjacent to the OCT-imaged blocks, corresponding to the *en face* imaging plane, were processed for hematoxylin and eosin (H&E) staining. The *en face* OCT images were spatially matched with the corresponding histologic sections, enabling direct correlation between OCT-based morphological features and histopathology (Fig. 1(a)). Based on this correlation, a board-certified pathologist (J.Y.S.)

Biomedical Optics EXPRESS

manually annotated regions corresponding to distinct histopathologic entities on the *en face* OCT images based on the 5th WHO classification [4], while also referencing the corresponding B-scan depth information to support consistent labeling. These annotations were subsequently mapped to the corresponding B-scan images and patch regions for use in model training. Annotation was performed at the tissue level, reflecting histologic diagnosis rather than individual cellular features, and was restricted to regions with unequivocal pathological correlation (Fig. 1(a)). A total of seven pathological labels were defined for model training. These comprised five thyroid carcinoma subtypes: papillary thyroid carcinoma (PTC; n = 11), follicular thyroid carcinoma (FTC; n = 6), oncocytic carcinoma (OCA; n = 3), ATC (n = 9), and medullary thyroid carcinoma (MTC; n = 12). In addition, non-carcinoma categories included benign thyroid tissue (Benign), such as thyroid follicular nodular disease and lymphocytic thyroiditis, and normal thyroid gland (NT). These non-carcinoma regions were extracted from areas identified on matched H&E histology slides and subsequently correlated with the corresponding OCT images from carcinoma specimens. Specifically, benign regions were sampled from a subset of carcinoma specimens (n = 10), while NT regions were sampled from all specimens (n = 41) [4]. This study was reviewed and approved by the Institutional Review Board (GCIRB2021-241, 9-2023-0090), and the requirement for informed consent was waived. All procedures were conducted in accordance with the Declaration of Helsinki.

The proposed model performs diagnostic classification using patch-based OCT B-scan inputs (Fig. 1(b)). Only B-scan images containing regions annotated by a pathologist were included in the training and testing datasets to ensure label accuracy. Because annotated regions varied in size across samples, B-scan images were cropped into fixed-size lateral patches based on typical thyroid follicular dimensions [31,32]. Patches with a width of 64 pixels, corresponding to approximately 380 µm along the A-scan direction, were used as inputs to the model, and randomly sampled from within annotated regions of each B-scan during training. Annotated regions smaller than the predefined patch size were excluded from training to avoid partial or ambiguous representation of pathological features. Along the depth axis, the patch size was selected to encompass the full range of diagnostically relevant signal by covering approximately 1.85 mm, corresponding to 550 pixels, to capture the typical penetration depth observed in thyroid OCT imaging [11,33,34].

Model generalization was evaluated using five-fold cross-validation. In each fold, data derived from specific specimens within each subtype were held out exclusively for testing, ensuring complete separation between training and test sets at the case level. Accordingly, the test set in each fold consisted of one specimen (i.e., one patient) per subtype (n = 1), while the remaining specimens were used for training. All images derived from the same specimen were confined to either the training or test set to prevent overlap between the two. To assess generalization across cases (i.e., specimens), different specimens from the same subtype were selected as the test specimens across folds. For subtypes with fewer cases than the number of folds, some specimens were inevitably used as test cases more than once; however, the combinations of test cases across all subtypes were configured to differ between folds. During training, a comparable number of B-scans was sampled across pathological categories to minimize class imbalance-related bias. The test set comprised all available data from the held-out cases and therefore varied in size across folds. The total number of frames used in each fold is summarized in Table 1.

2.3. Deep learning model and training strategy

A DL-based classification model was developed with reference to the EfficientNetV2 architecture, which offers a favorable balance between computational efficiency and classification performance [35,36]. An overview of the network architecture is summarized in Fig. 2. EfficientNetV2 employs a combination of Fused-Mobile Inverted Bottleneck Convolution (MBConv) blocks for early-stage feature extraction and MBConv blocks for deeper layers, enabling efficient

Biomedical Optics EXPRESS

Table 1. Summary of OCT B-scan frame counts used in the five-fold cross-validation

	fold 1		fold 2		fold 3		fold 4		fold 5	
	train	test	train	test	train	test	train	test	train	test
PTC	3440	42	3440	962	4916	588	3840	574	4886	977
FTC	3440	1186	3440	773	4916	440	3840	615	4886	317
OCA	3440	1792	3440	1792	4916	316	3840	1392	4886	346
ATC	3440	257	3440	1470	4916	1161	3840	2422	4886	2443
MTC	3440	1250	3440	767	4916	727	3840	1449	4886	572
Benign	3440	386	3440	809	4916	1581	3840	477	4886	253
NT	3440	814	3440	886	4916	658	3840	1004	4886	459

representation learning across multiple spatial scales (Fig. 2) [36]. Squeeze-and-Excitation modules are integrated within the MBConv blocks to adaptively recalibrate channel-wise feature responses, enhancing representational capacity with minimal computational overhead [36]. Because OCT B-scan images exhibit spatially anisotropic resolution along the axial and lateral directions, stride sizes within the convolutional layers were optimized to balance feature map dimensions. In particular, selected network blocks employed larger strides along the axial direction to progressively aggregate depth-wise information while maintaining stable lateral feature resolution (Fig. 2). Features extracted by the EfficientNetV2 backbone were subsequently integrated using adaptive average pooling, followed by two fully connected layers to produce probability estimates for seven pathological labels. Softmax activation was applied to each output node to yield probabilities ranging from 0 to 1, and the final diagnostic prediction was determined by selecting the label with the highest probability.

