American Journal of Smart Technology and Solutions (AJSTS)

ISSN: 2837-0295 (ONLINE)

VOLUME 5 ISSUE 2 (2026)



PUBLISHED BY
E-PALLI PUBLISHERS, DELAWARE, USA

American Journal of Smart Technology and Solutions (AJSTS)

Volume 5 Issue 2, Year 2026
ISSN: 2837-0295 (Online)
DOI: https://doi.org/10.54536/ajsts.v5i2.8842
https://journals.e-palli.com/home/index.php/ajsts

HCT-Mammo: A Compact Hybrid CNN-Transformer for Benign–Malignant Mammogram Classification with Uncertainty and Cross-Dataset Evaluation

Mst Kamrun Nahar Borsha[1], Md Shakibur Rahman[1], Md Sakawat Hossain[2*]

Article Information

Received: July 31, 2026

Accepted: October 16, 2026

Published: November 01, 2026

Keywords

Calibration, Cross-Dataset Evaluation, Hybrid CNN-Transformer, Mammography, Uncertainty Estimation

ABSTRACT

The difficulty in comparing public-dataset results for benign–malignant mammogram classification is due to patient overlap, lesion crops, and a lack of independent testing. This study uses 896×448 breast-cropped mammograms, and the evaluated model is a hybrid convolutional neural network (CNN)–Transformer (HCT-Mammo) with a truncated ResNet-18 stem, a 4-layer Transformer encoder, and 4.91 M parameters. Patient-level separation, training, and model selection were done on the Curated Breast Imaging Subset of DDSM (CBIS-DDSM), and the models were frozen for zero-shot evaluation on the Chinese Mammography Database (CMMD). Sigmoid Confidence, Temperature Scaling and Monte Carlo (MC) Dropout were evaluated with calibration and risk–coverage analyses, and Grad-CAM and Attention Rollout were compared to lesion masks and chance baselines. The mean area under the receiver operating characteristic curve (AUROC) of HCT-Mammo was 0.780 ± 0.005 across three seeds, with a median-seed AUROC of 0.781 (95% confidence interval 0.742–0.821). HCT-Mammo did not differ significantly from EfficientNet-B0 and ResNet-50 (DeLong $p = 0.59$ and $p = 0.33$), whereas ViT-S/16 was poorer ($p = 0.002$). On CMMD, AUROC decreased to 0.664, and the expected calibration error after temperature scaling increased from 0.059 to 0.121. MC-Dropout offered no obvious improvement in calibration or error ranking, internally or externally. The mass localisation hit rate for Grad-CAM was 35.5%, compared with 24.4% for Attention Rollout and 4.7% for a breast-uniform baseline. The compact hybrid was competitive in-domain, but its discrimination and calibration were worse across datasets. Both datasets contain only abnormal examinations; therefore, the results reflect lesion characterization rather than screening.

INTRODUCTION

Breast cancer is among the most frequently diagnosed cancers worldwide (Bray *et al.*, 2024), and mammography remains the main imaging tool for its detection. Deep learning systems have been evaluated for population-wide breast cancer screening (McKinney *et al.*, 2020; L. Shen *et al.*, 2019). A related but distinct task is associated with public benchmarks such as the Curated Breast Imaging Subset of DDSM (CBIS-DDSM) (Lee *et al.*, 2017) and the Chinese Mammography Database (CMMD) (Cai *et al.*, 2023). It consists of discriminating benign from malignant findings in lesions that have already been detected on a mammogram and confirmed by biopsy. Both datasets contain abnormal examinations only, so their results reflect lesion characterisation and should not be interpreted as screening performance.

Two technical difficulties need to be addressed in this task. First, findings such as small calcifications and spiculated margins occupy only a small part of a full mammogram, so aggressive downscaling to standard network input sizes can cause information loss, whereas processing high-resolution images requires substantial computational resources. Second, not only local features but also the surrounding context matter, since a finding is interpreted against the background of the adjacent parenchyma. Hybrid networks that place a convolutional stem in front of a Transformer encoder are a natural way to combine local feature extraction with long-range interaction (Dosovitskiy *et al.*, 2021; Xiao *et al.*, 2021), and several hybrid models have been developed for mammography (Aldawsari *et al.*, 2026; Badawi *et al.*, 2026; Hoang *et al.*, 2026; Jeny *et al.*, 2025).

However, evaluation practice has not advanced alongside architectural development. Results on CBIS-DDSM are hard to compare across studies because the unit of analysis (lesion crops versus whole images), the treatment of patients who appear in more than one partition, and the input resolution all differ. Predictive accuracy is the measure usually reported, but predictive confidence should also be examined: a classifier used as a decision aid should be judged by whether its probabilities are calibrated and whether its uncertainty ranks its own errors (Geifman & El-Yaniv, 2017; Guo *et al.*, 2017). Independent-dataset testing is also uncommon, even though the collections differ in acquisition technology and population, and dataset shift is well known to degrade both accuracy and uncertainty estimates (Ovadia *et al.*, 2019). In addition, saliency maps are often shown without any comparison to a chance baseline, while the faithfulness measures commonly used to justify attribution methods can fail to separate real attributions from random ones (Wu *et al.*, 2024).

[1] School of Computer Science & Technology, Nanjing University of Information Science & Technology, Nanjing, China
[2] School of Artificial Intelligence, Nanjing University of Information Science & Technology, Nanjing, China
[*] Corresponding author's e-mail: 20235346002@nuist.edu.cn

This paper therefore does not propose a new architecture. Instead, it presents a controlled evaluation of a compact hybrid CNN-Transformer, HCT-Mammo, and answers three questions. How does HCT-Mammo compare with standard convolutional neural network (CNN) and Transformer baselines trained under the same protocol at the same working resolution? Can Monte Carlo (MC) Dropout provide anything beyond sigmoid confidence and temperature scaling, both in-domain and on an independent dataset? How consistent are attention-based and gradient-based maps with lesion annotations once chance baselines are taken into account?

The contributions of this paper are fourfold:

- A compact single-view hybrid CNN-Transformer (4.91 M parameters and 17.9 Giga Multiply–Accumulate Operations (GMACs) at 896×448), compared with EfficientNet-B0, ResNet-50, and ViT-S/16 under an identical patient-level protocol, with an ablation of the Transformer stage, loss function, and preprocessing.
- A zero-shot cross-dataset test in which the CBIS-DDSM-trained model, decision threshold, and temperature are frozen before testing on CMMD, highlighting the differences in discrimination and calibration.
- An analysis of uncertainty and selective prediction that compares sigmoid confidence, temperature scaling, and MC-Dropout using calibration, error-detection, and risk–coverage measures on both datasets.
- A localisation-consistency comparison of Attention Rollout and Grad-CAM against CBIS-DDSM lesion masks, with uniform and random chance baselines.

LITERATURE REVIEW

Hybrid CNN-Transformer Models for Mammography

The original Vision Transformer work also introduced a hybrid variant in which CNN feature maps, instead of raw image patches, are passed to the Transformer encoder (Dosovitskiy *et al.*, 2021), and it has been demonstrated that convolutional early layers improve the stability and accuracy of Transformer training (Xiao *et al.*, 2021). In mammography, recent hybrid approaches involve the ConvNeXt–ViT architecture optimized via differential evolution using CBIS-DDSM and MIAS (Aldawsari *et al.*, 2026), a lesion-aware hybrid with weakly supervised evidence modeling based on VinDr-Mammo (Hoang *et al.*, 2026), a CNN-Transformer fusing past and present examinations (Jeny *et al.*, 2025), as well as hierarchical multi-scale, edge-aware Transformers with Grad-CAM and SHAP interpretation on several publicly available datasets (Badawi *et al.*, 2026). The comparison of 33 convolutional and Transformer architectures for whole-mammogram classification (Cantone *et al.*, 2023) and a weakly supervised classifier for high-resolution images (Y. Shen *et al.*, 2021) tackle the question of resolution and model choice from different perspectives, while a recent review highlights Transformer-based approaches in breast imaging (L. Wang, 2025). Chest radiograph classification and localisation tasks are solved with hybrid CNN-Transformers as well (Hsieh *et al.*, 2026), and deep learning has also been applied to detect renal abnormalities in CT images (Mitu, 2026). Considering the above, HCT-Mammo is developed following the hybrid recipe and is analysed in the current paper as an object of a controlled experiment rather than an architectural innovation.

