

GAN-Augmented Swin Transformer for Breast Cancer Detection in a Three-Tier IoMT Framework

Bo Li*, Aniket Mahanti*, Kalpdrum Passi[†], Ranesh Naha[‡], Arunkumar Arulappan[§]

*University of Auckland, New Zealand; [†]Laurentian University, Canada

[‡]Queensland University of Technology, Australia; [§]Vellore Institute of Technology, India

bli397@aucklanduni.ac.nz, a.mahanti@auckland.ac.nz

kpassi@laurentian.ca, ranesh.naha@qut.edu.au, arunkumar.a@vit.ac.in

Abstract—Automated analysis is increasingly used to support breast cancer screening where specialist radiologists are in short supply. Deploying such systems at point-of-care sites is held back by two practical limits: too little labelled data to train high-capacity classifiers, and the bandwidth needed to send raw mammograms to a central server. This paper presents an integrated detection pipeline that targets both limits and maps onto a three-tier Internet of Medical Things (IoMT) edge-fog-cloud architecture. A reserved-space Single Shot MultiBox Detector (SSD), scored by an Intersection over Prediction (IoP) criterion that requires full lesion enclosure, attains a 97.52% true positive rate on the CBIS-DDSM mass dataset. A Generative Adversarial Network with Efficient Multi-scale Attention (EMA-GAN) then synthesises mammographic patches that raise the classifier area under the curve (AUC) from 0.75 to 0.85. Among sixteen architectures, the Swin Transformer Base attains the highest AUC at 86.8%, ahead of all convolutional and Vision Transformer baselines. Over a 200-mammogram batch, distributing the pipeline across the three tiers reduces upstream data volume by 99.72% and, under an analytic model of the stated link parameters, communication time by 99.3% relative to a cloud-only baseline. These results indicate that accurate mammographic detection can be supported within the bandwidth limits typical of low-resource screening sites.

Index Terms—breast cancer detection, mammography, generative adversarial network, Swin Transformer, Internet of Medical Things, edge computing, CBIS-DDSM

I. INTRODUCTION

Deep learning has opened a practical pathway to expanding cancer screening in resource-constrained settings. Breast cancer is the most frequently diagnosed malignancy in women globally, with 2.3 million new cases and 666 000 deaths in 2022 [1]. Survival depends strongly on the stage at which the disease is detected [2], which makes timely screening important. Screening mammography is the established population-level modality, but its value depends on expert interpretation by a radiologist. Where radiologists are scarce, automated computer-aided detection (CADE) systems are needed to sustain screening capacity [3].

Deploying CADE in mobile health networks introduces two technical barriers that prior systems have not resolved together.

Data scarcity. The Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM) [4], [5] provides 1,696 annotated mass cases alongside a separate calcification subset. This work uses the mass cases, a volume insufficient for high-capacity classifiers to generalise

across the morphological diversity of breast masses. Standard geometric augmentation increases sample count without introducing pathological diversity, leaving a gap that generative synthesis is well-positioned to fill.

Bandwidth and latency. Rural and low-resource sites cannot transmit raw Digital Imaging and Communications in Medicine (DICOM) studies, often tens of megabytes each, to centralised servers within acceptable time. Three-tier IoMT architectures reduce this burden by offloading preprocessing to edge nodes and feature extraction to fog nodes [6], [7]. Few prior studies, however, integrate Region of Interest (ROI) extraction, synthetic augmentation, and classification within such a framework.

This paper addresses both barriers with a single pipeline. At the edge, full-resolution DICOM studies are preprocessed and compressed. At the fog, a reserved-space detector extracts lesion-centred ROIs. At the cloud, an EMA-GAN augments the limited training set and a Swin Transformer Base classifies each ROI as benign or malignant.

The contributions are:

- 1) A reserved-space SSD detector with an Intersection over Prediction (IoP) criterion that enforces complete lesion enclosure, reaching a 97.52% true positive rate (TPR).
- 2) An EMA-GAN augmentation method that raises the classification area under the curve (AUC) from 0.75 to 0.85, outperforming geometric augmentation.
- 3) A benchmark of sixteen architectures that identifies Swin Transformer Base as the strongest classifier (86.8% AUC).
- 4) A three-tier IoMT mapping that lowers upstream data volume by 99.72% and, under an analytic model of the stated link parameters, communication time by 99.3%.

