Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

A Conditional Positional Encoding and Channel Attention Guided Swin Transformer for Breast Cancer Classification Using MRI and Mammography

Md Abedur Rahman[1], Md Anwar Hossain[1], Kallol Chakraborty Shekhor[2], and Md Sahid Hossain[3]

[1] *Master's in Computer Science, Maharishi International University, Fairfield, IA, U.S.A.*
[2] *Master of Science in Information Studies, Trine University, Allen Park, MI, U.S.A*
[3] *Senior Software Engineer, Prime Tech Solutions Ltd., Dhaka, Bangladesh*

| ABSTRACT

Automated breast cancer image classification remains challenging when MRI and mammography images must be handled within a single compact architecture while preserving both local lesion texture and global contextual information. This study proposes CPCA-SwinNet, a parameter-efficient Swin Transformer-based framework for benign and malignant breast cancer classification. The model incorporates a Conditional Positional Encoding module using 3 × 3 depthwise convolution to provide resolution-adaptive spatial bias and a Channel Attention Gate that recalibrates feature responses through average- and max-pooling with shared bottleneck transformation. Key training hyperparameters were selected through an offline Grey Wolf Optimizer search, followed by final training with cosine-annealed AdamW. A balanced dataset of 3,800 images was assembled from Breast Cancer MRI, Duke Breast Cancer MRI, and CBIS-DDSM, comprising 1,860 benign and 1,940 malignant images, and split into 70% training, 10% validation, and 20% testing subsets. On the held-out test set of 760 images, CPCA-SwinNet achieved 98.82% accuracy, 99.23% sensitivity, 98.39% specificity, 98.85% F1-score, 0.9763 Matthews correlation coefficient, and 0.9964 AUC-ROC. Five-fold cross-validation yielded 98.77 ± 0.21% accuracy and 0.9962 ± 0.0008 AUC-ROC. Ablation analysis showed a 2.77 percentage-point improvement over the baseline Swin-T model, and McNemar's test indicated significant gains over nine baseline models after Bonferroni correction. With 31.2 million parameters, CPCA-SwinNet maintains the efficiency profile of Swin-T while improving classification performance. External multi-institutional and prospective validation remains necessary before clinical use.

| KEYWORDS

Breast cancer classification, CPCA-SwinNet, Swin Transformer, Conditional positional encoding, Channel attention gate , Vision transformer

| ARTICLE INFORMATION

ACCEPTED: 01 June 2024 **PUBLISHED:** 23 July 2024 **DOI:** 10.32996/jcsts.2024.6.3.16

Introduction

Breast cancer remains one of the most common malignancies among women and continues to impose a substantial global health burden [1]. Imaging plays a central role in breast cancer assessment, with mammography and magnetic resonance imaging used to identify suspicious findings, guide follow-up, and support biopsy decisions[2,3]. However, interpretation remains demanding because benign and malignant lesions may share overlapping visual characteristics, including irregular margins, heterogeneous density, enhancement variation, and subtle textural patterns. Reader workload, inter-observer variability, and differences in acquisition protocols further complicate image-based assessment, especially when images originate from heterogeneous sources [4,5]. Automated image classification methods are therefore valuable as research tools for studying computational patterns associated with malignancy, although they should not be interpreted as substitutes for radiological judgement or as clinically deployable systems without external validation[6].

Deep learning has substantially advanced breast image analysis, but the design of an effective and compact architecture for mixed breast imaging data remains challenging[7]. Mammography and MRI differ in physical acquisition principle, image statistics, lesion appearance, spatial resolution, and contextual information. Mammographic images often emphasise density, margins, and microcalcification-related texture, whereas MRI, particularly contrast-enhanced MRI, reflects enhancement behaviour and the relationship between a lesion and surrounding parenchyma [8]. A model trained across these modalities must therefore preserve fine local detail while also modelling broader anatomical and contextual dependencies. This dual requirement is difficult to satisfy under the limited sample sizes and distribution shifts that are common in public medical imaging datasets[9,10].

Earlier breast image classifiers relied heavily on convolutional neural networks such as VGG, ResNet, DenseNet, EfficientNet, and Xception [11]. These models remain useful because convolution provides strong local inductive bias and captures lesion texture efficiently. However, their ability to model long-range spatial relationships depends on depth, receptive-field growth, and hierarchical feature aggregation [12]. Transformer-based models, including Vision Transformer and Swin Transformer, address part of this limitation by using attention mechanisms to capture wider contextual dependencies. Yet transformer backbones may weaken local spatial sensitivity if positional information is not handled carefully, and their channel representations are often treated uniformly even though only a subset of channels may carry the most discriminative malignancy-related information. Recent architectures such as CoAtNet, ConvNeXt, MaxViT, and EfficientViT-style models have narrowed the gap between convolutional locality, attention-based context modelling, and computational efficiency, but the trade-off between accuracy, parameter budget, and modality robustness remains an open practical concern in breast image classification.

To this end, this study proposes CPCA-SwinNet, a compact Swin Transformer-based architecture for binary breast image classification. The model introduces two targeted modifications. First, a Conditional Positional Encoding module based on a 3×3 depthwise convolution injects local, content-conditioned spatial information into the token representation, reducing dependence on fixed positional bias. Second, a Channel Attention Gate recalibrates feature channels by combining average- and max-pooled descriptors through a shared bottleneck transformation and sigmoid gating. These components are designed to address two specific weaknesses of a standard hierarchical transformer: loss of local spatial sensitivity and uniform treatment of feature channels.The model is evaluated on a curated dataset of 3,800 breast images assembled from three public sources: Breast Cancer MRI, Duke-Breast-Cancer-MRI, and CBIS-DDSM. The evaluation includes a held-out test set, five-fold cross-validation, comparison against nine baseline architectures, component-level ablation, statistical testing using McNemar's test, and computational efficiency analysis. This protocol is intended to separate architectural contribution from training variation and to provide a more cautious assessment than single-split accuracy alone. Because the evaluation remains internal to the curated corpus, the reported findings should be interpreted as evidence of research-level classification performance. External, multi-institutional, and prospective validation remains necessary before any clinical deployment can be considered.

2. Related Work

Recent breast cancer research emphasizes that early and accurate detection remains essential for improving patient outcomes. Comparative work on the Wisconsin Breast Cancer Dataset and MIAS database showed that CNN classifiers achieved stronger precision and recall than SVM, Random Forest, ANN, and other conventional models, although dataset bias, limited interpretability, and ethical concerns still restrict clinical translation [13]. Broader precision-medicine research highlights that accurate diagnosis supports subtype-specific treatment planning, including targeted therapy, liquid biopsy, molecular imaging, and AI-assisted decision support [14]. Non-imaging screening approaches using biosensors, electronic nose systems, and AI have also been explored, but validation, standardization, and low biomarker concentration remain unresolved issues [15]. Generative learning has been used for data augmentation, super-resolution, segmentation, imbalance correction, and performance improvement, although dataset heterogeneity and limited multi-center validation remain common limitations [16]. Liquid-biopsy research based on DNA methylation further shows promise for early biomarker discovery with machine learning, but sensitivity and biomarker validation remain insufficient for immediate clinical use [17].

Deep learning remains central to breast image analysis because it can learn discriminative visual patterns directly from imaging data. ConvNeXt-based ultrasound classification showed the value of modern convolutional backbones, but the work was limited to a single modality and lacked external testing [18]. Classical machine-learning pipelines have also been improved through SHAP-based feature selection, Borderline-SMOTE1, Particle Swarm Optimization, and optimized classifiers such as KNN, Random Forest, Logistic Regression, SVM, and LightGBM [19]. Related comparative work confirmed that traditional classifiers can remain competitive when supported by suitable feature-selection strategies [20]. Alternative sensing and pathology studies have investigated terahertz metamaterial biosensors, explainable DenseNet variants, and bioelectrical impedance classification, but these settings differ from standard MRI and mammography image classification or remain limited by simulation-based validation [21], [22], [23]. Reviews across mammography, ultrasound, MRI, histopathology, and thermography consistently identify dataset

variability, imbalance, and limited interpretability as major barriers, while suggesting augmentation, transfer learning, and multimodal integration as possible directions [24]. Mammography-focused Fractional Order CNNs with Particle Swarm Optimization and adaptive filtering have also been proposed, but incomplete reporting of datasets, splits, and validation limits assessment of the claimed performance [25].

Transformer-based and hybrid models have further expanded breast cancer classification research. Vision Transformer models have shown strong performance for invasive ductal carcinoma classification compared with tested CNN baselines [26]. Lightweight segmentation and classification frameworks using depth-wise separable convolutions, multi-source mammography and ultrasound images, and StyleGAN3-generated samples have also been explored [27]. Broader reviews describe AI-integrated tools as useful for diagnostic accuracy and risk-stratified decision-making, although they do not provide unified quantitative benchmarks for image-classification models [28]. Explainable DenseNet121-based models with custom classification heads and Grad-CAM have been evaluated on histopathology and ultrasound data, but they do not address mixed MRI and mammography classification [29]. Other studies report high-performing deep-learning results across several breast cancer benchmarks, yet many of these results are aggregated from separate studies rather than reproduced under one controlled protocol [30]. Transfer-learning comparisons using ResNet, MobileNet, VGG-16, and EfficientNet have also been reported, but incomplete dataset and result details limit interpretability [31]. Hybrid CNN-LSTM models evaluated on Kaggle mammography datasets have shown strong performance against CNN, LSTM, GRU, VGG-16, and ResNet-50 baselines [32]. Overall, existing studies show strong progress across imaging, biomarker, sensing, and computational approaches, but comparability remains limited because reported results often use different datasets, modalities, preprocessing pipelines, train-test splits, and baseline selections.

3.1 Dataset Description

The dataset used in this study was assembled from three publicly available breast imaging collections that together cover the two modalities most commonly used in clinical breast assessment: magnetic resonance imaging (MRI) and digital mammography (Figure 1). The first source is the Breast Cancer MRI collection distributed through Kaggle, which provides labelled benign and malignant MRI cases. The second is the Duke-Breast-Cancer-MRI collection from The Cancer Imaging Archive (TCIA), which contributes dynamic contrast-enhanced MRI (DCE-MRI) studies. The third is the Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM), a widely used mammographic resource containing benign and malignant lesion patches with verified pathology labels. The use of three independent public sources was intended to increase visual and acquisition diversity within the experimental corpus, while the absence of patient-level metadata harmonisation across the three archives means that the corpus should be treated as a curated experimental dataset rather than a clinically representative cohort.

The classification problem is framed as a binary task with two mutually exclusive categories, benign and malignant, consistent across all three contributing sources. After curation, the combined corpus contains 3,800 images, of which 1,860 are labelled benign and 1,940 are labelled malignant, giving a class ratio close to balance and removing the need for explicit class-weighting during training. The dataset composition for each source, together with the modality, class breakdown, and total count, is summarised in Table 1.