Uncertainty Calibration and Selective Prediction

Overconfidence is typical for modern networks, and temperature scaling is a straightforward post hoc calibration method that is learned on the held-out data (Guo *et al.*, 2017). MC-Dropout is another popular approach to handling uncertainty, which treats the test-time dropout as approximate Bayesian inference (Gal & Ghahramani, 2016). Entropy and mutual information allow splitting total uncertainty into aleatoric and epistemic (Houlsby *et al.*, 2011; Kendall & Gal, 2017). Selective classification is concerned with the possibility of ranking the model's own errors according to the uncertainty score, which can be described by risk–coverage curves (Geifman & El-Yaniv, 2017), while uncertainty-driven referral has been investigated for retinal disease detection (Leibig *et al.*, 2017). Uncertainty estimation performance deteriorates under dataset shift, and none of the methods is universally reliable (Ovadia *et al.*, 2019). In breast imaging, approaches taking account of uncertainty have been proposed for the DDSM and breast ultrasound (Chegini & Mahloojifar, 2024b), for dual-view contrast-enhanced spectral mammography using a calibrated Transformer and MC-Dropout (Alkurdi & Sallow, 2026), and for molecular subtype prediction on the CMMD (Chegini & Mahloojifar, 2024a). MC-Dropout and selective predictions based on ensemble learning have been studied for brain-tumour MRI triage (Sharma, 2026) and thyroid nodule ultrasound in the setting of dataset shift (Sadib *et al.*, 2026). One of the unpublished open-source projects is an implementation of the MC-Dropout deferral strategy along with Grad-CAM grounding on the CBIS-DDSM masses using the VGG16 architecture (teo-tat, 2026).

Cross-Dataset Evaluation with CBIS-DDSM and CMMD

Gabbualoy *et al.* (2026) fine-tuned five different backbone models on CBIS-DDSM and tested them on CMMD and other datasets using paired DeLong tests, while the topological-conditioning experiment used CBIS-DDSM as the training dataset and evaluated the results on CMMD and INbreast (Fanning & Aktas, 2025). Notably, the lesion masks of CBIS-DDSM are known to be inaccurate; the authors of RBIS-DDSM documented these errors and released a re-curated version (Chakravarty *et al.*, 2022). Risk–coverage analysis of uncertainty was not performed in the context of CBIS-DDSM to CMMD transfer in any of the studies cited above, which makes the second question of the paper relevant.

Localisation Consistency of Attention and Gradient Maps

Gradient-based relevance methods for Transformers (Chefer *et al.*, 2021) extend gradient-weighted class activation mapping (Grad-CAM; Selvaraju *et al.*, 2017), whereas Attention Rollout propagates attention weights across all Transformer layers (Abnar & Zuidema, 2020). The pointing game (Zhang *et al.*, 2018) and its energy-based modification (H. Wang *et al.*, 2020) are two methods for measuring the agreement with annotated regions. Attention weights are not the only weights to generate faithful explanations (Jain & Wallace, 2019); there can be saliency methods that do not rely on model parameters (Adebayo *et al.*, 2018), and standard measurements for faithfulness might be unable to reveal spurious explanations (Wu *et al.*, 2024). Towards mammography, ExViT-Breast explored multiple variations of Grad-CAM, applied Transformer relevance propagation to the CBIS-DDSM breast cancer recognition task, included calibration, ensembling-based uncertainty quantification, and validated with INbreast (Kumar *et al.*, 2026); a review covered explanation method evaluation for mammography (Shifa *et al.*, 2025). In the present study, the explicit chance baselines are applied to a hybrid model to evaluate both Attention Rollout and Grad-CAM, in addition to comparing localisation analysis with MC-Dropout and applying it to frozen cross-dataset evaluation on CMMD.

MATERIALS AND METHODS

Overview

HCT-Mammo maps a breast-cropped mammogram to a probability of malignancy (Figure 1). For this purpose, a truncated ResNet-18 pretrained on ImageNet is used to extract local features, and the feature map is encoded in the form of a tokenised sequence. Subsequently, a small Transformer encoder is used to encode the interactions between the tokens within the breast region. The following three analyses are performed on the trained model with unchanged parameters: (i) MC-Dropout for uncertainty estimation, (ii) Attention Rollout and Grad-CAM for localisation consistency, and (iii) a zero-shot test on a separate dataset (Figure 2).

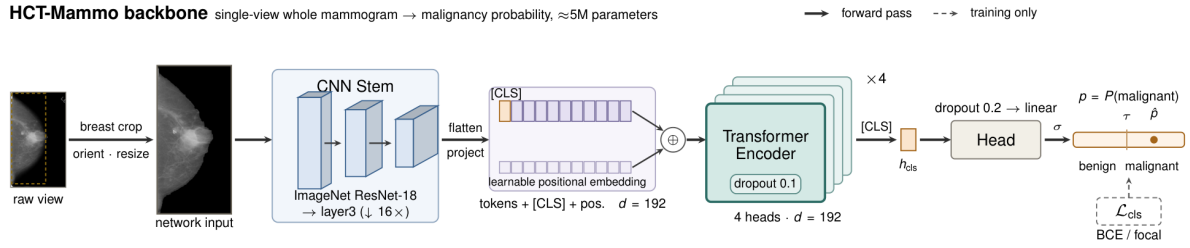


Figure 1: Architecture of HCT-Mammo: a ResNet-18 stem (up to layer3), a [CLS] token with a learnable positional embedding, and a four-layer Transformer encoder that outputs P(malignant).

CNN Stem and Tokenisation

Let $X \in R^{896\times448\times3}$ be a preprocessed input image. ϕ is composed of all ResNet-18 layers up to layer3 and returns a 16 × downscaled feature map,

$$F = \phi(X) \in R^{56\times28\times256} \quad (1)$$

The map is reshaped into N=1568 tokens and mapped linearly to d=192 model width. A learnable [CLS] token is prepended to it, as well as a learnable positional embedding over the 56 × 28 grid, leading to the initial encoding

$$Z_0=[t_{cls};\mathrm{Flatten}(F)W_p] + E_{pos}, \quad Z_0 \in R^{(N+1)\times d} \quad (2)$$

Transformer Encoder and Classification Head

The encoder comprises L=4 pre-norm layers, each with a multi-head attention (MHA), layer normalisation (LN), and a multilayer perceptron (MLP) with an expansion ratio of four. In addition, MHA uses four heads. Dropout is used with probability 0.1 within MLP blocks,

$$Z'_l = Z_{(l-1)} + \mathrm{MHA}(\mathrm{LN}(Z_{l-1})), \quad (3)$$

$$Z_l = Z'_{-l} + \mathrm{MLP}(\mathrm{LN}(Z'_l)), \quad l = 1, ..., L. \quad (4)$$

The [CLS] embedding $h = \mathrm{LN}(Z_L[0]) \in R^d$ is dropped out with probability 0.2 and mapped by a linear function,

$$z=w^T\mathrm{Drop}_{0.2}(h)+b,\ p=\sigma(z)=P(\text{malignant}|X). \quad (5)$$

Two training losses are compared. The binary cross-entropy (BCE) loss is formulated as $L_{BCE}=-[y\log p+1-y\log 1-p]$. The focal loss has the form of $L_{focal}=-\alpha_t (1 - p_t)^\gamma \log(p_t)$, where $p_t=p$ if y=1 and $p_t=1-p$ otherwise, $\gamma = 2$, and α_t is derived from the inverse class frequency. The total number of parameters of the entire neural network in the implementation is 4.91 M (Table 2).

Uncertainty Estimation and Selective Prediction

There are three sources of probabilities and uncertainties. Deterministic prediction $p=\sigma(z)$ produces sigmoid confidence, while predictive entropy is equal to $H(p)=-p\log p-1-p\log 1-p$. Temperature scaling replaces p with $\sigma(z/\varkappa)$, where k>0 and k is selected to minimise the validation negative log-likelihood (Guo *et al.*, 2017). In MC-Dropout, all dropout layers are kept active during inference and T=20 forward passes with averaging of predictions $\bar{p}$ are performed (Gal & Ghahramani, 2016); the mutual information is computed as

$$\bar{p} = \frac{1}{T}\sum_{t=1}^{T} p^{(t)}, \quad MI = H(\bar{p}) - \frac{1}{T}\sum_{t=1}^{T} H\left(p^{(t)}\right), \quad (6)$$

where MI (mutual information) is the epistemic component of the uncertainty, and the predictive entropy $H(\bar{p})$ is used as the MC-Dropout uncertainty score. Because positive temperature scaling preserves the ordering of the absolute logits, it also preserves the ordering of the binary predictive entropy; temperature-scaled entropy therefore ranks images exactly as sigmoid

entropy does and serves as a reference.

