Big Data and Information Analytics

https://www.aimspress.com/journal/BDIA

BDIA, 9: 385–399.
DOI:10.3934/bdia.2025019
Received: 24 November 2025
Revised: 08 December 2025
Accepted: 11 December 2025
Published: 16 December 2025

Research article

Exploring machine learning for breast cancer detection in X-ray imaging

Weitong Liao[1], Ligang He[1,*], Mohammed H. Alghamdi[2,3] and Hammam M. Alghamdi[3]

[1] Department of Computer Science, University of Warwick, Coventry, United Kingdom
[2] College of Computer Science, King Khalid University, Abha, Saudi Arabia
[3] College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia

* **Correspondence:** Email: ligang.he@warwick.ac.uk.

Abstract: Breast X-ray imaging is a primary method for early breast cancer screening. However, applying machine learning techniques to this task presents significant challenges, particularly in capturing fine-grained details from high-resolution medical images. This study explored multiple machine learning approaches for breast cancer detection, including image-level and patch-level CNN-based classification, diffusion model-based anomaly detection, and federated learning (FL) for training models across distributed datasets. We evaluated these methods on real-world breast cancer X-ray datasets. Our findings suggest that FL offers a promising balance between performance and privacy preservation, underscoring its potential for real-world medical AI applications.

Keywords: breast cancer detection; machine learning; federated learning; medical image analysis; mammography

1. Introduction

According to the GLOBOCAN 2020 study, an estimated 2.3 million new cases of female breast cancer were reported in 2020, accounting for 11.7% of all cancer cases, surpassing lung cancer as the most common cancer in women [1]. The medical imaging techniques commonly used to detect breast cancer in clinical practice include X-ray imaging, magnetic resonance imaging (MRI), nuclear medicine imaging (NMI), and ultrasound imaging (UI) [2]. Due to its low cost and ease of detection, breast X-ray imaging has an advantage over other diagnostic methods for early breast cancer detection [3, 4]. However, it also has limitations, such as the inability to accurately detect cancer in patients with high breast density, as dense breast tissue often appears similar to cancerous tissue in imaging [3]. This highlights the need for computer-aided systems or machine learning models to assist doctors in making more accurate diagnoses.

This paper explores the feasibility and potential of applying different kinds of machine learning

algorithms to breast cancer detection using X-ray images. Given the unique characteristics of different datasets, such as the imbalance in class distributions and the heterogeneity in data sources, we selected appropriate models and training approaches to optimize the performance of breast cancer detection tasks. By constructing real-world scenarios with different datasets across various institutions, our federated learning model achieved performance comparable to local training, even under conditions of data imbalance, non-IID distributions, and institutional data heterogeneity.

The remainder of this paper is organized as follows: Section 2 reviews the related work. Section 3 presents the employed machine learning approaches, detailing i) convolutional neural network (CNN)-based classification at both image and patch levels, ii) anomaly detection using diffusion models, and iii) the federated learning approach. Section 4 describes the experimental setup and datasets, followed by the experimental results. Finally, Section 5 concludes the paper.

2. Related work

2.1. Mammography

Mammography is also referred to as breast X-ray imaging. In a traditional clinical scenario, doctors manually assess X-ray images of patients, with the accuracy of diagnosis influenced by various factors such as the doctor's experience, image quality, and the patient's age. This diagnostic approach increases the risk of missed detections or misdiagnoses, potentially delaying early treatment or causing unnecessary burden on patients [5]. Furthermore, as patients' imaging data is considered private and not shared through medical institutions, this limits further research based on patient clinical data. Existing representative breast X-ray datasets such as CBIS-DDSM, CMMD, and VinDr-Mammo provide limited samples, with 2620, 1775, and 5000 patients, respectively [5–7]. The limited sample size poses a significant challenge for training machine learning models, as it may lead to overfitting and limit the generalization ability of the models.

Clinical data and the imaging reports are commonly annotated according to the standards set by the American College of Radiology (ACR) [8], using the BI-RADS (Breast Imaging-Reporting and Data System) categories. These categories are as follows: BI-RADS 0 (incomplete, additional imaging evaluation required), BI-RADS 1 (normal), BI-RADS 2 (typically benign), BI-RADS 3 (probably benign, follow-up recommended), BI-RADS 4 (suspicious), BI-RADS 5 (highly suggestive of malignancy), and BI-RADS 6 (known biopsy-proven) [5, 8]. The standards also define specific mammographic features, which are described in the BI-RADS lexicon, including densities and masses, micro/macrocalcifications, architectural distortions, special cases (such as ductal ectasia, intramammary lymph nodes, or focal asymmetric density), and associated findings (including skin or nipple retraction, skin thickening, cutaneous lesions, and axillary lymph nodes) [8]. However, the aforementioned datasets provide limited descriptions of each patient. For example, the CBIS-DDSM dataset only records benign (BI-RADS 2) and malignant cases (above BI-RADS 3), with each case's abnormality type classified as either a mass or a calcification.