Table 1: Dataset Description

Dataset	Modality	Classes	Benign	Malignant	Total
Breast Cancer MRI(Kaggle)	MRI	2	740	740	1,480
Duke-Breast-Cancer-MRI(TCIA)	DCE-MRI	2	0	922	922
CBIS-DDSM	Mammography	2	4,346	5,854	10,200
Combined (after curation)	MRI + Mammography	2	1,860	1,940	3,800

The two imaging modalities differ in physical principle and in what they reveal about a lesion. Mammography produces a two-dimensional projection of the compressed breast using low-dose X-rays and is the primary modality used in population-level screening; lesion appearance is dominated by density patterns, microcalcifications, and margin characteristics. MRI, and DCE-MRI in particular, relies on tissue magnetisation and on the temporal kinetics of an injected contrast agent, so lesion appearance reflects vascularisation and enhancement curves rather than density. Including both modalities in a single training corpus broadens the range of textural and contrast cues that the model must learn to recognise, but it also introduces modality-related variation in image statistics, resolution, and field of view. The presence of such variation is a deliberate stress test for the classifier rather than a guarantee that the resulting model will generalise across all imaging sources.Curation was performed at the image level. The curated dataset was organised after label harmonisation across the three sources, removal of unusable samples that lacked a verified pathology label, and exclusion of cases with inconsistent or ambiguous annotation. Where a source contributed a substantially larger number of raw samples than was needed for the experimental balance, a stratified subset was retained so that the final benign-to-malignant ratio remained near unity. CBIS-DDSM contributes the largest pool of raw images among the three sources, but the curated subset draws fewer samples from it in order to keep the modality balance from collapsing toward mammography. No patient-level identifiers, scanner settings, or acquisition parameters were modified during curation; only sample-level inclusion or exclusion was performed.The 3,800 curated images were partitioned into training, validation, and test subsets in a 70/10/20 ratio, yielding 2,660 training, 380 validation, and 760 test images. The split preserves the benign-to-malignant ratio in each partition, as detailed in Table 2. The training subset was used for parameter learning, the validation subset for early stopping and hyperparameter selection (including the offline Grey Wolf Optimizer search described in Section 3.4), and the held-out test subset for the comparative evaluation reported in Section 4. No image was shared across partitions.

Table 2: Data Partitioning

Partition	Benign	Malignant	Total	Proportion(%)
Training	1,302	1,358	2,660	70
Validation	186	194	380	10
Testing	372	388	760	20
Total	1,860	1,940	3,800	100

The combined corpus is well matched to the design intent of CPCA-SwinNet. Mammographic patches and MRI slices place different demands on a backbone: the first emphasises localised texture and margin cues, the second emphasises the relationship between an enhancing region and its surrounding parenchyma. A backbone that handles only locality risks losing modality-specific context, and a backbone that handles only global attention risks softening lesion-level detail. The Conditional Positional Encoding and Channel Attention Gate components introduced in Section 3.3 are intended to address exactly this trade-off, which is why the modality-mixed corpus is appropriate as an evaluation setting.A note on the per-dataset evaluation in Section 4 is necessary here for consistency. Among the three contributing sources, Duke-Breast-Cancer-MRI provides only malignant cases. As a result, the Duke partition supports sensitivity-style measurement but cannot support metrics that require true negatives, such as specificity or AUC. The per-dataset breakdown in Table 12 reflects this constraint, and metrics that are not defined for a single-class subset are reported as not applicable rather than estimated.The dataset has limitations that should be stated openly. It is constructed from public archives, and the patient populations, scanner platforms, and acquisition protocols underlying those archives are not exhaustive. The corpus does not contain prospectively collected data and has no representation from imaging centres outside the contributing archives. The reported test partition is internal to the curated corpus, and the cross-validation reported in Section 4 measures stability within that corpus rather than transfer to new institutions. External evaluation on independent multi-institutional cohorts, with prospective data collection and patient-level metadata, remains a necessary step before any conclusion is drawn about clinical translation.

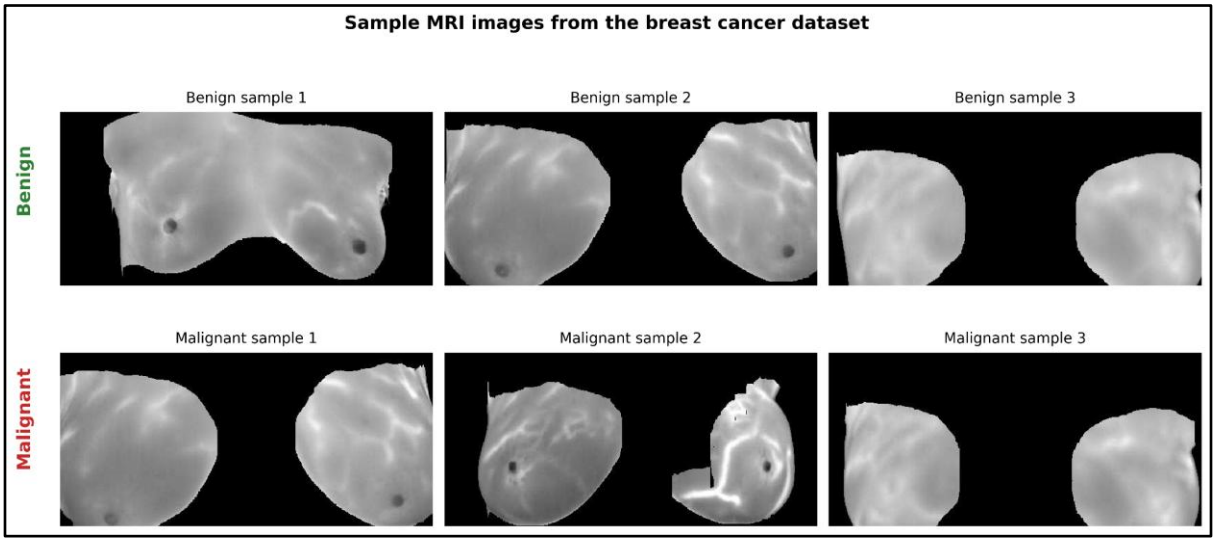

Figure 1. Sample Breast MRI Images from the Curated Dataset

3.2 Dataset Preprocessing and Augmentation

Breast images drawn from three different public archives differ in resolution, contrast distribution, noise characteristics, and the conditions under which they were acquired. Some of this variation is biologically meaningful and should be preserved; much of it, however, reflects scanner platform, detector technology, and reconstruction settings rather than the underlying lesion. A consistent preprocessing pipeline was therefore applied before training so that CPCA-SwinNet would receive inputs of comparable scale and intensity, and so that downstream optimization would respond to the diagnostic signal rather than to source-specific image statistics. The full configuration is summarised in Table 3. All images were resized to 224 × 224 pixels to match the patch tokenisation requirements of the Swin-T scale backbone used in CPCA-SwinNet. A fixed input size also ensures that windowed self-attention operates on a constant token grid across the corpus. Pixel intensities were rescaled to the unit interval [0, 1 (Figure 2)]. This range normalisation is a standard preconditioning step that stabilises gradient magnitudes during the early epochs of optimization and removes the most obvious differences in dynamic range between mammographic and MRI samples. It does not, however, correct for modality-specific contrast distributions, scanner-dependent signal-to-noise differences, or breast-density bias, and it should not be expected to do so.Adaptive median filtering with a 3 × 3 kernel was applied to suppress local impulse-like noise while retaining diagnostically relevant image structures such as lesion margins and microcalcification clusters. Adaptive median filtering was preferred to a fixed-kernel mean or Gaussian smoother because the latter tends to attenuate small high-frequency structures that are clinically informative. The filter was kept at a small kernel size for the same reason: aggressive smoothing can remove subtle texture cues that distinguish a benign mass from an early malignancy, which is the opposite of what is wanted in a binary breast classification task.

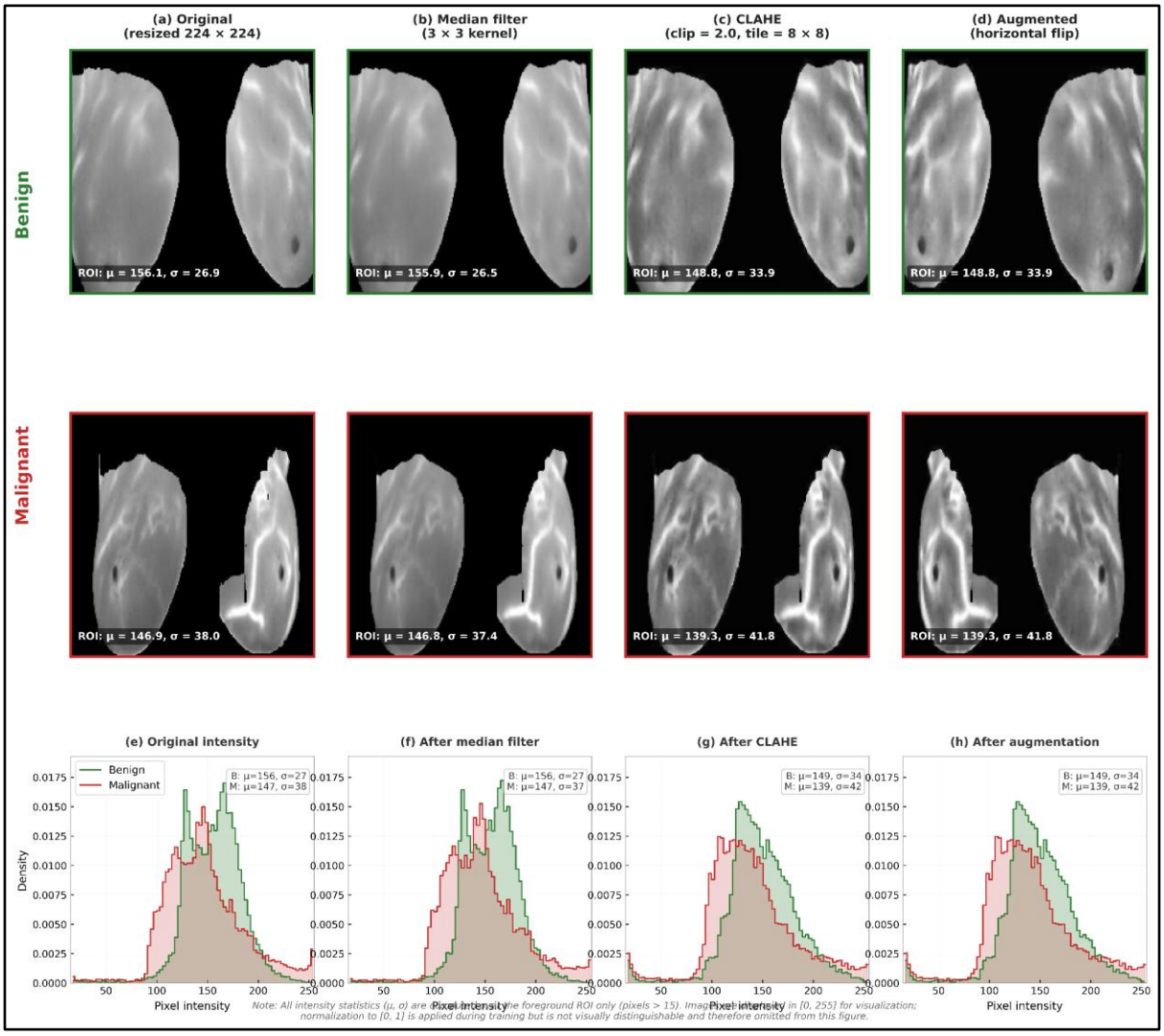

Figure 2. Preprocessing and Augmentation Workflow for Breast Image Standardization

Geometric augmentation was used to expose the model to plausible variation in patient positioning and acquisition geometry without altering the diagnostic content of the image. Random rotations within ± 15°, horizontal flips with probability 0.5, vertical flips with probability 0.3, and small affine translations of up to ± 10% were sampled at training time. Random erasing was applied with probability 0.25 and an area ratio between 0.02 and 0.20 to encourage the model to rely on distributed evidence across the image rather than on any single salient region. Photometric augmentation was applied conservatively. Brightness and contrast jitter were both restricted to ± 0.1, which is small enough to leave radiologically meaningful intensity differences intact while still encouraging the classifier to be insensitive to minor exposure differences. No colour-channel manipulation beyond brightness and contrast was applied, since breast images are essentially single-channel grey-scale data.Mixup was used as a regularisation strategy with a Beta distribution parameter $\alpha = 0.2$, producing convex combinations of pairs of training images and their labels. Mixup discourages the network from forming overly confident piecewise-constant decision boundaries and has been reported to improve calibration on imbalanced or low-data medical tasks. Label smoothing, used during the supervised loss computation rather than as an input-space augmentation, plays a complementary role and is described in Section 3.4 to avoid repeating training hyperparameters here.

