A Hybrid CNN-Transformer Deep Learning Model for Differentiating Benign and Malignant Breast Tumors Using Multi-View Ultrasound Images

Qi Zhang[1], Pan Tang[1], Miao Chen[1]*, Jianxing Zhang[1]*

[1]The Second Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou, 510120,China.

*Corresponding Author: Jian-xing Zhang, . Emai: venant@126.com; Miao Chen, . Emai: 297170770@qq.com

Abstract

Breast cancer is a leading malignancy threatening women's health globally, making early and accurate diagnosis crucial. Ultrasound is a key screening and diagnostic tool due to its non-invasive, real-time, and cost-effective nature. However, its diagnostic accuracy is highly dependent on operator experience, and conventional single-image analysis often fails to capture the comprehensive features of a lesion. This study introduces a computer-aided diagnosis (CAD) system that emulates a clinician's multi-view diagnostic process. We developed a novel hybrid deep learning model that integrates a Convolutional Neural Network (CNN) with a Transformer architecture. The model uses a pretrained EfficientNetV2 to extract spatial features from multiple, unordered ultrasound images of a single lesion. These features are then processed by a Transformer encoder, whose self-attention mechanism globally models and fuses their intrinsic correlations. A strict lesion-level data partitioning strategy ensured a rigorous evaluation. On an internal test set, our CNN-Transformer model achieved an accuracy of 0.93, a sensitivity of 0.92, a specificity of 0.94, and an Area Under the Curve (AUC) of 0.98. On an external test set, it demonstrated an accuracy of 0.93, a sensitivity of 0.94, a specificity of 0.91, and an AUC of 0.97. These results significantly outperform those of a baseline single-image model, which

achieved accuracies of 0.88 and 0.89 and AUCs of 0.95 and 0.94 on the internal and external test sets, respectively. This study confirms that combining CNNs with Transformers yields a highly accurate and robust diagnostic system for breast ultrasound. By effectively fusing multi-view information, our model aligns with clinical logic and shows immense potential for improving diagnostic reliability.

Keywords: Breast cancer, EfficientNetV2, Transformer, Hybrid model, Deep learning, Ultrasound imaging

1.Introduction

Breast cancer stands as the most frequently diagnosed malignancy among women worldwide, posing a significant threat to their lives and health(1). Clinical evidence overwhelmingly shows that early detection, diagnosis, and treatment are paramount for improving patient prognosis and achieving higher five-year survival rates(2). Among the various imaging modalities, breast ultrasound has emerged as an indispensable first-line tool for the screening, diagnosis, and follow-up of breast diseases, owing to its unique advantages of being non-invasive, radiation-free, highly repeatable, real-time, and cost-effective(3).

However, the traditional ultrasound diagnostic process has notable limitations. The accuracy of a diagnosis is heavily reliant on the sonographer's expertise, technical skill, and subjective judgment, which can lead to inter-observer variability and diagnostic discrepancies(4). In clinical practice, an experienced physician does not rely on a single, static image to assess a suspicious lesion. Instead, they perform a dynamic scan from multiple angles and planes to comprehensively observe multi-dimensional features such as the lesion's margins, shape, internal echo pattern, and posterior acoustic characteristics(5). This process allows them to build a

46 holistic, three-dimensional impression of the lesion before arriving at a final diagnosis. This
47 integrative analysis is central to diagnostic accuracy, and a key challenge in intelligent medical
48 imaging is to develop artificial intelligence (AI) that can effectively simulate and enhance this
49 complex cognitive process.

50 In recent years, AI technologies, particularly deep learning (DL), have made transformative
51 breakthroughs in medical image analysis(6, 7). Convolutional Neural Networks (CNNs), with
52 their powerful ability to automatically learn hierarchical feature representations, have become the
53 gold standard for processing grid-like data such as images(8-10). In the context of breast cancer
54 diagnosis, various classic CNN architectures like VGG and ResNet have been successfully
55 applied to classification tasks using mammography, histopathology, and ultrasound images,
56 demonstrating performance comparable to, or even superior to, that of human experts on specific
57 datasets(11-15). These models can autonomously learn subtle textural, morphological, and
58 boundary features associated with malignancy, reducing the reliance on manually engineered
59 features and paving the way for objective, efficient, and automated diagnosis.

60 Despite the remarkable success of CNN-based models, the majority of existing studies treat each
61 2D ultrasound image as an independent sample(16, 17). This approach fundamentally
62 disconnects the AI model from the clinical diagnostic reality. The question answered by a single-
63 image model is, "What is in this specific image?" whereas the essential clinical question is,
64 "How do all findings from all these images collectively point to a final diagnosis?" This
65 cognitive leap from single-point recognition to multi-dimensional integration represents a key
66 gap that current AI models must bridge. This study is motivated by the need to bridge this gap,
67 positing that the intelligent analysis of breast ultrasound should treat the complete set of multi-

view images from a single lesion as a holistic, information-rich sequence.

To model image sequences, some researchers have explored hybrid architectures, such as combining a CNN with a Recurrent Neural Network (RNN) or its variant, Long Short-Term Memory (LSTM)(18). While valuable for data with a clear temporal or spatial order, these models are ill-suited for the unordered sets of images typical in a breast ultrasound examination. The arbitrary acquisition order of these images conflicts with the inherent sequential dependency of RNNs. To overcome this, we introduce the Transformer model, originally developed for natural language processing(19, 20). The Transformer's core self-attention mechanism allows it to weigh the importance of all elements in a sequence simultaneously, regardless of their position. This property, known as permutation invariance, perfectly aligns with the clinical need to synthesize information from a collection of unordered images.

Based on this analysis, this study proposes a new intelligent diagnostic paradigm that more closely mirrors clinical logic. The primary contributions are:

- **Model Innovation:** An end-to-end hybrid model combining EfficientNetV2 and a Transformer for breast ultrasound classification.
- **Paradigm Innovation:** The novel treatment of multiple images from a single tumor as a sequence, leveraging the Transformer's global relationship modeling for comprehensive analysis.
- **Performance Validation:** Rigorous, tumor-level experiments demonstrating superior performance (AUC of 0.97 on an external test set) over traditional single-image models.
- **Clinical Potential:** A computationally effective solution to the core clinical challenge of multi-view information integration, showing immense promise for future high-precision,

reliable CAD systems.

2.Materials and Methods

This section provides a detailed description of the data sources, preprocessing pipeline, model architecture, experimental configuration, and performance evaluation standards to ensure the study's reproducibility. The overall workflow, illustrated in Fig 1, comprises three main stages: data preparation, model construction and training, and performance evaluation.

Fig 1: The overall workflow of the study.

2.1.Study Population and Ethical Approval

This retrospective study utilized datasets from two independent sources. The internal dataset was collected from The Second Affiliated Hospital of Guangzhou University of Chinese Medicine (Dade Road General Hospital) between January 2019 and December 2024. It comprised 832 breast tumors (430 benign, 402 malignant) from 488 female patients (age range: 11–88 years; mean age: 51.28 ± 13.55 years). The external validation dataset was collected from the University Town Branch of the same hospital over the same period, including 133 tumors (70 benign, 63 malignant) from 90 female patients (age range: 21–78 years; mean age: 53.62 ± 13.15 years).

Inclusion criteria were: (1) patients with complete clinical and final pathological results who underwent preoperative ultrasound; and (2) patients who had not received neoadjuvant chemotherapy or radiotherapy. Exclusion criteria were: (1) poor ultrasound image quality; (2) an interval between ultrasound and pathology greater than two weeks; and (3) prior surgery, biopsy, or chemotherapy.

The study was conducted in accordance with the Declaration of Helsinki and was approved by the Institutional Ethics Committee of The Second Affiliated Hospital of Guangzhou University of Chinese Medicine (Approval No. BE2025-207). The requirement for informed consent was waived due to the retrospective nature of the study.

2.2.Data Acquisition and Preprocessing

Ultrasound examinations were performed by experienced physicians using GE LOGIQ E9 and Philips EPIQ 7C systems with high-frequency linear probes (7.5–13.0 MHz). Standard images of each lesion were acquired and stored in the picture archiving and communication system (PACS).

The raw DICOM images were converted to JPG format. Two physicians, each with over five years of experience in breast ultrasound diagnosis, manually cropped the images to isolate the lesion and remove extraneous information, such as patient data and measurement markers (Fig 2). This step ensures the model focuses on relevant pathological features. All images were then resized to 299×299 pixels to match the model's input dimensions, and pixel values were normalized to a range of [0, 1].

Fig 2. Pre- and Post-processing of Ultrasound Images; Example of the image preparation workflow. (a) The original ultrasound image as exported from the PACS, containing patient information and measurement markers. (b) The image after being manually cropped by an expert physician to isolate the lesion of interest for model input.