To evaluate uncertainty scores on the same subset of errors, class prediction $\hat{y}=1[p\geq\tau]$ is obtained based on deterministic prediction with threshold τ selected on the validation set. For each score u in the set, the ordering of the images is done based on the increasing order of u. The selective risk at the given level of coverage c is defined as the error rate for c fraction of images having the lowest score and is represented by R(c). Area under the risk vs coverage graph (AURC) is obtained as follows for n test images,

$$AURC = \frac{1}{n}\sum_{k=1}^{n} R(k/n). \qquad (7)$$

The errors at levels of coverage of 80% and 90% are presented as the errors made for deferring 20% and 10% most uncertain images, respectively. Calibration error is expressed by the expected calibration error (ECE), which is obtained as

$$ECE = \sum_{b=1}^{B} \frac{n_b}{n} |\bar{y}_b - \bar{p}_b|$$

using B=15 bins. It is also evaluated using the Brier score (Brier, 1950) and the negative log-likelihood (NLL).

Localization Consistency

Two post hoc attribution maps are obtained from the trained network for the malignancy logit. Attention Rollout (Abnar & Zuidema, 2020) consists of averaging the heads of each layer, uses the residual connection by adding the identity function, and multiplies the matrices across layers as follows,

$$\hat{A}_l = \frac{1}{2}\bar{A}_l + \frac{1}{2}I, \quad R = \hat{A}_4\hat{A}_3\hat{A}_2\hat{A}_1, \quad m_{roll} = R[0, 1:], \qquad (8)$$

where $\bar{A}_l$ is the head-averaged attention of layer l, and m_{roll} is reshaped to the 56×28 token grid. Grad-CAM (Selvaraju *et al.*, 2017) is computed on the stem output F; gradients of the malignancy logit are backpropagated through the Transformer to the stem output,

$$\alpha_k = \frac{1}{56\cdot 28}\sum_{i,j}\frac{\partial z}{\partial F_{ijk}}, \quad m_{cam} = ReLU\left(\sum_k \alpha_k F_{:,:,k}\right). \qquad (9)$$

Both maps are bilinearly upsampled to 896×448. For an image i with lesion mask M_i, a pointing-game hit is recorded when the maximum of the map falls inside M_i dilated by 15 pixels, since the contours are not delineated precisely. Energy based pointing game is gin by

$$E_i = \sum_{x\in M_i} m_i^+(x) / \sum_x m_i^+(x)$$

(H. Wang *et al.*, 2020) and the Intersection Over Union *IoU@k* indicates overlap between Mi and top k% of the map in the breast. As chance references, the same lesion masks are also scored against two maps that carry no image information: a map that is uniform over the breast region and a random map over the whole image. Because neither Grad-CAM nor Attention Rollout is a segmentation method, IoU is treated as a secondary measure.

Frozen Cross-Dataset Protocol

All model selection was done on the CBIS-DDSM validation set alone. Network weights θ, decision threshold τ (Youden's J; Youden, 1950) and temperature k were fixed before any CMMD image was evaluated, and no parameter tuning or re-calibration was done for CMMD.

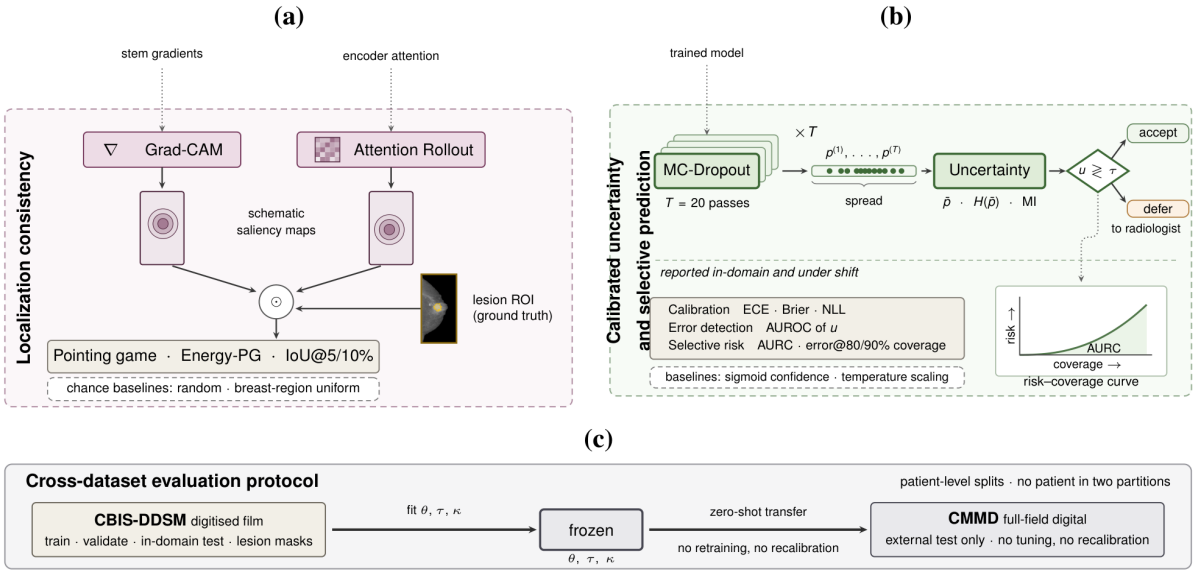


Figure 2: Analyses of the trained network: (a) Grad-CAM and Attention Rollout localisation, (b) MC-Dropout uncertainty and selective prediction, (c) zero-shot CMMD testing with θ, τ, and κ fixed on CBIS-DDSM.

Datasets and Partitions

The CBIS-DDSM is a carefully selected subset of the Digital Database for Screening Mammography, which comprises digital versions of the masses/calcifications in a scanned film mammography image with pathology annotations and binary region-of-interest (ROI) masks for each lesion (Lee *et al.*, 2017; Sawyer-Lee *et al.*, 2016). Calcification ROI is coarse, and for the mass masks,

a level-set-based approach, seeded by the original outlines, was used for cleaning. The CMMD dataset (Cai *et al.*, 2023; Cui *et al.*, 2021) consists of full-field digital mammograms of patients from China with biopsy-diagnosed labels, indicating either a benign or malignant lesion, for each breast side, but without lesion masks. Both of these datasets were made available through The Cancer Imaging Archive (Clark *et al.*, 2013). The CBIS-DDSM dataset is read in JPEG format, while the CMMD dataset is retrieved directly from the original DICOM images. CBIS-DDSM is used for training, validation, in-domain testing and localisation, while the CMMD is used only for out-of-domain testing.

Case description files for all four CBIS-DDSM were merged and deduplicated on each full mammogram file. From the pathological labelling, any image with at least one judgment of malignancy was labelled as malignant, and all others as benign (the latter includes all BENIGN and BENIGN_WITHOUT_CALLBACK images). The official test partition was not modified. To ensure the lack of patients in more than one partition (Table 1), patients appearing in the official test set were removed from the training set, and 15 per cent of patients not appearing in the official test set were removed and constituted the validation set. For the CMMD dataset, if a breast side in the clinical data was identified, then the image was retained; if two patient identifiers were present, the second one was used; and if two images were pixel-identical, the one to be retained was the second image. Of the 5,202 CMMD images, 1,474 images of breasts without clinical labels (the contralateral breast of patients whose clinical information covered only one side) were excluded, leaving 3,728 images; after removing three pixel-identical duplicates, 3,725 images from 1,774 patients remained (Table 1).

Table 1: Data partitions after patient-level separation; malignant percentages are image-level. 64 training-pool images (31 patients) overlapping the official test set were removed.

Dataset	Partition	Images	Patients	Malignant (%)	Role
CBIS-DDSM	Training	2,033	1,034	45.6	Model fitting
CBIS-DDSM	Validation	361	183	42.1	Model, loss and threshold selection; temperature
CBIS-DDSM	Test (official)	645	349	40.9	In-domain test; localisation
CMMD	External	3,725	1,774	62.7	Zero-shot external test only

Preprocessing and Augmentation

The same preprocessing pipeline was used for both datasets. The intensity values of all images were rescaled to [0, 1]. The breast region was found by Otsu thresholding (Otsu, 1979), keeping the largest connected component, and the image was cropped to the breast bounding box, which removes film labels and scanner artefacts. Each image was then flipped to have the chest wall on the left side, resized to 896×448, replicated into three channels, and normalised using ImageNet statistics (Deng *et al.*, 2009). These same transformations were applied consistently to the ROI masks from CBIS-DDSM.

