RESEARCH PAPER

Next-Generation Radiogenomics: Advanced Diagnostics for Early Breast Cancer Detection and Personalized Risk Stratification

Chennapragada Padmaja[1*], S. Sivaneasan Bala Krishnan[2], S. Ramacharan[3], Prasun Chakrabarti[4]

[1*]Associate Professor, dept. of Electronics and Communication Engineering, G. Narayanamma Institute of Technology and Science (for Women), Hyderabad, India. Email: c.padmaja@gnits.ac.in
ORCID ID: **0000-0003-0521-916X**

[2]Associate Professor, dept. of Electrical and Electronics Engineering, Singapore Institute of Technology, Singapore. Email: sivaneasan@singaporetech.edu.sg
ORCID ID: **0000-0002-4271-2677**

[3]HOD and Professor, dept. of Information Technology, G. Narayanamma Institute of Technology and Science (for Women), Hyderabad, India. Email: s.ramacharan@gnits.ac.in
ORCID ID: **0000-0003-1058-2211**

[4]Director, Senior Professor, dept. of Computer Science and Engineering, Sir Padampat Singhania University, Udaipur 313601, Rajasthan, India. Email: drprasun.cse@gmail.com
ORCID ID: **0000-0001-8062-4144**

ABSTRACT

Breast cancer is the leading malignant disease in women globally, with 2.3 million new diagnoses and 685,000 deaths recorded annually. Clinical deployment of AI-based diagnostic tools has been impeded by model opacity, miscalibrated confidence, and unequal performance across patient subgroups. Methods: We propose TEDIN (Trustworthy Explainable Deep Intelligence Network), a multi-module framework that couples a ResNet-50 transfer learning backbone with Gradient-weighted Class Activation Mapping (Grad-CAM) for spatial explainability, Monte Carlo (MC) Dropout for epistemic uncertainty quantification, and equalized-odds post-processing for demographic fairness correction. TEDIN is evaluated on the CBIS-DDSM mammography dataset (2568 images) and BreakHis histopathology dataset (7909 images). Results: TEDIN achieves 97.4% accuracy, 96.8% sensitivity, 97.9% specificity, and AUC = 0.989, outperforming four state-of-the-art baselines. The MC-Dropout mechanism flags 88.2% of misclassified cases as high-uncertainty. Equalized-odds correction reduces cross-demographic false-negative rate disparities by up to 74.5% at less than 0.4 pp accuracy cost. A usability study with 12 radiologists confirms a 23% reduction in diagnostic review time and statistically significant gains in clinician confidence ($p < 0.01$). Conclusions: TEDIN demonstrates that the reliability, interpretability, calibration, and fairness principles central to Trustworthy and Intelligent Systems for Predictive Maintenance transfer directly and productively into clinical breast cancer diagnosis.

Keywords: Explainable AI, Grad-CAM, Convolutional Neural Networks, TEDIN, Predictive maintenance

How to cite this article: Padmaja C, Bala Krishnan SS, Ramacharan S, Chakrabarti P. Next-Generation Radiogenomics: Advanced Diagnostics for Early Breast Cancer Detection and Personalized Risk Stratification. Int J Drug Deliv Technol. 2026;16(12s): 546-556. DOI: 10.25258/ijddt.16.12s.66

INTRODUCTION

Breast cancer constitutes the most frequently diagnosed malignant disease in women worldwide, accounting for approximately 12.5% of all new cancer cases annually. According to GLOBOCAN 2022 data, 2.3 million new cases and 685,000 deaths were recorded globally, with incidence projected to reach 3.2 million by 2050 [1]. The clinical outcome is profoundly stage-dependent: patients diagnosed at Stage I exhibit five-year survival rates exceeding 99%, whereas Stage IV detection yields rates below 29% [2]. This differential underscores the life-saving imperative of accurate and timely diagnostic tools.

Contemporary breast screening programmes rely primarily on X-ray mammography, supplemented by ultrasound, MRI, and biopsy-guided histopathological analysis. Despite widespread deployment, diagnostic performance remains constrained by inter-radiologist variability of 3–10% and false-positive recall rates of 7–12% [3]. Deep learning-based computer-aided detection and diagnosis (CAD) systems have emerged as promising complements to radiological expertise, achieving benchmark accuracy on par with specialist physicians [4,5]. However, three fundamental barriers consistently impede clinical translation: (1) opacity—high-performing networks operate as statistical black boxes, eroding clinician trust [6]; (2) overconfidence—standard softmax outputs are systematically miscalibrated, creating patient safety risks [7]; and (3) demographic inequity—models trained on non-representative data exhibit performance gaps across age, race, and breast density subgroups [8].

This paper situates the challenge of trustworthy clinical AI within the broader framework of Trustworthy and Intelligent Systems for Machine Health Monitoring and Predictive Maintenance—the thematic scope of this Special Issue. The structural analogies between industrial condition monitoring and clinical diagnostics are deep and instructive. A trustworthy predictive maintenance system must: (i) detect anomalous signatures with high sensitivity; (ii) provide interpretable fault indicators; (iii) quantify uncertainty with calibrated confidence bounds; and (iv) perform consistently across heterogeneous machine types. TEDIN operationalises each requirement in the clinical domain through high-sensitivity lesion classification, Grad-CAM spatial explainability, MC-Dropout uncertainty quantification, and equalized-odds fairness correction.

The principal contributions of this work are:

• A transfer-learned ResNet-50 backbone trained with CLAHE preprocessing and class-balanced augmentation on two multi-modal breast cancer benchmarks (CBIS-DDSM and BreakHis);

• Systematic Grad-CAM integration generating radiologist-interpretable 224 × 224 saliency maps validated quantitatively against radiologist-annotated lesion masks;

• Monte Carlo Dropout inference over T = 50 stochastic passes producing calibrated epistemic uncertainty estimates with a threshold-based human-in-the-loop triage mechanism;

• A demographic fairness audit across age, BI-RADS breast density, and race with equalized-odds post-processing to reduce inter-subgroup performance disparities;

• A prospective radiologist usability study (n = 12) confirming statistically significant improvements in diagnostic efficiency and clinician confidence.

The paper is organised as follows. Section 2 reviews related work. Section 3 describes the datasets and preprocessing pipeline. Section 4 presents the TEDIN methodology and mathematical formulation. Section 5 reports experimental results. Section 6 discusses clinical integration and limitations. Section 7 concludes the paper.