2.3.Dataset Partitioning Strategy

A scientifically rigorous dataset partitioning strategy is fundamental to the objective and reliable evaluation of AI models in medical imaging. A common pitfall is the random allocation of multiple images from the same patient's tumor into the training, validation, and test sets. This practice leads to severe data leakage, as the model may learn to identify patient-specific artifacts rather than generalizable pathological features. This results in an artificially inflated performance on the test set, which plummets when the model is exposed to new data in a real-world setting.

To address this issue, this study employed a strict per-lesion partitioning methodology. Specifically, all images associated with a single tumor were treated as an indivisible unit and assigned exclusively to one of the three sets: training, validation, or test. The internal dataset was partitioned using a 70%:15%:15% ratio, ensuring complete independence at the lesion level. The entire external dataset was held out as a separate, independent test set to evaluate the model's generalization performance. The data distribution is detailed in Table 1.

Model Type	**Data Unit**	**Training Set**	**Validation Set**	**Test Set**	**External Test Set**	**Total**
Baseline	Images	2657	590	539	663	4449
Hybrid	Lesions	572	120	122	136	950

Table 1: Dataset distribution for the baseline and hybrid models.

2.4.Model Architectures

This study designed and compared two distinct models: a single-image classification baseline and the proposed CNN-Transformer hybrid model.

2.4.1.Baseline Model: EfficientNetV2 Single-Image Classifier

The baseline model was implemented using a standard transfer learning paradigm to establish a performance benchmark for a conventional single-image approach.

- **Backbone:** The efficientnet_v2_s model, pre-trained on the ImageNet dataset, was selected as the feature extractor due to its excellent balance between performance and efficiency.
- **Feature Extraction:** The parameters of the convolutional layers in the backbone were frozen, allowing it to function as a fixed feature extraction tool.
- **Classification Head:** The original fully connected layer was replaced with a custom head consisting of a linear layer (1280 input dimensions) and a single output neuron with a sigmoid activation function to produce a binary probability of malignancy.

2.4.2.Proposed Hybrid Model: CNN-Transformer Multi-View Classifier

This model uses a two-stage architecture to integrate information from multiple views of a lesion.

- **Stage 1:** Feature Extraction with EfficientNetV2 For a given tumor with N images, I_1,I_2,...,I_N, each image is passed through the EfficientNetV2 backbone to generate a 1280-dimensional feature vector f_iinmathbbR1280. This transforms the set of images into a sequence of feature vectors, S=[f_1,f_2,...,f_N].
- **Stage 2:** Feature Fusion with a Transformer Encoder
  - **Input Embedding:** A learnable classification token, [CLS], is prepended to the feature sequence. The input to the Transformer is this augmented sequence: S_aug=[textCLS,f_1,f_2,...,f_N]. The final state of the [CLS] token is used for classification.
  - **Positional Encoding:** Crucially, no positional encoding is used. This is a deliberate design choice to make the model permutation-invariant, forcing it to rely on the features

themselves, not their order. This aligns the model's logic with the clinical reality of analyzing an unordered set of images.

- **Self-Attention Structure:** The core of the model is a stack of 4 encoder layers. Key hyperparameters include a feature dimension of 1280, 8 attention heads, and a dropout rate of 0.1 for regularization.
- **Classification Head:** The final output vector corresponding to the [CLS] token is passed through a linear layer and a sigmoid function to predict the probability of malignancy for the entire lesion.

2.5.Experimental Settings and Training Parameters

All experiments were conducted in the following environment:

- **Hardware Platform:** NVIDIA GeForce RTX 4060Ti GPU
- **Software Environment:** Python 3.10, PyTorch 2.7.1, torchvision, scikit-learn, pandas, numpy

To ensure the reproducibility of the experiments, the detailed hyperparameters are provided in Table 2.

Parameter	**Baseline Model**	**Hybrid Model**
Optimizer	AdamW	AdamW
Initial Learning Rate	1×10^{-4}	5×10^{-5}
Learning Rate Scheduler	Cosine Annealing	Cosine Annealing
Batch Size	32 (images)	4 (lesions)

Epochs	100 (with Early Stopping)	100 (with Early Stopping)
Early Stopping Patience	5	5
Monitored Metric	Validation AUC	Validation AUC
Loss Function	Binary Cross-Entropy	Binary Cross-Entropy

Table 2: Hyperparameter configuration for model training.

2.6.Evaluation Metrics

Model performance was assessed using five standard binary classification metrics. Let TP, TN, FP, and FN represent True Positives, True Negatives, False Positives, and False Negatives, respectively.

- **Accuracy:** Measures the proportion of all predictions that are correct.

$$\text{Accuracy}=TP+TN+FP+FNTP+TN$$

- **Sensitivity:** Measures the model's ability to correctly identify malignant samples.

$$\text{Sensitivity}=TP+FNTP$$

- **Specificity:** Measures the model's ability to correctly identify benign samples.

$$\text{Specificity}=TN+FPTN$$

- **AUC (Area Under the ROC Curve):** Measures the model's overall discriminative ability across all classification thresholds. An AUC closer to 1 indicates better performance.

3.Results

3.1.Overall Model Performance Comparison

The proposed CNN-Transformer hybrid model demonstrated a marked improvement in diagnostic performance over the conventional single-image baseline model across all evaluation metrics. This advantage was consistent on both the internal test set and, crucially, on the independent external validation set, underscoring the model's robustness and generalization capabilities.

On the internal test set, which was partitioned at the tumor level to prevent data leakage, the hybrid model achieved an accuracy of 0.93 and an AUC of 0.98. When challenged with the external test set, comprising data acquired with different equipment, the model's performance remained remarkably stable, yielding an accuracy of 0.93 and an AUC of 0.97. In contrast, the baseline EfficientNetV2 model, which classifies each image independently, achieved lower metrics on both test sets. These results, detailed in Table 3, provide strong evidence that the fusion of multi-view information through the Transformer architecture leads to a significant leap in diagnostic power.

Model	**Test Set**	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**
Baseline Model (EfficientNetV2)	**Internal**	0.88	0.86	0.91	0.95
	External	0.89	0.91	0.82	0.94
Proposed Hybrid Model (CNN-Transformer)	**Internal**	0.93	0.92	0.94	0.98
	External	0.93	0.94	0.91	0.97

Table 3: Performance comparison of the two models on the test sets.

3.2.Analysis of Generalization Ability

To assess the model's learning dynamics and potential for overfitting, the Area Under the Curve (AUC) was tracked for both the training and validation sets throughout the training process (Fig 3). The training and validation AUC curves rose in tandem and remained closely aligned, ultimately converging at a high level of performance (>0.96). This indicates that the model did not overfit to the training data; instead, it learned generalizable features applicable to unseen data.

Fig 3. Model Training and Validation Curves (AUC vs. Epoch).

The ultimate test of generalization is performance on entirely new data. The Receiver Operating Characteristic (ROC) curves for the validation, internal test, and external test sets are presented in Fig 4. The curves are nearly superimposed, with high AUC values of 0.965, 0.982, and 0.968, respectively. This remarkable consistency across three independent datasets strongly suggests that the model has learned robust, universally applicable diagnostic features, validating both the architectural design and the rigorous data partitioning strategy.

Fig 4. ROC Curves on Validation, Internal Test, and External Test Sets.

3.3.Detailed Classification Performance

A detailed breakdown of the model's classification performance is presented via confusion matrices for the internal and external test sets (Fig 5) and a comprehensive classification report (Table 4).

On the internal test set, the confusion matrix (Fig 5a) shows the model correctly identified 61 of 65 benign lesions and 56 of 61 malignant lesions. As detailed in Table 4, this performance corresponds to a recall of 0.94 (specificity) and a precision of 0.92 for benign cases. For the crucial task of identifying malignancy, the model achieved a recall of 0.92 (sensitivity) and a precision of 0.93. The excellent F1-scores of 0.93 for both classes highlight a strong and desirable balance between precision and recall.

The model's robustness was further validated on the external test set (Fig 5b). It achieved a recall (specificity) of 0.91 and a precision of 0.94 for benign lesions. For malignant lesions, it demonstrated an even higher recall (sensitivity) of 0.94, with a corresponding precision of 0.91. This indicates a very low rate of missed malignant cases. The resulting F1-scores of 0.93 for benign and 0.92 for malignant classes confirm the model's strong diagnostic capability on unseen data.

In summary, the consistently high performance across both datasets, highlighted by macro-averaged F1-scores of 0.93 (internal) and 0.92 (external), affirms the model's reliability and strong generalization potential.