II. RELATED WORK

A. Mammographic Mass Detection

Anchor-based object detectors have become the practical choice for mass localisation owing to their speed and accuracy trade-off at task-relevant resolutions. Faster R-CNN [8] introduced region proposal networks that achieve strong detection sensitivity. YOLO-v3 [9] and YOLO-v4 [10] offer single-pass inference suited to edge deployment. A persistent shortcoming of Intersection over Union (IoU) based evaluation is that a

prediction partially overlapping a lesion can be scored as a true positive. The IoP criterion introduced here requires the predicted bounding box to *fully enclose* the annotated mass, enforcing a stricter, screening-oriented standard of localisation.

B. Synthetic Augmentation for Medical Imaging

Generative Adversarial Networks [11] address data scarcity in radiology by synthesising plausible training samples [12]. A common limitation of such models is the loss of fine-scale detail, which is diagnostically important in mammography. The Efficient Multi-scale Attention (EMA) module [13] mitigates this by applying attention across global context, local texture, and pixel-level detail. Integrating EMA into a GAN generator for the CBIS-DDSM mass category has not been previously reported.

C. Transformer-Based Classification

The Vision Transformer (ViT) [14] demonstrated that patch-level self-attention is competitive with convolutional networks at ImageNet scale. Swin Transformer [15] introduced a hierarchical, shifted-window attention mechanism that recovers local inductive bias while retaining global receptive fields. This design suits high-resolution medical images better than the isotropic attention of a plain ViT. Transformer and hierarchical-attention models have been shown to benefit mammographic classification [3]. However, few studies have combined a Swin Transformer with GAN-based augmentation in an IoMT-oriented framework.

D. IoMT Architectures for Medical Imaging

Three-tier IoMT deployments distribute preprocessing to the edge, feature extraction to the fog, and inference to the cloud. This reduces upstream bandwidth by transmitting features rather than raw images. Siddiqui et al. [6] reported 97.81% accuracy for IoMT-based breast cancer stage prediction. Rawas and Tafran [7] achieved 86.32% accuracy with an IoMT-cloud fusion model. Neither result is directly comparable to ours, because the prior studies evaluate different tasks and report different metrics. Both also transmit full imaging data to the cloud, which is impractical in bandwidth-limited sites. Our system confines raw DICOM handling to the edge tier and transmits only 256×256 ROI patches upstream.

III. METHODOLOGY

Section II identified three gaps in prior work. IoU-based detection can score a partially covered lesion as a true positive, existing augmentation adds little pathological diversity to the CBIS-DDSM mass category, and few prior systems integrate ROI extraction, augmentation, and classification within one IoMT-oriented framework. The three-component architecture below addresses each gap. Fig. 1 shows how each stage maps to an IoMT tier.

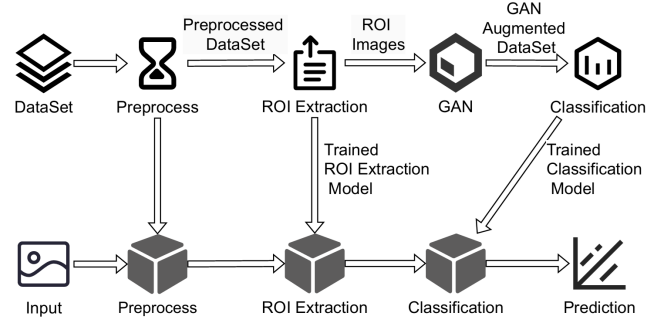


Fig. 1. End-to-end pipeline within the three-tier IoMT architecture. *Edge*: DICOM ingestion, CLAHE, Gaussian smoothing, and JPEG compression reduce the per-batch payload from ~ 6 GB to 7.36 MB. *Fog*: reserved-space SSD with IoP scoring extracts 256×256 mass ROIs. *Cloud*: EMA-GAN augmentation (training) and Swin-Base inference.