2.2. Medical image analysis

Machine learning methods are widely applied in the field of medical imaging analysis, covering various tasks such as classification [9–12], anomaly detection [13, 14], segmentation [6, 15, 16], and annotation [16].

Most research in the field of medical image analysis focuses on tissue-level analysis. For example, [9] used datasets such as CAMELYON-16 [17] and TCGA-BRCA [18], which consists of tissue slices. Due to the high resolution of tissue images, entire tissue slides are typically divided into smaller patches and then fed into a pre-trained CLIP model [19] to generate embeddings, which helps save computational resources during training. The downstream models can utilize these dimensionality-reduced embeddings for tasks such as prediction, segmentation, and others. However, for breast X-ray images, which are generally smaller in size and require consideration of the entire image for accurate diagnosis, dividing them into patches may lead to the loss of valuable information.

In the area of breast X-ray image analysis, [11] focused on the CBIS-DDSM breast X-ray dataset, using complete images and applying CNN for patient classification. [10] proposed case-level breast cancer prediction by combining the patient's craniocaudal view (CC view) and mediolateral oblique view (MLO view) X-ray images for more accurate prediction. [12] applied multiple machine learning models on the CBIS-DDSM dataset, providing a transparent baseline for subsequent research.

References [13] and [14] approached the diagnostic task as an anomaly detection problem, where the data distribution of normal cases differs from that of abnormal cases. Machine learning models such as variational autoencoders models (VAE) and diffusion models are trained to capture these differences, enabling the identification of anomalies or the reconstruction of abnormal regions. Reference [13] applied a diffusion model on the BRATS2020 brain tumor MRI dataset to reconstruct abnormal regions into normal tissue, enabling precise segmentation of tumor areas and demonstrating the potential of diffusion models in medical imaging. However, for breast X-ray images, abnormal features are often less distinct than brain tumors and can be small in size, such as tiny calcifications. Therefore, the effectiveness of this approach in breast cancer detection requires further investigation.

2.3. Medical applications of federated learning

In real-world clinical scenarios, the issue of data silos is significant, as medical institutions are restricted from sharing data due to privacy regulations [20]. Additionally, data heterogeneity is prevalent—benign cases naturally outnumber malignant cases, leading to highly imbalanced labels. Moreover, differences in data collection and annotation protocols across institutions further complicate the application of machine learning [21]. As a result, individual institutions often struggle to train high-performance models due to limited and biased data availability. Federated learning offers a solution to these challenges by enabling each medical institution to train local models on their private data without sharing it [22]. After each training round, a central server aggregates the locally trained models to produce a more robust global model [22]. This approach allows for effective utilization of distributed data while preserving patient privacy and addressing issues related to data silos and heterogeneity. Reference [23] proposed an inside and outside model personalization federated learning (IOP-FL) framework, which enhances personalized federated learning by distinguishing between global and local gradients; this framework achieved performance improvements of 2.63% and 2.79%, respectively, compared to FedAvg [22].

Most existing studies artificially partition the same dataset across different clients to evaluate federated learning performance. However, in real-world scenarios, such an approach fails to fully capture the heterogeneity of medical data. Research on federated learning with truly heterogeneous datasets distributed across different clients remains limited, particularly in the domain of breast cancer detection.

3. Methodology

This study applies various machine learning approaches, including image-level and patch-level CNN-based classification, diffusion model-based anomaly detection, and federated learning, to conduct breast cancer detection using real-world datasets. We used two publicly available datasets, VinDr-Mammo [5] and CMMD [7], both annotated by multiple expert radiologists. A comprehensive analysis was conducted using these machine learning approaches.

Additionally, due to performance constraints, the input to neural networks is often resized, which results in a loss of fine details compared to the original images. To investigate whether the approach of partitioning images into smaller patches, commonly used in tissue analysis, is applicable to breast cancer detection, this study leverages pre-annotated suspicious regions from the CBIS-DDSM dataset [6]. These regions are considered as small patches and used for classification tasks with machine learning models. Then, to simulate a real-world scenario, we distributed the VinDr-Mammo dataset and the CMMD dataset across separate clients and evaluated the performance of the aggregated global model after federated training.

3.1. CNN-based classification task

3.1.1. Whole image-level classification

For image classification tasks, the most straightforward approach is to feed the entire image into a neural network, minimizing dataset requirements. Non-end-to-end methods, however, require manual annotation of regions of interest (ROIs), which is time-consuming, costly, and dependent on doctors' expertise [11]. Additionally, the limited availability of annotated datasets and the high cost of annotations make transferring models to other datasets more challenging [11].

Several CNN models are explored in this research. The vision Transformer (ViT) [24], which incorporates an attention mechanism [25], is one of the options. ViT splits the entire image into patches, generates features through an embedding layer, and then processes them through transformer encoder layers with attention mechanisms, ultimately generating predictions [24]. However, as shown in [24] and supported by our experiments, although ViT performs better on large datasets due to its complex multi-head attention mechanism and a greater number of parameters, it lacks the inherent inductive bias of CNNs, making it less effective on smaller datasets [24]. Therefore, ViT is not suitable for this study. Instead, a ResNet [26] variant with residual connection mechanism, ResNet-101, is chosen as the model is suitable for the classification task.