Fig 5. Confusion Matrices on the (a) Internal and (b) External Test Sets

Test Set	Class	Precision	Recall	F1-Score	Support
Internal	**Benign**	0.92	0.94	0.93	65
	Malignant	0.93	0.92	0.93	61
	Macro Avg	0.93	0.93	0.93	126
External	**Benign**	0.94	0.91	0.93	70

Malignant	0.91	0.94	0.92	63
Macro Avg	0.92	0.93	0.92	133

Table 4: Detailed Classification Report on the Internal and External Test Sets.

3.4.Prediction Confidence Analysis

Beyond correctness, a model's confidence in its predictions is an important dimension of its performance. Fig 6 shows the distribution of the model's predicted probabilities for all benign and malignant samples in the test sets. For benign samples (blue), the prediction probabilities are highly concentrated near 0.0. For malignant samples (red), the probabilities are overwhelmingly clustered near 1.0. The minimal overlap between the two distributions visually explains the model's high accuracy and AUC values, demonstrating that it not only makes correct judgments but does so with a high degree of certainty.

Fig 6:Histogram of predicted probability distributions on the (a) internal and (b) external test sets.

3.5.Model Interpretability using Class Activation Mapping

To investigate the model's decision-making basis, Class Activation Mapping (CAM) was used to visualize the regions of interest the CNN feature extractor focused on. Although the final diagnosis results from the Transformer's fusion of features, CAM provides insight into whether the foundational feature extraction is clinically relevant.

As shown in Fig 7, when the model correctly classifies a lesion as malignant, the high-attention areas (heatmaps) precisely overlap with clinically significant pathological features, such as irregular margins, internal microcalcifications and angular contours. This visualization confirms

that the CNN component effectively learns to identify meaningful biomarkers, providing a reliable and interpretable foundation for the Transformer's subsequent global analysis. This enhances trust in the model's decisions by demonstrating that its reasoning is aligned with established clinical diagnostic criteria.

Figure 7. Visualization of Class Activation Mapping (CAM) for model interpretability. (a) Color Doppler and (b) 2D grayscale ultrasound images of a malignant tumor. (c) and (d) are the corresponding CAM heatmaps, where high-activation areas (red) are concentrated on diagnostically significant regions of the lesion.

4.Discussion

The results of this study conclusively demonstrate that the proposed CNN-Transformer hybrid model is significantly superior to conventional single-image CNN approaches for classifying breast tumors in ultrasound images. The core reason for this performance gain lies in an architecture that more faithfully emulates the cognitive process of clinical experts.

The effectiveness of multi-view fusion is the central finding. A single-image model processes each view in isolation, whereas the Transformer encoder in our hybrid model functions as a "virtual expert panel." Its self-attention mechanism allows the model to dynamically weigh and integrate information from the entire set of images for a given lesion(21, 22). For instance, it can learn to assign greater importance to a view that clearly depicts spiculated margins while down-weighting another view obscured by acoustic artifacts. This dynamic, global evidence integration is a powerful analogy for expert diagnostic reasoning and is the key driver behind the model's performance leap. This architecture also confers greater robustness to the variations inherent in

real-world clinical data, such as differences in image quality or the number of available views, as the model can learn to prioritize high-quality, informative images(23, 24).

The clinical implications of the model's performance are substantial. The high specificity achieved on both internal (94%) and external (91%) test sets indicates a strong ability to correctly identify benign lesions. This could translate to a significant reduction in false-positive findings, potentially sparing many women from the anxiety, physical discomfort, and cost of unnecessary biopsies. Concurrently, the model's high sensitivity (92% internal, 94% external) ensures a low rate of missed malignant tumors, which is critical for facilitating early diagnosis and treatment.

An interesting observation is the slight trade-off between sensitivity and specificity on the external dataset, where sensitivity increased at the cost of a minor drop in specificity. This may reflect subtle differences in patient populations or imaging protocols between the two institutions. It also highlights the clinical utility of an adjustable decision threshold. While the model currently uses a default threshold of 0.5, this could be lowered in a screening setting to maximize sensitivity and minimize the risk of missing a cancer, or raised in a diagnostic setting to maximize specificity and reduce false positives. This flexibility allows the model to be adapted to different clinical needs.

Despite its promising results, this study has several limitations. First, while an external validation set was used, testing on data from a wider variety of institutions, encompassing different geographic regions and equipment vendors, is necessary to fully establish the model's generalizability. Second, the computational cost of the Transformer architecture is considerable, which could pose a challenge for deployment in resource-constrained environments. Finally,

while CAM provides some insight, the "black box" nature of deep learning remains a barrier to full clinical adoption, and further work on enhancing model interpretability is crucial for building trust with clinicians.

Future work will focus on addressing these limitations. A large-scale, multi-center external validation study is the highest priority. We also plan to explore model optimization techniques, such as knowledge distillation and quantization, to create a lightweight version of the model suitable for deployment on standard clinical workstations. Ultimately, the goal is to integrate this technology into a user-friendly CAD system and conduct prospective clinical trials to evaluate its real-world impact on diagnostic accuracy, workflow efficiency, and patient outcomes.

5.Conclusion

This study successfully developed and validated a novel CNN-Transformer hybrid deep learning model for breast ultrasound diagnosis. By synergistically combining the powerful local feature extraction of a CNN with the sophisticated global information fusion capabilities of a Transformer, the model effectively simulates the multi-view diagnostic process of clinical experts. Using a rigorous, tumor-level evaluation methodology, the proposed model demonstrated outstanding classification performance on an internal test set (AUC ≈ 0.98) and proved its excellent generalization on an independent external test set (AUC ≈ 0.97). The consistently high accuracy, sensitivity, and specificity across diverse datasets significantly exceed those of traditional single-image methods. This work not only validates a superior technical architecture but also underscores the methodological importance of multi-view analysis and robust external validation. The model shows immense potential as a powerful and reliable tool to assist clinicians in making earlier and more accurate diagnoses of breast cancer, with the

ultimate aim of improving patient health and well-being.

Acknowledgments

This work was funded by the Guangdong Provincial Administration of Traditional Chinese Medicine, under grant No. (20242032). Therefore, the authors thank the Guangdong Provincial Administration of Traditional Chinese Medicine for its financial support.

References

1.Xiong X, Zheng LW, Ding Y, Chen YF, Cai YW, Wang LP, et al. Breast cancer: pathogenesis and treatments. Signal Transduct Target Ther. [Journal Article; Review]. 2025 2025-02-19;10(1):49.

2.Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. [Journal Article]. 2021 2021-05-01;71(3):209-49.

3.Wang Y, Li Y, Song Y, Chen C, Wang Z, Li L, et al. Comparison of ultrasound and mammography for early diagnosis of breast cancer among Chinese women with suspected breast lesions: A prospective trial. THORAC CANCER. [Clinical Trial; Journal Article; Research Support, Non-U.S. Gov't]. 2022 2022-11-01;13(22):3145-51.

4.Wang H, Yang X, Ma S, Zhu K, Guo S. An Optimized Radiomics Model Based on Automated Breast Volume Scan Images to Identify Breast Lesions: Comparison of Machine Learning Methods: Comparison of Machine Learning Methods. J Ultrasound Med. [Journal Article]. 2022 2022-07-01;41(7):1643-55.

5.Evans A, Trimboli RM, Athanasiou A, Balleyguier C, Baltzer PA, Bick U, et al. Breast ultrasound: recommendations for information to women and referring physicians by the

356 European Society of Breast Imaging. Insights Imaging. [Journal Article]. 2018 2018-08-
357 01;9(4):449-61.
358 6.Rajpurkar P, Lungren MP. The Current and Future State of AI Interpretation of Medical
359 Images. N Engl J Med. [Journal Article; Review]. 2023 2023-05-25;388(21):1981-90.
360 7.Assié G, Allassonnière S. Artificial Intelligence in Endocrinology: On Track Toward Great
361 Opportunities. J Clin Endocrinol Metab. [Comment; Journal Article]. 2024 2024-05-
362 17;109(6):e1462-7.
363 8.Rawat W, Wang Z. Deep Convolutional Neural Networks for Image Classification: A
364 Comprehensive Review. NEURAL COMPUT. [Journal Article; Research Support, Non-U.S.
365 Gov't]. 2017 2017-09-01;29(9):2352-449.
366 9.Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, et al. Man against
367 machine: diagnostic performance of a deep learning convolutional neural network for
368 dermoscopic melanoma recognition in comparison to 58 dermatologists. ANN ONCOL.
369 [Comparative Study; Journal Article; Observational Study; Validation Study]. 2018 2018-08-
370 01;29(8):1836-42.
371 10.Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep Convolutional Neural
372 Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and
373 Transfer Learning. IEEE Trans Med Imaging. [Journal Article]. 2016 2016-05-01;35(5):1285-98.
374 11.Qu X, Lu H, Tang W, Wang S, Zheng D, Hou Y, et al. A VGG attention vision transformer
375 network for benign and malignant classification of breast ultrasound images. MED PHYS.
376 [Journal Article]. 2022 2022-09-01;49(9):5787-98.
377 12.An F, Li X, Ma X. Medical Image Classification Algorithm Based on Visual Attention
378 Mechanism-MCNN. OXID MED CELL LONGEV. [Journal Article]. 2021 2021-01-