Augmentations for the training set included rotations up to ±10°, scales and translations up to ±10% and minor manipulations with brightness, contrast, and gamma correction. Horizontal and vertical flips were not used since either one of them inverted the orientation of the chest wall or resulted in anatomically unreasonable pictures. Contrast-limited adaptive histogram equalisation (CLAHE) was only considered in the ablation study.

Baselines and Training Protocol

To compare with the proposed architecture, three baseline cases were also trained each with the same resolution, train/validation split, augmentations, loss, optimizer and stopping criterion: EfficientNet-B0 (Tan & Le, 2019) with a similar number of parameters to that of the proposed model, ResNet-50 (He *et al.*, 2016) to serve as a standard CNN, and ViT-S/16 (Dosovitskiy *et al.*, 2021) as a pure Transformer, using 16-pixel patches (a 56×28 patch grid at 896×448) with interpolated position embeddings. The weights for all the backbones were set to those pre-trained using the ImageNet dataset, and a single-logit head was employed in all cases. Optimization through: AdamW (Loshchilov & Hutter, 2019); weight decay 0.05; initial learning rates of 10^{-4} for pretrained layers and 3×10^{-4} for newly initialised layers. A learning schedule of cosine decay was used with two warm-up epochs, followed by up to 40 epochs. The early stopping on validation AUROC was set to 8 epochs. Mixed precision and a batch size of 16 were used for training. HCT-Mammo, each ablation variant and each of the three baselines were trained with three random seeds. The loss function (binary cross-entropy and focal loss) and the CLAHE option for HCT-Mammo were selected based on the validation AUROC, while the median-seed model was selected based on the validation AUROC for analysis of uncertainty, statistical testing and external data. Training was done on a single NVIDIA A100 40 GB PCIe GPU with all of the models in PyTorch (Paszke *et al.*, 2019). Parameter counts and multiply–accumulate operations (MACs) were measured at 896×448 resolution and are reported in millions (M) and GMACs, respectively; for Transformer models, the MACs include the attention matrix multiplication (QK^T) and the attention-weighted sum over V. The inference performance was evaluated with the same GPU using mixed precision (batch size 32) with 20 warm-up runs followed by an average of 100 runs.

Evaluation Metrics and Statistics

Discrimination was assessed using the primary measure AUROC and the area under the precision–recall curve

(AUPRC), while sensitivity, specificity, and F1 score were calculated at the Youden threshold (Youden, 1950), defined on the validation set. Calibration and selective-prediction measures are defined in the Uncertainty Estimation and Selective Prediction subsection above. Ninety-five per cent confidence intervals were calculated using patient-level bootstrap resampling with 1,000 iterations, which guarantees co-resampling of images from the same patient. The same patient-level bootstrap was utilised for the confidence intervals for the AUROC and AUPRC on CBIS-DDSM and CMMD as well as for the hit rate confidence intervals and for the differences in calibration and selective-prediction scores. The point estimates are presented both as the average value for the three seeds and as the median-seed model (Table 2); confidence intervals and tests are based on the median-seed model. For a subset of comparisons between median-seed models in terms of AUROC, DeLong's test (DeLong *et al.*, 1988) was used without multiplicity correction. The attention maps provided by the Attention Rollout and Grad-CAM were compared using a paired exact McNemar test in the pointing game and a Wilcoxon signed-rank test in energy score and IoU evaluation.

The test images provided as part of the mass and calcification masses in the localisation are 361 test images of masses and 284 test images of calcifications. The main analysis is mass localisation, as the mass masks are more reliable than the calcification ROIs. Note that F1 and AUPRC are sensitive to the prevalence of malignancy and require careful interpretation independently of each other for two different datasets. In many cases, sensitivity and specificity are conditional to class and cannot be equated with prevalence, but differences in the kind of cases considered (case mixture) and the fixed cutoff for the decision rule make cross-dataset comparisons difficult. The evaluation includes 5 experiments: comparisons with the 3 baselines, an ablation experiment, an in-domain uncertainty experiment, frozen CMMD evaluation for both HCT-Mammo and EfficientNet-B0, and localisation consistency.

RESULTS AND DISCUSSION

Main Comparison on CBIS-DDSM

As can be seen from Table 2, HCT-Mammo achieved numerically higher mean AUROC (0.780 ± 0.005) and mean AUPRC (0.729) than EfficientNet-B0 (AUROC 0.770; AUPRC 0.704) and ResNet-50 (AUROC 0.764; AUPRC 0.711), whereas ViT-S/16 had the lowest mean AUROC (0.717). The means of the three top-performing models are not very divergent from each other when compared to the sampling uncertainty of a test set with 645 images. Similarly, the 95% confidence interval of HCT-Mammo (0.742–0.821) overlapped those of EfficientNet-B0 (0.731–0.810) and ResNet-50 (0.720–0.801), and the DeLong tests between the median-seed models were not statistically significant ($p = 0.59$ and $p = 0.33$, respectively). The comparison with ViT-S/16 favoured HCT-Mammo (ΔAUROC = 0.062; $p = 0.002$). A difference in the operating points of the models can be explained by a difference in the threshold that was set on the relatively small validation data set: The sensitivity (0.52 to 0.70) and specificity (0.69 to 0.78) were different for the models (mean over 3 seeds), which yielded F1 scores of 0.56 to 0.65. The ROC curves of the results obtained are shown in Figure 3a.

HCT-Mammo comprises 4.91 million (M) parameters, about one-fifth that of ResNet-50 (23.51 M) and ViT-S/16 (22.19 M), but larger than EfficientNet-B0 (4.01 M). At 896×448 resolution, HCT-Mammo uses 17.9 GMACs, which is just c. 55% of the 32.7 GMACs used by ResNet-50 and 5.8 times as much as EfficientNet-B0's 3.1 GMACs. The speed for inference on a single NVIDIA A100 40 GB PCIe GPU in mixed-precision mode (520 images/s) is 1.5 and 2.7 times faster than ResNet-50 (340 images/s) and ViT-S/16 (190 images/s), respectively, but slower than EfficientNet-B0 (640 images/s). So, the term compact holds only relative to ResNet-50 and ViT-S/16, not EfficientNet-B0.

Table 2: Official CBIS-DDSM test set at 896 × 448. Upper block: mean ± SD over three seeds. Lower block: median-seed model, used for CIs, tests and later analyses. Thresholds are validation-selected; GMACs include attention.

Model	Params (M)	GMACs	AUROC	95% CI	AUPRC	Sens.	Spec.	F1
Mean over three seeds								
HCT-Mammo	4.91	17.9	0.780 ± 0.005	–	0.729	0.63	0.78	0.64
EfficientNet-B0	4.01	3.1	0.770 ± 0.007	–	0.704	0.67	0.71	0.64
ResNet-50	23.51	32.7	0.764 ± 0.008	–	0.711	0.70	0.69	0.65
ViT-S/16	22.19	56.5	0.717 ± 0.014	–	0.649	0.52	0.76	0.56
Median-seed model								
HCT-Mammo	–	–	0.781	0.742–0.821	0.714	0.59	0.82	0.64
EfficientNet-B0	–	–	0.771	0.731–0.810	0.690	0.58	0.81	0.62
ResNet-50	–	–	0.762	0.720–0.801	0.711	0.74	0.64	0.65
ViT-S/16	–	–	0.719	0.678–0.762	0.658	0.51	0.78	0.56

Ablation

Minimal ablation (Table 3) showed that replacing the Transformer stage with global average pooling within the same stem leads to lower mean AUROC (from 0.780 to 0.757; −0.023), while all three HCT-Mammo runs exceed three stem-only runs; nevertheless, the DeLong test of median-seed models does not reach statistical significance ($p=0.22$). This transformer adds 2.13 million

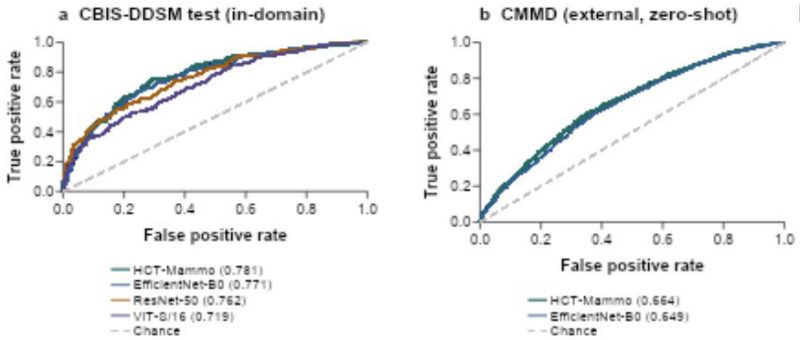


Figure 3: ROC curves of the median-seed models on (a) the CBIS-DDSM test set and (b) CMMD without adaptation; legend values are AUROC.

more parameters and 6.7 GMACs to the stem. Validation AUROC shows preference for BCE loss over focal loss (0.792 versus 0.790) and absence of CLAHE over CLAHE (0.792 versus 0.787), while the differences on the test set were insignificant (−0.002 and −0.005). The BCE condition without CLAHE was kept. Downscaling the input size to 448×224 decreased the mean AUROC by 0.039 (p=0.046), which is consistent with the motivation for a high input resolution but does not prove it.