For each sample in the dataset, two views of the same side breast for each patient are captured, namely the MLO view and the CC view. To maximize the use of the dataset, both views are treated as separate samples and fed into the neural network during training. Due to its unique angle, the MLO view provides a comprehensive view of both the breast tissue and the underlying muscle tissue, often demonstrating a stronger correlation with the patient's actual condition compared to the CC views [5, 27]. Therefore, for evaluation, we combine the two views to leverage the additional information provided by both perspectives. Let $p_{\mathrm{CC}}^{(1)}$ and $p_{\mathrm{CC}}^{(0)}$ represent the predicted probabilities for class 1 (malignant) and class 0 (benign) from the craniocaudal (CC) view, respectively. Similarly, let $p_{\mathrm{MLO}}^{(1)}$ and $p_{\mathrm{MLO}}^{(0)}$ represent the predicted probabilities for class 1 (malignant) and class 0 (benign) from the mediolateral oblique (MLO) view. To combine the predictions for class 1 and class 0, we use a

weighted average with a factor α, where $0 \leq \alpha \leq 1$:
For class 1 (malignant):

$$p_{\text{final}}^{(1)} = \alpha \cdot p_{\text{CC}}^{(1)} + (1 - \alpha) \cdot p_{\text{MLO}}^{(1)}$$

For class 0 (benign):

$$p_{\text{final}}^{(0)} = \alpha \cdot p_{\text{CC}}^{(0)} + (1 - \alpha) \cdot p_{\text{MLO}}^{(0)}$$

3.1.2. Patch-level classification

In the tissue classification task, an image is divided into multiple patches. The image-level label is determined by the labels of its patches under the following condition. Let an image I be composed of N patches, where each patch i has an individual classification label y_i, where $y_i \in \{0, 1\}$ (0 for negative, 1 for positive). The image-level classification label Y is given by:

$$Y = \max_{i \in \{1,2,\ldots,N\}} y_i$$

where $y_i = 1$ if patch i is classified as positive, and $y_i = 0$ if patch i is classified as negative [28]. Thus, the image is classified as positive ($Y = 1$) if at least one patch is positive; otherwise, it is classified as negative ($Y = 0$) [28].

To explore the feasibility of this approach for breast X-ray images, we applied CNNs at the patch level to examine whether the model could distinguish between benign and malignant tissue without access to global contextual information. To achieve this, we utilized the CBIS-DDSM dataset [6], which includes expert-annotated ROI regions for each case. Each ROI is labeled with pathology (benign or malignant) and abnormality type (calcification or mass).

3.2. Diffusion model-based anomaly detection task

Diffusion models have shown strong performance in brain tumor MRI tasks [13]. Since diffusion models inherently model data distributions, they learn the distribution of normal samples and can generate normal samples from random noise. This suggests their potential for transforming unseen abnormal samples into normal ones, enabling anomaly detection by comparing differences. However, as generative models, their reliability and stability in medical applications remain a concern [29].

In this research, the denoising diffusion probabilistic model (DDPM) proposed by [30] is applied to the breast cancer detection domain, and its performance is evaluated. The diffusion process in this study follows the DDPM formulation, where Gaussian noise is progressively added to the sample using the forward process:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t;\ \sqrt{\alpha_t}x_{t-1}, (1 - \alpha_t)I) \tag{3.1}$$

where t is a hyperparameter chosen from the range $[0, T]$, controlling the noise level. x_t is the image at time step t, obtained by progressively adding noise to the original image over t steps. The model learns to denoise by estimating the noise at a given timestep t, following the reverse process:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{3.2}$$

where θ represents the parameters of the noise prediction model (U-Net). The U-Net takes the noisy image x_t and the current time step t as input, and generates a predicted noise $\epsilon_\theta(x_t, t)$ [30]. The goal is for the model to predict the noise such that subtracting it from the noisy image x_t results in a cleaner image. The diffusion loss is computed as defined in DDPM:

$$L(\theta) = \mathbb{E}_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2\right]. \tag{3.3}$$

This loss function compares the predicted noise ϵ_θ with the actual noise ϵ that was added during the forward process.

In the evaluation phase, a hyperparameter N is set to control the number of noise steps added to the test samples. This helps transform the potential malignant case distribution into a neutral distribution. The model, trained on benign cases, then uses the reverse process to restore the image regions. The complete process can be seen in Figure 1.