379 20;2021:6280690.
380 13.Yousif M, van Diest PJ, Laurinavicius A, Rimm D, van der Laak J, Madabhushi A, et al.
381 Artificial intelligence applied to breast pathology. VIRCHOWS ARCH. [Journal Article;
382 Review]. 2022 2022-01-01;480(1):191-209.
383 14.Harrison P, Hasan R, Park K. State-of-the-Art of Breast Cancer Diagnosis in Medical Images
384 via Convolutional Neural Networks (CNNs). J Healthc Inform Res. [Journal Article; Review].
385 2023 2023-12-01;7(4):387-432.
386 15.Sechopoulos I, Teuwen J, Mann R. Artificial intelligence for breast cancer detection in
387 mammography and digital breast tomosynthesis: State of the art. SEMIN CANCER BIOL.
388 [Journal Article; Review]. 2021 2021-07-01;72:214-25.
389 16.Songpan W, Chaianunporn T, Lomthaisong K, Keithmaleesatti S. Automated classification in
390 turtles genus Malayemys using ensemble multiview image based on improved YOLOv8 with
391 CNN. Sci Rep. [Journal Article]. 2024 2024-10-23;14(1):24993.
392 17.Xiang K, Shang D. Construction of a Multi-View Deep Learning Model for the Severity
393 Classification of Acute Pancreatitis. DISCOV MED. [Journal Article]. 2025 2025-01-
394 01;37(192):73-92.
395 18.Kaddes M, Ayid YM, Elshewey AM, Fouad Y. Breast cancer classification based on hybrid
396 CNN with LSTM model. Sci Rep. [Journal Article]. 2025 2025-02-05;15(1):4409.
397 19.Wu G, Zheng WS, Lu Y, Tian Q. PSLT: A Light-Weight Vision Transformer With Ladder
398 Self-Attention and Progressive Shift. IEEE Trans Pattern Anal Mach Intell. [Journal Article].
399 2023 2023-09-01;45(9):11120-35.
400 20.Maziarka Ł, Majchrowski D, Danel T, Gaiński P, Tabor J, Podolak I, et al. Relative molecule
401 self-attention transformer. J Cheminform. [Journal Article]. 2024 2024-01-03;16(1):3.

21.Li J, Chen J, Tang Y, Wang C, Landman BA, Zhou SK. Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives. MED IMAGE ANAL. [Journal Article; Review]. 2023 2023-04-01;85:102762.

22.Liu X, Zhu X, Li M, Wang L, Tang C, Yin J, et al. Late Fusion Incomplete Multi-View Clustering. IEEE Trans Pattern Anal Mach Intell. [Journal Article]. 2019 2019-10-01;41(10):2410-23.

23.Kumar A, Yadav SP, Kumar A. An improved feature extraction algorithm for robust Swin Transformer model in high-dimensional medical image analysis. COMPUT BIOL MED. [Journal Article; Review]. 2025 2025-04-01;188:109822.

24.Gong P, Cheng L, Zhang Z, Meng A, Li E, Chen J, et al. Multi-omics integration method based on attention deep learning network for biomedical data classification. Comput Methods Programs Biomed. [Journal Article]. 2023 2023-04-01;231:107377.

Author Contributions

Z. Q. and P.T. conceived the experiment(s). Z.Q. conducted the experiment(s) and wrote the initial manuscript draft. C.M. and J.Z. supervised the project and provided critical revisions. All authors analysed the results and reviewed the final manuscript.

Funding

This work was funded by the Guangdong Provincial Administration of Traditional Chinese Medicine, under grant No. (20242032).

Competing Interests

The authors declare no competing interests.

Ethical Statement

The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. This study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The medical ethics committee of the Second Affiliated Hospital of Guangzhou University of Chinese Medicine approved the study and waived the requirement for written informed consent due to the retrospective nature of the study.

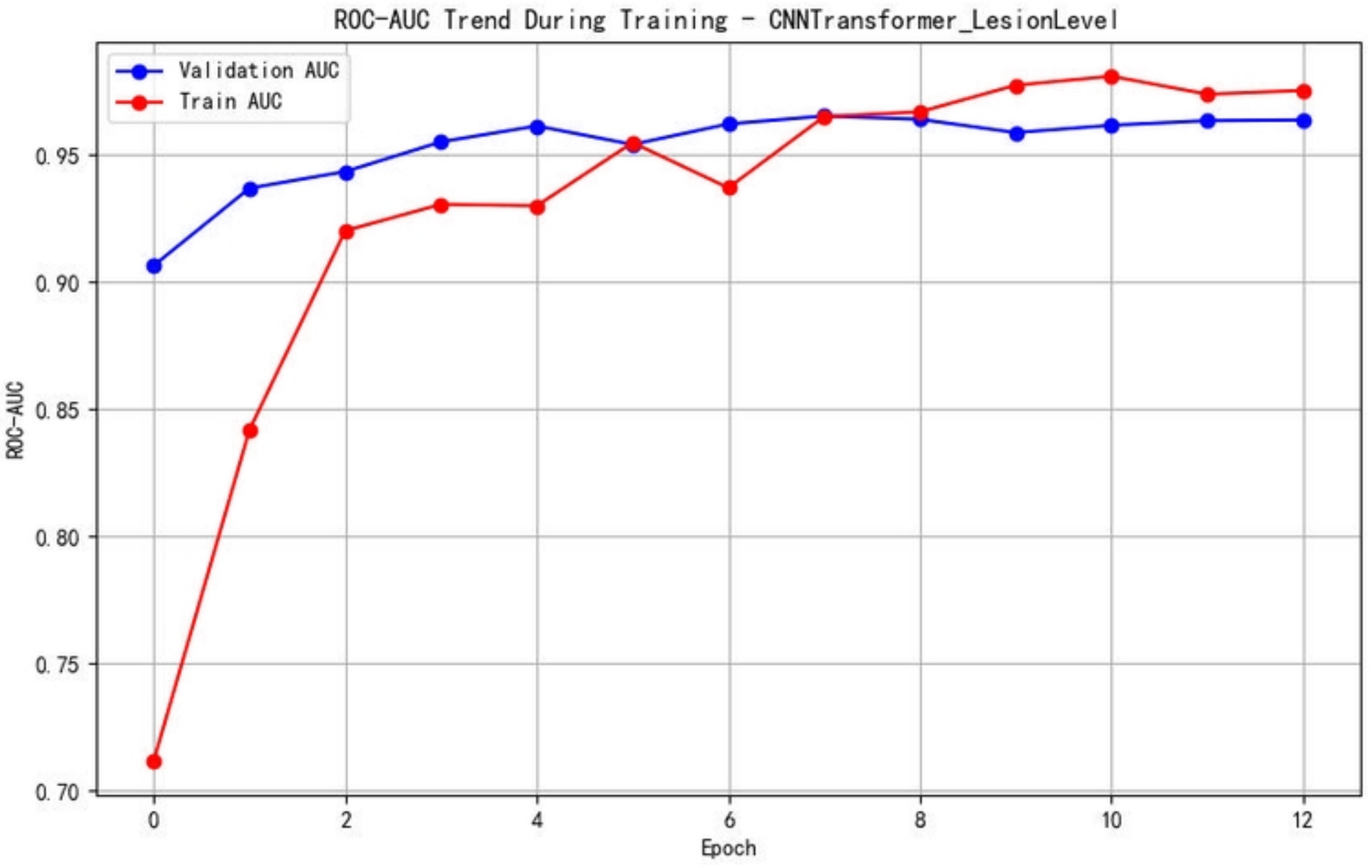

ROC-AUC Trend During Training - CNNTransformer_LesionLevel
Validation AUC
Train AUC
ROC-AUC
0.95
0.90
0.85
0.80
0.75
0.70
0
2
4
6
8
10
12
Epoch