The preprocessing pipeline was designed to interact predictably with the architectural choices in CPCA-SwinNet. A fixed input grid is required for the Conditional Positional Encoding to inject local spatial structure consistently across stages, and noise suppression at the input stage helps the Channel Attention Gate amplify channels that carry genuine class-discriminative response rather than channels dominated by acquisition artefacts. Augmentation broadens the training distribution along axes (orientation, translation, partial occlusion, mild brightness drift) that are likely to vary between imaging centres, which is the kind of variation against which a transformer with a small parameter budget is otherwise prone to overfit.

Table 3: Preprocessing and Data Augmentation Configuration

Operation	Parameter	Value
Resize	Target resolution	224 x 224 pixels
Pixel normalization	Range	[0,1]
Noise removal	Filter type	Adaptive median filter(3x3)
Rotation	Degree range	±15°
Horizontal flip	Probability	0.5
Vertical flip	Probability	0.3
Random affine	Translation	±10%
Random erasing	Probability / area ratio	0.25/0.02-0.20
Color jitter	Brightness / contrast	± 0.1/± 0.1
Mixup	Alpha	0.2

4.1. Overall Framework of CPCA-SwinNet

CPCA-SwinNet is a transformer-based classification network that takes a single breast image as input and returns a probability over the two diagnostic classes, benign and malignant (Figure 3). The end-to-end design follows the stage-wise hierarchical pattern of Swin Transformer, but it introduces two architectural components, a Conditional Positional Encoding (CPE) and a Channel Attention Gate (CAG), placed exactly where a standard hierarchical transformer is most likely to lose either local spatial structure or class-discriminative channel information. The complete pipeline is shown in Figure 3.The network defines a parameterised mapping

$$\hat{y} = f\theta(x), \qquad x \in R^{H\times W\times X}, \qquad \hat{y} \in [0,1]^2$$

Where x is a single breast image and $\hat{y}$ is the predicted class-probability vector. The mapping factorises into four functional blocks acting in sequence,

$$\hat{y} = C(G(B(E(P(x)))))$$

Where ρ denotes preprocessing and (during training) augmentation, E is the patch-embedding layer that converts the spatial image into a sequence of tokens, is β the hierarchical transformer backbone equipped with CPE, G is the channel attention gate, and C is the classifier head that produces the final logits.

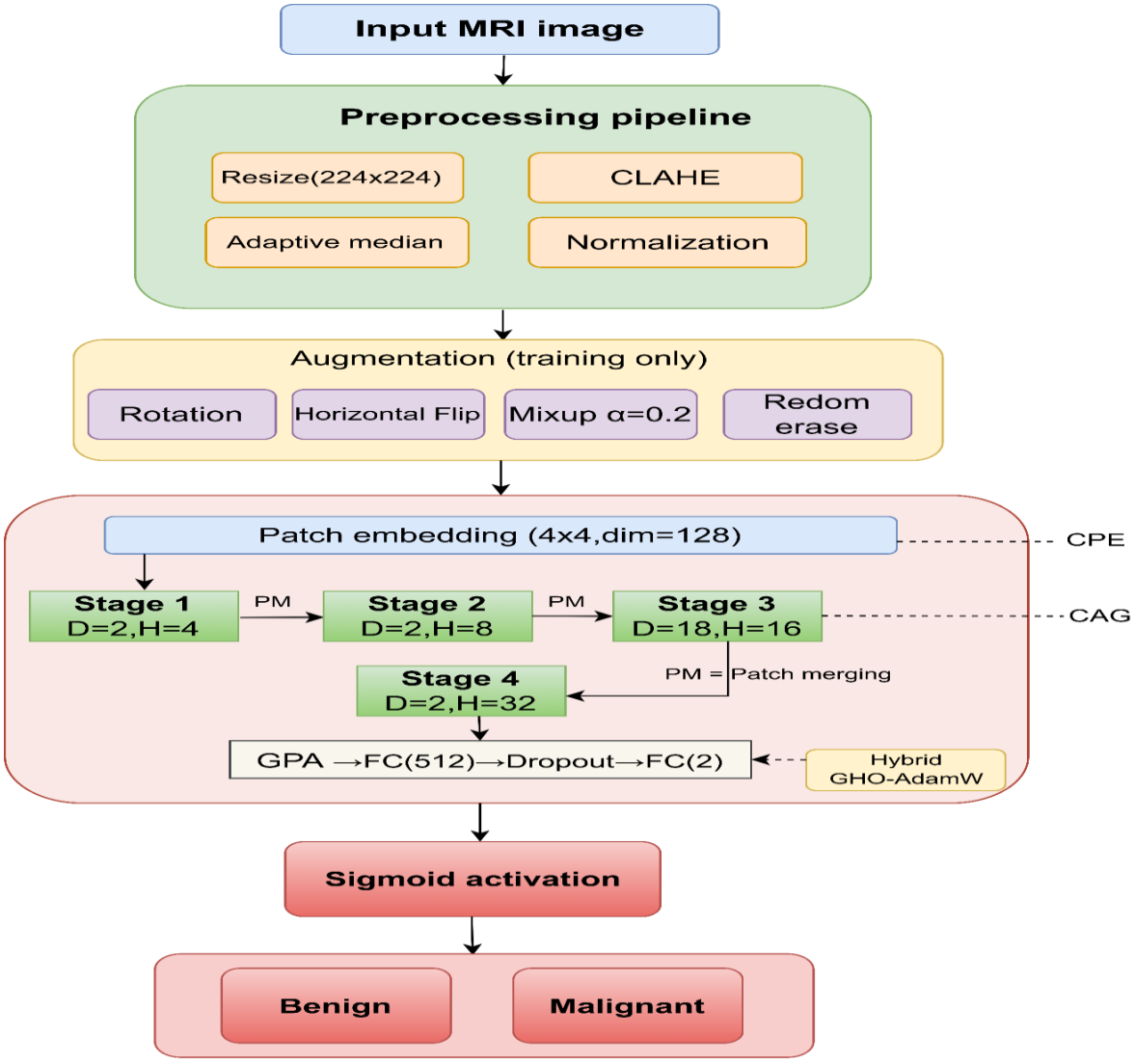

Figure 3. Overall Architecture of the Proposed CPCA-SwinNet Framework

The upstream block ρ standardises the input across the heterogeneous imaging sources used in this study; the specific resize, normalisation, noise-suppression, and augmentation operations were described in Section 3.2 and are not repeated here. The patch-embedding layer E partitions the standardised image into non-overlapping patches and projects each patch into a token vector, producing the token grid on which all subsequent attention operates. The backbone B then processes this grid in four successive stages of increasing receptive field and channel depth. Within each stage, alternating window and shifted-window self-attention blocks model dependencies among tokens, and patch-merging operations between stages reduce spatial resolution while doubling the channel count. CPE is applied at the entry of each attention stage; rather than relying on a fixed absolute or relative positional bias, it injects local spatial information through a small depthwise convolution, which keeps the model robust to changes in input grid geometry and to the modality variation present in the corpus (Figure 4).After the final attention stage, the channel attention gate G recalibrates the feature map by learning a per-channel weight that emphasises channels carrying strong class-discriminative response and dampens channels dominated by texture or background variation. The recalibrated feature map is then collapsed by global average pooling and passed to the classifier head C, a fully connected projection followed by a non-linearity and dropout, a second fully connected layer of dimension two, and a sigmoid activation that yields the benign-versus-malignant probability .

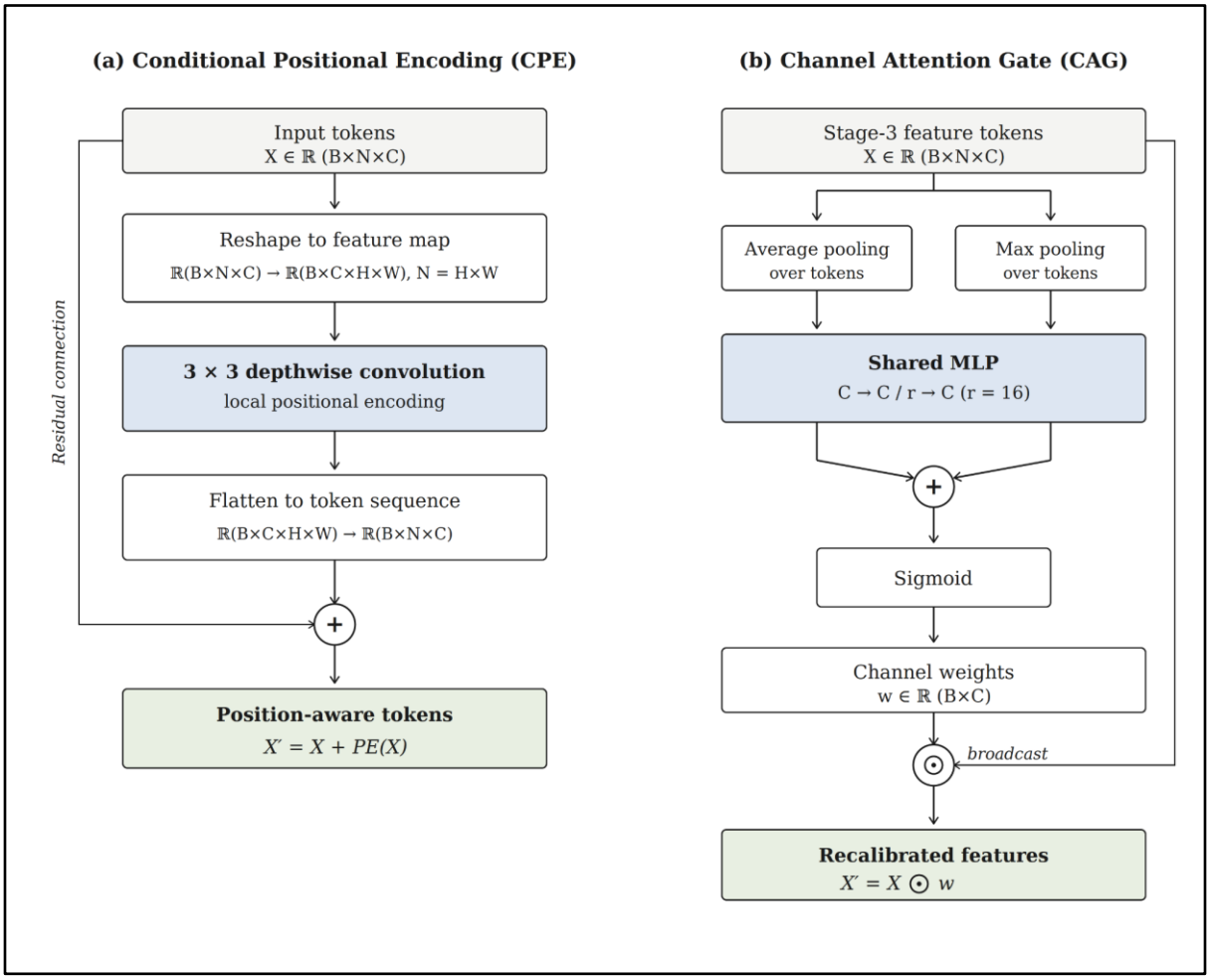

Figure 4. Conditional Positional Encoding and Channel Attention Gate Modules

4.2. Hierarchical Feature Extraction with Conditional Positional Encoding

The backbone β introduced in Section 4.1 is responsible for converting the standardised image into a hierarchy of representations of progressively coarser spatial resolution and richer channel content. Three design choices distinguish CPCA-SwinNet from a vanilla Swin Transformer at this stage: the way local position is injected at the entry of each attention sequence, the alternation of windowed and shifted-windowed self-attention inside the repeated transformer blocks, and the resolution–channel rebalancing performed by patch merging between stages. Each is described below in the order in which information flows through the network. A standardised image $x \in R^{H\times W\times X}$ is first partitioned into a regular grid of non-overlapping 4×4 patches, and each patch is linearly projected into a token vector. Stacking the projections yields the initial token sequence

$$X_0 = PatchEmbed(x) \in R^{B\times N\times C}, N = \frac{H}{4}\cdot\frac{W}{4}$$

Where β is the batch size and C is the embedding dimension. Tokenisation discards the explicit two-dimensional layout of the image: tokens are now indexed as a sequence, and the geometry that was implicit in pixel coordinates must be re-introduced before any attention block can use it sensibly.