A. Preprocessing and ROI Extraction

The edge tier ingests full-resolution DICOM studies and standardises them before transmission. Each 16-bit DICOM image is converted to 8-bit JPEG. Contrast Limited Adaptive Histogram Equalisation (CLAHE) enhances local contrast to reveal fine detail in dense tissue. A Gaussian filter with a small kernel suppresses noise while preserving microcalcification-scale features. Full mammograms are resized to 384×512 pixels, preserving the native 3:4 aspect ratio, to provide uniform input to the detector. The reserved-space ROIs produced at the fog tier (below) are subsequently standardised to 256×256 pixels for GAN augmentation and classification.

At the fog tier, a Single Shot MultiBox Detector (SSD) [16] with a VGG-16 backbone detects mass candidates. A reserved-space strategy fixes each bounding box to a 256×256 region centred on the lesion, which retains the surrounding tissue context that rigid boxes discard. Standard IoU-based scoring accepts detections that partially overlap a lesion. The IoP metric is instead defined as

$$\text{IoP} = \frac{|P \cap G|}{|G|} \quad (1)$$

where P is the predicted bounding box and G is the ground-truth region. A detection is a true positive only when $\text{IoP} = 1.0$, meaning the prediction fully contains the ground truth. This is stricter than IoU, which accepts partial overlaps. In screening, a partially covered lesion is equivalent to a miss, so full enclosure is the application-motivated standard [5].

B. EMA-GAN Synthetic Augmentation

Training data scarcity is addressed at the cloud tier by a GAN with Efficient Multi-scale Attention. Fig. 2 shows the architecture. The generator is a U-Net Encoder-Decoder that synthesises 256×256 ROI patches from input masks. An Efficient Multi-scale Attention (EMA) module [13] is inserted at each encoder and decoder stage. The EMA module captures features at multiple scales, which helps the generator preserve microcalcifications and tissue texture. The discriminator is a