Figure3

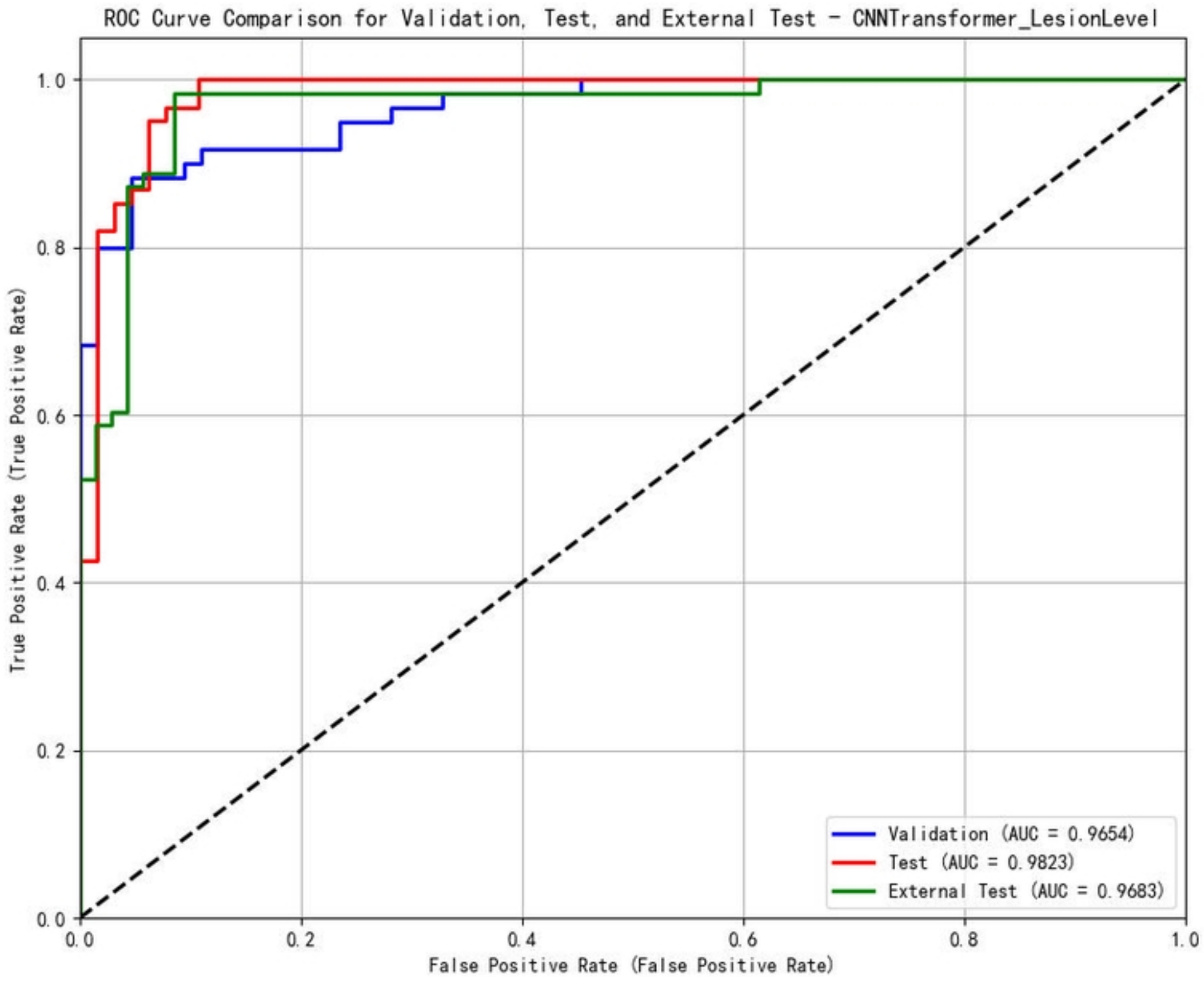
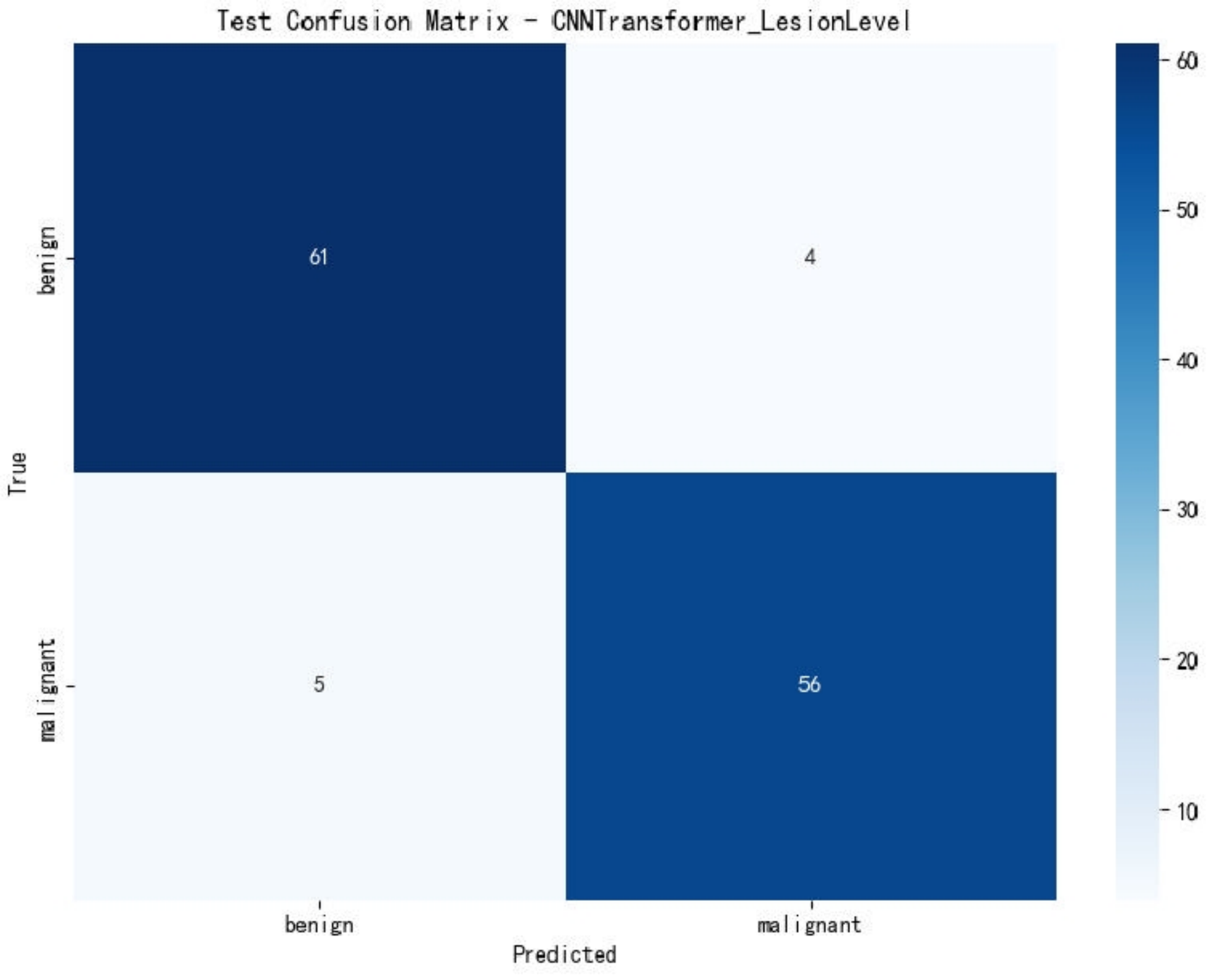
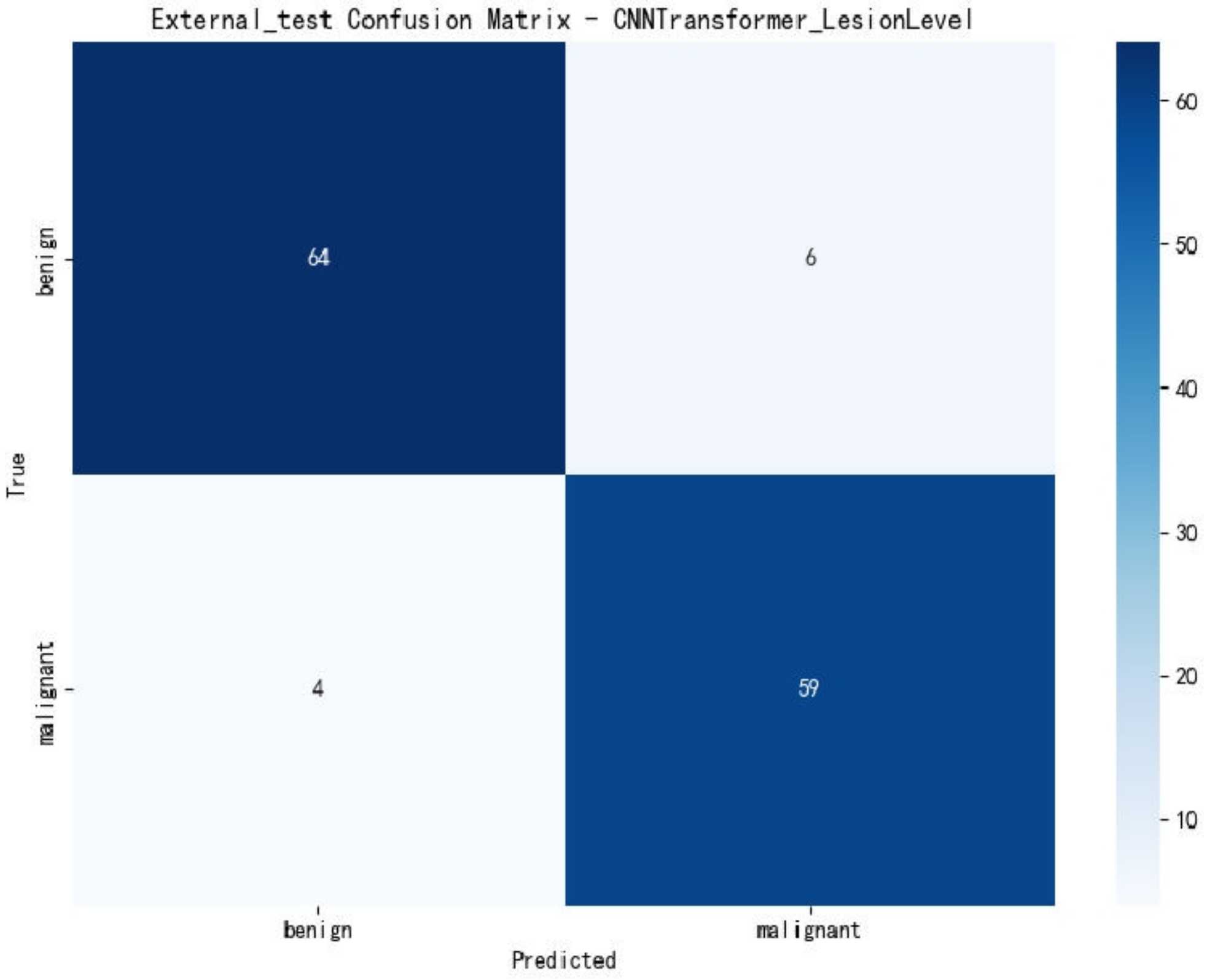