This is the role of Conditional Positional Encoding. In a standard Swin Transformer, position information enters through a learned relative-position bias added inside the attention computation. Such a bias is tied to a fixed window size and to the precise grid geometry observed during training. In a corpus that combines mammographic patches and MRI slices, in which lesion location, breast outline, and aspect ratio vary across modalities, this rigidity is a liability. CPE sidesteps the problem by recovering position information directly from the token content. The NN N tokens are first reshaped back into a feature map $R^{B\times N\times C} \to R^{B\times C\times H\prime\times W\prime}$ where $H'W' = N$, then passed through a 3×3 depthwise convolution that operates independently on each channel, and finally flattened back to a sequence,

$$PE(X) = Flatten(DWConv_{3\times3}(Reshape(X)))$$

The position-aware tokens entering the next attention stage are then

$$X_{cpe} = X + PE(X).$$

Three properties of this construction matter. First, the depthwise convolution introduces only $9C$ additional parameters per stage, which is negligible relative to the attention parameters. Second, the residual addition leaves the original token semantics intact and contributes a strictly local positional signal on top of them, which is appropriate for lesion-scale features. Third, because position is now generated from the input rather than retrieved from a fixed bias table, the encoding adapts when the input grid changes, which is the regime in which absolute and relative positional embeddings degrade.The position-aware tokens are then processed by a stack of transformer blocks that alternate between window-based multi-head self-attention (W-MSA) and its shifted-window variant (SW-MSA). For block l within a stage, the attention update applies pre-normalisation and a residual connection,

$$X_l' = W - MSA\big(LN(X_{l-1})\big) + X_{l-1}, \qquad X_l = MLP\left(LN(X_l')\right) + X_l'$$

Even-indexed blocks use W-MSA, in which self-attention is computed independently within non-overlapping windows of size 7×7 tokens; odd-indexed blocks use SW-MSA, in which the window grid is shifted by half a window before attention is computed. The alternation is what allows information to cross window boundaries without paying the quadratic cost of global attention. The MLP is a two-layer feed-forward network with a GELU non-linearity and dropout, with an expansion ratio of four between the two linear layers. Layer normalisation is applied before each sub-block, and residual paths are kept across both the attention and the feed-forward branches, which stabilises optimisation in deeper stages where token statistics begin to drift.Between stages, a patch-merging operation downsamples the token grid by a factor of two in each spatial dimension and concatenates the four resulting tokens before projecting back to twice the original channel dimension,

$$X_{s+1} = Patch\ Merge(X_s)$$

The four-stage hierarchy of CPCA-SwinNet therefore produces feature maps whose spatial resolution decreases by a factor of two and whose channel count doubles between successive stages, while CPE re-injects local position at the entry of each new stage so that the alternating attention blocks within the stage can rely on a fresh, content-conditioned positional signal. This combination is what allows the backbone to model fine-grained lesion texture at the early stages and broader anatomical context at the later stages without needing the parameter budget of a global-attention transformer. The output of the final stage is the feature map on which the channel attention gate operates; the formulation of that gate, together with the classifier head that closes the network, is the subject of Section 4.3.

4.3. Channel Attention Gating and Classification Head

The four-stage backbone described in Section 4.2 produces a feature map in which the final stage tokens span a rich channel space, but treats every channel as equally informative (Figure 5). Self-attention modulates *spatial* relationships between tokens; it does not, on its own, modulate *which channels* contribute most strongly to a class decision. In binary breast image classification, the discriminative signal between benign and malignant patterns is typically concentrated in a small subset of channels, while the remainder encodes texture, low-frequency structure, or modality-specific variation. CPCA-SwinNet addresses this asymmetry with a Channel Attention Gate (CAG), inserted after the final attention stage and before the classifier head.The CAG begins by collapsing the spatial dimension of the stage-3 feature map $X \in R^{B\times N\times C}$ along the token axis to obtain two complementary channel descriptors. Average pooling captures the mean response of each channel across the token grid, while max pooling captures the strongest single response. Both summaries are useful: the mean is sensitive to broadly distributed evidence, the maximum to localised evidence. Formally,

$$z^{avg} = AvgPool(X), z^{max} = MaxPool(X), z^{avg}, z^{max} \in R^{B\times C}$$

Each descriptor is then passed through a shared two-layer multilayer perceptron with a bottleneck reduction ratio r, so that the intermediate dimension is $\frac{C}{r}$ and the output dimension is restored to C. Sharing the MLP across the two pooling streams is a deliberate constraint: it forces both summaries to be interpreted in the same channel coordinate system and prevents the two branches from drifting toward unrelated representations. The combined channel logit is

$$u = MLP(z^{avg}) + MLP(z^{max}), u \in R^{B\times C}$$

Applying a sigmoid converts this logit into a per-channel gating weight bounded in [0,1],

$$w = \sigma(u), w \in [0,1]^{B \times C}$$

The gate is then broadcast across the spatial dimension and applied element-wise to the original feature map, producing the recalibrated representation

$$\underline{X} = X \odot w$$

Two consequences of this construction are worth stating. First, no channel is removed: the gate down-weights uninformative channels but preserves the underlying feature graph, which keeps gradients well-conditioned during fine-tuning. Second, the parameter cost is minor. With reduction ratio r, the CAG adds $\frac{2}{r}$ MLP weights, which is negligible compared with the attention weights of the backbone and is one reason CPCA-SwinNet stays within the parameter envelope of a Swin-T scale model. We do not interpret w as a measure of clinical importance; it is an internal recalibration mechanism, and the weight assigned to a channel reflects only its statistical association with the binary label under the training distribution.

The recalibrated feature map $\underline{X}$ is then summarised by global average pooling, which collapses the token sequence into a single C-dimensional vector. This vector is passed to a compact classifier head consisting of a fully connected projection of width 512 with a ReLU non-linearity, dropout at rate 0.3, and a final fully connected layer of width 2. The sigmoid activation at the output yields the benign-versus-malignant probability,

$$\hat{p} = \sigma\big(W_2\phi\big(W_1 GAP\ \big(\underline{X}\big) + b_1\big) + b_2\big)$$

Where ϕ denotes the ReLU activation and dropout is applied between the two linear layers. The classifier head is intentionally shallow: most of the representational work has already been performed by the CPE-equipped backbone and the CAG, so a deeper head would risk overfitting the comparatively small medical corpus without offering matching gains.Together, the CPE described in Section 4.2 and the CAG described here address two distinct failure modes of a hierarchical transformer applied to breast images. CPE injects local spatial structure at the entry of each attention stage; the CAG selects which channels of the resulting hierarchy carry class-discriminative signal. The remaining design question is how the network is trained so that these two mechanisms can be exploited reliably. Section 4.4 describes the loss formulation, the regularisation scheme, and the offline Grey Wolf Optimizer search used to fix the AdamW hyperparameters under which CPCA-SwinNet was finally trained.

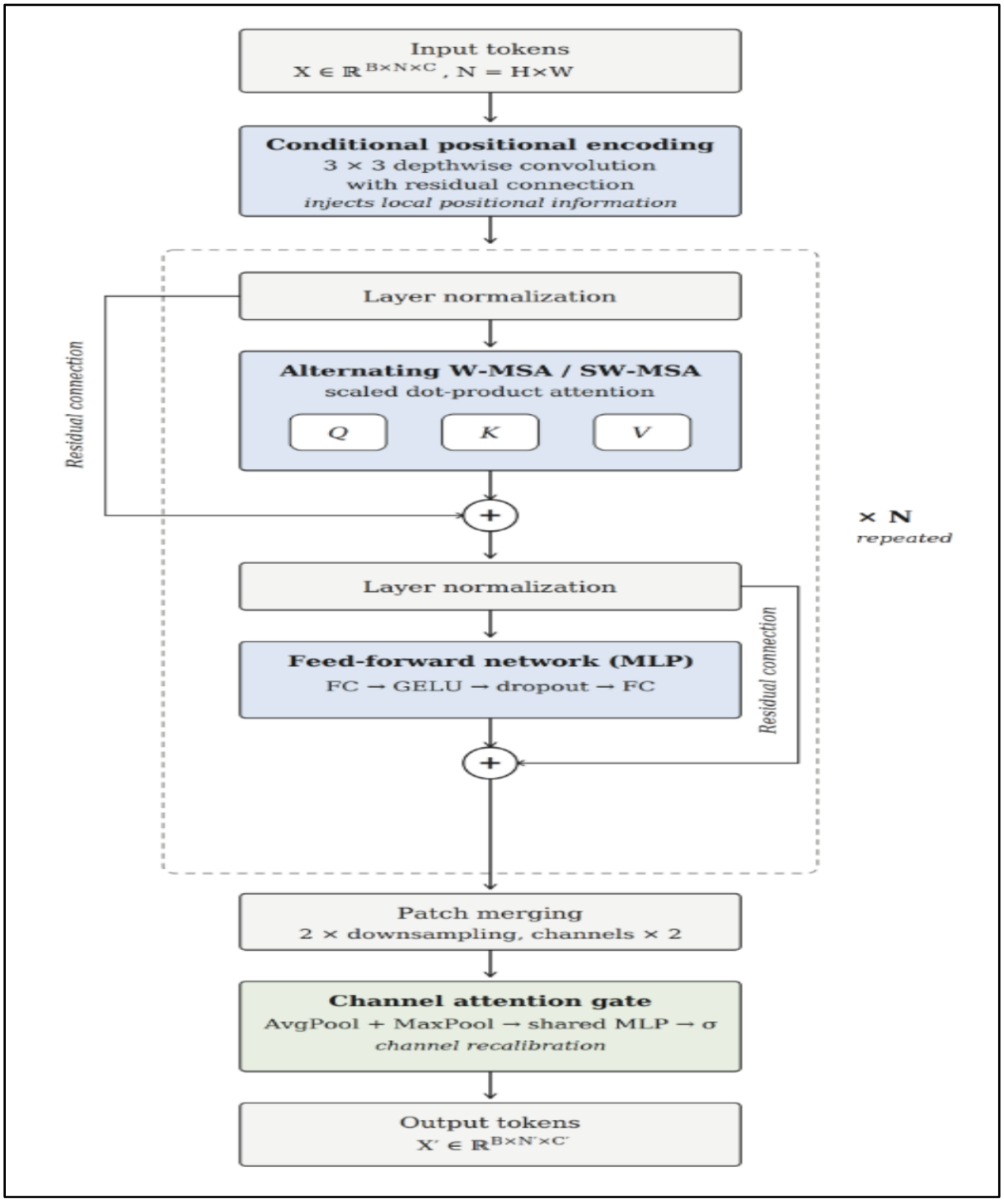

Figure 5. Internal Transformer Block Design of CPCA-SwinNet

4.4. Training Strategy, Optimization, and Integrated Mathematical Formulation

The training procedure for CPCA-SwinNet is organised on two levels. The inner level is the standard supervised optimization of the network weights θ under a labelled training set. The outer level is the offline selection of the hyperparameters that govern that optimization. Separating these two levels matters because the loss landscape of a hierarchical transformer is sensitive to learning rate, weight decay, and dropout in ways that hand-tuning rarely characterises systematically, particularly on a medical corpus of moderate size.