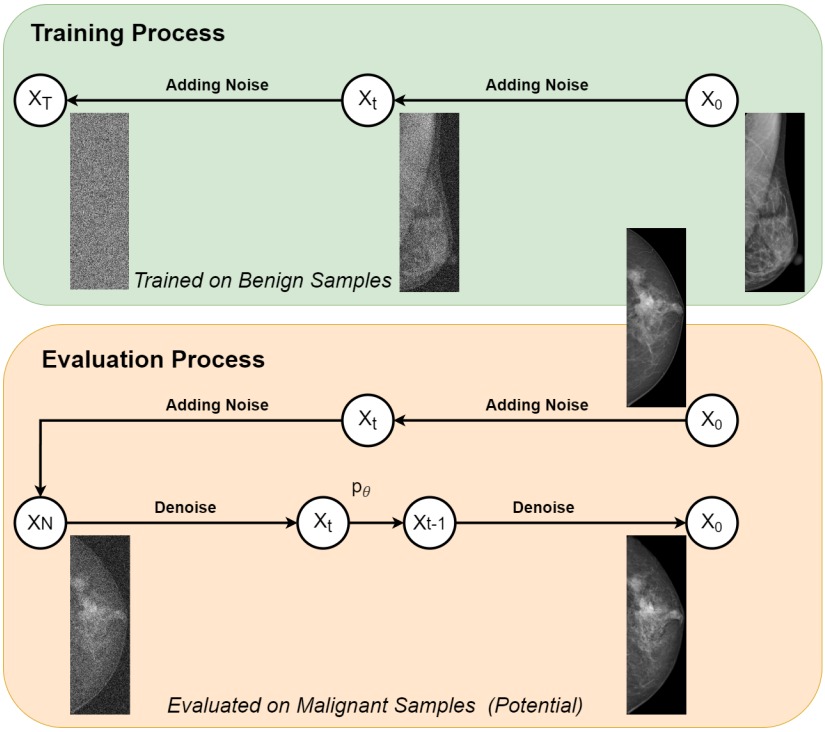

Figure 1. Overview of the diffusion process for anomaly detection in breast cancer imaging.

3.3. Federated learning for breast cancer detection

Finally, this research examines the use of federated learning to train models on data collected from multiple institutions. The study incorporates two datasets, VinDr-Mammo and CMMD, each assigned to a separate client. Each client only trains a local model using its respective dataset. The specific FL process is outlined in Algorithm 1.

Algorithm 1 Federated learning for breast cancer detection

Require: Initial global model θ_0, number of communication rounds T
Require: Clients A and B with datasets D_A and D_B
for each round $t = 1, 2, \ldots, T$ **do**
  for each client $i \in \{A, B\}$ **do**
    Sample dataset D_i from client i
    Train local model $\theta_i^{(t)}$ by solving:

$$\theta_i^{(t)} = \arg\min_{\theta_i} \left(\mathcal{L}_i(\theta_i) + \frac{\mu}{2} \|\theta_i - \theta_{t-1}\|^2 \right)$$

    Send local model update $\theta_i^{(t)}$ to the server
  end for
  Aggregate local updates at the server:

$$\theta_t = Agg(\theta_A^{(t)}, \theta_B^{(t)})$$

  Update global model $\theta_t \leftarrow \theta_t$
end for

At the beginning of the FL process, a global model is initialized and distributed to clients as the local model for round 0. As discussed in Section 3.1.1, we also selected ResNet-101 as the model used by each client. During each communication round in the FL process, the server aggregates the local models from the clients according to the aggregation algorithm and sends the updated model back to all clients for the next round of training. Additionally, the VinDr-Mammo and CMMD datasets exhibit differences in label distributions, which is to be explained in Section 4.1. To address the imbalance in data distribution, FedProx [31] is used to optimize the training process. FedProx incorporates a proximal term into each client's local updates to mitigate the impact of data distribution discrepancies [31].

4. Experiments, results, and discussion

4.1. Dataset

4.1.1. VinDr-Mammo dataset

The VinDr-Mammo dataset is a publicly available large-scale dataset from Vietnam, which comprises X-ray images of both side breasts (MLO and CC views) of 5000 patients, totaling 10,000 samples [5]. Three radiologists evaluated each image to determine the level of BI-RADS, which ranges from 1 to 5 [5]. While the dataset is relatively large compared to others, it still suffers from an imbalanced label distribution. The label distribution is shown in Figure 2. It can be seen that limited amount malignant samples are available in the dataset.

For different tasks, we used varying data selection strategies. For the classification task, since the CMMD dataset only includes benign and malignant labels, we aligned the training labels with the