ROC Curve Comparison for Validation, Test, and External Test - CNNTransformer_LesionLevel
True Positive Rate (True Positive Rate)
False Positive Rate (False Positive Rate)
0.0
0.2
0.4
0.6
0.8
1.0
Validation (AUC = 0.9654)
Test (AUC = 0.9823)
External Test (AUC = 0.9683)

Figure4


Test Confusion Matrix - CNNTransformer_LesionLevel
61
4
5
56
benign
malignant
True
benign
malignant
Predicted
60
50
40
30
20
10

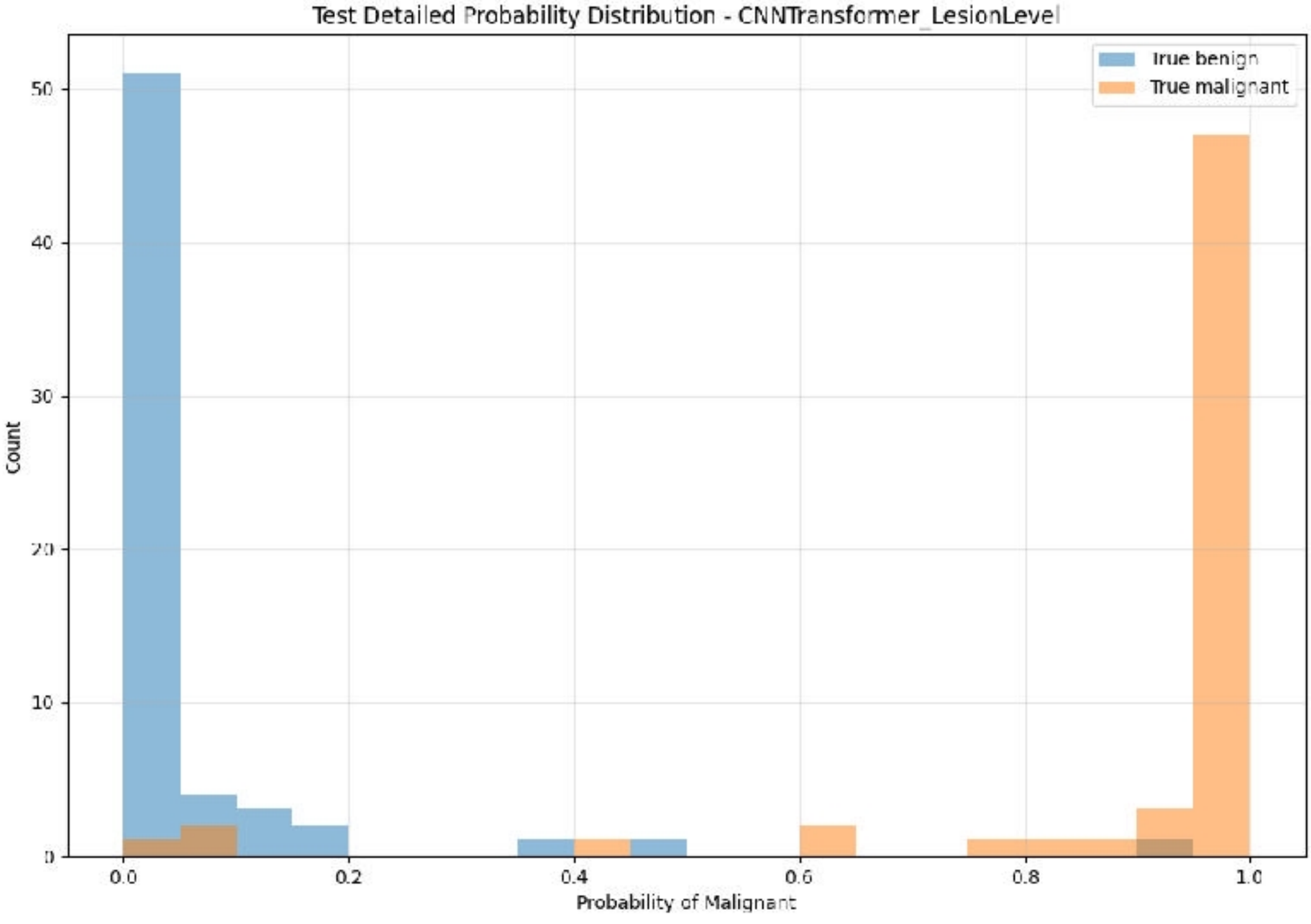
a

Figure5 a


External_test Confusion Matrix - CNNTransformer_LesionLevel
64
6
4
59
benign
malignant
True
benign
malignant
Predicted
60
50
40
30
20
10