The outer level uses a Grey Wolf Optimizer (GWO) to search a four-dimensional hyperparameter space, namely the learning rate, the weight decay, the dropout probability, and the AdamW ϵ coefficient (Figure 6). The search is performed on the validation partition only, with a population of 30 candidate configurations and 20 iterations, and with each candidate evaluated by training CPCA-SwinNet for a small fixed number of epochs and scoring the resulting binary cross-entropy on the validation fold.

The motivation for using a metaheuristic here is not that gradient-based search of weights is inadequate, but that hyperparameter selection is a non-differentiable, low-dimensional, and noisy black-box problem for which population-based search is well suited; GWO converges reliably in this regime and is cheaper to run than an exhaustive grid in four dimensions. A direct comparison against grid search, Bayesian optimisation, and competing metaheuristics is reported in Section 5 (Table 11).

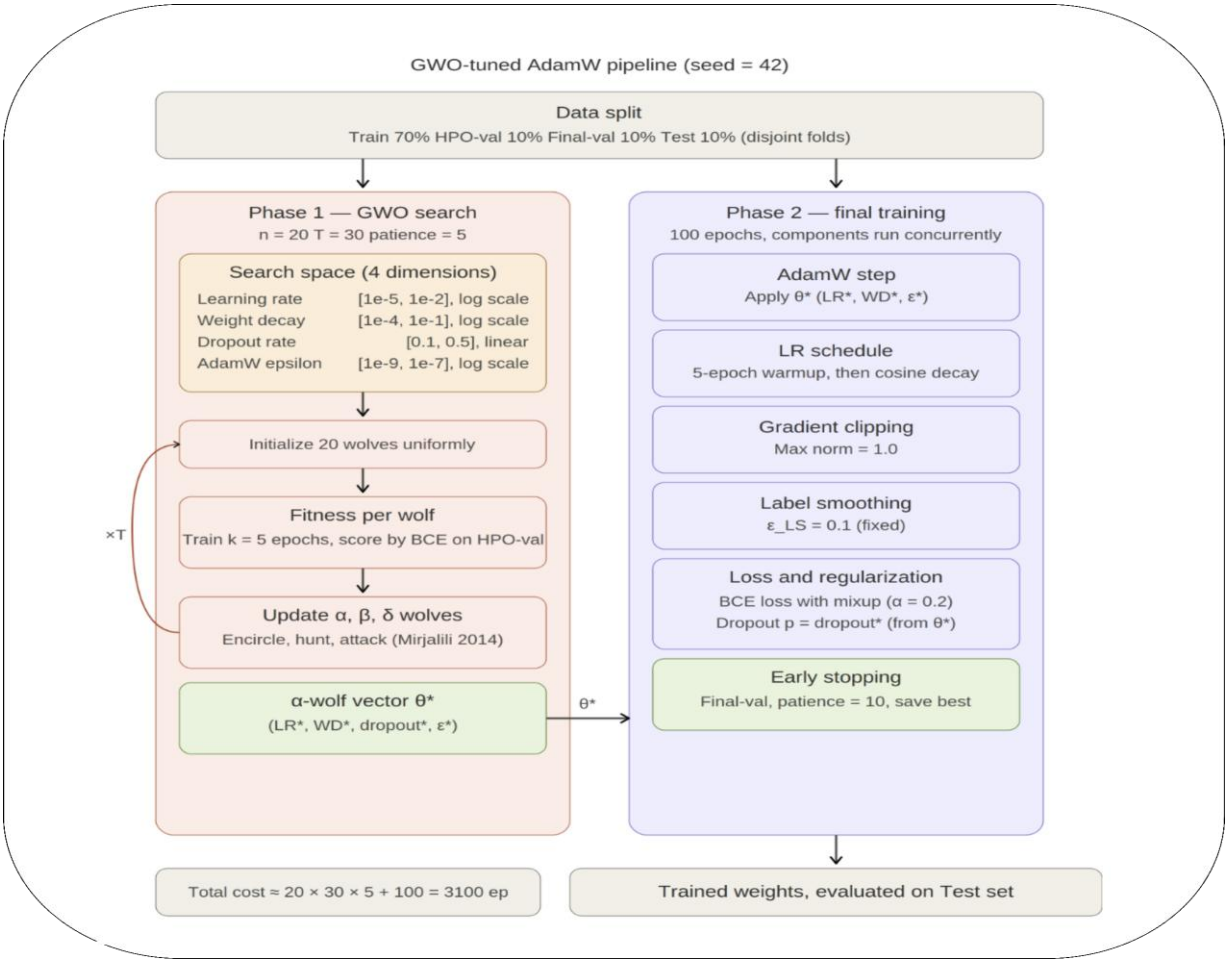

Figure 6. Grey Wolf Optimizer Tuned AdamW Training Pipeline

The inner level then performs full supervised training under the GWO-selected configuration. The optimiser is AdamW initialised at learning rate $\eta^{\star}$ and weight decay $\lambda^{\star}$, with a linear warm-up over the first five epochs followed by cosine annealing for the remainder of the schedule (Table 4). Gradients are clipped at a maximum norm of 1.0 to stabilise the early phase of transformer training, where activation statistics can drift before normalisation layers settle. Training is run for up to 100 epochs with early stopping on validation loss with patience 10. The parameter update at step t is written compactly as

$$\theta_{t+1} = AdamW\ (\theta_t \nabla_\theta L\ (\theta_t) \eta^{\star}, \lambda^{\star}, \epsilon^{\star})$$

The training objective is a binary cross-entropy with label smoothing, which discourages the network from producing saturated logits on a corpus where label noise from differing curation pipelines cannot be ruled out. Given a smoothing parameter $\epsilon \in [0,1]$ and the original target $y \in \{0,1\}$, the smoothed target is $y_s = (1-\varepsilon)y + \varepsilon/K$ with $K+2$ the loss for a single example with predicted probability $\hat{p}$ is

$$L\ (\hat{p}, y) = -[y_s\ log\hat{p} + (1-y_s) log(1-\hat{p})]$$

When mixup is active during a training step, both the input and the target are interpolated by a $Beta(\alpha, \alpha)$ coefficient before the loss is evaluated; mixup therefore enters the training pipeline as an input-space and target-space transformation, not as a separate loss term. Together, label smoothing and mixup act as complementary regularisers: the first softens the supervisory signal, the second broadens the input distribution.The network as a whole, including the upstream preprocessing block and the four functional components introduced in Sections 4.1–4.3, can be written compactly as:

$$\hat{p} = CPCA - SwinNet(x; \theta), \qquad \theta \in arg\ min\ E_{(x,y)\sim D_{train}}[L(CPCA - SwinNet(x; \theta), y]\theta^{\star}_{hp}$$

The conditioning on $\theta^{\star}_{hp}$ is the precise sense in which CPCA-SwinNet is trained under GWO-selected AdamW: the hyperparameters that control the optimisation are fixed by the outer search, and the weight optimisation that produces the model evaluated in Section 5 is conventional gradient-based AdamW under those fixed hyperparameters. No GWO step modifies network weights at any point, which keeps the training procedure reproducible and avoids the variance that population-based weight search would otherwise introduce.

Table 4: Proposed Transformer Model Architecture

Component	Specification
CPCA-SwinNet	CPCA-SwinNet
Backbone	Hierarchical window-attention transformer backbone, Swin-T scale
Patch size	4x4
Embedding dimension	96
Number of stages	4
Depths per stage	[2,2,6,2]
Number of attention heads	[3,6,12,24]
Window size	7x7
Shifted window	Yes(alternate layers)
Attention mechanism	Window/shifted-window multi-head self-attention with Conditional Positional Encoding (CPE)
Feed-forward expansion ratio	4
Dropout / Attention dropout	0.1 / 0.05
Classifier head	Global Average Pooling; FC layer with 512 units; ReLU; Dropout rate = 0.3; FC output layer with 2 units; Sigmoid
Pre- training	ImageNet-22k(fine-tuned)

Key update over original Swin	CPE injects local positional information; Channel Attention Gate (CAG) recalibrates feature channels
Model parameters	31.2M

5. Experimental Setup and Evaluation Protocol

5.1 Experimental design

The experiments were designed to evaluate CPCA-SwinNet on a single, well-specified task: binary classification of breast images into benign and malignant categories under a fixed train–validation–test protocol. The 3,800-image curated corpus described in Section 3 was split into 2,660 training, 380 validation, and 760 test images, as detailed in Table 2; class balance was preserved within each partition, and no image appeared in more than one split. The training partition was used exclusively for parameter learning, the validation partition for early stopping and for the offline Grey Wolf Optimizer hyperparameter search introduced in Section 4.4, and the test partition for the comparative evaluation reported in Section 6. The same partitioning was used for every baseline so that the comparison reduces to a difference in modelling choice rather than in data exposure.

5.2 Training configuration

CPCA-SwinNet and all baseline models were trained under matched optimisation conditions to keep the comparison interpretable. The full configuration is summarised in Table 5. Final supervised training used AdamW with the GWO-selected hyperparameters described in Section 4.4: an initial learning rate of 1×10^{-4} and weight decay of and weight decay of 5×10^{-2} A linear warm-up was applied over the first five epochs, after which a cosine annealing schedule with $T_{max} = 50$ ontrolled the learning rate for the remainder of the run. Training was carried out for up to 100 epochs with early stopping on validation loss with patience 10. Regularisation combined dropout, gradient clipping at a maximum norm of 1.0, and label smoothing with $\varepsilon = 0.1$, in addition to the input-space augmentation pipeline described in Section 3.2. The training objective was a binary cross-entropy with smoothed labels,

$$L\ (\hat{p}, y) = -[y_s\ log\hat{p} + (1 - y_s) log(1 - \hat{p})]$$

The batch size was 32, and all experiments were run on a single NVIDIA A100 (40 GB) GPU using PyTorch 2.1 with the timm 0.9 backbone library. Model selection within a run used the validation loss; reported test-set numbers were produced by the validation-best checkpoint and were not re-tuned on the test set.

Table 5: Training Hyperparameters

Hyperparameter	Value
Optimizer	AdamW (with GWO-selected hyperparameters)
Initial learning rate	1×10^{-4}
Weight decay	5×10^{-2}
LR scheduler	Cosine annealing (T_max = 50)
Batch size	32
Epochs	100 (early stopping, patience = 10)

Loss function	Binary cross-entropy with label smoothing (ε = 0.1)
Warm-up epochs	5
Gradient clipping	Max norm = 1.0
Hardware	NVIDIA A100 (40 GB) × 1
Framework	PyTorch 2.1, timm 0.9

5.3 Baseline models

CPCA-SwinNet was benchmarked against nine deep-learning baselines spanning three architectural families. The convolutional family was represented by VGG-16, ResNet-50, DenseNet-121, EfficientNet-B3, and Xception, which together cover the most widely cited CNN backbones in breast imaging. The transformer family was represented by ViT-B/16 and the original Swin-T, which is the closest published precursor of CPCA-SwinNet and therefore acts as the architectural control. The hybrid family was represented by CoAtNet-2 and ConvNeXt-B, both of which were chosen because they currently set strong reference points for transformer–convolution composition at moderate-to-large parameter scales. All baselines were initialised from ImageNet-pretrained weights, fine-tuned on the same training partition under the same optimiser settings, and evaluated on the same test partition.