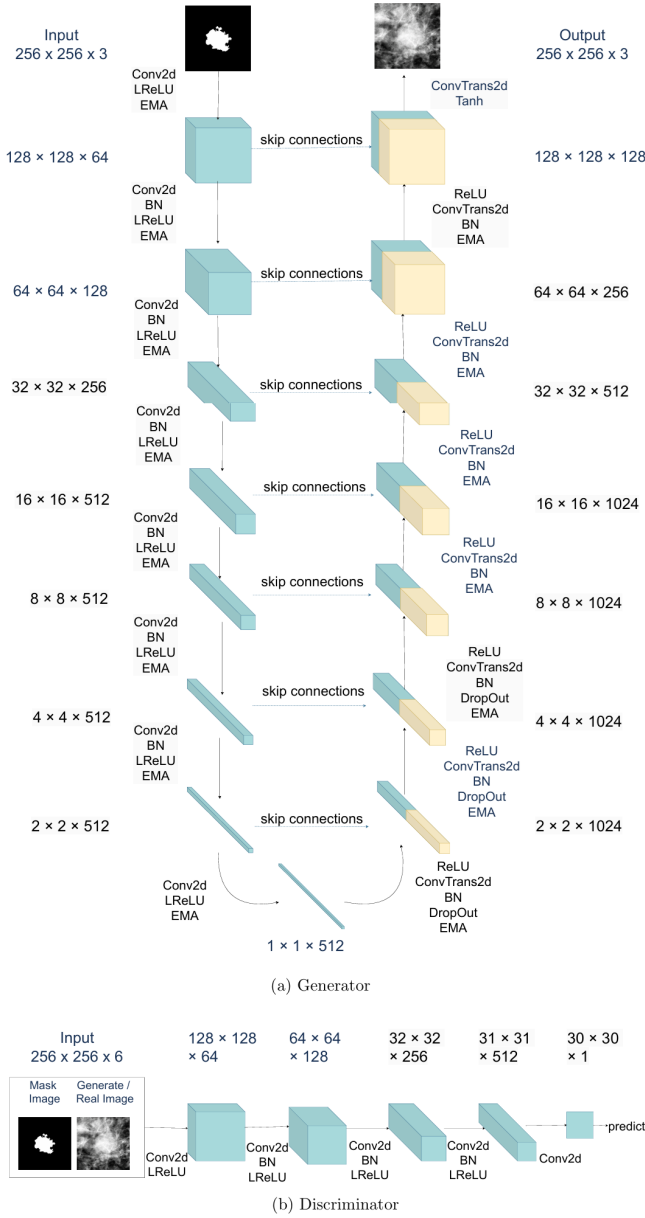


Fig. 2. EMA-GAN architecture. *Top*: U-Net generator with skip connections and EMA attention modules (Conv2d, BN, LReLU, EMA) at each encoder/decoder stage. Output is a 256×256 ROI patch. *Bottom*: PatchGAN discriminator evaluating local realism across overlapping receptive fields.

PatchGAN [17] that evaluates realism locally across overlapping regions, improving texture fidelity compared with image-level discrimination.

The generator objective combines a Hinge adversarial term with L1 and VGG-19 perceptual reconstruction terms [18]:

$$\mathcal{L}_G = -\mathbb{E}[D(G(\mathbf{m}), \mathbf{m})] + \lambda_1 \|G(\mathbf{m}) - \mathbf{x}\|_1 + \lambda_2 \|\phi(G(\mathbf{m})) - \phi(\mathbf{x})\|_1 \quad (2)$$

where $\mathbf{m}$ is the input lesion mask, $\mathbf{x}$ the target ROI, and $\phi(\cdot)$ denotes feature maps from a pretrained VGG-19 [18]. The discriminator is trained on the corresponding Hinge objective

$\mathbb{E}[\max(0, 1 - D(\mathbf{x}, \mathbf{m}))] + \mathbb{E}[\max(0, 1 + D(G(\mathbf{m}), \mathbf{m}))]$, taking the image and mask concatenated as input. The scalars λ_1 and λ_2 weight the L1 and perceptual terms against the adversarial term. Their numerical values are not recorded in the available results. The Hinge adversarial loss provides stable, margin-based gradients, the L1 term enforces pixel-level fidelity, and the perceptual term preserves semantic structure. The generator and discriminator are optimised alternately, with no freeze schedule. The EMA-GAN is trained on the CBIS-DDSM mass training split and used to augment the dataset to 200% of its original size.

C. Swin Transformer Classification

The cloud-tier classifier is a Swin Transformer Base [15] pretrained on ImageNet-22K. It is fine-tuned on the EMA-GAN-augmented CBIS-DDSM mass subset in two phases. In the first phase (50 epochs), the backbone is frozen and only the classification head is trained. In the second phase (300 epochs), all parameters are jointly optimised. This staged approach prevents early overfitting on the limited mammographic dataset.

Swin Transformer computes self-attention within local windows that shift between layers, capturing fine-grained local features and broader context without the quadratic cost of global attention. Fine-tuning uses SGD with momentum 0.9, an initial learning rate of 0.01 under cosine annealing, weight decay 0.0005, and a batch size of 32. The objective is cross-entropy loss.

D. Three-Tier IoMT Deployment

The IoMT framework assigns workloads to tiers by resource requirement:

Edge tier (Raspberry Pi-class device): DICOM decompression, Gaussian smoothing, histogram equalisation, and JPEG compression. Raw DICOM (about 6,000 MB per 200-image batch) is reduced to 7.36 MB before upstream transmission.

Fog tier (mid-range server): SSD-based ROI detection with IoP scoring. Only 256×256 ROI patches (9.41 MB per batch) are forwarded to the cloud.

Cloud tier (GPU server): Swin-Base inference and class score return.

Link parameters reflect representative wireless conditions: edge-fog at 50 ms and 15 Mbps, fog-cloud at 30 ms and 50 Mbps, and a cloud-only baseline at 100 ms and 15 Mbps. Fig. 3 shows the tier assignment.