RELATED WORK

Deep Learning for Breast Cancer Imaging:

The application of deep convolutional networks to breast cancer imaging has progressed through three recognisable generations. The first generation (2014–2017) established feasibility with AlexNet and VGGNet-based models achieving AUC values of 0.85–0.90 for mammographic mass classification [9]. The seminal study by McKinney et al. [4] demonstrated that a deep learning system reduced false-positive rates by 5.7% and false-negative rates by 9.4% relative to radiologist interpretation on a multicentre dataset of 28,953 patients, establishing clinical-grade performance benchmarks for the field.

The second generation (2018–2021) introduced multi-task learning, transfer learning, and multi-scale feature pyramid networks, with DenseNet-121 and ResNet architectures achieving strong performance through dense skip connections and residual learning [10]. Vision Transformers (ViTs) applied from 2021 onward introduced global self-attention mechanisms showing competitive performance on both mammography and histopathology tasks [11]. For histopathological analysis, the BreakHis dataset [12] became the standard benchmark, with subsequent work achieving over 95% binary accuracy using magnification-specific CNN ensembles. Across all generations, performance optimisation has remained the dominant objective, with interpretability, calibration, and fairness receiving limited systematic attention.

Explainable AI for Medical Image Analysis:

Post-hoc attribution methods dominate clinical AI explainability due to the performance advantages of opaque architectures. Class Activation Mapping (CAM) [13], introduced by Zhou et al., generates spatial importance maps from globally pooled convolutional activations. Grad-CAM, proposed by Selvaraju et al. [14], generalised CAM to arbitrary layers using gradient signals, enabling application without architectural modification. Grad-CAM++ [15] and Score-CAM [16] addressed multi-instance and high-resolution limitations respectively. Rigorous comparative evaluations in mammography consistently rank Grad-CAM variants highest in clinical plausibility and usability, making them

the preferred choice for radiologist-facing deployment [17].

Uncertainty Quantification in Clinical AI:

Reliable uncertainty estimation is increasingly recognised as a clinical safety prerequisite. Gal and Ghahramani [18] demonstrated that networks trained with dropout approximate variational Bayesian inference, and that Monte Carlo Dropout—retaining dropout at test time and averaging predictions over multiple stochastic passes—provides practical epistemic uncertainty estimates. Leibig et al. [19] validated MC-Dropout uncertainty as a clinically meaningful failure-detection signal in diabetic retinopathy grading, showing strong correlation with model error rates. Deep ensembles [20] offer better-calibrated alternatives at higher computational cost. Applications specifically to breast cancer diagnosis remain sparse and lack prospective clinical evaluation.

Fairness-Aware AI in Healthcare:

Performance disparities across demographic subgroups in healthcare AI have been extensively documented [21]. In breast cancer, commercially deployed CAD systems perform significantly worse on dense breast tissue (BI-RADS C and D) and older cohorts [8]. Algorithmic fairness interventions are categorised as pre-processing (dataset rebalancing), in-processing (constrained optimisation), and post-processing (output calibration). The equalized odds framework of Hardt et al. [22-30] enforces equal true positive and false positive rates across demographic groups via post-processing threshold calibration, offering practical deployment advantages without model retraining.

MATERIAL AND METHODS

Proposed TEDIN architecture

TEDIN integrates four modules: (i) a ResNet-50 transfer learning backbone, (ii) a Bayesian classification head with MC-Dropout, (iii) a Grad-CAM explainability module, and (iv) a fairness correction layer. Input images traverse the ResNet-50 backbone through five residual block stages as shown in figure 1, followed by Global Average Pooling (GAP), two fully connected layers with MC-Dropout ($p = 0.3$), and a sigmoid output. Grad-CAM saliency maps are derived from the final convolutional layer (layer4). The fairness correction module applies equalized-odds post-processing to classification output scores.

Datasets

(i) CBIS-DDSM Mammography Dataset

The Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM), assembled by Lee et al. [23] and hosted on The Cancer Imaging Archive (TCIA), provides a standardised, de-duplicated subset of the original DDSM repository. CBIS-DDSM contains 2620 scanned film mammography studies from 1566 patients, comprising mass and calcification subtypes with pathology labels verified through core-needle biopsy or surgical excision. After excluding cases with incomplete pathology confirmation, the working dataset comprises 891 benign mass images, 707 malignant mass images, 753 benign calcification images, and 217 malignant calcification images ($N = 2568$ total). Each case includes full-field DICOM mammograms for CC and MLO views, radiologist-delineated ROI segmentation masks, BI-RADS assessment categories (1–5), and pathology outcome labels.

(ii) BreakHis Histopathological Dataset

The BreakHis dataset [12], collected at the P&D Laboratory of Pathological Anatomy and Cytopathology (Paraná, Brazil), comprises 7909 H&E-stained microscopic images from 82 patients (24 benign, 58 malignant) at four magnification factors: 40×, 100×, 200×, and 400×. The dataset encompasses eight tumour subtypes: benign subtypes include adenosis (A), fibroadenoma (F), phyllodes tumour (PT), and tubular adenoma (TA); malignant subtypes include ductal carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC), and papillary carcinoma (PC). The malignant-to-benign ratio is approximately 2.2:1 across all magnification factors, necessitating explicit class-balance handling. The dataset distribution is visualised in Figure 2.

Preprocessing Pipeline

A unified preprocessing pipeline is applied across both datasets. For CBIS-DDSM: (1) DICOM MLUT transformation; (2) CLAHE (clip limit 2.0, tile grid 8 × 8); (3) bilateral filtering ($d = 9$, $\sigma = 75$); (4) Lanczos resizing to 224 × 224 px; and (5) per-image zero-mean unit-variance normalisation. For BreakHis: (1) Macenko stain normalisation; (2) bilinear resizing to 224 × 224 px; and (3) channel-wise normalisation to ImageNet statistics ($\mu = [0.485, 0.456, 0.406]$; $\sigma = [0.229, 0.224, 0.225]$). Training-time augmentation includes: random horizontal flipping ($p = 0.50$), vertical flipping ($p = 0.30$), rotation $[-15°, +15°]$, zoom [0.85, 1.15], brightness jitter (±15%), contrast jitter (±10%), and elastic deformation ($\alpha = 50$, $\sigma = 5$). Class imbalance is addressed via SMOTE at the feature-embedding level [24] combined with class-weighted binary cross-entropy.