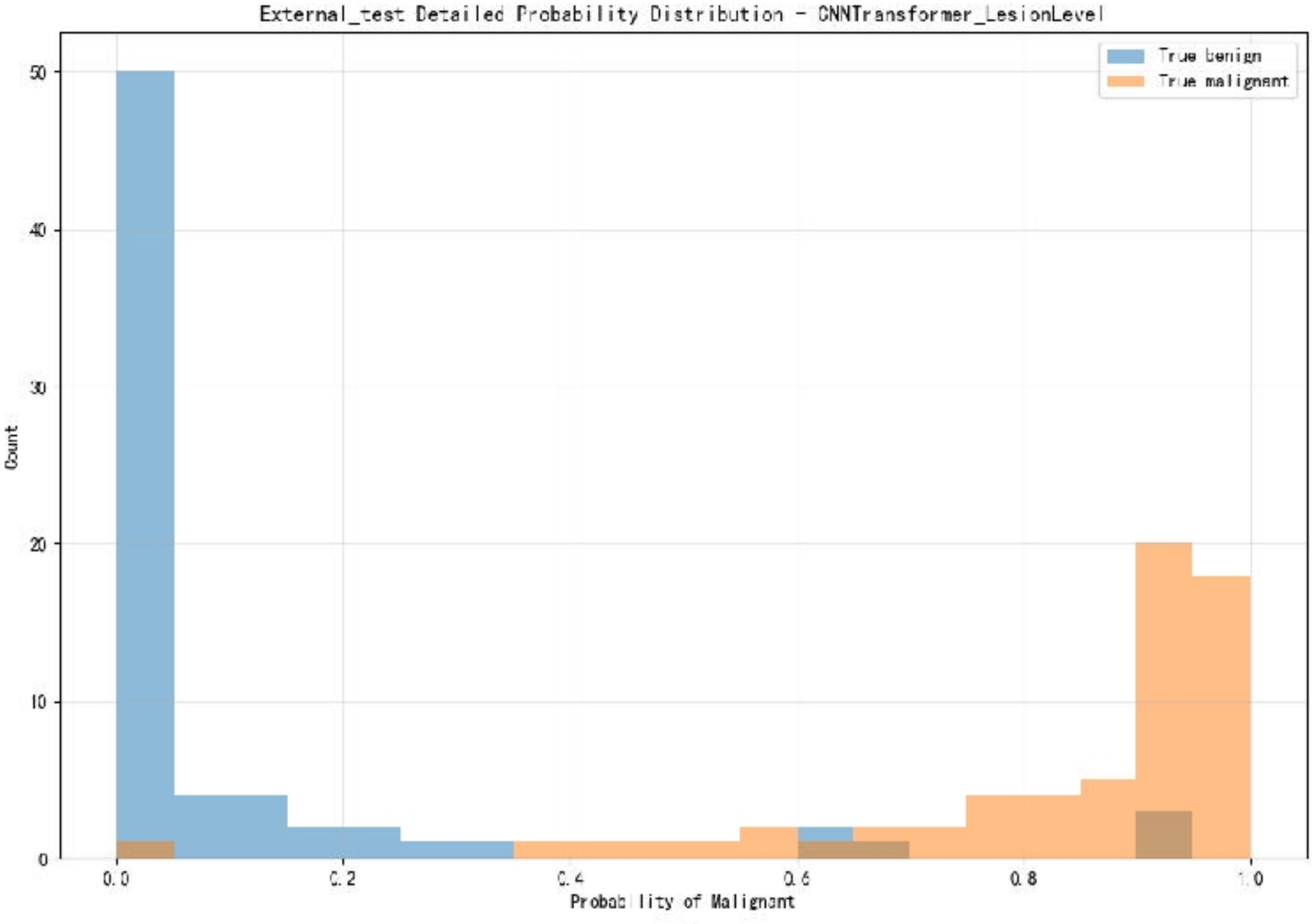
b

Figure5 b


Test Detailed Probability Distribution - CNNTransformer_LesionLevel
True benign
True malignant
Count
Probability of Malignant
0
10
20
30
40
50
0.0
0.2
0.4
0.6
0.8
1.0

a

Figure6 a


External_test Detailed Probability Distribution - CNNTransformer_LesionLevel
True benign
True malignant
Count
Probability of Malignant
0
10
20
30
40
50
0.0
0.2
0.4
0.6
0.8
1.0