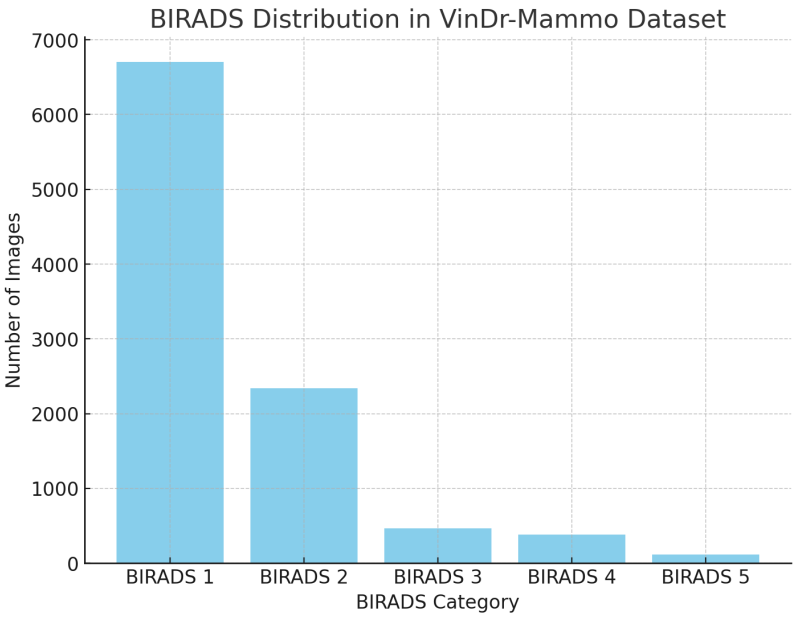

Figure 2. The data distribution of VinDr-Mammo dataset.

CMMD dataset. Specifically, BI-RADS 2 was selected as benign, and BI-RADS 4 and 5 were selected as malignant. BI-RADS 3 was typically excluded in the classification task, as it is often considered unreliable, based on the guidelines provided by the VinDr-Mammo dataset [5]. Therefore, for the classification task using the VinDr-Mammo dataset, there are 494 positive samples and 2338 negative samples. For the anomaly detection task, the model is only trained using normal samples as input, resulting in 5362 training samples (BI-RADS 1 only) and 494 abnormal test samples (BI-RADS 4 and BI-RADS 5).

To prevent precision loss, the raw data of VinDr-Mammo dataset, which has a 16-bit depth, is loaded with the same 16-bit depth. Additionally, each image is processed using computer graphics techniques to crop the blank areas outside the tissue regions.

4.1.2. CMMD dataset

The Chinese Mammography Database (CMMD) is collected from the Sun Yat-sen University Cancer Center in Guangzhou and the Nanhai Affiliated Hospital of Southern Medical University in Foshan [7]. For each subject, CC view and MLO view images were obtained and stored in 8-bit format [7]. The dataset includes 556 benign cases and 1316 malignant cases. Before fed into training, each image undergoes computer graphics algorithms to remove blank areas.

4.1.3. CBIS-DDSM dataset

The CBIS-DDSM dataset, which includes accurate region-of-interest (ROI) annotations by professional radiologists, is utilized in this study to analyze the feasibility of patch-level classification for breast cancer detection. In addition to labeling patients with suspicious anomaly as benign or malignant, the CBIS-DDSM dataset also provides annotations for abnormal region subtypes (see Table 1). This allows patch-level classification to be conducted from multiple perspectives.

Table 1. Data distribution of the CBIS-DDSM dataset.

	Benign	Malignant	**Total**
Calcification	1199	673	1872
Mass	912	784	1696
Total	2111	1457	3568

Table 2. Comparison of performance between different methods and datasets.

Method	**Dataset**	**Acc**	**Macro-F1**	**AUC**
Kim et al. (2018)	CBIS	–	–	0.64 ± 0.00
ES-Attside	VinDr	–	–	0.83 ± 0.00
ResNet-101	VinDr	0.81 ± 0.03	0.65 ± 0.02	0.75 ± 0.03
ResNet-101	CMMD	0.68 ± 0.02	0.63 ± 0.01	0.77 ± 0.02
ResNet-101	VinDr+CMMD	0.79 ± 0.02	0.79 ± 0.01	0.85 ± 0.02
FL Approach	VinDr+CMMD	0.78 ± 0.02	0.77 ± 0.01	0.82 ± 0.02

4.2. CNN-based classification task

4.2.1. Whole image-level classification results

In the experiment, we evaluated the performance of neural networks on two datasets locally, as shown in Table 2. We then selected existing works in the field of breast cancer classification, namely [10] and [32], for comparison. Most of the current work primarily focuses on the AUC metric, and therefore, only the AUC values are shown. From the first two ResNet-101 records in the table, it can be observed that although CNN achieves high accuracy on certain datasets, due to the imbalance in the dataset, the neural network struggles to differentiate the minority class effectively, as seen from the F1 score and AUC. The experimental results indicate that, under conditions of data imbalance, models trained by a single medical institution using only local private data may perform poorly. However, in a federated learning scenario, despite challenges such as data heterogeneity and non-IID [33] issues, the availability of a global model for training guidance has the potential to alleviate the data silo problem. The result of the FL approach will be illustated in Section 4.4.

4.2.2. Patch-level classification results

In the study of patch-level classification tasks on the CBIS-DDSM dataset, significant variations in model performance were observed depending on the task, as shown in Table 3. In the benign vs. malignant classification task, challenges arise due to the absence of global image context. When determining the patient's condition solely based on suspicious regions, the model may struggle. For instance, in cases of patients with high breast density, such as obese individuals, suspicious regions tend to exhibit similarly high density. This can lead the model to erroneously classify the patient as malignant based on the large density of the patch. This highlights certain limitations of using breast X-rays for breast cancer screening.