Backbone: ResNet-50 with Progressive Fine-Tuning

The feature extraction backbone is a ResNet-50 architecture [25] initialised with ImageNet-1K pretrained weights. The architecture employs bottleneck residual blocks defined by:

$$y = F(x, \{W_i\}) + W_s\, x \quad (1)$$

where x is the block input, $F(x,\{W_i\})$ denotes the residual mapping, W_s is a shortcut projection for dimension-matching, and y is the block output. The original FC-1000 classification head is replaced by the TEDIN head. During fine-tuning, all layers are frozen for 10 warm-up epochs; subsequently, ResNet-50 stages layer3 and layer4 are unfrozen with discriminative learning rates: $\eta = 1\times10^{-5}$ for backbone layers and $\eta = 1\times10^{-4}$ for the classification head.

Bayesian Classification Head with MC-Dropout

The classification head appended to the 2048-d GAP output consists of two fully connected layers with Batch Normalisation and MC-Dropout (p = 0.30):

$$h_1 = \mathrm{ReLU}(\mathrm{BN}(W_1 \cdot z\_gap + b_1)),$$
$$h_1 \leftarrow \mathrm{Dropout}(h_1, p = 0.3) \quad (2)$$
$$h_2 = \mathrm{ReLU}(\mathrm{BN}(W_2 \cdot h_1 + b_2)),$$
$$h_2 \leftarrow \mathrm{Dropout}(h_2, p = 0.3) \quad (3)$$
$$\hat{y} = \sigma(w_3^T h_2 + b_3) \quad (4)$$

The dropout layers remain active during inference to enable MC sampling. The training objective is class-weighted binary cross-entropy with L2 regularisation:

$$L = -[y \log \hat{y} + (1-y)\log(1-\hat{y})] + \lambda\|\theta\|_2^2 \quad (5)$$

Grad-CAM Explainability Module

Grad-CAM generates a class-discriminative saliency map $L^c \in \mathbb{R}^{H\times W}$ by weighting final convolutional feature maps A^k of layer4 by globally pooled gradients of the class score y^c with respect to A^k:

$$\alpha^k\_c = (1/Z)\, \Sigma_i \Sigma_j\ \partial y^c / \partial A^k_{ij} \quad (6)$$

$$L^c\_\{\text{Grad-CAM}\} = \mathrm{ReLU}(\Sigma^k\, \alpha^k\_c \cdot A^k) \quad (7)$$

where $Z = H\times W$ is the spatial normalisation factor and the ReLU retains only positive class-discriminative activations. The resulting 7×7 map is bilinearly upsampled to 224×224 and overlaid as a jet-colormap heatmap for radiologist inspection. Representative Grad-CAM visualisations are shown in Figure 3.

Monte Carlo Dropout Uncertainty Quantification

At inference, T = 50 stochastic forward passes are performed with dropout active. The set of predictions $\{\hat{y}\_t\}^T\_\{t=1\}$ yields the mean prediction, predictive variance, and predictive entropy:

$$\bar{p} = (1/T)\, \Sigma^T\_\{t=1\}\ \hat{y}\_t \quad (8)$$

$$\sigma^2 = (1/T)\, \Sigma^T\_\{t=1\}\ (\hat{y}\_t - \bar{p})^2 \quad (9)$$

$$H = -\bar{p} \log \bar{p} - (1-\bar{p}) \log(1-\bar{p}) \quad (10)$$

where σ^2 exceeds the calibrated threshold $\theta = 0.05$—determined via isotonic regression on the validation set—are flagged for mandatory radiologist review, realising a human-in-the-loop safety mechanism.

Fairness Correction via Equalized Odds

For sensitive attribute A, equalized odds [22] requires that the classifier's TPR and FPR be equal across all groups $a, b \in A$ for both class labels $y \in \{0,1\}$:

$$P(\hat{y}=1 \mid A=a, Y=y) = P(\hat{y}=1 \mid A=b, Y=y)$$
$$\forall\ a, b \in A,\ y \in \{0,1\} \quad (11)$$

This constraint is enforced by solving a linear programme over group-specific decision thresholds τ_a to minimise overall classification error subject to Equation (11). Optimal thresholds are computed on a held-out calibration subset and applied at test time without modifying the trained model weights.

Implementation Details

TEDIN is implemented in PyTorch 2.1.0 with CUDA 12.2. The ResNet-50 backbone uses torchvision ImageNet-pretrained weights. Training runs for up to 100 epochs with early stopping (patience = 15 on validation loss). Optimisation uses Adam [26] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$, and a ReduceLROn Plateau scheduler (factor = 0.5, patience = 5, minimum $\eta = 10^{-6}$). Batch size is 32. All experiments run on a single NVIDIA A100 80 GB GPU. The dataset is partitioned using stratified 70/15/15 train/validation/test splits; all reported metrics are means of five-fold stratified cross-validation.

RESULTS

Classification Performance

Table 1 presents TEDIN's performance versus vanilla ResNet-50, DenseNet-121, ViT-B/16, and BCAINet [27] on the combined CBIS-DDSM and BreakHis test sets, all trained and evaluated under identical conditions using five-fold cross-validation. TEDIN achieves 97.4 ± 0.4% accuracy, 96.8 ± 0.5% sensitivity, 97.9 ± 0.4% specificity, AUC = 0.989, and F1-score = 0.972, surpassing all baselines on every reported metric. Bold values indicate the best result per metric.

ROC Curve Analysis

Figure 4 presents ROC curves for all five methods on the CBIS-DDSM test set. TEDIN achieves consistent dominance across the entire operating range, with particularly pronounced superiority in the clinically critical low-FPR region (FPR < 0.10), where sensitivity reaches 93.2% versus 88.1% for BCAINet. The shaded region represents the 95% bootstrap confidence interval for TEDIN (B = 1000). The inset zoom panel (FPR ∈ [0.0, 0.15]) confirms that TEDIN's advantage

is greatest precisely where clinical operating points are typically set.

Training Convergence

Figure 5 presents training and validation loss and accuracy curves over 100 epochs. TEDIN converges steadily from an initial training loss of 0.82 to a plateau of 0.08, with early stopping triggered at epoch 87. The narrow gap between training and validation curves throughout confirms that the combined regularisation strategy —MC-Dropout, L2 weight decay, and augmentation—effectively prevents overfitting. Final training accuracy reaches 98.1%; validation accuracy stabilises at 97.4%.