E. Implementation Details

All models were implemented in PyTorch 2.2.2 with CUDA 12.1 under Python 3.11. Training and evaluation ran on a workstation with an NVIDIA GeForce RTX 4070 Ti GPU (12 GB), an Intel Core i7-13700F CPU, and 32 GB of RAM. The 200% EMA-GAN-augmented dataset was partitioned into training, validation, and test subsets in an 8:1:1 ratio with class distribution preserved. The order of augmentation and partitioning is not fully specified in the available results, a potential source of leakage discussed in Section V. The

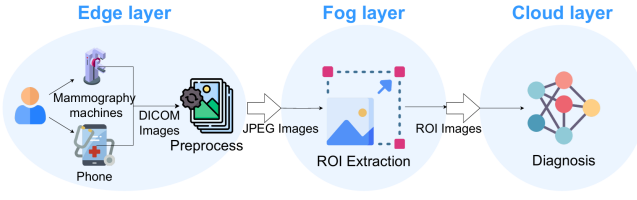


Fig. 3. Three-tier IoMT edge-fog-cloud architecture: DICOM acquisition and preprocessing at the edge, ROI extraction at the fog, and classification at the cloud. Per-link payloads and link parameters are stated in the text and in Table IV.

TABLE I

ROI DETECTION: TPR (%) UNDER IoP CRITERION ACROSS DETECTORS AND INPUT RESOLUTIONS. SSD AT 256×256 IS THE PROPOSED METHOD.

Method	Input Size	TPR (%)
Faster R-CNN [8]	128	82.07
Faster R-CNN [8]	256	97.20
SSD [16]	128	86.31
SSD [16]	256	97.52
YOLO-v3 [9]	128	87.99
YOLO-v3 [9]	256	96.08
YOLO-v4 [10]	128	87.99
YOLO-v4 [10]	256	95.52

detector and classifier each used a two-phase schedule, with the backbone frozen for the first 50 epochs and then unfrozen for joint optimisation. A cosine annealing schedule controlled the learning rate in both phases.

IV. EXPERIMENTAL RESULTS

Each subsection evaluates one component of the pipeline in order, followed by the end-to-end IoMT characterisation. All experiments use the CBIS-DDSM mass subset [5], comprising the 1,696 mass cases with radiologist-annotated bounding boxes and pathology labels.

A. ROI Detection

Under the IoP=1.0 criterion, the proposed SSD achieves 97.52% TPR at 256×256 input resolution. Table I shows the full comparison. Because IoP=1.0 rejects all partial overlaps, this TPR figure is a stricter claim than one measured with IoU. SSD at 256×256 leads all detectors, outperforming Faster R-CNN (97.20%), YOLO-v3 (96.08%), and YOLO-v4 (95.52%). All detectors improve from 128 to 256, so 256×256 is used throughout.

B. EMA-GAN Augmentation

The EMA-GAN trained with the hybrid Hinge+L1+Perceptual loss achieves a Fréchet Inception Distance (FID) of 15.2, a Peak Signal-to-Noise Ratio (PSNR) of 24.80 dB, and a Structural Similarity Index Measure (SSIM) of 0.63 on held-out CBIS-DDSM mass images. Fig. 4 shows that this configuration achieves the lowest and most stable FID among the four evaluated loss variants, which indicates that the perceptual term is the main contributor to image quality. Table II quantifies the

TABLE II
EFFECT OF AUGMENTATION STRATEGY ON SWIN-BASE CLASSIFICATION (CBIS-DDSM HELD-OUT TEST SET).

Dataset	Prec. (%)	Rec. (%)	F1 (%)	AUC
Original	70.34	67.45	68.68	0.75
Traditional (100%)	70.52	68.21	69.35	0.75
Traditional (200%)	71.46	69.31	70.37	0.77
GAN (100%)	74.71	71.74	73.20	0.81
GAN (200%)	76.89	73.66	75.24	0.85