Model training employed a composite loss function consisting of cross-entropy loss and center loss, the latter introduced to promote compact intra-class feature distributions and enhanced inter-class separability in the learned feature space [37]. A weighting factor of 0.01 was applied to the center loss term when computing the total training loss. Optimization was performed using the AdamW optimizer [38]. The learning rate was initialized at 0.01 and reduced by a factor of 0.5 when the training loss plateaued. To enhance model generalization through data variability, random horizontal flipping was applied as an additional augmentation to the randomly cropped patch inputs during training. Prior to being fed into the network, all input data were also normalized using mean and standard deviation. The model was trained from scratch without using pretrained weights, as OCT images are substantially different from natural-image data. Models were trained for a total of 200 epochs with a batch size of 128, corresponding to an overall training time of approximately 4 hours. All implementations were carried out using PyTorch version 1.13.1 on a GPU server equipped with four NVIDIA RTX 4090 GPUs and CUDA 12.1. Patch-based inference on individual B-scan images required 0.03 seconds per frame.

2.4. Evaluation of diagnostic classification performance

Diagnostic classification performance was first evaluated using a binary formulation distinguishing carcinoma from non-carcinoma. Carcinoma cases included PTC, FTC, OCA, ATC, and MTCs, whereas non-carcinoma comprised Benign and NT. Binary classification performance was quantitatively evaluated using receiver operating characteristic (ROC) analysis, which characterizes the trade-off between sensitivity and specificity across varying decision thresholds. The area under the ROC curve (AUC) was used as a summary measure of overall discriminative performance. In addition, standard classification metrics were computed, including accuracy, precision, sensitivity, specificity, and F1 score [39].

Biomedical Optics EXPRESS

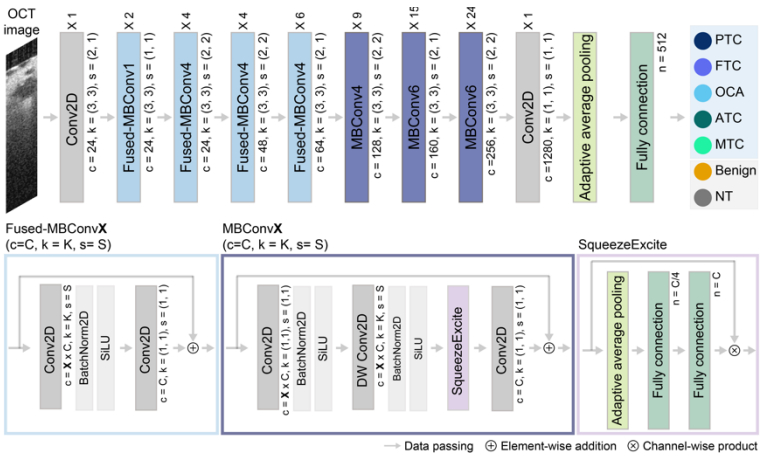

Fig. 2. Architecture of the proposed OCT-based diagnostic classification model. The model is based on the EfficientNetV2 architecture and takes patch-based grayscale OCT B-scan images as input. The number of repeated blocks in each stage is indicated at the top of the corresponding module, while the output channel size (c), kernel size (k), and stride (s) are specified for each block. Detailed structures of the Fused-MBConv, MBConv, and Squeeze-and-Excitation modules are illustrated at the bottom. In both Fused-MBConv and MBConv blocks, the numeric suffix denotes the channel expansion ratio within the block. Extracted feature maps are compressed via adaptive average pooling, reducing the spatial dimensions to one while preserving the channel dimension. The pooled features are then passed through two fully connected layers to generate outputs corresponding to seven pathological labels. Softmax activation is applied to produce probability values between 0 and 1 for each label, and the final diagnostic classification is determined by selecting the label with the highest predicted probability.

Following binary evaluation, diagnostic performance was further assessed using the same trained model under a multi-class formulation encompassing the full spectrum of thyroid nodule pathology. All seven pathological labels were retained, and the model was evaluated for its ability to perform subtype-level diagnostic classification across these categories. Overall classification performance was assessed using a confusion matrix, which summarizes prediction outcomes across all classes and enables evaluation of class-wise accuracy and misclassification patterns. In addition, feature representations learned by the model were examined using t-distributed stochastic neighbor embedding (t-SNE) [40]. The t-SNE visualization was applied to high-dimensional feature maps extracted from the network to assess whether samples from different pathological categories formed separable clusters in the learned feature space, providing qualitative insight into the discriminative structure of the model representations. This analysis was conducted using standard parameter settings (perplexity = 45, Euclidean distance).