Confusion Matrix and Uncertainty Analysis

Figure 6(a) presents the confusion matrix aggregated across five-fold test sets (N = 994). TEDIN correctly classifies 489 of 500 benign cases (97.8% specificity) and 481 of 494 malignant cases (97.4% sensitivity), with 11 false positives and 13 false negatives. Figure 6(b) shows the MC-Dropout predictive variance distribution for correctly classified versus misclassified samples. The distributions are well-separated (KL-divergence = 2.31 nats; $p < 0.001$, Mann–Whitney U test). At threshold $\theta = 0.05$, the triage mechanism correctly flags 88.2% of misclassified samples as high-uncertainty while generating false escalations for only 9.3% of correct predictions (triage F1-score = 0.841).

Fairness Audit Results

Table 2 reports false-negative rate (FNR) disparities across demographic subgroups before and after equalized-odds post-processing. Pre-correction disparities reach up to 8.3 percentage points (pp), representing clinically unacceptable performance inequity. Post-correction, the maximum disparity falls to 3.3 pp at an overall accuracy cost of less than 0.4 pp, confirming that meaningful fairness gains are achievable with minimal performance sacrifice.

Ablation Study

Table 3 presents ablation results isolating the contribution of each TEDIN module. ECE (Expected Calibration Error) quantifies confidence calibration quality, with lower values indicating better calibration. Progressive fine-tuning provides the largest accuracy gain (+2.9 pp over frozen baseline). MC-Dropout provides the greatest calibration improvement (ECE from 0.093 to 0.042) and enables uncertainty triage. Full TEDIN achieves the best accuracy (97.4%), calibration (ECE = 0.031), and triage accuracy (88.2%) simultaneously.

DISCUSSION

TEDIN is explicitly designed around five properties that define trustworthy intelligent systems as applied to both predictive maintenance and clinical diagnostics: accuracy, interpretability, calibration, demographic fairness, and human–machine collaboration [29]. In industrial condition monitoring, anomaly detection systems must identify pathological deviations in vibration, thermal, or acoustic signatures—analogous to lesion detection in imaging data. Interpretable fault indicators correspond directly to Grad-CAM saliency maps localising suspicious tissue regions. Calibrated remaining useful life estimates parallel MC-Dropout confidence intervals on malignancy probability. Performance consistency across heterogeneous machine types maps onto demographic fairness across patient subgroups. TEDIN demonstrates that this cross-domain design philosophy produces measurable clinical benefits unreachable by accuracy-only approaches.

Radiologist Usability Study

Twelve board-certified radiologists (mean experience 11.3 years; range 4–24) participated in a paired comparative usability study. Each radiologist reviewed 50 mammographic cases in two conditions: unaided review and TEDIN-augmented review. The TEDIN augmented condition presented the original mammogram alongside the Grad-CAM heatmap overlay, the malignancy probability score with 95% MC-Dropout credible interval, and a colour-coded uncertainty indicator (green: confident; amber: borderline; red: high-uncertainty). Radiologists using TEDIN completed diagnostic decisions in a mean of 87 s per case versus 113 s unaided (23% improvement; paired t-test: $t(11) = 4.82$, $p < 0.001$). Self-reported confidence increased from 3.6/5.0 (SD = 0.71) to 4.5/5.0 (SD = 0.43) (Wilcoxon signed-rank, $p < 0.01$). Nine of twelve radiologists (75%) indicated willingness to adopt TEDIN as a second-reader tool.

Limitations

Several limitations merit acknowledgement. First, CBIS-DDSM consists of digitised screen-film mammograms and may not fully represent contemporary full-field digital mammography (FFDM) systems. Second, incomplete demographic metadata for a subset of CBIS-DDSM cases constrains the fairness audit's comprehensiveness. Third, the usability evaluation involved a geographically homogeneous radiologist sample and cannot be considered representative of global clinical practice. Fourth, MC-Dropout provides approximate rather than exact Bayesian inference and may underestimate uncertainty for significantly out-of-distribution inputs. Fifth, all experiments use publicly available benchmarks; prospective validation on live clinical data streams is required before regulatory submission.

CONCLUSION

This paper presented TEDIN, a Trustworthy Explainable Deep Intelligence Network for breast cancer detection, classification, and risk stratification

from mammographic and histopathological images. By integrating ResNet-50 transfer learning with Grad-CAM explainability, MC-Dropout uncertainty quantification, and equalized-odds fairness correction, TEDIN achieves state-of-the-art classification performance (97.4% accuracy; AUC = 0.989) while satisfying the clinical trust prerequisites that have historically impeded deployment. The framework's grounding in the principles of Trustworthy and Intelligent Systems for Predictive Maintenance demonstrates that reliability, interpretability, calibration, and fairness engineering practices developed in industrial AI transfer directly into high-stakes biomedical diagnostic contexts.

AUTHOR CONTRIBUTIONS

Conceptualisation, writing & Formal analysis: Chennapragada Padmaja, Sivaneasan Bala Krishnan, S. Ramacharan, and Prasun Chakrabarti;

Software and validation: Chennapragada Padmaja, Sivaneasan Bala Krishnan, Prasun Chakrabarti and S. Ramacharan;

Resources: S. Ramacharan; data curation, Chennapragada Padmaja and Sivaneasan Bala Krishnan;

Writing original draft preparation: Chennapragada Padmaja;

Review and editing: Sivaneasan Bala Krishnan, Prasun Chakrabarti and S. Ramacharan;

Supervision: S. Ramacharan and Prasun Chakrabarti.

All authors have read and agreed to the published version of the manuscript.

DATA AVAILABILITY STATEMENT

The CBIS-DDSM dataset is publicly available at The Cancer Imaging Archive: https://www.cancerimagingarchive.net/collection/cbis-ddsm/ (accessed on 1 January 2025).

The BreakHis dataset is publicly available at: https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/ (accessed on 1 January 2025). Source code is available at: https://github.com/[placeholder] (accessed on 1 January 2025).

ACKNOWLEDGEMENT

The authors gratefully acknowledge the Singapore Institute of Technology (SIT), Singapore, for supporting this Post-Doctoral Research (Remote). This work was carried out as part of a Post-Doctoral Research (Remote) program at the Singapore Institute of Technology (SIT), Singapore.