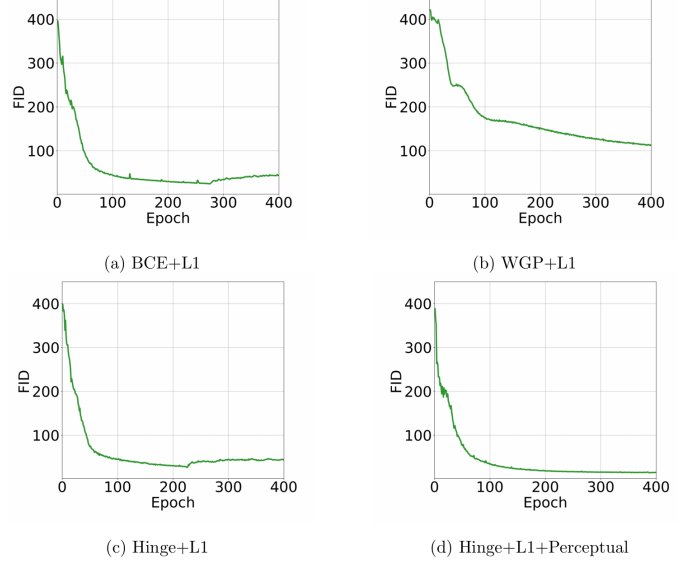


Fig. 4. FID convergence curves for four GAN loss configurations on CBIS-DDSM. The Hinge+L1+Perceptual combination (bottom right) achieves the lowest stabilised FID (≈ 15), outperforming BCE+L1, WGAN-GP+L1, and Hinge+L1. Lower FID indicates that synthetic patches more closely match the real data distribution, reducing domain shift in the augmented training set.

downstream classification impact across five augmentation strategies. The step from no augmentation (AUC=0.75) to 200% EMA-GAN augmentation (AUC=0.85) gives a 10 percentage-point gain, whereas traditional augmentation at 200% reaches only 0.77. This shows that the improvement comes from the pathological diversity of the synthetic data, not from the larger sample count alone. The corresponding Receiver Operating Characteristic (ROC) curves in Fig. 5 show consistently higher true positive rates for the GAN-augmented configurations across the operating range.

C. Architecture Comparison

With augmentation fixed at 200% EMA-GAN, Table III compares five representative models from a sixteen-architecture sweep. Swin-Base reaches 86.8% AUC, the highest of all evaluated architectures. ViT-B/16 (84.5%) beats every Convolutional Neural Network (CNN) model, which points to a general advantage for attention-based architectures here. Swin-Base adds a further 2.3 percentage points over ViT-B/16. Table II and Table III come from separate experiment series in the source study and their Swin-Base entries disagree on

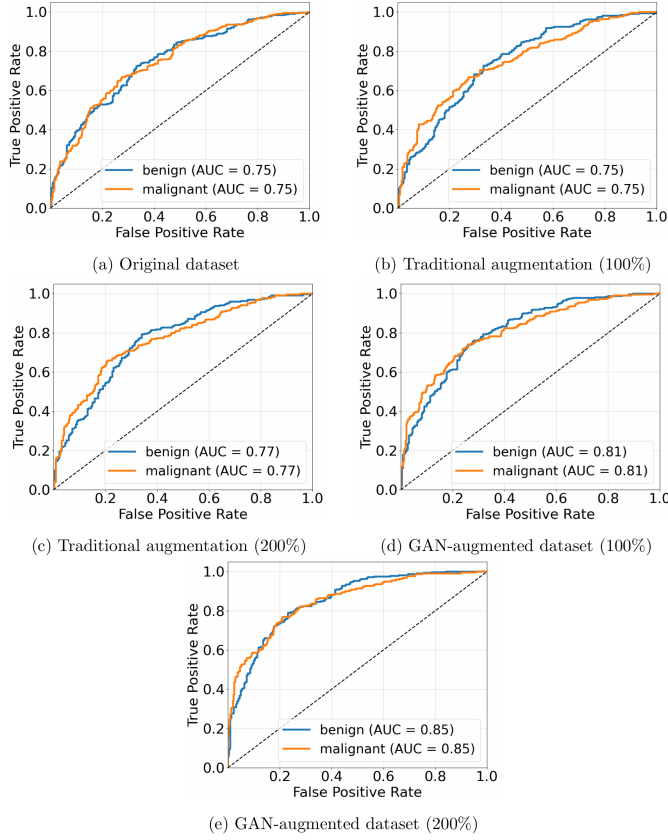


Fig. 5. ROC curves for five augmentation strategies on the CBIS-DDSM mass test set (Swin-Base classifier). GAN-augmented datasets progressively outperform the original and traditionally augmented baselines.