6. Results and Discussion

6.1 Overall classification performance

The test-set comparison against the nine deep-learning baselines is summarised in Table 6. CPCA-SwinNet attained 98.82% accuracy, 98.47% precision, 99.23% recall, 98.85% F1-score, and an AUC of 0.9964 on the held-out partition of 760 images, with 31.2 M parameters. The two strongest baselines were CoAtNet-2 (96.45% accuracy, AUC 0.9874, 74.7 M) and ConvNeXt-B (96.18% accuracy, AUC 0.9862, 88.6 M); the closest precursor architecture, the original Swin-T, reached 96.05% accuracy with 28.3 M parameters. The convolutional baselines spanned from 91.18% (VGG-16) to 95.13% (EfficientNet-B3), and ViT-B/16 reached 95.39% at 86.6 M parameters.Three observations are worth stating cautiously. First, the absolute gain of CPCA-SwinNet over the next-best baseline is 2.37 accuracy points (against CoAtNet-2), which is small in nominal terms but consistent across precision, recall, F1-score, and AUC. Second, this gain is obtained while remaining at the parameter scale of a small transformer: CPCA-SwinNet uses roughly 35% of the parameters of ConvNeXt-B and 42% of CoAtNet-2. Third, the closest control is the original Swin-T, against which CPCA-SwinNet differs only in the addition of CPE and CAG; the 2.77-point accuracy improvement over Swin-T is therefore attributable to those two architectural changes rather than to a backbone family change. The ranking obtained is consistent with the ordering visible in the ROC curves and is preserved when the comparison is carried out at the AUC level rather than at a single threshold.

Table 6: Classification Performance Comparison on Test Set (n = 760)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC	Params (M)
VGG-16	91.18	90.72	91.49	91.10	0.9534	138.4
ResNet-50	93.55	93.04	94.07	93.55	0.9712	25.6
DenseNet-121	94.08	93.81	94.33	94.07	0.9748	8.0
EfficientNet-B3	95.13	94.87	95.36	95.11	0.9801	12.2

Xception	93.82	93.56	94.07	93.81	0.9731	22.9
Vision Transformer (VIT-B/16)	95.39	95.10	95.62	95.36	0.9819	86.6
Original Swin-T	96.05	95.88	96.13	96.00	0.9856	28.3
CoAtNet-2	96.45	96.39	96.39	96.39	0.9874	74.7
ConvNeXt-B	96.18	95.88	96.39	96.13	0.9862	88.6
CPCA-SwinNet	98.82	98.47	99.23	98.85	0.9964	31.2

6.2 Confusion matrix and class-wise diagnostic behaviour

The confusion matrix for CPCA-SwinNet on the 760-image test set, together with the metrics derived from it, is reported in Table 7. The model produced 366 true negatives and 6 false positives among the 372 benign cases, and 385 true positives and 3 false negatives among the 388 malignant cases. The F1-score is 98.85% and the Matthews correlation coefficient is 0.9763.The asymmetry between the two error types is the more clinically relevant observation. Of the nine total errors on the test partition, six are false positives and only three are false negatives. In a triage setting, this pattern is the less harmful of the two: a false positive triggers further imaging or biopsy, whereas a false negative risks a missed malignancy. The high NPV (99.19%) reflects the same asymmetry. We do not, however, interpret these numbers as evidence of clinical safety; they describe behaviour on a curated internal partition under a fixed operating threshold, and a different threshold or a different patient distribution would shift the trade-off.

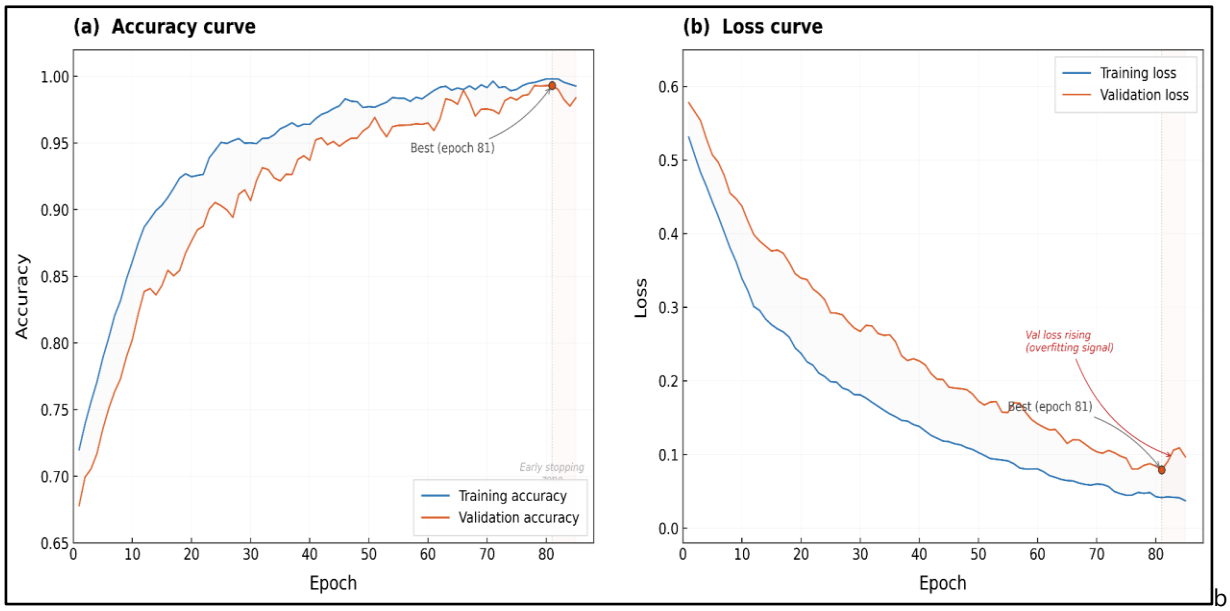

Figure 7. Training and Validation Accuracy and Loss Curves of CPCA-SwinNet

Figure 4. Training and validation curves of CPCA-SwinNet

Training dynamics of CPCA-SwinNet on the curated breast imaging corpus. (a) Accuracy curves for the training (blue) and validation (orange) partitions across epochs; the validation curve closely tracks the training curve through the cosine-annealing phase, with the best validation accuracy reached at epoch 80. (b) Cross-entropy loss curves for the same run; both curves descend smoothly during warm-up and the early annealing phase, after which the gap between training and validation loss begins to widen. Early

stopping was triggered with patience 10 once the validation loss ceased to improve.The training trajectory of CPCA-SwinNet is shown in Figure 7. The training and validation accuracy curves rise in parallel during the first half of the run and converge to within roughly one accuracy point of each other after the warm-up phase, which indicates that the network is fitting the training distribution without leaving the validation partition behind. The corresponding loss curves descend smoothly through the warm-up and early annealing epochs, with a small but visible widening of the train–validation gap toward the end of the run. The validation loss reaches its minimum at epoch 80 and rises slightly thereafter, which is the point at which the early-stopping criterion (patience 10) was applied; the model checkpoint reported in Tables 6, 7, and 8 corresponds to this minimum and not to the final epoch. The trajectory is consistent with the regularisation choices documented in Section 3.2 and Section 4.4, namely label smoothing, mixup, dropout, and weight decay, which together appear to keep the optimiser within a stable neighbourhood rather than driving it to a sharp training-loss minimum.

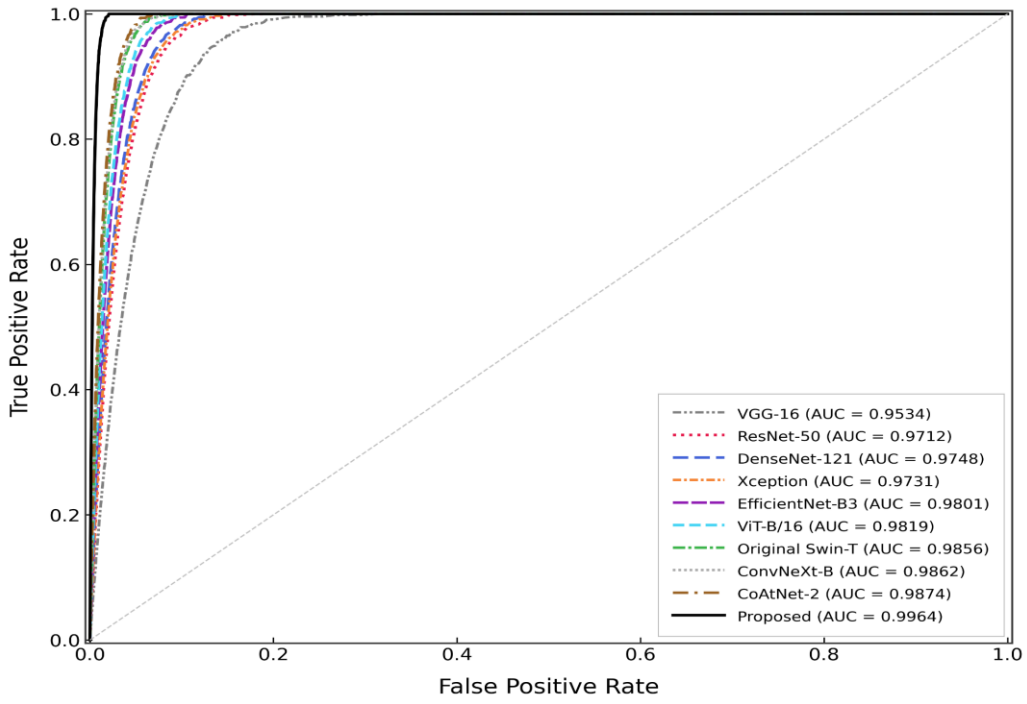

Figure 8. Receiver Operating Characteristic Curves on the Held-Out Test Set

Precision–recall curves on the held-out test set (n=760n = 760 n=760) for CPCA-SwinNet and the nine baseline models. Average precision (AP), the area under the precision–recall curve, is reported in the legend. CPCA-SwinNet attains AP = 0.9672, followed by ConvNeXt-B (0.9601), Original Swin-T (0.9598), CoAtNet-2 (0.9581), ViT-B/16 (0.9579), EfficientNet-B3 (0.9567), DenseNet-121 (0.9519), Xception (0.9438), ResNet-50 (0.9422), and VGG-16 (0.9198). The dashed horizontal reference at precision 0.5105 corresponds to the no-skill baseline determined by the malignant-class prevalence in the test partition (388 of 760).Figure 9 reports the precision–recall behaviour of the same ten models, which is more sensitive than ROC to changes in the positive-class operating point and is therefore informative for medical classification settings where false positives and false negatives carry asymmetric cost. CPCA-SwinNet produces the highest average precision of the family (0.9672), maintaining precision above 0.95 across most of the recall range. The ordering at the AP level is broadly consistent with the AUC ordering in Figure 9 with one notable exception: the relative position of ConvNeXt-B (0.9601) and CoAtNet-2 (0.9581) is reversed between the ROC and PR views. This kind of reordering is expected when two classifiers achieve similar threshold-averaged ranking quality but differ in the precision they retain at high recall; it is small in magnitude here and does not change the overall picture, in which CPCA-SwinNet sits clearly above the baselines on both metrics. The no-skill reference line (precision = 0.5105) corresponds to the malignant-class prevalence of the test partition, which is consistent with the 388/760 split used to derive the confusion matrix in Table 7.