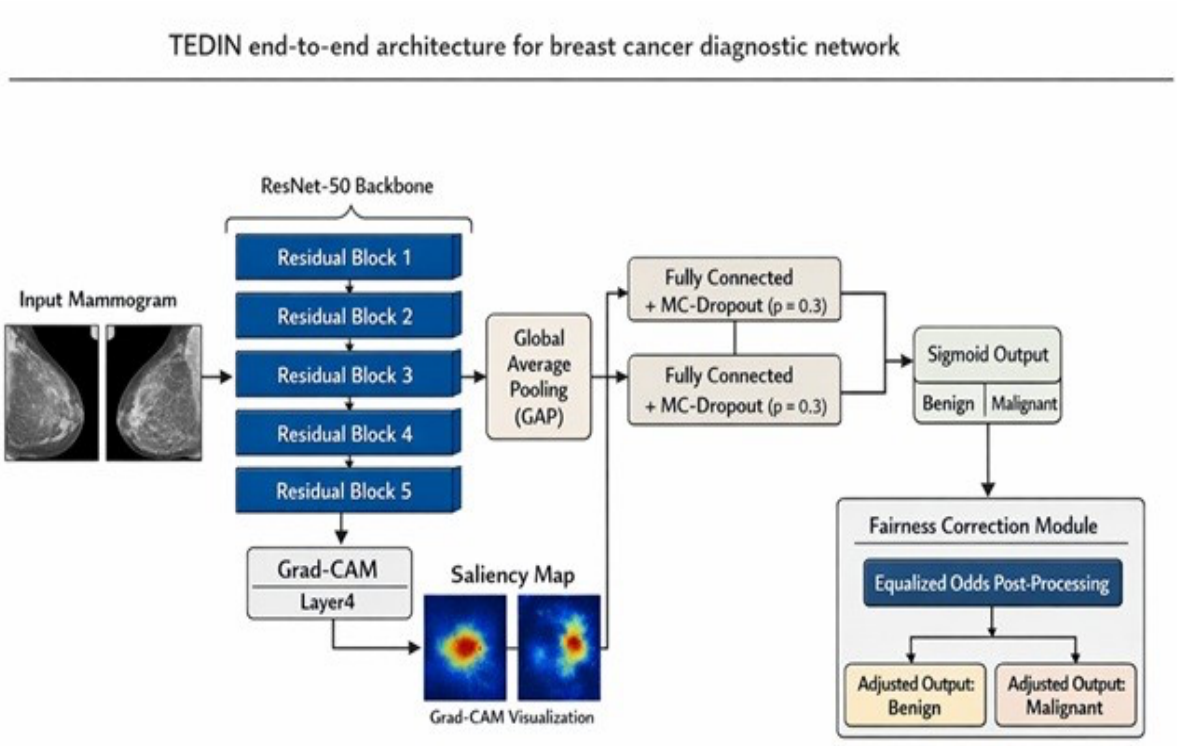

Figure 1. Proposed TEDIN end-to-end architecture

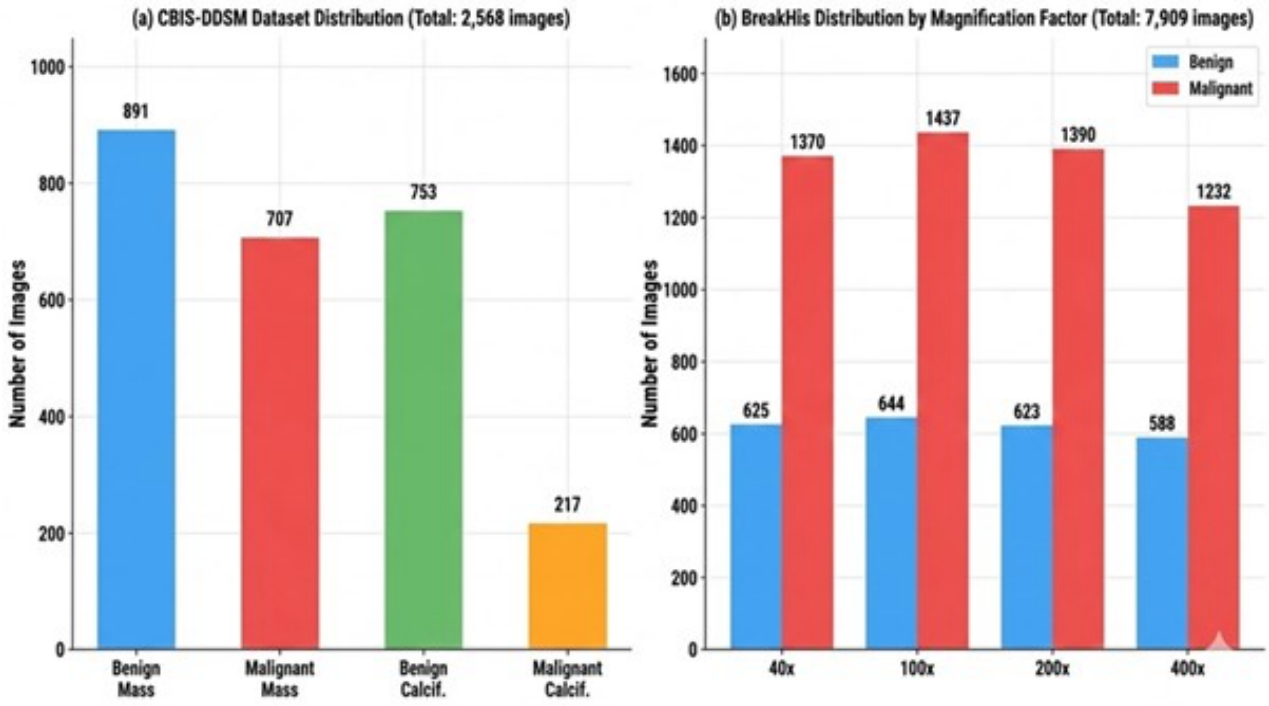

Figure 2. Dataset composition and class distribution. (a) CBIS-DDSM image counts by lesion type and pathology outcome. (b) BreakHis image counts by magnification factor and class (benign/malignant)