All quantitative performance evaluations were systematically conducted using five-fold cross-validation to assess the consistency and generalizability of the model performance. Given the limited number of cases, all quantitative metrics were assessed by averaging results across all folds to reduce potential case-specific bias. The final metrics were computed by aggregating patch-level predictions across all folds. Performance metrics were reported as the mean ± standard deviation across all folds.

2.5. *Visualization of diagnostic classification results*

To facilitate intuitive interpretation of diagnostic classification outcomes, model predictions were visualized with spatial correspondence to the OCT images (Fig. 1(c)). Patch-based classification

Biomedical Optics EXPRESS

results were first mapped onto the corresponding OCT B-scan images by assigning a distinct color to each pathological label. The color intensity at each patch location was scaled according to the predicted probability of the label with the highest confidence. The same visualization strategy was subsequently applied to the full three-dimensional OCT datasets. Patch-level predictions were combined and rendered as color-coded *en face* images to provide an overview of diagnostic patterns across the imaged tissue volume. To achieve smoother spatial continuity in the *en face* visualization, patch extraction was performed with a 50% overlap relative to the patch size (effective step size of 32 pixels). This overlap was introduced to provide a denser spatial representation based on the selected patch size, while maintaining a reasonable computational cost. For regions with overlapping predictions, probability values were averaged prior to visualization. Along the slow-axis scanning direction, classification results were averaged over the number of frames corresponding to the effective patch step size used for B-scan classification (25 frames for 50% overlapping patches), enabling the rendered *en face* results to be displayed as approximately square tiled units with matched spatial dimensions. When the full field of view was not exactly divisible by the effective step size, minor adjustments were applied at the image boundaries by slightly adjusting the size of the final tile. While model training was restricted to pathologist-annotated regions with definitive histopathologic correlation, visualization was performed on the entire OCT volume without prior masking. This allowed assessment of model behavior across the full field of view and provided comprehensive qualitative insight into the spatial distribution of diagnostic predictions.

3. Results

3.1. Binary classification between carcinoma and non-carcinoma

Binary diagnostic performance for distinguishing carcinoma from non-carcinoma was evaluated using the trained model. All performance assessments were conducted exclusively on test sets that were fully separated from the training data at the case level.

The confusion matrix summarizing binary classification results is shown in Fig. 3(a). Across all folds, the model achieved a high overall accuracy of 98.37%, indicating robust distinction between carcinoma and non-carcinoma. Misclassifications were rare and evenly distributed between carcinoma and non-carcinoma, suggesting that the model did not exhibit systematic bias toward either diagnostic category. ROC analysis further demonstrated strong diagnostic performance, with an AUC of 0.997 (Fig. 3(b) and Table 2). This near-unity AUC reflects excellent separability across a wide range of decision thresholds, suggesting that the model reliably distinguishes malignant from non-malignant tissue with minimal trade-off between sensitivity and specificity. This indicates that high diagnostic performance can be maintained without requiring fine-tuning of the operating threshold.

Table 2. Summary of five-fold cross-validation results for binary carcinoma versus non-carcinoma classification

	Binary classification
Accuracy [%]	98.37 ± 0.87
AUC	0.997 ± 0.003
Precision [%]	99.38 ± 0.30
Sensitivity [%]	98.57 ± 0.97
Specificity [%]	97.74 ± 1.35
F1 score [%]	98.97 ± 0.50

Biomedical Optics EXPRESS

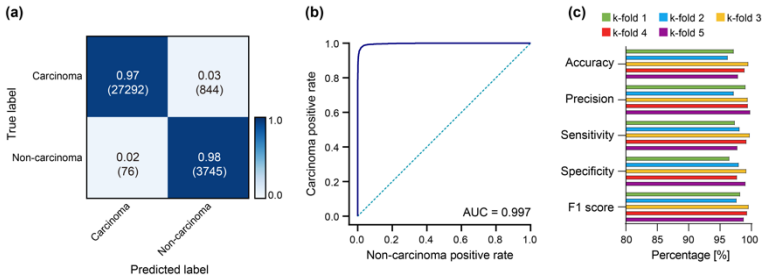

Fig. 3. Binary diagnostic classification performance for carcinoma versus non-carcinoma. (a) Confusion matrix summarizing binary classification results across all folds. The x-axis represents the predicted labels, and the y-axis represents the true labels. Prediction outcomes are normalized and displayed using a colormap ranging from 0 to 1. The numbers shown in parentheses below each normalized value indicate the corresponding number of patches before normalization. (b) Receiver operating characteristic (ROC) curve evaluating binary classification performance. (c) Quantitative results of five-fold cross-validation, demonstrating consistent performance across folds.

Quantitative evaluation using standard binary classification metrics consistently supported these findings, confirming effective separation between the two categories (Table 2). Importantly, comparable performance was observed across all folds in the five-fold cross-validation analysis, indicating stable generalization rather than fold-specific optimization (Fig. 3(c), Table 2, and Supplement 1 Table S1).