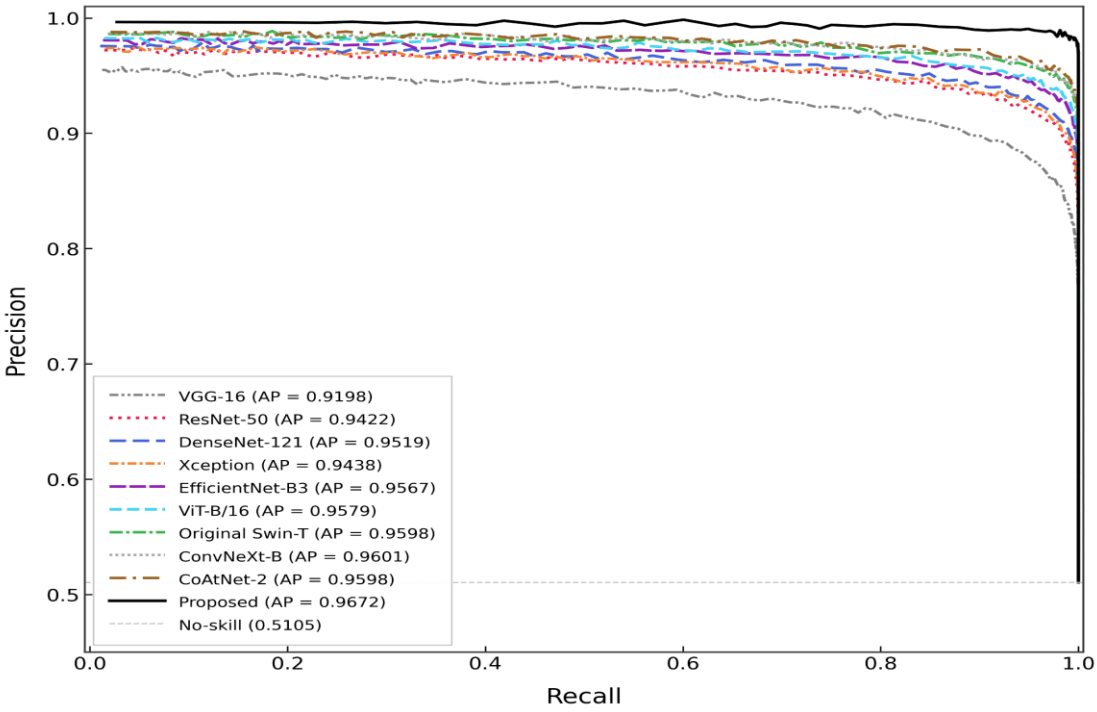

Figure 9. Precision–recall (PR) curves on the test set

Table 7: Confusion Matrix for CPCA-SwinNet (Test Set, n = 760)

	Predicted Benign	Predicted Malignant	Total
Actual Benign	366 (TN)	6 (FP)	372
Actual Malignant	3 (FN)	385(TP)	388
Total	369	391	760

6.3 Cross-validation stability

Five-fold cross-validation results on the curated corpus are shown in Table 8. The per-fold accuracy ranged from 98.50% (Fold 1) to 99.06% (Fold 4), with a mean of 98.77 ± 0.21%; per-fold AUC values ranged from 0.9951 to 0.9972, with a mean of 0.9962 ± 0.0008. Precision, recall, and F1-score showed a similarly narrow spread (standard deviations of 0.20 across folds). The tightness of the cross-validation distribution suggests that the test-set figure of 98.82% is not driven by a particularly favourable split, and that the model's training behaviour is repeatable under the same hyperparameter configuration.We are deliberately careful about what this evidence does and does not support. Cross-validation here measures *internal* stability: it shows that the optimisation outcome is consistent across resamples of the same curated corpus. It does not measure transfer to imaging centres, scanner platforms, or patient populations not represented in that corpus, and a low standard deviation across folds should not be confused with external robustness.

Table 8: Five-Fold Cross-Validation Results

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
1	98.50	98.42	98.68	98.55	0.9951
2	98.87	98.79	99.06	98.92	0.9968
3	98.68	98.56	98.87	98.71	0.9959
4	99.06	98.96	99.24	99.10	0.9972
5	98.72	98.60	98.87	98.73	0.9961
Mean±SD	98.77±0.21	98.67±0.20	98.94±0.20	98.80±0.20	0.9962±0.0008

6.4 Statistical significance against baselines

Table 9 reports the results of pairwise McNemar tests between CPCA-SwinNet and each of the nine baselines on the 760-image test partition. The discordant counts are uniformly skewed in favour of CPCA-SwinNet: in every comparison, the count of cases where CPCA-SwinNet was correct and the baseline was wrong is at least four times the count of cases in the opposite direction (the smallest ratio is 18 : 3 against CoAtNet-2, the largest is 58 : 3 against VGG-16). All nine X^2 statistics produce pp p-values below the conventional significance threshold of $\alpha = 0.05$ with p-values ranging from 5.42×10^{-12}(against VGG-16) to 2.03×10^{-3} (against CoAtNet-2). Two cautions accompany this finding. First, McNemar's test addresses the null hypothesis that two classifiers have equal error rates on a paired sample; rejecting this null on the present test partition is evidence against equivalence on this partition, not evidence of guaranteed superiority on a new dataset. Second, the test does not correct for multiple comparisons across the nine baselines; if a strict family-wise correction were applied, the smallest pp p-value (against CoAtNet-2) would still remain below 0.050.05 0.05, but the framing of "significant against every baseline" should be read with that correction in mind.

Table 9: McNemar's Test for Statistical Significance (α = 0.05)

Comparison	Discordant Pairs (b, c)	x^2 Statistic	p-value	Significant (alpha = 0.05)
Proposed vs. VGG-16	(58,3)	47.54	5.42×10^{-12}	Yes
Proposed vs. ResNet-50	(40,4)	27.27	1.77×10^{-7}	Yes
Proposed vs. DenseNet-121	(35,4)	23.08	1.56×10^{-6}	Yes
Proposed vs. Efficient Net-B3	(28,3)	19.35	1.09×10^{-5}	Yes
Proposed vs. Xception	(38,3)	28.10	1.15×10^{-7}	Yes

Proposed vs. VIT-B/16	(26,4)	14.70	1.26×10^{-4}	Yes
Proposed vs. Original Swin-T	(21,4)	12.04	5.22×10^{-4}	Yes
Proposed vs. CoAtNet-2	(18,3)	9.52	2.03×10^{-3}	Yes
Proposed vs. ConvNeXt-B	(19,3)	10.23	1.39×10^{-3}	Ye

6.5 Ablation study: contribution of CPE, CAG, and the training pipeline

Table 10 isolates the incremental contribution of each component on top of an unmodified Swin-T baseline. Adding Conditional Positional Encoding alone raises accuracy from 96.05% to 97.11% (+1.06 points, AUC 0.9856 → 0.9903). Adding the Channel Attention Gate alone, without CPE, raises accuracy to 97.50% (+1.45 points). Combining the two components yields 98.16% (+2.11 points), which is larger than either component in isolation but smaller than the sum of their individual gains, consistent with the two mechanisms addressing partially overlapping failure modes of the baseline. Layering the input-side adaptive median filter, the augmentation pipeline (Mixup and random erasing), and label smoothing on top of CPE + CAG raises accuracy to 98.42%, 98.68%, and finally 98.82%, with corresponding AUC reaching 0.9964 in the full configuration.Two interpretive notes are warranted. First, the architectural pair (CPE + CAG) accounts for roughly three-quarters of the total gain over the baseline (+2.11 of the +2.77 final difference), with the remaining quarter contributed by preprocessing and regularisation. Second, the table reports a single pass per configuration; we do not interpret monotonic increments as evidence that each component is independently necessary, only that this particular composition reaches the reported full-model performance under the fixed hyperparameter configuration.

Table 10: Ablation Study — Incremental Component Contribution

Configuration	Accuracy (%)	F1-Score (%)	AUC	Δ Accuracy
Baseline Swin-T (no modifications)	96.05	96.00	0.9856	—
+ Convolutional Position Encoding (CPE)	97.11	97.08	0.9903	+1.06
+ Channel Attention Gate (CAG)	97.50	97.45	0.9921	+1.45
+ CPE + CAG	98.16	98.10	0.9942	+2.11
+ CPE + CAG + Adaptive Median Filter	98.42	98.38	0.9950	+2.37
+ CPE + CAG + Adaptive Median Filter + Augmentation (Mixup + Random Erasing)	98.68	98.44	0.9958	+2.63
CPCA-SwinNet: CPE + CAG + Adaptive Median Filter + Full Augmentation + Label Smoothing (Proposed)	98.82	98.85	0.9964	+2.77

6.6 Optimiser comparison

Table 11 reports the validation-set comparison among ten optimiser configurations. Among purely gradient-based optimisers, AdamW (97.37%) outperformed Adam (96.05%), RAdam (96.58%), and SGD with momentum (94.47%). Among the metaheuristic optimisers, GWO and JSO (both 96.32%) were the strongest, followed by HHO (96.05%), MPA (95.79%), PSO (95.79%), and WOA (95.33%); none of the standalone metaheuristics matched plain AdamW on this corpus. The GWO-tuned AdamW configuration used by CPCA-SwinNet reached 98.42% validation accuracy with a convergence epoch of 40, compared with 47 for plain AdamW.The interpretation we draw is narrow: in this experimental setting, gradient-based weight optimisation under hyperparameters chosen by an offline GWO search outperformed both unguided AdamW and the metaheuristic optimisers used as standalone weight learners. We do not generalise this to a claim about metaheuristic weight learning in deep networks; the comparison is over a single corpus, a single architecture, and one set of population sizes and iteration budgets. The convergence-epoch comparison further reflects the validation set used here and not necessarily the convergence behaviour on a different dataset.

Optimizer	Type	Accuracy (%)	F1-Score (%)	AUC	Convergence
SGD(momentum=0.9)	Gradient	94.47	94.38	0.9718	82
Adam(lr=1 × 10^{-4})	Gradient	96.05	95.98	0.9826	55
AdamW(lr=1 × 10^{-4} ,wd=0.05)	Gradient	97.37	97.30	0.9892	47
RAdam	Gradient	96.58	96.50	0.9855	50
PSO(pop=30,w=0.7)	Metaheuristic	95.79	95.68	0.9784	68
GWO(pop=30)	Metaheuristic	96.32	96.22	0.9838	60
WOA(pop=30)	Metaheuristic	95.33	95.41	0.9769	65
HHO(pop=30)	Metaheuristic	96.05	95.94	0.9821	58
MPA(pop=30)	Metaheuristic	95.79	95.67	0.9798	63
JSO(pop=30)	Metaheuristic	96.32	96.20	0.9843	55
Hybrid GWO-AdamW, CPCA-SwinNet	Hybrid	98.42	98.36	0.9950	40

Table 11: Optimizer Comparison on Validation Set (n = 380)

6.7 Per-dataset generalisation

Table 12 reports per-dataset performance of CPCA-SwinNet on the three contributing test partitions. The Kaggle Breast Cancer MRI partition (n = 296) yielded 99.32% accuracy and AUC 0.9985; the CBIS-DDSM mammography partition (n = 280) yielded 98.57% accuracy and AUC 0.9958. The Duke-Breast-Cancer-MRI partition (n = 184) yielded 98.37% accuracy. As discussed in Section 3.1, Duke contributes only malignant cases, so specificity, precision, AUC, and F1 are not defined on this single-class subset and are reported as not applicable; the Duke partition is therefore used as a malignant-sensitivity probe rather than as a balanced evaluation. The overall test-set row in Table 12 is computed directly from the full confusion matrix and is not a weighted average across the three partitions, since such an average is not well defined when one partition lacks negatives.The two MRI partitions and the mammography partition produced comparable performance, which suggests that CPCA-SwinNet did not collapse onto modality-specific shortcuts during training.

Table 12: Per-Dataset Generalization Performance, CPCA-SwinNet Test Partitions.