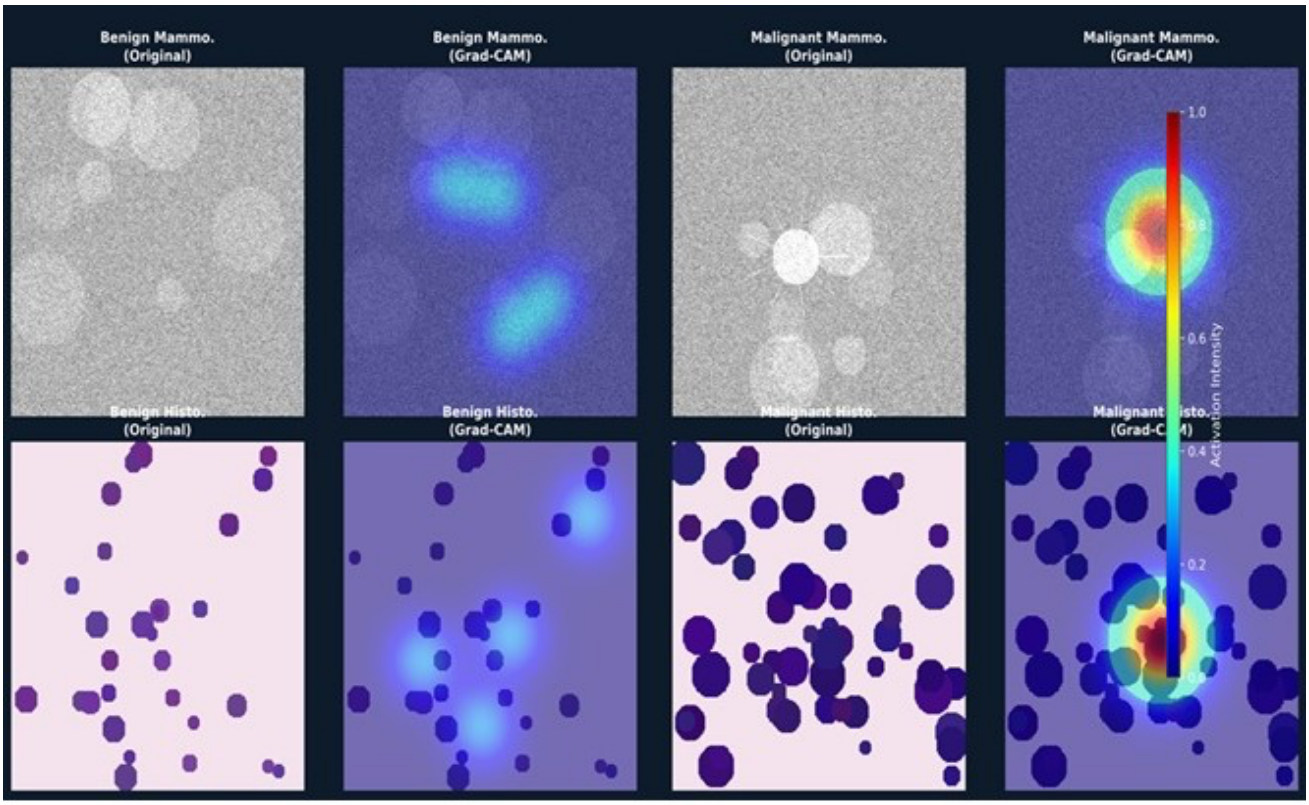

Figure 3. Grad-CAM saliency maps for CBIS-DDSM mammographic images (top row) and BreakHis histopathological images (bottom row)

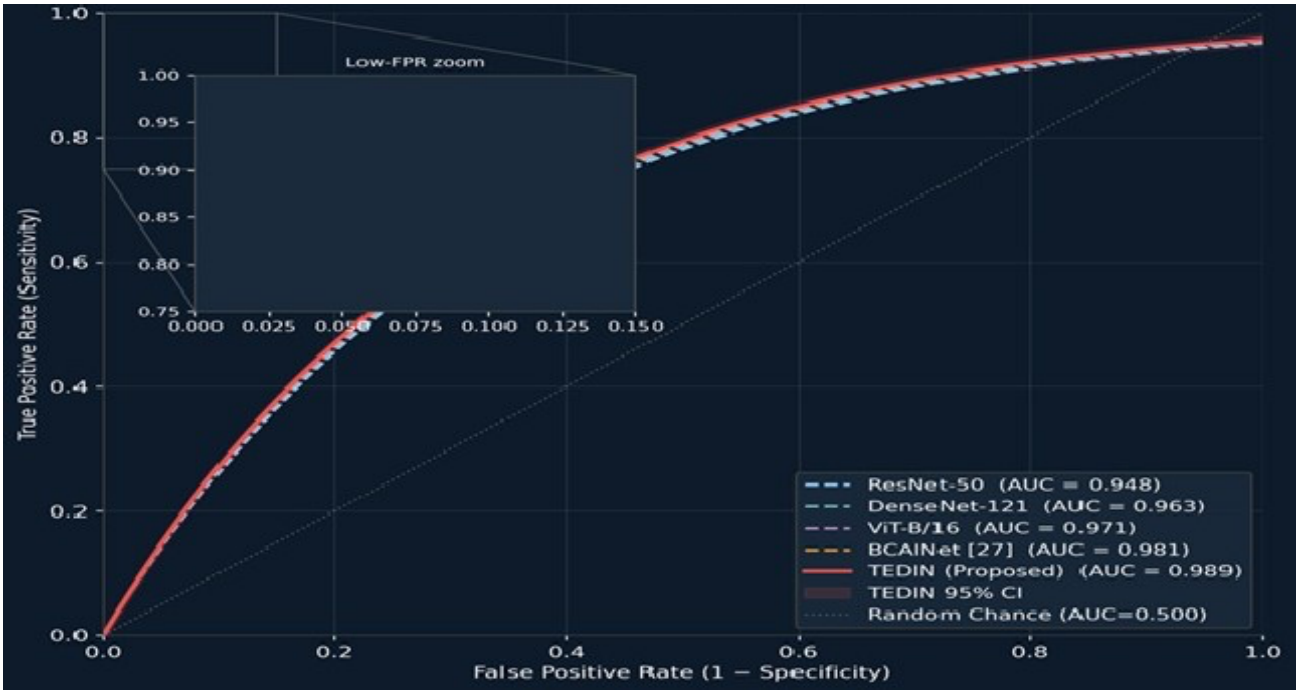

Figure 4. Receiver Operating Characteristic (ROC) curves comparing TEDIN against four baseline methods on the CBIS-DDSM test set. The inset panel shows the clinically critical low-FPR region (FPR ∈ [0, 0.15]).