Table 3: Ablations on CBIS-DDSM (mean ± SD over three seeds). Δ is the AUROC difference from the default model, computed from the mean AUROCs; p is a DeLong test on median-seed models.

Variant	Params (M)	Val. AUROC	Test AUROC	AUPRC	Δ AUROC	p
HCT-Mammo (default)	4.91	0.792	0.780 ± 0.005	0.729	–	–
ResNet-18 stem + GAP, no Transformer	2.78	0.769	0.757 ± 0.009	0.709	−0.023	0.22
Focal loss instead of BCE	4.91	0.790	0.778 ± 0.007	0.723	−0.002	0.87
CLAHE preprocessing on	4.91	0.787	0.775 ± 0.007	0.710	−0.005	0.74
Input 448 × 224	4.69	0.753	0.741 ± 0.010	0.675	−0.039	0.046

Uncertainty and Calibration on CBIS-DDSM

The median-seed model misclassified 0.276 of the test images on the threshold chosen during validation (p ≥ 0.54). Table 4 summarises in-domain and external discrimination. Table 5 shows the probability quality and selective predictions for the three methods, all having the same decision criterion. The sigmoid probability distribution showed signs of overconfidence (ECE 0.081), with the fitted temperature of k=1.93. Scaling the temperature lowered ECE to 0.059 and the NLL from 0.567 to 0.560. MC-Dropout resulted in an intermediate improvement in the metrics (ECE 0.070; NLL 0.563). The bootstrap CI of the difference in ECE between MC-Dropout and temperature scaling contained zero (95% CI −0.034 to 0.052), as well as the CI of the difference between MC-Dropout and sigmoid (−0.033 to 0.020), thus failing to prove any in-domain superiority of one calibration method over the other.

All three scores ranked errors better than chance. The AUROC of the error detection was 0.670 for sigmoid entropy and, by definition, for temperature scaling, 0.673 for MC-Dropout (difference −0.006 to 0.011). The AURC values obtained were 0.173 for sigmoid entropy and temperature scaling, and 0.171 for MC-Dropout (difference −0.007 to 0.002). Deferring the classification of the 20% most uncertain images led to a decrease in error from 0.276 to 0.236 for sigmoid entropy and to 0.236 for MC-Dropout, and an increase in error from 0.109 in the lowest to 0.434 in the highest uncertainty quintile for MC-Dropout (Figure 4f).

External Evaluation on CMMD

Table 4 and Figure 3b describe the evaluation of the frozen models on CMMD (3,725 images acquired from 1,774 patients, where 62.7% were malignant). The AUROC decreased from 0.781 to 0.664 (95% CI 0.644–0.684) for

Table 4: In-domain and external evaluation (median-seed models; CBIS-DDSM rows repeat the lower block of Table 2). The threshold is fixed on the CBIS-DDSM validation set. Pred. malig. is the fraction classified as malignant.

Model	Test set	AUROC	95% CI	AUPRC	Sens.	Spec.	F1	Pred. malig. (%)
HCT-Mammo	CBIS-DDSM test	0.781	0.742–0.821	0.714	0.59	0.82	0.64	35.0
HCT-Mammo	CMMD	0.664	0.644–0.684	0.756	0.56	0.68	0.64	47.3
EfficientNet-B0	CBIS-DDSM test	0.771	0.731–0.810	0.690	0.58	0.81	0.62	34.6
EfficientNet-B0	CMMD	0.649	0.628–0.668	0.744	0.57	0.65	0.64	49.2

Table 5: Probability quality and selective prediction of the median-seed HCT-Mammo. All methods share one-class decision. ECE uses 15 bins; Err. AUROC detects misclassified images; E@80 and E@90 are errors at 80% and 90% coverage.

Test set	Method	ECE	Brier	NLL	Err. AUROC	AURC	E@80	E@90
CBIS-DDSM test	Sigmoid	0.081	0.189	0.567	0.670	0.173	0.236	0.250
CBIS-DDSM test	Temperature scaling	0.059	0.189	0.560	0.670	0.173	0.236	0.250
CBIS-DDSM test	MC-Dropout (T = 20)	0.070	0.188	0.563	0.673	0.171	0.236	0.253
Error at full coverage: 0.276								
CMMD	Sigmoid	0.166	0.253	0.734	0.608	0.308	0.362	0.370
CMMD	Temperature scaling	0.121	0.233	0.658	0.608	0.308	0.362	0.370
CMMD	MC-Dropout (T = 20)	0.156	0.247	0.708	0.614	0.300	0.361	0.375
Error at full coverage: 0.394								

HCT-Mammo and from 0.771 to 0.649 (0.628–0.668) for EfficientNet-B0, which constitutes drops of 0.117 and 0.122, respectively. The AUROC difference (0.015) between the two models on CMMD falls within their overlapping intervals. With a fixed decision threshold, the operating point became less discriminating: HCT-Mammo reached sensitivity of 0.56 and specificity of 0.68 on CMMD compared to its in-domain performance of 0.59 and 0.82, respectively (median-seed model, Table 2), with 47.3% of cases considered malignant. The area under the precision– recall curve (AUPRC) and F1 score depend on the prevalence of the positive class (62.7%

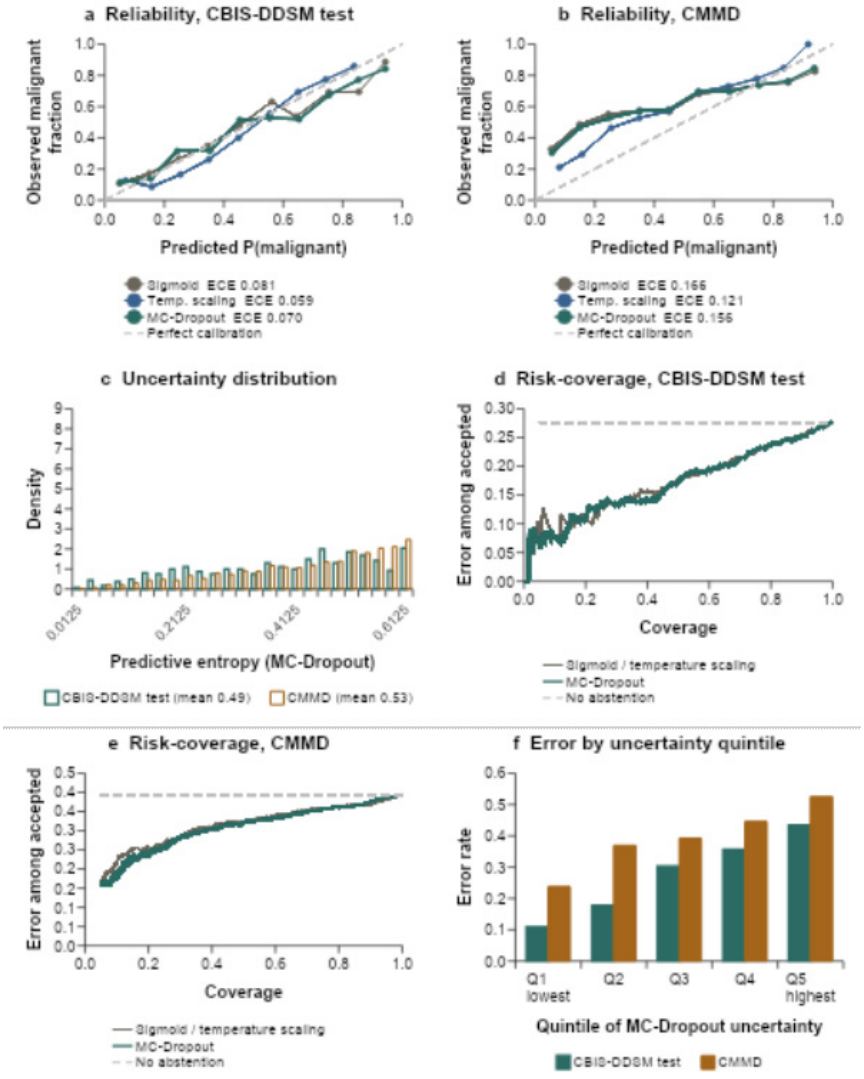


Figure 4: Uncertainty of the median-seed HCT-Mammo: (a,b) reliability diagrams (10 bins shown; ECE uses 15), (c) MC-Dropout entropy, (d,e) risk–coverage curves, (f) error by uncertainty quintile.

in CMMD compared to 40.9% in the CBIS-DDSM test set); thus, it is not appropriate to compare them across different datasets.