We also conducted a study on tissue subtype classification, where the model achieved an AUC of 0.92, which is much higher than that of the previous task. This suggests that one potential direction

for future research is to use a patch-level classifier to first distinguish tissue subtypes and then perform further classification tasks downstream. For example, when the patch-level classifier identifies regions with clear calcifications, which are relatively easier to be classified, a more confident classifier could be applied. For regions with masses, a downstream classifier with higher redundancy could be utilized to achieve a lower false negative rate as dense tissues such as adipose may obstruct the X-ray [1], leading to missed diagnoses.

Table 3. Performance of different classification tasks at the patch level.

Task	Acc	F1	AUC
Benign vs malignant	0.65 ± 0.03	0.60 ± 0.02	0.71 ± 0.01
Calcification vs mass	0.83 ± 0.01	0.83 ± 0.01	0.92 ± 0.01

4.3. Diffusion model results

The diffusion model has shown good performance in other medical imaging domains. However, its performance in breast cancer anomaly detection tasks is not ideal. We present both quantitative and qualitative analyses to illustrate the experimental results. In the qualitative analysis, we selected test data to reconstruct the abnormal regions and visualized the reconstruction results using a difference heatmap. Ideally, the reconstructed abnormal regions should show significant differences compared to the original image. However, the experimental results did not reveal clear differences, with only some subtle variations observed at the contours of the abnormal regions, which can be seen in Figure 3. When using the reconstruction difference as a feature for classification, the model achieved only a 0.55 F1 score, which is significantly lower than the baseline performance. A possible reason for this is that while being capable of generating high-quality synthetic breast X-ray samples from random noise, the U-Net model used for noise prediction is not sufficient to capture the differences between the abnormal and normal distributions. Additionally, the task of reconstructing regional details is inherently challenging. A more feasible approach could be to focus on reconstructing the mask of the abnormal regions [13].

4.4. Federated learning approach

As shown in Table 2, we evaluate the performance of the FL approach by distributing different data sets between different clients for federated learning. The results indicate that even when data distribution is imbalanced within a single client, the aggregated global model performs well in breast cancer classification tasks. In particular, our FL model delivers results that closely parallel the performance of the centralized ResNet-101 model trained on the combination of the two datasets. This suggests that FL can effectively mitigate the challenges of data non-sharing across institutions in medical image recognition.

However, despite efforts to simulate inter-institutional heterogeneity, the study has limitations, primarily due to the small number of available datasets, which constrains baseline model performance. As shown in Figure 4, after a certain number of training epochs, the model's performance begins to degrade while the training loss continues to decrease, indicating severe overfitting due to the limited dataset size. This highlights the critical need for larger and more diverse datasets in medical image analysis.