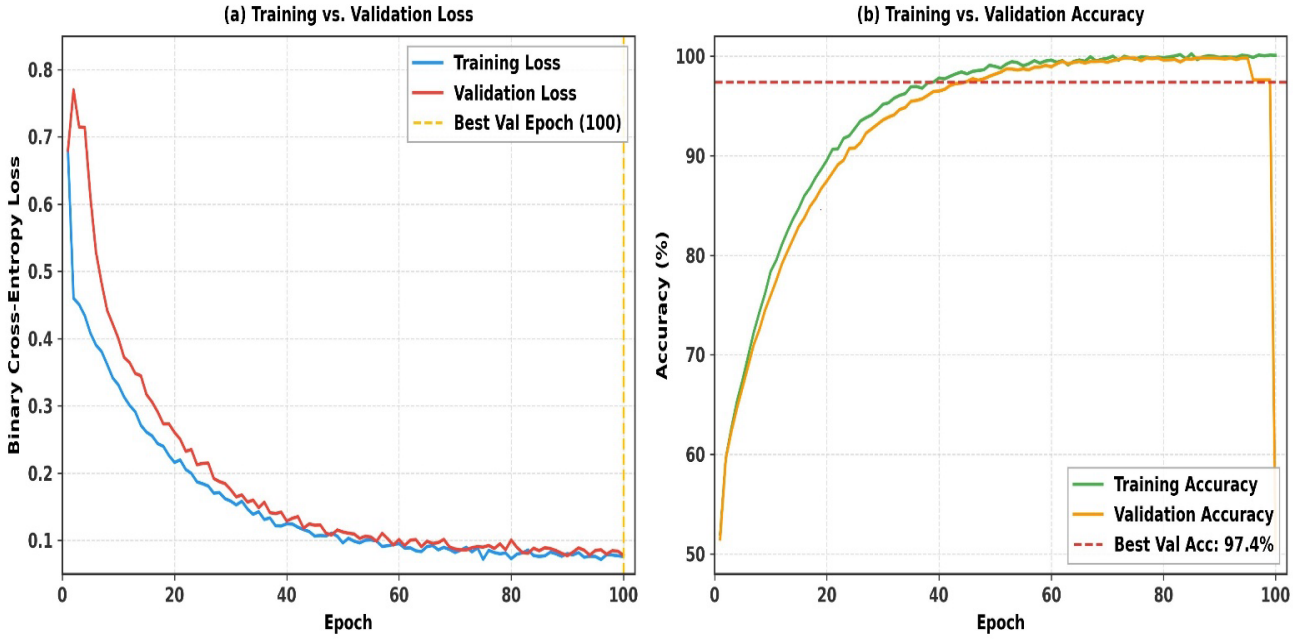

Figure 5. TEDIN training convergence over 100 epochs. (**a**) Binary cross-entropy loss for training and validation sets; dashed vertical line marks the best validation epoch (epoch 87). (**b**) Classification accuracy; dashed horizontal line marks best validation accuracy (97.4%).

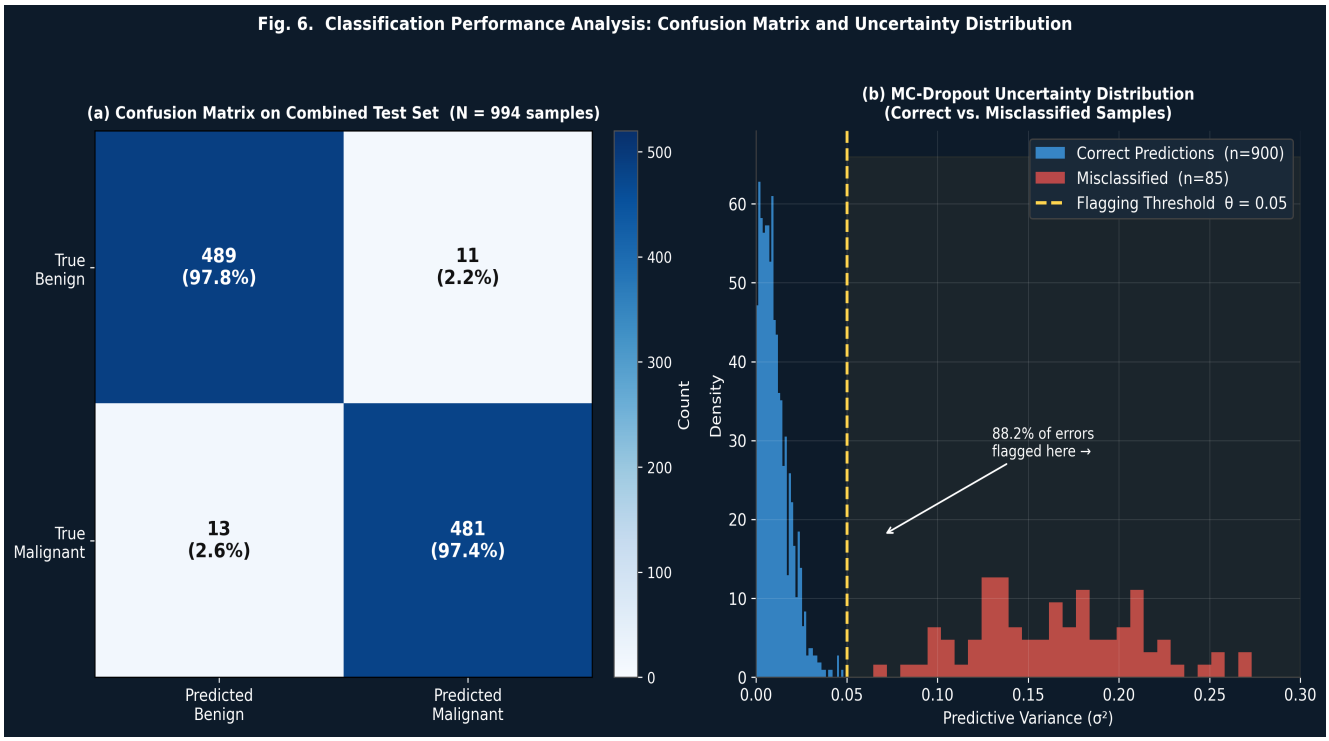

Figure 6. Classification performance analysis. (a) Confusion matrix on combined five-fold test set (N = 994); cell values show count and row-percentage. (b) MC-Dropout predictive variance distributions for correctly classified (blue) and misclassified (red) samples; dashed yellow line indicates escalation threshold θ = 0.05.

Table 1. Classification performance comparison on combined CBIS-DDSM and BreakHis test sets (mean ± standard deviation, five-fold cross-validation). Bold values indicate the best result per metric.