b

Figure6 b

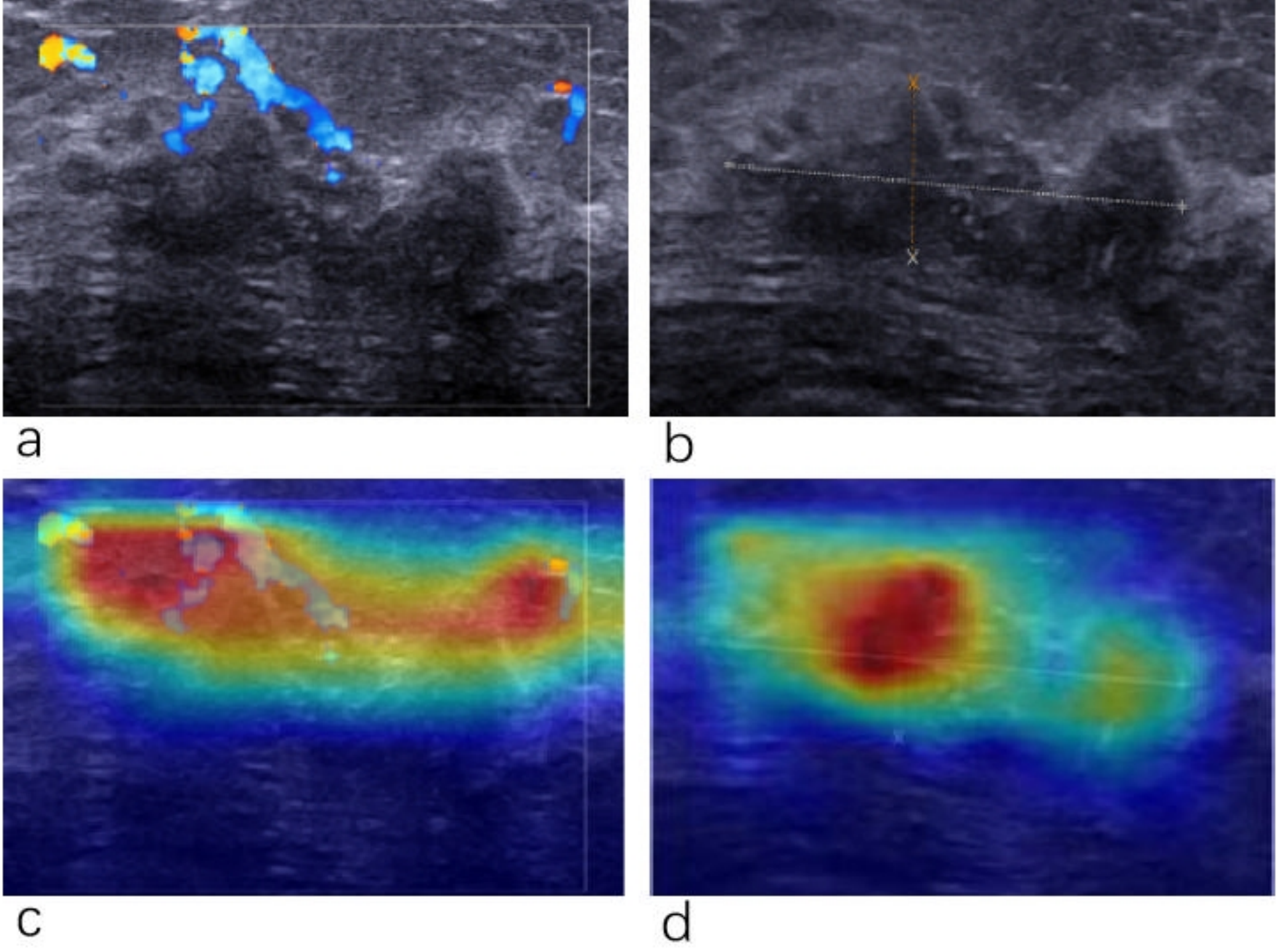


Figure7

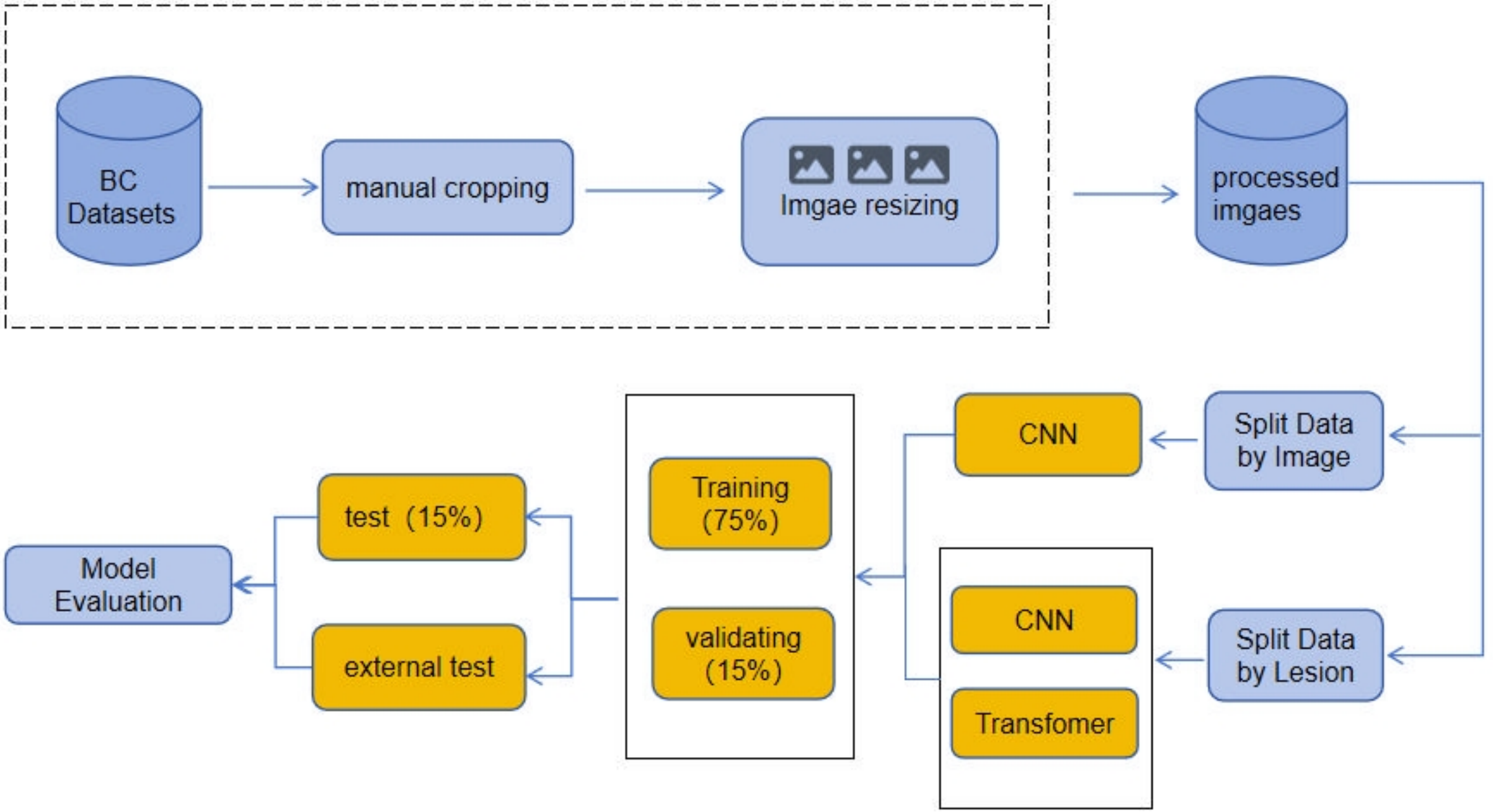
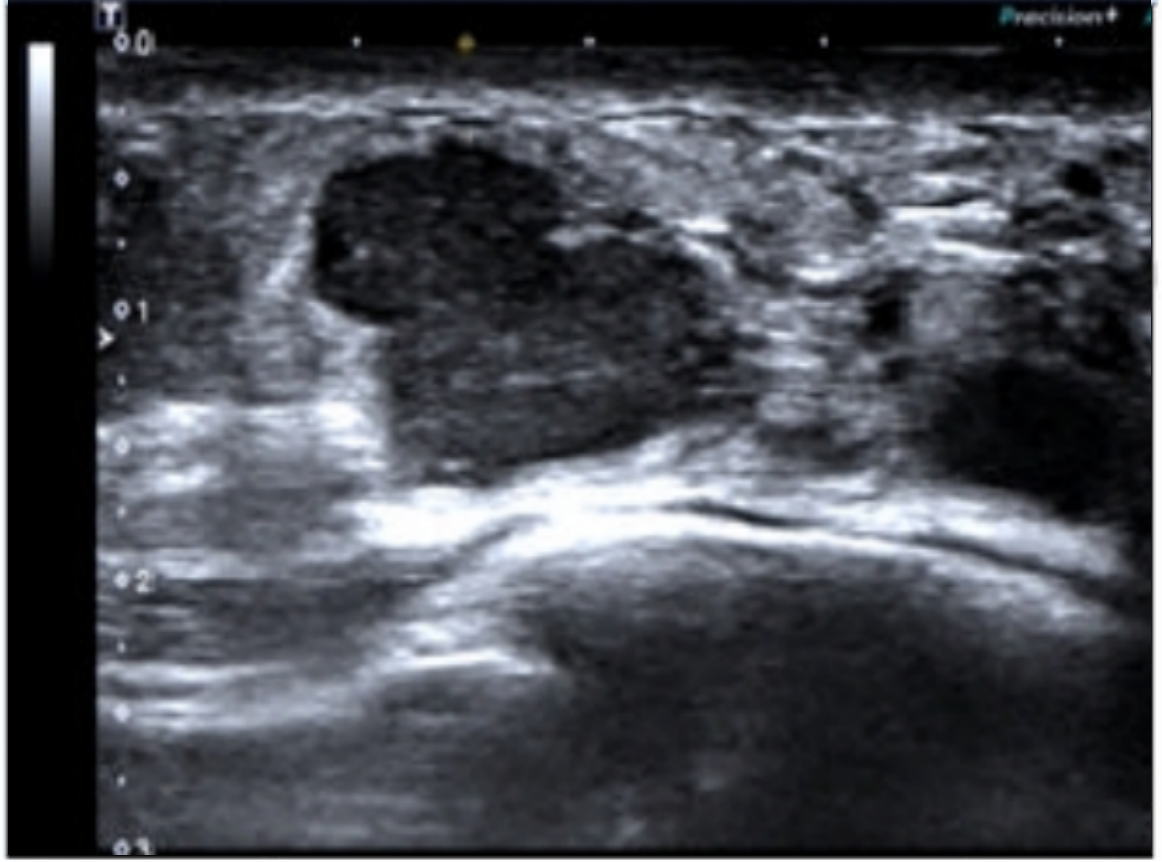
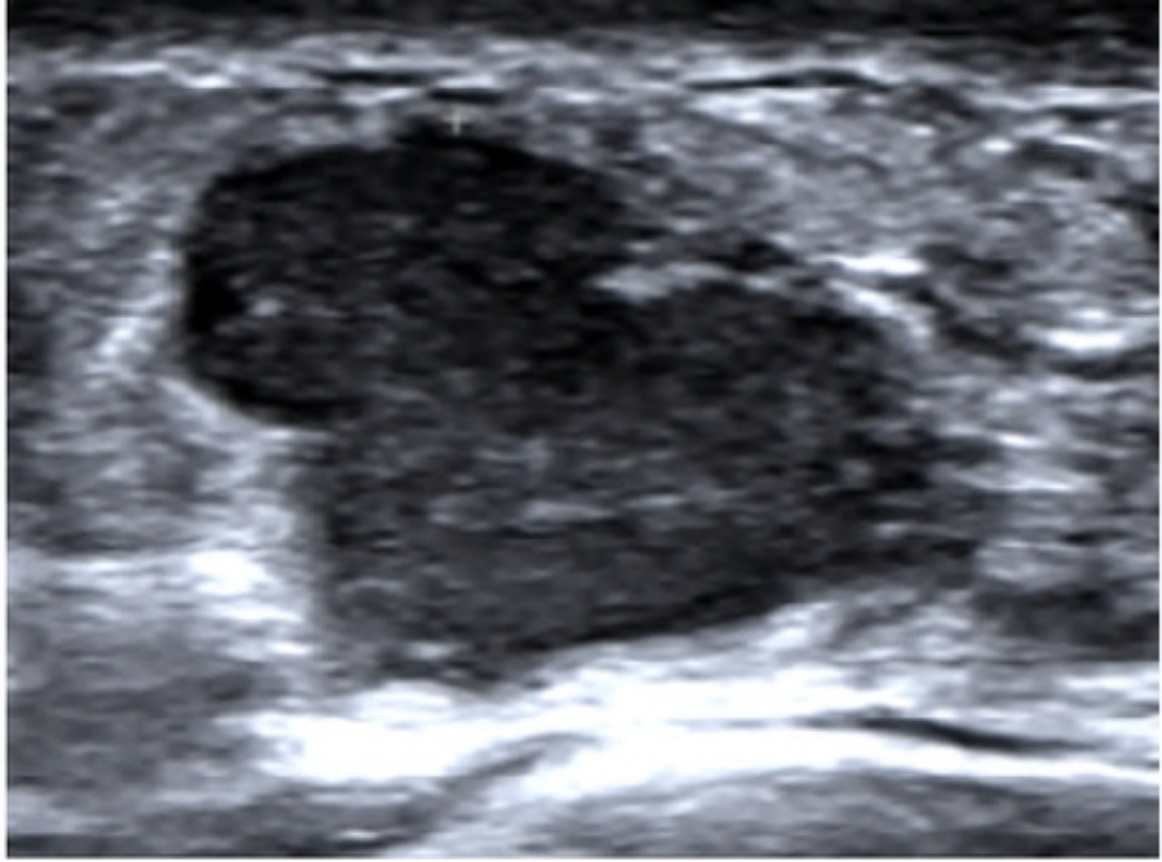

BC Datasets
manual cropping
Imgae resizing
processed imgaes
Split Data by Image
CNN
Split Data by Lesion
CNN
Transfomer
Training (75%)
validating (15%)
test (15%)
external test
Model Evaluation

Figure1



a



b

Figure2