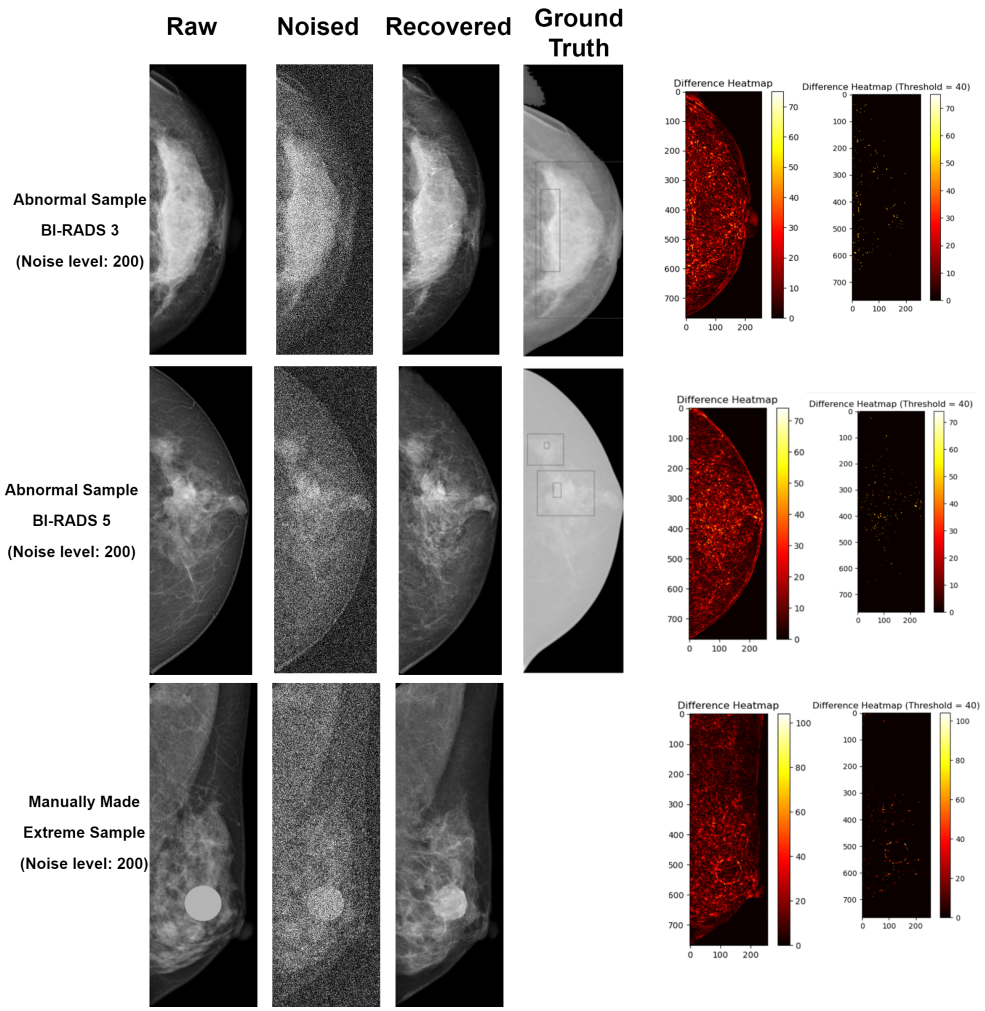

Figure 3. Examples of qualitative results from the diffusion model.

5. Conclusions

This study applies multiple machine learning approaches for breast cancer detection using multiple medical image datasets. Our experiments demonstrated that FL can improve detection performance while addressing data privacy concerns and non-IID distributions across institutions. By aggregating local models from different clients, FL enabled the creation of a global model that achieved higher performance. Nonetheless, several challenges remain in breast cancer detection. Since X-rays serve primarily as an early screening tool, suspicious tissues often lack clear signs of abnormality. Additionally, crucial diagnostic factors, such as patient age, are considered private information and are unavailable in research datasets, further impacting model performance. Despite these limitations, FL has shown significant promise in breast X-ray imaging. Given that breast X-rays are a cost-effective and widely accessible screening method [4], vast amounts of potential data exist across medical institutions. FL offers a promising solution to harness this distributed data, facilitating the development of more accurate and efficient diagnostic models for breast cancer detection.

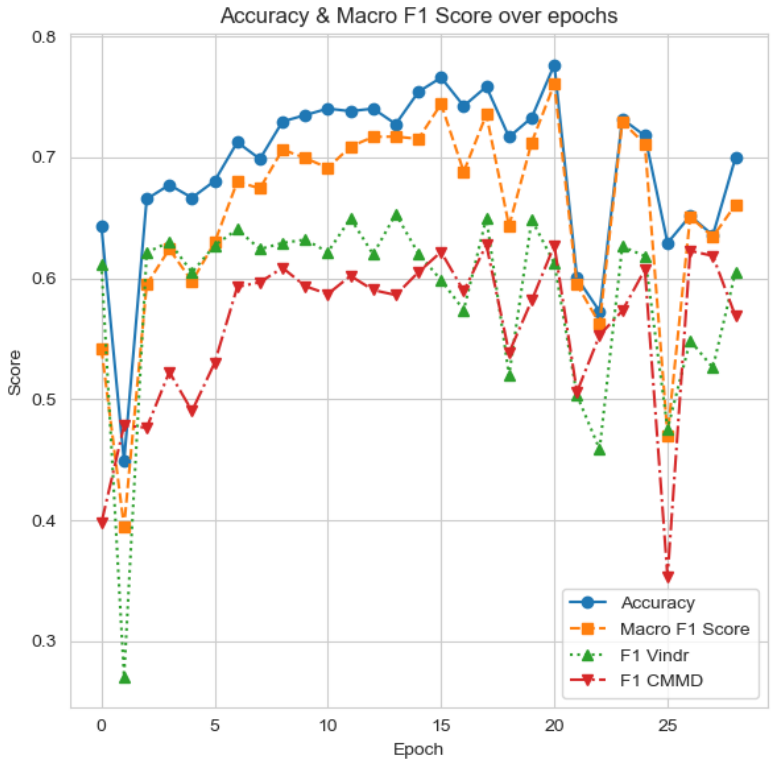

Figure 4. Overfitting problem is observed through training process.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work is partly supported by the British Council under the UK-Saudi Challenge Fund.

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. Karellas A, Vedantham S, (2008) Breast cancer imaging: A perspective for the next decade. *Med Phys* 35: 4878–4897. https://doi.org/10.1118/1.2986144
2. Zhang Y, Xia K, Li C, Wei B, Zhang B, (2021) Review of breast cancer pathologigcal image processing. *BioMed Res Int* 2021: 1994764. https://doi.org/10.1155/2021/1994764
3. Fishman MDC, Rehani MM, (2021) Monochromatic X-rays: The future of breast imaging. *Eur J Radiol* 144: 109961. https://doi.org/10.1016/j.ejrad.2021.109961

4. Coleman C, (2017) Early detection and screening for breast cancer, In: *Seminars in Oncology Nursing*, 33: 141–155. https://doi.org/10.1016/j.soncn.2017.02.009