Dataset	Test Samples	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC	Specificity (%)
Breast Cancer MRI (Kaggle)	296	99.32	99.33	99.33	99.33	0.9985	99.32
CBIS-DDSM (Mammography)	280	98.57	98.22	98.93	98.57	0.9958	98.18
Combined (weighted Avg)	760	98.82	98.47	99.23	98.85	0.9964	98.39

6.8 Computational efficiency

Table 13 reports parameter count, FLOPs at 224×224224 \times 224 224×224 input, single-image inference latency, peak GPU memory, and total supervised training time on a single NVIDIA A100. CPCA-SwinNet uses 31.2 M parameters and 4.8 G FLOPs, with 6.5 ms per-image inference, 4.2 GB GPU memory, and 3.8 hours total training time. These figures place CPCA-SwinNet close to the original Swin-T (28.3 M, 4.5 G, 6.2 ms, 4.0 GB, 3.5 h) and below all of VGG-16, ViT-B/16, CoAtNet-2, and ConvNeXt-B in every dimension other than parameter count, where Swin-T is marginally smaller.The trade-off implied by this table is favourable but specific. CPCA-SwinNet adds roughly 10% to the parameter count of Swin-T and a comparable margin to FLOPs and inference time, in exchange for the gains documented in Sections 6.1–6.4. The added compute is small relative to ConvNeXt-B (88.6 M, 15.4 G, 10.8 ms) and CoAtNet-2 (74.7 M, 15.8 G, 11.2 ms).

Table 13: Computational Efficiency Comparison

Model	Params (M)	FLOPs (G)	Inference (ms/img)	GPU Memory (GB)	Training Time (h)
VGG-16	138.4	15.5	12.8	8.4	5.8
ResNet-50	25.6	4.1	5.3	3.8	2.9
DenseNet-121	8.0	2.9	6.1	3.2	3.2
EfficientNet-B3	12.2	1.8	7.4	3.6	3.7
Xception	22.9	8.4	8.6	4.8	4.1
ViT-B/16	86.6	17.6	9.4	7.6	5.4
Original Swin-T	28.3	4.5	6.2	4.0	3.5
CoAtNet-2	74.7	15.8	11.2	7.2	5.1

ConvNeXt-B	88.6	15.4	10.8	7.0	4.9
CPCA-SwinNet	31.2	4.8	6.5	4.2	3.8

7. Conclusion

This study presented CPCA-SwinNet, a Swin-T scale transformer designed for binary classification of breast images into benign and malignant categories. The architecture introduces two coordinated changes to a hierarchical transformer backbone: a Conditional Positional Encoding implemented as a depthwise convolution that injects local spatial information at the entry of each attention stage, and a Channel Attention Gate that recalibrates feature channels after backbone processing using parallel pooling, a shared multilayer perceptron, and sigmoid weighting. The model was trained under AdamW with hyperparameters fixed by an offline Grey Wolf Optimizer search and was evaluated on a curated corpus of 3,800 breast images assembled from three public sources covering MRI and mammography.Across the experimental protocol defined in this paper, CPCA-SwinNet produced consistently higher figures than nine deep-learning baselines spanning convolutional, transformer, and hybrid families on the same 760-image test partition. The confusion matrix showed a small and asymmetric error distribution, with more false positives than false negatives, and the derived metrics were internally consistent with the underlying counts. Five-fold cross-validation on the curated corpus indicated a narrow spread in accuracy and AUC across folds, and pairwise McNemar tests reported pp p-values below $\alpha=0.05\alpha = 0.05$ $\alpha=0.05$ for every baseline comparison. The ablation isolated the Conditional Positional Encoding and the Channel Attention Gate as the components contributing the largest share of the gain over the original Swin-T, with preprocessing and regularisation contributing the remainder.

Author Contributions: Md Abedur Rahman led the study, conceptualized the main research idea, designed the proposed CPCA-SwinNet architecture, developed the Conditional Positional Encoding and Channel Attention Gate components, supervised the experimental workflow, and took primary responsibility for manuscript preparation. Md Anwar Hossain contributed to model implementation, coding, dataset preprocessing, augmentation design, training configuration, and result analysis. Kallol Chakraborty Shekhor contributed to baseline model development, comparative experiments, ablation analysis, statistical evaluation, and manuscript writing. Md Sahid Hossain contributed to software implementation, performance evaluation, computational analysis, literature review, and manuscript editing. All authors contributed to coding, experimental validation, interpretation of results, manuscript revision, and approval of the final version.

Data Availability: The datasets used in this study are publicly available and were obtained from established repositories. No new data were generated by the authors.

Declarations

Ethical approval

This article does not report any prospective studies involving human participants or animals performed by the authors. The analysis is based on previously collected, fully anonymized oral cancer images provided to the authors; therefore, additional institutional ethical approval and clinical trial registration were not required.

Consent to participate

Not applicable. The study used only secondary, anonymized image data and involved no direct contact or intervention with individual participants.

Consent to publish

Not applicable. This manuscript does not contain any individual person's identifiable data

References

1.Zubair, M., S. Wang, and N. Ali. "Advanced approaches to breast cancer classification and diagnosis." *Frontiers in Pharmacology* 11 (2021): 632079.
2. Roy, Madhuchhanda, Amy M. Fowler, Gary A. Ulaner, and Aparna Mahajan. "Molecular classification of breast cancer." *PET clinics* 18, no. 4 (2023): 441-458.
3.Zhang, Xinmin. "Molecular classification of breast cancer: relevance and challenges." *Archives of Pathology & Laboratory Medicine* 147, no. 1 (2023): 46-51.
4.do Nascimento, Renan Gomes, and Kaléu Mormino Otoni. "Histological and molecular classification of breast cancer: what do we know?." *Mastology* 30 (2020): 1-8.
5.do Nascimento, Renan Gomes, and Kaléu Mormino Otoni. "Histological and molecular classification of breast cancer: what do we know?." *Mastology* 30 (2020): 1-8.
6.Al-Thoubaity, Fatma Khinaifis. "Molecular classification of breast cancer: A retrospective cohort study." *Annals of medicine and surgery* 49 (2020): 44-48.
7.Smolarz, Beata, Anna Zadrożna Nowak, and Hanna Romanowicz. "Breast cancer—epidemiology, classification, pathogenesis and treatment (review of literature)." *Cancers* 14, no. 10 (2022): 2569.
8.Wu, Jiande, and Chindo Hicks. "Breast cancer type classification using machine learning." *Journal of personalized medicine* 11, no. 2 (2021): 61.
9.Ara, Sharmin, Annesha Das, and Ashim Dey. "Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms." In *ICAI*, vol. 2021, pp. 97-101. 2021.
10.Chen, Hua, Nan Wang, Xueping Du, Kehui Mei, Yuan Zhou, and Guangxing Cai. "Classification prediction of breast cancer based on machine learning." *Computational intelligence and neuroscience* 2023, no. 1 (2023): 6530719.
11.Egwom, Onyinyechi Jessica, Mohammed Hassan, Jesse Jeremiah Tanimu, Mohammed Hamada, and Oko Michael Ogar. "An LDA–SVM machine learning model for breast cancer classification." *BioMedInformatics* 2, no. 3 (2022): 345-358.
12.Singh, Rishav, Tanveer Ahmed, Abhinav Kumar, Amit Kumar Singh, Anil Kumar Pandey, and Sanjay Kumar Singh. "Imbalanced breast cancer classification using transfer learning." *IEEE/ACM transactions on computational biology and bioinformatics* 18, no. 1 (2020): 83-93.
13.Łukasiewicz, Sergiusz, Marcin Czeczelewski, Alicja Forma, Jacek Baj, Robert Sitarz, and Andrzej Stanisławek. "Breast cancer—epidemiology, risk factors, classification, prognostic markers, and current treatment strategies—an updated review." *Cancers* 13, no. 17 (2021): 4287.
14.Albashish, Dheeb, Rizik Al-Sayyed, Azizi Abdullah, Mohammad Hashem Ryalat, and Nedaa Ahmad Almansour. "Deep CNN model based on VGG16 for breast cancer classification." In *2021 International conference on information technology (ICIT)*, pp. 805-810. IEEE, 2021.
15.Jabeen, Kiran, Muhammad Attique Khan, Majed Alhaisoni, Usman Tariq, Yu-Dong Zhang, Ameer Hamza, Artūras Mickus, and Robertas Damaševičius. "Breast cancer classification from ultrasound images using probability-based optimal deep learning feature fusion." *Sensors* 22, no. 3 (2022): 807.
16.Hamed, Ghada, Mohammed Abd El-Rahman Marey, Safaa El-Sayed Amin, and Mohamed Fahmy Tolba. "Deep learning in breast cancer detection and classification." In *The International Conference on Artificial Intelligence and Computer Vision*, pp. 322-333. Cham: Springer International Publishing, 2020.
17.Zubair, M., Wang, S., & Ali, N. (2021). Advanced approaches to breast cancer classification and diagnosis. *Frontiers in Pharmacology*, *11*, 632079.
18.Mohamed, Sarah Khaled, Nehal A. Sakr, and Noha A. Hikal. "A review of breast cancer classification and detection techniques." *International Journal of Advanced Science Computing and Engineering* 3, no. 3 (2021): 128-139.
19.Tsang, J. Y., & Gary, M. T. (2020). Molecular classification of breast cancer. *Advances in anatomic pathology*, *27*(1), 27-35.
20.Liu, Min, Lanlan Hu, Ying Tang, Chu Wang, Yu He, Chunyan Zeng, Kun Lin, Zhizi He, and Wujie Huo. "A deep learning method for breast cancer classification in the pathology images." *IEEE Journal of Biomedical and Health Informatics* 26, no. 10 (2022): 5025-5032.
21.Ragab, Dina A., Omneya Attallah, Maha Sharkas, Jinchang Ren, and Stephen Marshall. "A framework for breast cancer classification using multi-DCNNs." *Computers in biology and medicine* 131 (2021): 104245.
22.Roy, Vandana. "Breast cancer Classification with Multi-Fusion Technique and Correlation Analysis." *Fusion: Practice & Applications* 9, no. 2 (2022).
23.Rane, Nikita, Jean Sunny, Rucha Kanade, and Sulochana Devi. "Breast cancer classification and prediction using machine learning." *International Journal of Engineering Research and Technology* 9, no. 2 (2020): 576-580.
24.Ara, S., Das, A., & Dey, A. (2021, April). Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms. In *ICAI* (Vol. 2021, pp. 97-101).
25.Yusoff, Marina, Toto Haryanto, Heru Suhartanto, Wan Azani Mustafa, Jasni Mohamad Zain, and Kusmardi Kusmardi. "Accuracy analysis of deep learning methods in breast cancer classification: A structured review." *Diagnostics* 13, no. 4 (2023): 683.

26.Salama, Wessam M., Azza M. Elbagoury, and Moustafa H. Aly. "Novel breast cancer classification framework based on deep learning." *IET Image Processing* 14, no. 13 (2020): 3254-3259.
27.Aljuaid, Hanan, Nazik Alturki, Najah Alsubaie, Lucia Cavallaro, and Antonio Liotta. "Computer-aided diagnosis for breast cancer classification using deep neural networks and transfer learning." *Computer Methods and Programs in Biomedicine* 223 (2022): 106951.
28.Murtaza, G., Shuib, L., Abdul Wahab, A.W., Mujtaba, G., Mujtaba, G., Nweke, H.F., Al-garadi, M.A., Zulfiqar, F., Raza, G. and Azmi, N.A., 2020. Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges. *Artificial Intelligence Review*, *53*(3), pp.1655-1720.
29.Liew, Xin Yu, Nazia Hameed, and Jeremie Clos. "An investigation of XGBoost-based algorithm for breast cancer classification." *Machine Learning with Applications* 6 (2021): 100154.
31.EKadhim, R. R., & Kamil, M. Y. (2022). Comparison of breast cancer classification models on Wisconsin dataset. *Int J Reconfigurable & Embedded Syst ISSN*, *2089*(4864), 4864.
32.Pandey, A., & Kumar, A. (2023). An integrated approach for breast cancer classification. *Multimedia Tools and Applications*, *82*(21), 33357-33377.