Accurate differentiation between carcinoma and non-carcinoma is central to pathological evaluation and clinical decision-making. In combination with the real-time, non-invasive capability of OCT, the strong binary classification performance observed in this study underscores its potential utility for immediate pathological assessment in preoperative and intraoperative settings.

3.2. Multi-class classification across diverse thyroid nodule pathologies

After confirming robust performance in binary distinction between carcinoma and non-carcinoma, the proposed framework was further evaluated for subtype-level diagnostic classification across diverse thyroid nodule pathologies. This analysis encompassed all seven pathological categories, including five carcinoma subtypes together with Benign and NT, to assess the model's ability to perform comprehensive pathological differentiation.

The confusion matrix summarizing multi-class classification results is shown in Fig. 4(a). Across all folds, the model achieved a high overall accuracy of $93.66 \pm 1.00\%$, demonstrating reliable differentiation across the full pathological spectrum. Comparable performance was consistently observed across individual folds (Supplement 1 Table S2). Misclassification was infrequent across most categories, indicating stable diagnostic performance at the subtype level. Consistent performance was observed across five-fold cross-validation, as reflected by stable classification accuracy and limited inter-fold variability, supporting robust model generalization. To further clarify the classification performance for each subtype, subtype-versus-others evaluations were additionally performed for the multi-class task using the same binary classification metrics applied in the carcinoma versus non-carcinoma analysis (Table 3). These results demonstrated consistently strong diagnostic discrimination across all subtypes. Notably, a modest tendency of misclassification was observed involving FTC and Benign labels. FTC cases were occasionally misclassified as OCA or NT, which may reflect the architectural similarity of follicular-patterned FTC to oncocytic or normal thyroid tissue in OCT images, particularly in normo-follicular growth patterns where follicular units remain relatively preserved [3,4].

Biomedical Optics EXPRESS

Likewise, Benign lesions showed a tendency to be misclassified as NT, potentially attributable to overlapping OCT features such as preserved follicular architecture that resemble normal thyroid parenchyma.

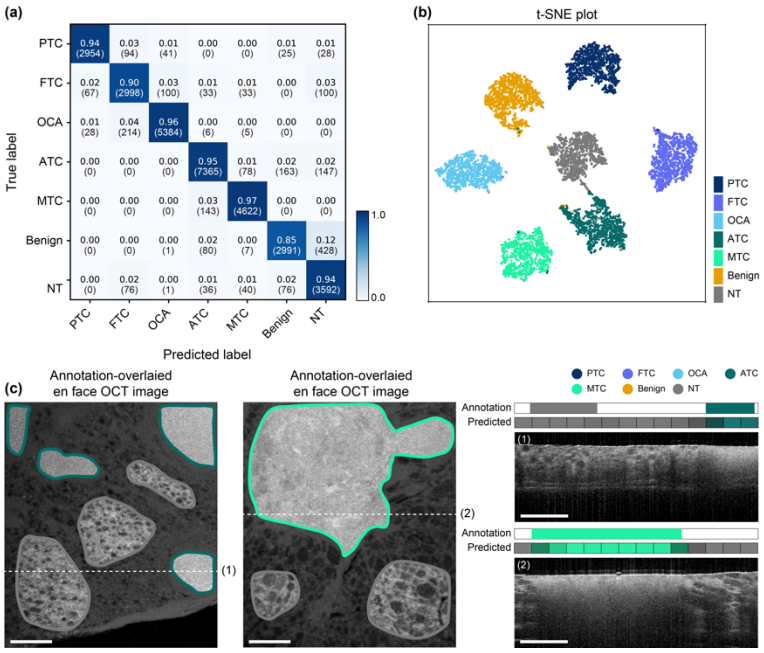

Fig. 4. Diagnostic classification across the broad spectrum of thyroid nodule pathology. (a) Confusion matrix summarizing multi-class classification results across all pathological labels, computed using predictions pooled from all folds. Values are normalized to represent the proportion of predictions for each class. The numbers shown in parentheses below each normalized value indicate the corresponding number of patches before normalization. (b) t-SNE visualization of feature representations learned by the model. The plot is generated from feature maps extracted prior to adaptive pooling, with each color indicating a distinct pathological category as labeled on the right. The spatial separation of feature clusters reflects discriminative representation learning across pathological entities. (c) Representative visualization of diagnostic classification results. The two left panels show annotation-overlaid *en face* OCT images from ATC and MTC cases. Corresponding B-scans at locations indicated by white dotted lines are displayed on the right, with annotation labels and model-predicted results shown above each image. Classification results are color-coded by pathological category, with color intensity indicating predicted probability. All scale bars are 1 mm.