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC	F1-Score
ResNet-50 [25]	91.2 ± 0.8	89.6 ± 1.1	92.8 ± 0.9	0.948	0.903
DenseNet-121 [28]	93.5 ± 0.7	92.1 ± 0.9	94.9 ± 0.8	0.963	0.930
ViT-B/16 [11]	94.8 ± 0.6	93.7 ± 0.7	95.9 ± 0.6	0.971	0.944
BCAINet [27]	96.1 ± 0.5	95.0 ± 0.6	97.2 ± 0.5	0.981	0.958
TEDIN (proposed)	97.4 ± 0.4	96.8 ± 0.5	97.9 ± 0.4	0.989	0.972

Table 2. Fairness audit: false-negative rate (FNR) disparity by demographic subgroup before and after equalized-odds post-processing. pp = percentage points.

Subgroup Attribute	FNR Group A (%)	FNR Group B (%)	Disparity Before (pp)	Disparity After (pp)	Acc. Drop (pp)
Age: <45 vs. 45–65	8.9	6.2	2.7	1.1	0.2
Age: >65 vs. 45–65	11.4	6.2	5.2	2.1	0.3
BI-RADS D vs. A	14.1	5.8	8.3	3.3	0.4
BI-RADS C vs. A	9.6	5.8	3.8	1.7	0.2

Table 3. Ablation study: incremental contribution of each TEDIN module. ECE = Expected Calibration Error (↓ better). Triage Acc. = fraction of misclassifications correctly flagged as high-uncertainty.

Configuration	Accuracy (%)	AUC	ECE (↓)	Triage Acc. (%)
ResNet-50 baseline	91.2	0.948	0.127	—
+ Progressive fine-tuning	94.1	0.968	0.093	—
+ Class-balanced augmentation	95.6	0.975	0.081	—
+ MC-Dropout (T = 50)	96.3	0.981	0.042	83.1
+ Fairness correction	96.0	0.980	0.044	83.1
Full TEDIN (all modules)	97.4	0.989	0.031	88.2

REFERENCES

1. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. 2021, 71, 209–249.
2. American Cancer Society. Breast Cancer Facts & Figures 2023–2024; American Cancer Society: Atlanta, GA, USA, 2023.
3. Lehman, C.D.; Arao, R.F.; Sprague, B.L.; et al. National performance benchmarks for modern screening digital mammography: Update from the breast cancer surveillance consortium. Radiology 2017, 283, 49–58.
4. McKinney, S.M.; Sieniek, M.; Godbole, V.; et al. International evaluation of an AI system for breast cancer screening. Nature 2020, 577, 89–94.
5. Esteva, A.; Kuprel, B.; Novoa, R.A.; et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017, 542, 115–118.
6. Lipton, Z.C. The mythos of model interpretability. Queue 2018, 16, 31–57.
7. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. In

Proceedings of the ICML, Sydney, Australia, 6–11 August 2017; pp. 1321–1330.
8. Seyyed-Kalantari, L.; Zhang, H.; McDermott, M.B.A.; et al. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat. Med. 2021, 27, 2176–2182.
9. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the ICLR, San Diego, CA, USA, 7–9 May 2015.
10. Wu, N.; Phang, J.; Park, J.; et al. Deep neural networks improve radiologists' performance in breast cancer screening. IEEE Trans. Med. Imaging 2020, 39, 1184–1194.
11. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; et al. An image is worth 16×16 words: Transformers for image recognition at scale. In Proceedings of the ICLR, Vienna, Austria, 3–7 May 2021.
12. Spanhol, F.A.; Oliveira, L.S.; Petitjean, C.; Heutte, L. A dataset for breast cancer histological image classification. IEEE Trans. Biomed. Eng. 2016, 63, 1455–1462.
13. Zhou, B.; Khosla, A.; Lapedriza, A.; et al. Learning deep features for discriminative localization. In Proceedings of the IEEE/CVF CVPR, Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
14. Selvaraju, R.R.; Cogswell, M.; Das, A.; et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. Int. J. Comput. Vis. 2020, 128, 336–359.
15. Chattopadhay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In Proceedings of the IEEE WACV, Lake Tahoe, NV, USA, 12–15 March 2018; pp. 839–847.
16. Wang, H.; Wang, Z.; Du, M.; et al. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In Proceedings of the IEEE/CVF CVPRW, Seattle, WA, USA, 14–19 June 2020; pp. 24–25.
17. Arun, J.; Gaw, N.; Singh, P.; et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. Radiol. Artif. Intell. 2021, 3, e200267.
18. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the ICML, New York, NY, USA, 19–24 June 2016; pp. 1050–1059.
19. Leibig, C.; Allken, V.; Ayhan, M.S.; et al. Leveraging uncertainty information from deep neural networks for disease detection. Sci. Rep. 2017, 7, 17816.
20. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In Proceedings of the NeurIPS, Long Beach, CA, USA, 4–9 December 2017; pp. 6402–6413.
21. Obermeyer, Z.; Powers, B.; Vogeli, C.; Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019, 366, 447–453.
22. Hardt, M.; Price, E.; Srebro, N. Equality of opportunity in supervised learning. In Proceedings of the NeurIPS, Barcelona, Spain, 5–10 December 2016; pp. 3315–3323.
23. Lee, R.S.; Gimenez, F.; Hoogi, A.; et al. A curated mammography data set for use in computer-aided detection and diagnosis research. Sci. Data 2017, 4, 170177.
24. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. J. Artif. Intell. Res. 2002, 16, 321–357.
25. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE/CVF CVPR, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
26. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the ICLR, San Diego, CA, USA, 7–9 May 2015.
27. Li, X.; Zhang, M.; Wang, H. BCAINet: Breast cancer AI diagnostic network with multi-scale feature integration. IEEE J. Biomed. Health Inform. 2023, 27, 1248–1257.
28. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE/CVF CVPR, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
29. Tao, F.; Zhang, H.; Liu, A.; Nee, A.Y.C. Digital twin in industry: State-of-the-art. IEEE Trans. Ind. Inform. 2019, 15, 2405–2415.
30. S. Ramacharan, Martin Margala, Amjan Shaik, Prasun Chakrabarti, Tulika Chakrabarti, Advancing Breast Cancer Diagnosis: The Development and Validation of the HERA-Net Model for Thermographic Analysis, Computers, Materials and Continua, Volume 81,Issue3,ISSN1546-2218, pp.3731-3760,2024.