The classification error went from 0.276 to 0.394, representing a 43% increase, with the ECE of sigmoid probabilities increasing from 0.081 to 0.166, the ECE of temperature scaling increasing from 0.059 to 0.121, and the ECE of MC-Dropout increasing from 0.070 to 0.156, respectively (Table 5, lower block). On CMMD, temperature scaling remained the best-calibrated method: MC-Dropout had a lower ECE than sigmoid (95% CI of the difference −0.016 to −0.005) but a higher ECE than temperature scaling (0.019 to 0.048), and even temperature scaling did not restore in-domain calibration (ECE 0.121 vs. 0.059). The magnitude of this increase in entropy was relatively small under the dataset shift, and the average MC-Dropout entropy increased by 8% (0.49 vs. 0.53), which is much less than the increase in error, while the average mutual information increased by 83% (0.024 vs. 0.044). Although it has increased, error-detection performance has been moderate. The error rankings were marginally higher than chance (error detection AUROC 0.608 for sigmoid entropy and 0.614 for MC-Dropout), while the AURC was 0.308 for sigmoid entropy and 0.300 for MC-Dropout. The absolute error reduction from using deterministic entropy for deferring the most uncertain 20% of cases was 0.032, which is close to the error reduction of 0.040 achieved from using in-domain entropy. The error increased by a factor of about 2.2 from the lowest quintile (0.238) to the highest (0.525).

Localization Consistency

Table 6 compares the two maps with respect to the lesion ROIs of the CBIS-DDSM test images. With respect to masses (n = 361), the Grad-CAM pointing-game accuracy was 35.5% (95% CI 30.5–40.8), compared to 24.4% (19.8–28.9) for Attention Rollout, 4.7% for the breast-uniform reference map, and 2.9% for random maps. Accordingly, both methods performed better than chance by ~7.6 and 5.2 times, respectively. When compared with Rollout, Grad-CAM achieved more pointing-game hits (78 versus 38 discordant images, exact McNemar $p < 0.001$) as well as a better energy score (0.069 versus 0.046, Wilcoxon $p < 0.001$) and IoU at 10% (0.097 versus 0.061, $p < 0.001$). The absolute numbers remain low, though: Only one-third of the masses have been intersected by the Grad-CAM max, and only 0.069 of the energy of the positive map is inside the mask.

Regarding calcifications (n = 284), the hit rates were 37.0% (31.5–43.1) for Grad-CAM and 33.5% (27.8–38.9) for Rollout, against a baseline of 9.9% (uniform). The two methods agreed more closely with each other than for masses (McNemar p=0.31; energy score p=0.038; IoU at 10%, p=0.79) and had smaller enrichments (3.7 and 3.4 times), in line with the greater size of calcification ROIs. Table 7 shows a subdivision of mass results by prediction result and pathology. The Grad-CAM hit rate did not differ between correctly and incorrectly classified images (35.9% vs. 34.3%) and between malignant and benign lesions (35.2% vs. 35.6%). The Rollout hit rate was lower for malignant than for benign lesions (18.6% vs. 28.2%) and for incorrect than correct classifications (19.6% vs. 26.3%); with sample sizes from 102 to 259 images, this finding is exploratory. Examples of successful and unsuccessful predictions are shown in Figure 5.

Table 6: Localisation of HCT-Mammo maps against CBIS-DDSM lesion ROIs. PG: pointing-game hit rate (15 px, 95% CI); Energy: energy-based pointing game; IoU: top 5% or 10% of the map. Chance: uniform and random maps.

Lesion (n)	Map	PG (%) (95% CI)	Energy	IoU@5%	IoU@10%
Masses (361)	Grad-CAM	35.5 (30.5–40.8)	0.069	0.076	0.097
Masses (361)	Attention Rollout	24.4 (19.8–28.9)	0.046	0.045	0.061
Masses (361)	Breast-uniform (chance)	4.7	0.027	0.016	0.021
Masses (361)	Random (chance)	2.9	0.017	0.010	0.012
Calcifications (284)	Grad-CAM	37.0 (31.5–43.1)	0.131	0.085	0.122
Calcifications (284)	Attention Rollout	33.5 (27.8–38.9)	0.124	0.080	0.120
Calcifications (284)	Breast-uniform (chance)	9.9	0.070	0.028	0.040
Calcifications (284)	Random (chance)	6.2	0.044	0.019	0.025

Table 7: Pointing-game hit rate (%) and energy score for mass images by prediction outcome and pathology.

Group	n	Grad-CAM PG (%)	Grad-CAM energy	Rollout PG (%)	Rollout energy
Correctly classified	259	35.9	0.068	26.3	0.050
Misclassified	102	34.3	0.073	19.6	0.036
Malignant	145	35.2	0.070	18.6	0.035
Benign	216	35.6	0.069	28.2	0.053

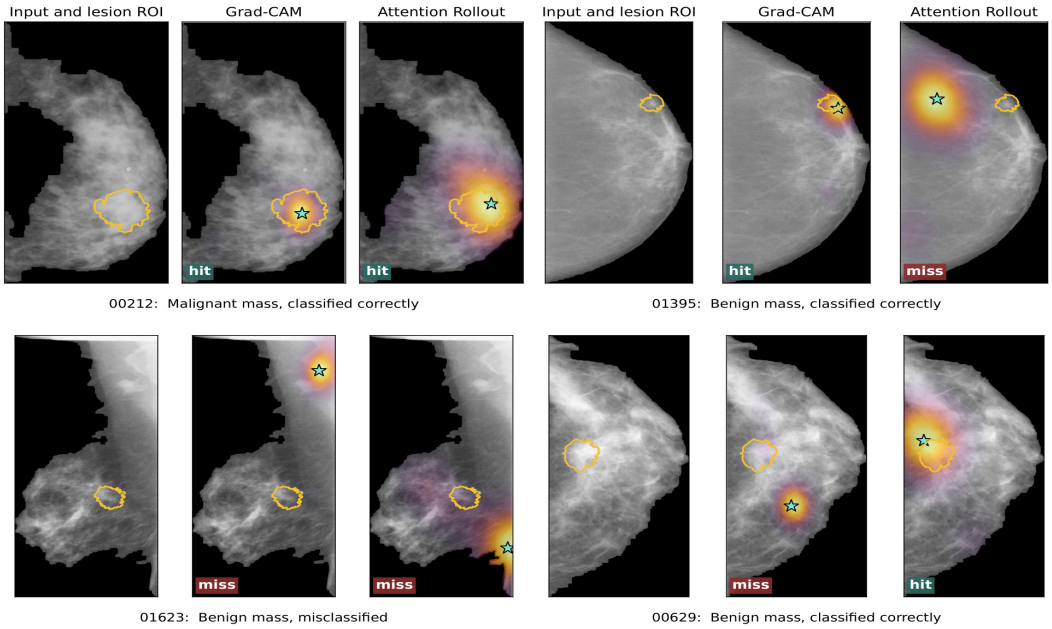


Figure 5: Test masses with lesion ROI (gold contour), Grad-CAM and Rollout maps, and the pointing-game maximum (star) as hit or miss; the last case is a Grad-CAM miss and a Rollout hit.