Table 3. Quantitative classification performance for each subtype versus other classes

	PTC	FTC	OCA	ATC	MTC	Benign	NT
Accuracy [%]	99.11	97.76	98.76	97.84	99.04	97.56	97.09
Precision [%]	96.88	88.64	97.45	96.06	96.63	91.89	83.63
Sensitivity [%]	94.02	90.00	95.49	95.00	97.00	85.29	94.03
Specificity [%]	99.67	98.66	99.46	98.75	99.40	99.07	97.50
F1 score [%]	95.43	89.32	96.46	95.53	96.82	88.47	88.53

To further examine whether the model learned discriminative representations across pathological categories, feature embeddings extracted from the network were visualized using t-SNE

Biomedical Optics EXPRESS

(Fig. 4(b)), based on a representative cross-validation fold (fold 1). Feature maps obtained prior to adaptive average pooling were projected into a two-dimensional space. The resulting visualization demonstrated clear separation among pathological categories, suggesting that the model effectively captured subtype-specific OCT features in its learned representation space. This separation suggests that the observed classification performance is supported by consistently encoded, pathology-relevant features embedded within the imaging data of each subtype across diverse cases.

3.3. Visual representation of diagnostic classification results

Figure 4(c) presents representative visual examples of the diagnostic classification results. After model training and performance evaluation based on pathologist-annotated data, the trained model was applied across the full spatial extent of each three-dimensional OCT volume to generate diagnostic predictions beyond region-level annotations.

In Fig. 4(c), the two left *en face* images correspond to annotated ATC and MTC cases, respectively. The B-scan locations indicated by white dotted lines in each *en face* image are shown on the right, where the corresponding classification results are visualized. In each panel, the upper color bar represents the annotation-based pathological labels, whereas the lower color bar displays the model-predicted results. Predicted labels are encoded by category-specific colors, with color intensity scaled according to the associated prediction probability.

Within annotated regions, the model predictions were consistent with the corresponding pathological labels in both cases. Importantly, in regions without annotation, the model generated predictions that were concordant with the underlying OCT structural features, suggesting that the classification results capture diagnostically relevant patterns beyond the explicitly annotated areas. These examples illustrate the potential of the proposed approach to provide spatially resolved diagnostic insight across the entire imaged tissue volume.

3.4. Diagnostic classification across entire en face OCT images

The proposed DL-based framework was further applied throughout each three-dimensional OCT volume to examine diagnostic classification performance at the whole-*en face* image level. As illustrated in Fig. 4(c), patch-based classification results were mapped using category-specific colors and overlaid onto the corresponding *en face* OCT images, enabling spatially resolved visualization of pathological predictions across the full imaging field.

Figure 5 presents representative results from PTC (Fig. 5(a)) and FTC (Fig. 5(b)) cases. For each case, the *en face* OCT image, the corresponding histologic section, the annotation-overlaid *en face* OCT image, and the color-mapped classification result are shown in sequence (Supplement 1 Fig. S1). In the PTC case, the model accurately identified regions exhibiting characteristic papillary architectural features, demonstrating strong concordance with the annotated pathological regions. Importantly, in areas without annotation, the predicted distribution remained consistent with OCT-observed structural patterns, providing a coherent separation between normal thyroid parenchyma and carcinoma.

In the FTC case, despite the presence of follicular architecture that partially resembles NT, the model successfully differentiated FTC from NT regions. The diagnostic predictions closely matched the annotated regions and extended logically into adjacent unannotated areas, reflecting pathology-consistent classification rather than fragmented or random labeling. Consistent qualitative performance was also observed across additional representative cases spanning multiple pathological categories (Supplement 1 Fig. S2).

Collectively, these results demonstrate that the proposed approach enables spatially continuous diagnostic interpretation across entire *en face* OCT images. Although the present analysis was conducted using *ex vivo*, formalin-fixed thyroid specimens acquired from a single center, the strong alignment between predicted results and histologic annotation, together with consistent