TABLE III
CLASSIFICATION PERFORMANCE ON THE EMA-GAN 200%-AUGMENTED CBIS-DDSM MASS SUBSET. FIVE REPRESENTATIVE ARCHITECTURES FROM A SIXTEEN-MODEL SWEEP, INCLUDING ALL THREE SWIN VARIANTS.

Model	Prec. (%)	Rec. (%)	F1 (%)	AUC (%)
ResNet-152 [19]	78.2	75.6	76.8	83.1
Swin-Tiny [15]	78.5	76.1	77.3	83.9
Swin-Small [15]	77.9	79.0	78.4	80.1 [†]
ViT-B/16 [14]	79.0	76.8	77.9	84.5
Swin-Base [15]	81.2	79.5	80.3	86.8

[†] Swin-Small AUC is lower than Swin-Tiny despite a higher F1 score, a pattern present in the source data and not analysed further.

every metric, so the two tables should be read independently rather than compared row to row.

D. IoMT Communication Cost

Table IV summarises the communication cost of processing a 200-mammogram batch. Two successive payload reductions drive it: edge-tier JPEG compression reduces the 6 GB DICOM batch to 7.36 MB transmitted to the fog, and fog-tier ROI extraction transmits 9.41 MB of 256×256 patches to the cloud, for a combined 16.77 MB against 6,000 MB for cloud-only transfer, a 99.72% reduction.

TABLE IV
IoMT COMMUNICATION COST: THREE-TIER VS. CLOUD-ONLY (200 CBIS-DDSM MAMMOGRAMS)

Metric	Cloud-Only	Three-Tier	Reduction
Transmitted volume (MB)	6,000	16.77	99.72 %
Communication time (s)	3,220	21.4	99.3 %

Link parameters: edge-fog 50 ms and 15 Mbps, fog-cloud 30 ms and 50 Mbps, cloud-only 100 ms and 15 Mbps. Communication time follows from (3) and excludes per-tier computation.

Communication time is derived from the payload sizes and the link parameters above. For a batch of N images crossing a link of bandwidth B (Mbps) and one-way latency ℓ (s) carrying S megabytes,

$$T = \frac{8S}{B} + N\ell. \quad (3)$$

With $N = 200$, the cloud-only baseline costs $8(6000)/15 + 200(0.100) = 3,220$ s, while the two three-tier links cost $8(7.36)/15 + 200(0.050) + 8(9.41)/50 + 200(0.030) = 21.4$ s, a 99.3% reduction. These figures cover transmission and link latency only. Per-tier computation is excluded because every stage of the source study ran on a single workstation rather than on the edge and fog hardware the tiers describe, so no representative per-tier processing time is available.

V. DISCUSSION

A. Interpretation of Results

IoP scoring sets a strict threshold: a detection counts only when the prediction fully contains the lesion, so 97.52% TPR under IoP = 1.0 is a stronger claim than the same figure under IoU. Adding EMA modules lowers FID from 23.50 to 15.2 at a fixed loss configuration, and the 200% EMA-GAN dataset raises AUC by 10 percentage points over no augmentation. Swin-Base leads ViT-B/16 by 2.3 percentage points of AUC, which is consistent with the shifted-window design suiting the irregular boundaries of breast masses better than isotropic global attention at the same model scale.

B. IoMT Deployment Implications

The payload reduction matters operationally. On a 15 Mbps backhaul representative of rural 4G, the 6 GB batch occupies the uplink for roughly 53 minutes, whereas the 7.36 MB the edge tier produces clears in under 4 s. A screening site that uploads between patients can absorb the second figure and not the first. Because (3) assumes fixed bandwidth and latency with no contention, packet loss, or retransmission, the reported times are a lower bound on what a deployment would incur.

C. Limitations

Five constraints bound the current results.

First, the order of augmentation and partitioning is not fully specified in the available results. If augmentation preceded partitioning, synthetic samples derived from training images could appear in the validation and test sets, which may inflate the reported classification metrics. Augmenting only

the training partition would remove this ambiguity and yield a cleaner estimate of generalisation.