Discussion

The proposed work assesses the proposed compact hybrid CNN-Transformer model in the patient-level protocol along with further research into uncertainty and localisation, and is evaluated on a frozen test on an external dataset. Finally, it is observed that the in-domain discrimination performance is similar to the traditional baselines, and the loss of discrimination in HCT-Mammo and EfficientNet-B0 with the external dataset is almost the same. Under distributional shift, the calibration of the HCT-Mammo model deteriorated, and the mean predictive entropy rose, but only slightly.

Discrimination and Efficiency

On CBIS-DDSM, HCT-Mammo achieved the highest mean AUROC, but this improvement did not exist outside of the sampling variability range of test sets for the other CNNs, and there was no statistically significant evidence that HCT-Mammo clearly surpassed the other CNNs. The mean AUROC of the ViT-S/16 baseline was lower than the mean AUROC of HCT-Mammo under the same resolution and training conditions (0.717 vs. 0.780). This is in line with the idea that a convolutional stem is significantly helpful when faced with a small training set (Xiao *et al.*, 2021). The Transformer module was then excluded; this reduced the AUROC slightly but not significantly by 0.023; therefore, it cannot be concluded statistically whether or not long-range interaction additionally contributes on top of the stem in this sample. The AUROC values obtained are difficult to compare with the reported AUROC values in the papers concerning CBIS-DDSM because there are differences with regard to the unit of analysis, splitting the data set into subsets, as well as the input resolution and the usage of one-view input images. For this reason, the present findings need to be considered in light of the leakage-controlled protocol followed. In terms of computational efficiency, HCT-Mammo has fewer parameters and MACs than ResNet-50 and ViT-S/16; however, EfficientNet-B0 is smaller and lighter, so the compactness claim is valid only when compared with ResNet-50 and ViT-S/16.

Uncertainty and Calibration

The probabilities were sigmoid, which led to over-confident estimates, and the single temperature tuned on the validation set resulted in the best in-domain estimates of ECE and NLL. Regarding the error ranking, there was no clear advantage to using MC-Dropout as compared to the entropy (without dropout) approach in the CBIS-DDSM. The reason may be that the dropout probability for each MLP block and before the head layer is 0.1, while that for the remaining layer before the head layer is 0.2, which means the small part of the network which is being affected by the dropout has small errors, which may be due to the appearance of the lesions that are difficult to classify; and without the stochastic passes, they cannot give much insight other than how far the logit is from the threshold. In CMMD, the calibration of both MC-Dropout and sigmoid was poor, with MC-Dropout being a little better; its improvement in error ranking was only slight, and temperature scaling remained the best-calibrated method. Most critically, the accuracy loss increased by 43%, and the average entropy evaluated from MC-Dropout sampling only increased by 8%: the loss in accuracy was not captured by the increase in error; thus, fitting of temperature in the source domain alone was insufficient to bring confidence up to the level of the accuracy loss. The size of the error reduction due to deferral was of similar magnitude in both datasets, but both forms of analysis were retrospective in terms of abnormal samples and do not prove the viability of such a strategy; it is merely an analysis of what was observed.

Localization Consistency

Both maps tended to overestimate the area which would normally be expected to be dilated by chance. Grad-CAM and Attention Rollout maps were more consistent for masses, and the differences were less for the calcifications. One reason is that Attention Rollout is calculating the attention flows for all four layers in the Attention Mechanism, and so it will produce dispersed maps, while Grad-CAM is using the gradient of the malignancy logit to assign weights to the feature channels of the stem. While the scores are indeed small, the small ROI of calcifications and non-ideal masks for masses (Chakravarty *et al.*, 2022) also reduces the utility of the scores. The results demonstrate that localisation agrees with annotated ROIs, but do not capture faithfulness to the model's decision-making process since attention weights do not correspond to explanations (Jain & Wallace, 2019), the saliency can be insensitive to model parameters (Adebayo *et al.*, 2018), and faithfulness metrics are not able to separate meaningful explanations from random noise (Wu *et al.*, 2024).

Limitations

Several restrictions are evident. Only abnormal examinations are in both datasets, so the task is lesion characterisation, not screening. The model applies to a single view and doesn't use mediolateral-oblique (MLO) or craniocaudal (CC) views to fuse them. Localisation performance is impacted by the quality of masks in the CBIS-DDSM database and the coarse calcification ROIs in the database, as well as labels (breast sides) in the CMMD dataset. The CBIS-DDSM images employed are 8-bit JPEG images, not 16-bit. The comparison of the performance between CMMD and CBIS-DDSM datasets is complicated given the use of different equipment during the scans, the different population and the prevalence of malignant cases, and annotation practices – and so the drop in performance can't be attributed to one of these factors alone. Just three seeds were applied, which provides a general indication of the training variability; there were no corrections for Multiple Testing for DeLong tests and only one external dataset. Lastly, the validation set single decision threshold selected was not stable and resulted in an imbalance between sensitivity and specificity for some models. Two views as input is part of further research that needs to be done, as are other public external datasets, recalibration on a small labelled sample of the target domain or separately validated, unsupervised adaptation, and prospective evaluation with radiologists.

CONCLUSION

The hybrid CNN-Transformer (HCT-Mammo) discriminated in-domain classes on the CBIS-DDSM dataset comparably with EfficientNet-B0 and ResNet-50 and outperformed only the pure Transformer architecture (ViT-S/16), while having roughly 1/5th the number of parameters of ResNet-50 and ViT-S/16. When applied to the CMMD dataset, the frozen model showed a loss in performance of approximately 0.117 AUROC and degraded calibration. Temperature scaling and MC-Dropout could not bring the calibration on CMMD back to the in-domain level. There was no difference in error ordering between MC-Dropout and sigmoid confidence in-domain, and only a small improvement out-of-domain. Grad-CAM was found to be more consistent with the lesion masks for masses than Attention Rollout. Results are limited to the categorisation of lesions presented in an abnormal population and are not necessarily applicable to screening performance or clinical effectiveness.

Data Availability Statement

Both datasets are available through The Cancer Imaging Archive. CBIS-DDSM: https://www.cancerimagingarchive.net/collection/cbis-ddsm/ (https://doi.org/10.7937/K9/TCIA.2016.7O02S9CY), with an alternative JPEG dataset available at https://www.kaggle.com/datasets/awsaf49/cbis-ddsm-breast-cancer-image-dataset; CMMD: https://www.cancerimagingarchive.net/collection/cmmd/ (https://doi.org/10.7937/tcia.eqde-4b16). Code used for training and evaluation is available from the corresponding author upon reasonable request.

REFERENCES

Abnar, S., & Zuidema, W. (2020). Quantifying attention flow in transformers. In P*roceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4190–4197). https://doi.org/10.18653/v1/2020.acl-main.385

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems* (Vol. 31). https://arxiv.org/abs/1810.03292

Aldawsari, M. A., Aldosari, S. J., Ismail, A., & Emam, M. M. (2026). A hybrid ConvNeXt-ViT framework with differential evolution optimization for breast cancer classification. *Scientific Reports, 16,* Article 23540. https://doi.org/10.1038/s41598-026-61594-4

Alkurdi, A. A. H., & Sallow, A. B. (2026). Calibrated transformer fusion for dual-view low-energy CESM classification. *Journal of Imaging, 12*(1), Article 41. https://doi.org/10.3390/jimaging12010041

Badawi, M. A., Arshad, E., Ali, A., Saidani, O., Saidani, T., Yang, Z., & Nam, Y. (2026). Explainable breast cancer detection using hierarchical multi-scale and edge-aware transformer networks. B*ioengineering, 13(*6), Article 657. https://doi.org/10.3390/bioengineering13060657

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians, 74*(3), 229–263. https://doi.org/10.3322/caac.21834

Brier, G. W. (1950). Verification of forecasts expressed in

terms of probability. *Monthly Weather Review, 78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2

Cai, H., Wang, J., Dan, T., Li, J., Fan, Z., Yi, W., Cui, C., Jiang, X., & Li, L. (2023). An online mammography database with biopsy-confirmed types. *Scientific Data, 10*, Article 123. https://doi.org/10.1038/s41597-023-02025-1

Cantone, M., Marrocco, C., Tortorella, F., & Bria, A. (2023). Convolutional networks and transformers for mammography classification: An experimental study. *Sensors, 23*(3), Article 1229. https://doi.org/10.3390/s23031229

Chakravarty, A., Sarkar, T., Sathish, R., Sethuraman, R., & Sheet, D. (2022). *Re-curated breast imaging subset DDSM dataset* (RBIS-DDSM) [Data set]. IEEE DataPort. https://doi.org/10.21227/nqp1-sp19