Biomedical Optics EXPRESS

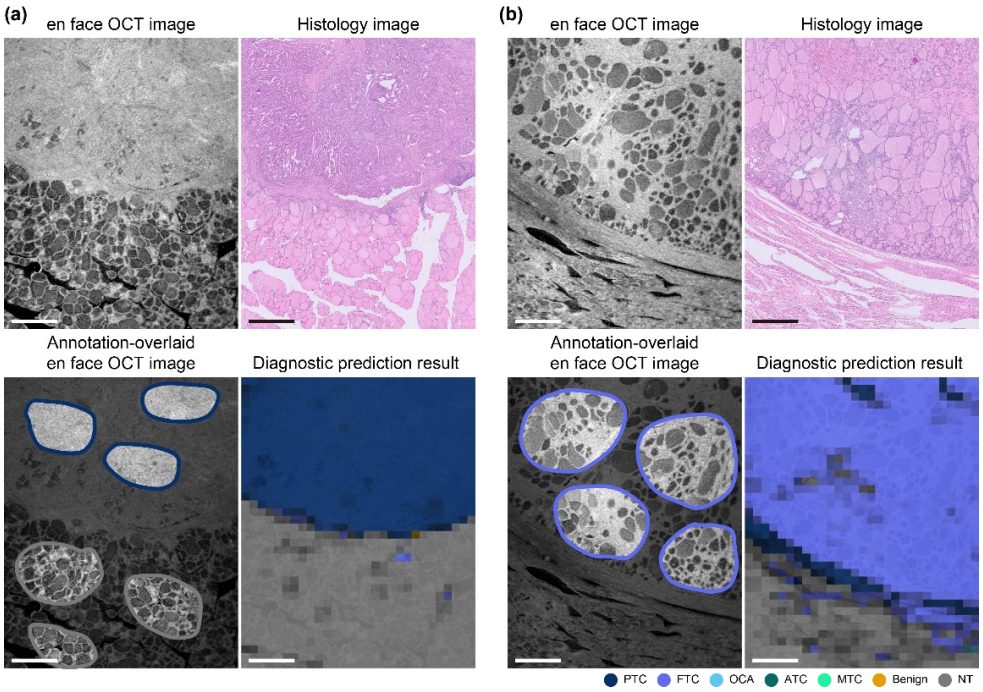

Fig. 5. Diagnostic classification applied to three-dimensional OCT volumes. (a) PTC case. (b) FTC case. For each case, results are presented in the following order: *en face* OCT image, corresponding histologic section, annotation-overlaid *en face* OCT image, and diagnostic prediction map. All scale bars are 1 mm.

classification patterns in non-annotated regions, suggests that the model captures pathology-related patterns when applied to whole-volume OCT data. These findings support the feasibility of this approach for comprehensive pathological assessment, while further validation in broader and clinically relevant datasets is needed.

4. Discussion and conclusion

Thyroid nodular disease presents diagnostic challenges due to overlapping morphological features among biologically distinct entities [1]. Conventional pathological workflows, which rely on invasive sampling and time-consuming histopathologic tissue process, offer limited support for rapid and detailed diagnostic stratification [6]. In this context, OCT has attracted interest as a non-invasive, label-free, and real-time imaging modality for thyroid tissue assessment [7,8]; however, its use for detailed pathological differentiation has remained limited, as it primarily reflects tissue-level architecture [9–11]. In this study, we address this limitation by proposing a DL-based framework that enables diagnostic prediction across the broad spectrum of thyroid nodule pathology using OCT images.

Binary evaluation demonstrated that the proposed framework reliably distinguished carcinoma from non-carcinoma, achieving an accuracy of 98.37%. This result highlights the potential of OCT-based analysis to support a clinically fundamental distinction that underlies pathological interpretation and downstream clinical decision-making. Extending beyond binary classification, the model achieved an overall accuracy of 93.66% across the full spectrum of thyroid nodule pathology, including carcinoma subtypes that are difficult to differentiate based on qualitative OCT image interpretation. Compared with prior OCT studies largely limited to normal-versus-diseased

Biomedical Optics EXPRESS

differentiation [27,29], these results demonstrate the feasibility of subtype-level pathological differentiation, supporting broader applicability in practical diagnostic workflows.

Model predictions were visualized in direct correspondence with the OCT images to facilitate intuitive interpretation. The visualizations demonstrated strong agreement with annotated regions and pathology-consistent predictions even in non-annotated areas. By contextualizing tissue heterogeneity within the structural framework of OCT imaging, these outputs support holistic pathological interpretation.

Importantly, the proposed framework was designed to operate at the B-scan level, reflecting the raw data stream generated during OCT acquisition. This approach enables diagnostic inference prior to full three-dimensional reconstruction, supporting the potential for real-time pathological assessment. In the current implementation, inference for a single B-scan required approximately 0.03 seconds without patch overlay. While this inference throughput does not yet fully match the acquisition rate of typical high-speed OCT systems, it is sufficiently fast to indicate that real-time deployment is achievable.

Despite the overall strong performance, occasional misclassifications were observed, particularly in cases that are also challenging from a pathological standpoint. For example, ATC with extensive inflammatory infiltration, or disrupted architectural organization may present OCT features that overlap with benign lymphocytic thyroiditis regions; follicular patterned tumor such as FTC or OCA with benign thyroid lesion of thyroid follicular nodular disease or relatively homogeneous normal tissue. These overlaps reflect the inherent complexity of interpreting subtle architectural variations in tissue-level OCT images, which can complicate both human and algorithmic interpretation. In addition, because benign regions in the dataset were derived from carcinoma specimens rather than benign-only cases, this dataset composition may have introduced a potential source of bias in pathological interpretation. Future improvements could be achieved by refining model architectures and expanding complementary data through the inclusion of larger and more diverse cohorts.

The present analysis was based on *ex vivo*, formalin-fixed thyroid specimens acquired from a single center, which may influence tissue optical properties and limit direct generalization to *in vivo* or multi-institutional settings [41]. In addition, the scope of the dataset and its pathological distribution may not fully capture the breadth of real-world diagnostic variability. Accordingly, further evaluation using fresh tissue imaging and larger, multi-center cohorts will be important to assess robustness and support broader clinical translation. Furthermore, although efforts were made to maximize the correlation between *en face* OCT images and the matched histologic sections during the labeling process, some degree of uncertainty in the spatial matching could still remain. This also highlights the need for further refinement in ground truth preparation to support more accurate training data. Future studies should also incorporate more rigorous specimen- or patient-level performance evaluation, enabled by larger cohorts and more comprehensive per-case data, to further support clinical applicability.