Second, Tables II and III both report Swin-Base on the 200% EMA-GAN-augmented subset but disagree on every metric: 76.89 against 81.2 precision, 73.66 against 79.5 recall, 75.24 against 80.3 F1, and 85.0 against 86.8 AUC. The two tables come from separate experiment series in the source study, which does not reconcile them. Comparisons are valid within each table and not across them.

Third, every reported figure comes from a single training run. No repeated runs, confidence intervals, or significance tests are available, so gaps of one to three AUC points between the leading architectures should not be read as established orderings.

Fourth, evaluation covers only the CBIS-DDSM mass category, and the EMA-GAN was trained under CBIS-DDSM annotation protocols. Neither is expected to transfer without adaptation to microcalcification screening, which needs native-resolution processing, or to images from different equipment and screening programmes.

Fifth, the IoMT figures are analytic rather than measured. No component ran on edge or fog hardware, and (3) omits computation time, link contention, packet loss, and retransmission. Measuring preprocessing latency on a representative edge device and characterising sensitivity to bandwidth and loss are needed before the deployment claim can be considered supported.

VI. CONCLUSION

This paper presented a three-component pipeline for mammographic mass detection deployed within a three-tier IoMT edge-fog-cloud architecture. The pipeline combines a reserved-space SSD with IoP scoring, an EMA-GAN for synthetic augmentation, and a Swin Transformer Base classifier. On CBIS-DDSM, it achieves a 97.52% TPR for detection and 86.8% AUC for classification, the best of sixteen evaluated architectures. Distributing the pipeline across the three tiers lowers transmitted volume by 99.72% and, under an analytic model of the stated link parameters, communication time by 99.3% relative to cloud-only transfer. These results suggest that the proposed pipeline may reduce communication overhead for mammographic AI workflows in bandwidth-constrained IoMT settings, subject to hardware deployment and multi-centre validation. Future work includes federated learning across hospital sites to avoid centralising patient data, on-device inference through quantisation or distillation to remove the cloud round-trip, and multi-modal fusion with complementary imaging to improve early-stage sensitivity.

REFERENCES

- [1] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 74, no. 3, pp. 229–263, 2024.
- [2] R. L. Siegel, A. N. Giaquinto, and A. Jemal, "Cancer statistics, 2024," *CA: A Cancer Journal for Clinicians*, vol. 74, no. 1, pp. 12–49, 2024.
- [3] I. Sechopoulos, J. Teuwen, and R. Mann, "Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: State of the art," *Seminars in Cancer Biology*, vol. 72, pp. 214–225, 2021.
- [4] R. S. Lee, F. Gimenez, A. Hoogi, and D. L. Rubin, "Curated breast imaging subset of DDSM (CBIS-DDSM)," 2016, data set.
- [5] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Scientific Data*, vol. 4, p. 170177, 2017.
- [6] S. Y. Siddiqui, A. Haider, T. M. Ghazal, M. A. Khan, I. Naseer, S. Abbas, M. U. Rahman, J. A. Khan, M. Ahmad, M. K. Hasan, and K. Ateeq, "IoMT cloud-based intelligent prediction of breast cancer stages empowered with deep learning," *IEEE Access*, vol. 9, pp. 146 478–146 491, 2021.
- [7] S. Rawas and C. Tafran, "Next-gen breast cancer diagnosis: IEMBC as an IoMT-enabled cloud computing solution," *Multimedia Tools and Applications*, vol. 84, no. 25, pp. 29 391–29 418, 2025.
- [8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 28, 2015.
- [9] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.
- [10] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.
- [11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, 2014.
- [12] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Synthetic data augmentation using GAN for improved liver lesion classification," in *Proc. IEEE 15th Int. Symp. Biomedical Imaging (ISBI)*, 2018, pp. 289–293.
- [13] D. Ouyang, S. He, G. Zhang, M. Luo, H. Guo, J. Zhan, and Z. Huang, "Efficient multi-scale attention module with cross-spatial learning," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021.
- [15] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2021, pp. 9992–10 002.
- [16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. European Conf. Computer Vision (ECCV)*, 2016, pp. 21–37.
- [17] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1125–1134.
- [18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2015.
- [19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.