Chefer, H., Gur, S., & Wolf, L. (2021). Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 782–791). https://doi.org/10.1109/CVPR46437.2021.00084

Chegini, M., & Mahloojifar, A. (2024a). *Reliable breast cancer molecular subtype prediction based on uncertainty-aware Bayesian deep learning by mammography* [Preprint]. arXiv. https://arxiv.org/abs/2412.11953

Chegini, M., & Mahloojifar, A. (2024b). Uncertainty-aware deep learning-based CAD system for breast cancer classification using ultrasound and mammography images. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 12*(1), Article 2297983. https://doi.org/10.1080/21681163.2023.2297983

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L., & Prior, F. (2013). The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository. J*ournal of Digital Imaging, 26*(6), 1045–1057. https://doi.org/10.1007/s10278-013-9622-7

Cui, C., Li, L., Cai, H., Fan, Z., Zhang, L., Dan, T., Li, J., & Wang, J. (2021). *The Chinese mammography database (CMMD): An online mammography database with biopsy-confirmed types for machine diagnosis of breast* [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/tcia.eqde-4b16

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics, 44*(3), 837–845. https://doi.org/10.2307/2531595

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Li, F.-F. (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). https://doi.org/10.1109/CVPR.2009.5206848

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *In International Conference on Learning Representations*. https://arxiv.org/abs/2010.11929

Fanning, C., & Aktas, M. E. (2025). *Topological conditioning for mammography models via a stable wavelet-persistence vectorization* [Preprint]. arXiv. https://arxiv.org/abs/2512.10151

Gabbualoy, S., Phasukkit, P., & Tungjitkusolmun, S. (2026). Cross-sensor and cross-population generalization of deep learning models for digital mammography: A controlled four-country benchmark of five backbone architectures with statistical significance testing. *Sensors, 26*(12), Article 3911. https://doi.org/10.3390/s26123911

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In M. F. Balcan & K. Q. Weinberger (Eds.), *Proceedings of the 33rd International Conference on Machine Learning* (Vol. 48, pp. 1050–1059). PMLR. https://proceedings.mlr.press/v48/gal16.html

Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems* (Vol. 30). https://arxiv.org/abs/1705.08500

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 1321–1330). PMLR. https://proceedings.mlr.press/v70/guo17a.html

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90

Hoang, D. Q., Cao, V. K., Nguyen, T. N., & Nguyen, N. S. (2026). LENS: A mammography-specific hybrid CNN-Transformer with lesion-aware evidence modelling. *PLOS ONE, 21*(9), Article e0350720. https://doi.org/10.1371/journal.pone.0350720

Houlsby, N., Huszár, F., Ghahramani, Z., & Lengyel, M. (2011). *Bayesian active learning for classification and preference learning* [Preprint]. arXiv. https://arxiv.org/abs/1112.5745

Hsieh, C.-F., Peng, H.-H., Tsai, Y.-H., Chang, C.-C., Juan, C.-H., Hsu, H.-H., & Juan, C.-J. (2026). An adaptive deep learning framework for multi-label chest X-ray diagnosis using a hybrid CNN–Transformer architecture and class-wise ensemble fusion. *Diagnostics, 16*(8), Article 1227. https://doi.org/10.3390/diagnostics16081227

Jain, S., & Wallace, B. C. (2019). Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3543–3556). https://doi.org/10.18653/v1/N19-

1357

Jeny, A. A., Hamzehei, S., Jin, A., Baker, S. A., Van Rathe, T., Bai, J., Yang, C., & Nabavi, S. (2025). Hybrid transformer-based model for mammogram classification by integrating prior and current images. *Medical Physics, 52*(5), 2999–3014. https://doi.org/10.1002/mp.17650

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems* (Vol. 30). https://arxiv.org/abs/1703.04977

Kumar, A., Khan, M., Mittal, G., Husain, I., Shukla, S., Ali, S. A., Sayal, A., Jha, J., & Ziar, R. A. (2026). Explainable vision transformer framework for breast lesion classification with Grad-CAM support. *Cancer Imaging.* Advance online publication. https://doi.org/10.1186/s40644-026-01097-7

Lee, R. S., Gimenez, F., Hoogi, A., Miyake, K. K., Gorovoy, M., & Rubin, D. L. (2017). A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific Data, 4,* Article 170177. https://doi.org/10.1038/sdata.2017.177

Leibig, C., Allken, V., Ayhan, M. S., Berens, P., & Wahl, S. (2017). Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports, 7*, Article 17816. https://doi.org/10.1038/s41598-017-17876-z

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations.* https://arxiv.org/abs/1711.05101

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., … Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature, 577*(7788), 89–94. https://doi.org/10.1038/s41586-019-1799-6

Mitu, M. Y. (2026). Deep learning-based detection of renal abnormalities from CT images: A study on a Bangladeshi cohort. *American Journal of Smart Technology and Solutions, 5*(2), 52–61. https://doi.org/10.54536/ajsts.v5i2.8226

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics, 9*(1), 62–66. https://doi.org/10.1109/TSMC.1979.4310076

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems* (Vol. 32). https://arxiv.org/abs/1906.02530

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* (Vol. 32). https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

Sadib, M. S. H., Mukit, M. M., Sarker, R. K. R., Tamim, T. A., & Shimul, M. M. H. (2026). *Calibrated selective prediction using deep ensembles for ROI-based thyroid nodule ultrasound classification under dataset shift: A retrospective evaluation* [Preprint]. arXiv. https://arxiv.org/abs/2607.12075

Sawyer-Lee, R., Gimenez, F., Hoogi, A., & Rubin, D. (2016). *Curated breast imaging subset of digital database for screening mammography (CBIS-DDSM)* [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2016.7O02S9CY

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618–626). https://doi.org/10.1109/ICCV.2017.74

Sharma, M. (2026). *Monte Carlo dropout uncertainty and entropy-thresholded selective prediction for architecture-agnostic brain tumour MRI triage* [Preprint]. arXiv. https://arxiv.org/abs/2607.16317

Shen, L., Margolies, L. R., Rothstein, J. H., Fluder, E., McBride, R., & Sieh, W. (2019). Deep learning to improve breast cancer detection on screening mammography. *Scientific Reports, 9*, Article 12495. https://doi.org/10.1038/s41598-019-48995-4

Shen, Y., Wu, N., Phang, J., Park, J., Liu, K., Tyagi, S., Heacock, L., Kim, S. G., Moy, L., Cho, K., & Geras, K. J. (2021). An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Medical Image Analysis, 68*, Article 101908. https://doi.org/10.1016/j.media.2020.101908

Shifa, N., Saleh, M., Akbari, Y., & Al Maadeed, S. (2025). A review of explainable AI techniques and their evaluation in mammography for breast cancer screening. *Clinical Imaging, 123*, Article 110492. https://doi.org/10.1016/j.clinimag.2025.110492

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th International Conference on Machine Learning* (Vol. 97, pp. 6105–6114). PMLR. https://proceedings.mlr.press/v97/tan19a.html

teo-tat. (2026). *ddsm-dl-screening: CNN-based deep learning system for breast cancer detection on the CBIS-DDSM* [Computer software]. GitHub. https://github.com/teo-tat/ddsm-dl-screening

Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., & Hu, X. (2020). Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition Workshops (pp. 111–119). https://doi.org/10.1109/CVPRW50498.2020.00020

Wang, L. (2025). Self-supervised learning and transformer-based technologies in breast cancer imaging. *Frontiers in Radiology, 5*, Article 1684436. https://doi.org/10.3389/fradi.2025.1684436

Wu, J., Kang, W., Tang, H., Hong, Y., & Yan, Y. (2024). On the faithfulness of vision transformer explanations. In P*roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10936–10945). https://doi.org/10.1109/CVPR52733.2024.01040

Xiao, T., Singh, M., Mintun, E., Darrell, T., Dollár, P., & Girshick, R. (2021). Early convolutions help transformers see better. In *Advances in Neural Information Processing Systems* (Vol. 34). https://arxiv.org/abs/2106.14881

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer, 3*(1), 32–35. https://doi.org/10.1002/1097-0142(1950)3:1%3C32::AID-CNCR2820030106%3E3.0.CO;2-3

Zhang, J., Bargal, S. A., Lin, Z., Brandt, J., Shen, X., & Sclaroff, S. (2018). Top-down neural attention by excitation backprop. *International Journal of Computer Vision, 126*(10), 1084–1102. https://doi.org/10.1007/s11263-017-1059-x