In conclusion, this study demonstrates that DL-based analysis of OCT images enables subtype-level diagnostic classification across a broad spectrum of thyroid nodule pathology. By combining non-invasive, real-time OCT imaging with data-driven interpretation and intuitive visualization, this approach highlights the potential of OCT to support thyroid-specific pathological assessment, including intraoperative tissue evaluation. We believe that these findings lay a foundation for the future translation of this approach into thyroid-focused diagnostic and point-of-care applications.

Funding. National Research Foundation of Korea (2021R1F1A1052586, RS-2024-00401786, RS-2026-25470369); Korea Health Industry Development Institute (RS-2021-KH113146); Yonsei University College of Medicine (6-2023-0154).

Acknowledgments. This work was supported by the National Research Foundation of Korea (NRF), funded by the Korea government (MSIP; Ministry of Science, ICT and Future Planning) [grant numbers: 2021R1F1A1052586, RS-2024-00401786, and RS-2026-25470369]; in part by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic

Biomedical Optics EXPRESS

of Korea [grant number: RS-2021-KH113146]; and by a faculty research grant of the Yonsei University College of Medicine [grant number: 6-2023-0154].

Disclosures. The authors declare that there are no financial interests, commercial affiliations, or other potential conflicts of interest that could have influenced the objectivity of this research or the writing of this paper.

Data availability. Code and data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

Supplemental document. See Supplement 1 for supporting content.

References

1. F. Bray, M. Laversanne, H. Sung, *et al.*, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA A Cancer J Clinicians **74**(3), 229–263 (2024).
2. K. C. Bible, E. Kebebew, J. Brierley, *et al.*, "2021 American thyroid association guidelines for management of patients with anaplastic thyroid cancer: American thyroid association anaplastic thyroid cancer guidelines task force," Thyroid **31**(3), 337–386 (2021).
3. B. R. Haugen, "2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: what is new and what has changed?" Cancer **123**(3), 372–381 (2017).
4. W. C. O. T. E. Board, *Endocrine and Neuroendocrine Tumours: Who Classification of Tumours* (World Health Organization, 2025).
5. S. Z. Ali and P. A. VanderLaan, *The Bethesda System for Reporting Thyroid Cytopathology: Definitions, Criteria, and Explanatory Notes* (Springer Nature, 2023).
6. S. Estebe, C. Montenat, A. Tremoureux, *et al.*, "Limitation of intraoperative frozen section during thyroid surgery," Eur. Arch. Otorhinolaryngol. **274**(3), 1671–1676 (2017).
7. D. Huang, E. A. Swanson, C. P. Lin, *et al.*, "Optical coherence tomography," Science **254**(5035), 1178–1181 (1991).
8. P. H. Tomlins and R. K. Wang, "Theory, developments and applications of optical coherence tomography," J. Phys. D: Appl. Phys. **38**(15), 2519–2535 (2005).
9. C. Zhou, Y. Wang, A. D. Aguirre, *et al.*, "Ex vivo imaging of human thyroid pathology using integrated optical coherence tomography and optical coherence microscopy," J. Biomed. Opt. **15**(01), 1 (2010).
10. L. Pantanowitz, P. L. Hsiung, T. H. Ko, *et al.*, "High-resolution imaging of the thyroid gland using optical coherence tomography," Head & Neck **26**(5), 425–434 (2004).
11. Y. Shao, Q. Xu, C. Feng, *et al.*, "Optical coherence tomography for the qualitative analysis of thyroid tissue images: Feasibility, features, and clinical potential," Transl. Res. **279**, 27–39 (2025).
12. N. Yang, C. Boudoux, E. De Montigny, *et al.*, "Rapid head and neck tissue identification in thyroid and parathyroid surgery using optical coherence tomography," Head & Neck **41**(12), 4171–4180 (2019).
13. M. A. Vasilescu and M. L. Macovei, "The Perspective of Using Optical Coherence Tomography in Ophthalmology: Present and Future Applications," Diagnostics **15**(4), 402 (2025).
14. S. Kim, H. S. Nam, D. O. Kang, *et al.*, "Intracoronary Structural-Molecular Imaging for Multitargeted Characterization of High-Risk Plaque: First-in-Human OCT-FLIm," JAMA Cardiol. **10**(7), 708–717 (2025).
15. J. H. Kim, J. W. Song, Y. H. Kim, *et al.*, "Multimodal Imaging-Assisted Intravascular Theranostic Photoactivation on Atherosclerotic Plaque," Circ. Res. **135**(5), e114–e132 (2024).
16. W. Lee, S. Kwon, and H. S. Nam, "Optical Coherence Tomography for Broad Spectrum of Thyroid Carcinoma: A Qualitative and Quantitative Analysis," Biomed. Eng. Lett., in press (2026).
17. H. S. Nam, D. O. Kang, J. Han, *et al.*, "Deep Learning-Based Saturation Compensation for High Dynamic Range Multispectral Fluorescence Lifetime Imaging," IEEE Trans. Biomed. Eng. **72**(9), 2635–2646 (2025).
18. W. Lee, H. S. Nam, J. Y. Seok, *et al.*, "Deep learning-based image enhancement in optical coherence tomography by exploiting interference fringe," Commun. Biol. **6**(1), 464 (2023).
19. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature **521**(7553), 436–444 (2015).
20. T. Jo, K. Nho, and A. J. Saykin, "Deep learning in Alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data," Front. Aging Neurosci. **11**, 220 (2019).
21. W. Lee, M. A. Jang, H. S. Nam, *et al.*, "Self-supervised denoising of dynamic fluorescence images via temporal gradient-empowered deep learning," PhotoniX **6**(1), 15 (2025).
22. I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial networks," Commun. ACM **63**(11), 139–144 (2020).
23. Z. Li, F. Liu, W. Yang, *et al.*, "A survey of convolutional neural networks: analysis, applications, and prospects," IEEE Trans. Neural Netw. Learning Syst. **33**(12), 6999–7019 (2022).
24. S. Misra, S. Na, K. Park, *et al.*, "Deep learning based label-free virtual staining and classification of human tissues using digital slide scanner," Med. Image Anal. **108**, 103865 (2026).
25. K. Y. Foo, K. Newman, Q. Fang, *et al.*, "Multi-class classification of breast tissue using optical coherence tomography and attenuation imaging combined via deep learning," Biomed. Opt. Express **13**(6), 3380–3400 (2022).
26. A. Butola, D. K. Prasad, A. Ahmad, *et al.*, "Deep learning architecture "LightOCT" for diagnostic decision support using optical coherence tomography images of biological samples," Biomed. Opt. Express **11**(9), 5017–5031 (2020).
27. H. Wang, D. Won, and S. W. Yoon, "A deep separable neural network for human tissue identification in three-dimensional optical coherence tomography images," IISE Trans. Healthc. Syst. Eng. **9**(3), 250–271 (2019).

Biomedical Optics EXPRESS

28. I. Leandro, B. Lorenzo, M. Aleksandar, *et al.*, "OCT-based deep-learning models for the identification of retinal key signs," Sci. Rep. **13**(1), 14628 (2023).
29. I. E. Tampu, A. Eklund, K. Johansson, *et al.*, "Diseased thyroid tissue classification in OCT images using deep learning: Towards surgical decision support," J. Biophotonics **16**(2), e202200227 (2023).
30. H. S. Nam, W. J. Kang, M. W. Lee, *et al.*, "Multispectral analog-mean-delay fluorescence lifetime imaging combined with optical coherence tomography," Biomed. Opt. Express **9**(4), 1930–1947 (2018).
31. J. L. Jackson, "The shape and size of the human thyroid follicle in health and disease," Anat. Rec. **48**(2), 219–239 (1931).
32. A. Faggiano, J. Coulot, N. Bellon, *et al.*, "Age-dependent variation of follicular size and expression of iodine transporters in human thyroid tissue," Int. J. Nucl. Med. **45**(2), 232–237 (2004).
33. A. W. Sainter, T. A. King, and M. R. Dickinson, "Effect of target biological tissue and choice of light source on penetration depth and resolution in optical coherence tomography," J. Biomed. Opt. **9**(1), 193–199 (2004).
34. J. Welzel, E. Lankenau, R. Birngruber, *et al.*, "Optical coherence tomography of the human skin," J. Am. Acad. Dermatol. **37**(6), 958–963 (1997).
35. Y. Mellak, A. Achim, A. Ward, *et al.*, "A machine learning framework for the quantification of experimental uveitis in murine OCT," Biomed. Opt. Express **14**(7), 3413–3432 (2023).
36. M. Tan and Q. Le, "EfficientNetV2: Smaller Models and Faster Training," in *Proc. Int. Conf. Mach. Learn*139, 10096–10106 (2021).
37. Y. Wen, K. Zhang, Z. Li, *et al.*, "A discriminative feature learning approach for deep face recognition," in *Comput. Vis.–ECCV 2016* LNCS 9911, 499–515 (2016).
38. I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," arXiv, arXiv:1711.05101 (2017).
39. Ž. Vujović, "Classification model evaluation metrics," Int. J. Adv. Comput. Sci. Appl. **12**(6), 599–606 (2021).
40. L. V. D. Maaten and G. Hinton, "Visualizing data using t-SNE," J. Mach. Learn. Res. **9**, 2579–2605 (2008).
41. P.-L. Hsiung, P. R. Nambiar, and J. G. Fujimoto, "Effect of tissue preservation on imaging using ultrahigh resolution optical coherence tomography," J. Biomed. Opt. **10**(6), 064033 (2005).