5. Nguyen HT, Nguyen HQ, Pham HH, Lam K, Le LT, Dao M, et al. (2023) VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *Sci Data* 10: 277. https://doi.org/10.1038/s41597-023-02100-7

6. Lee RS, Gimenez F, Hoogi A, Miyake KK, Gorovoy M, Rubin DL, (2017) A curated mammography data set for use in computer-aided detection and diagnosis research. *Sci Data* 4: 1–9. https://doi.org/10.1038/sdata.2017.177

7. Cai H, Wang J, Dan T, Li J, Fan Z, Yi W, et al. (2023) An online mammography database with biopsy confirmed types. *Sci Data* 10: 123. https://doi.org/10.1038/s41597-023-02025-1

8. Balleyguier C, Ayadi S, Van Nguyen K, Vanel D, Dromain C, Sigal R, (2007) BIRADS™ classification in mammography. *Eur J Radiol* 61: 192–194. https://doi.org/10.1016/j.ejrad.2006.08.033

9. Tang W, Zhou F, Huang S, Zhu X, Zhang Y, Liu B, (2024) Feature re-embedding: Towards foundation model-level performance in computational pathology, In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11343–11352. https://doi.org/10.1109/CVPR52733.2024.01078

10. Pathak S, Schlötterer J, Geerdink J, Veltman J, van Keulen M, Strisciuglio N, et al. (2023) Case-level breast cancer prediction for real hospital settings, preprint, arXiv:2310.12677. https://doi.org/10.48550/arXiv.2310.12677

11. Shen L, (2017) End-to-end training for whole image breast cancer diagnosis using an all convolutional design, preprint, arXiv:1711.05775. https://doi.org/10.48550/arXiv.1711.05775

12. Liao L, Aagaard EM, (2024) An open codebase for enhancing transparency in deep learning-based breast cancer diagnosis utilizing CBIS-DDSM data. *Sci Rep* 14: 27318. https://doi.org/10.1038/s41598-024-78648-0

13. Wolleb J, Bieder F, Sandkühler R, Cattin PC, (2022) Diffusion models for medical anomaly detection, In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 35–45. https://doi.org/10.1007/978-3-031-16452-1_4

14. Zhu J, Ding C, Tian Y, Pang G, (2024) Anomaly heterogeneity learning for open-set supervised anomaly detection, In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17616–17626. https://doi.org/10.1109/CVPR52733.2024.01668

15. Rahman A, Valanarasu JMJ, Hacihaliloglu I, Patel VM, (2023) Ambiguous medical image segmentation using diffusion models, In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11536–11546. https://doi.org/10.1109/CVPR52729.2023.01110

16. Ribli D, Horváth A, Unger Z, Pollner P, Csabai I, (2018) Detecting and classifying lesions in mammograms with deep learning. *Sci Rep* 8: 4165. https://doi.org/10.1038/s41598-018-22437-z

17. Bandi P, Geessink O, Manson Q, Van Dijk M, Balkenhol M, Hermsen M, et al. (2018) From detection of individual metastases to classification of lymph node status at the patient level: The camelyon17 challenge. *IEEE Trans Med Imaging* 38: 550–560. https://doi.org/10.1109/TMI.2018.2867350

18. Lingle W, Erickson BJ, Zuley ML, Jarosz R, Bonaccio E, Filippini J, et al. (2016) The cancer genome atlas breast invasive carcinoma collection (TCGA-BRCA). *Cancer Imaging Arch* 2016. https://doi.org/10.7937/k9/tcia.2016.ab2nazrp

19. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. (2021) Learning transferable visual models from natural language supervision, In: *International Conference on Machine Learning*, 139: 8748–8763.

20. Xu A, Li W, Guo P, Yang D, Roth HR, Hatamizadeh A, et al. (2022) Closing the generalization gap of cross-silo federated medical image segmentation, In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20866–20875. https://doi.org/10.1109/CVPR52688.2022.02020

21. Guan H, Yap PT, Bozoki A, Liu M, (2024) Federated learning for medical image analysis: A survey. *Pattern Recognit* 2024: 110424. https://doi.org/10.1016/j.patcog.2024.110424

22. McMahan B, Moore E, Ramage D, Hampson S, y Arcas BA, (2017) Communication-efficient learning of deep networks from decentralized data, In: *Artificial Intelligence and Statistics*, 54: 1273–1282.

23. Jiang M, Yang H, Cheng C, Dou Q, (2023) IOP-FL: Inside-outside personalization for federated medical image segmentation. *IEEE Trans Med Imaging* 42: 2106–2117. https://doi.org/10.1109/TMI.2023.3263072

24. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. (2020) An image is worth 16x16 words: Transformers for image recognition at scale, preprint, arXiv:2010.11929.

25. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. (2017) Attention is all you need. *Adv Neural Inf Process Syst* 30.

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

27. Gupta S, Zhang D, Sampat MP, Markey MK, (2006) Combining texture features from the MLO and CC views for mammographic CADx, In: *Medical Imaging 2006: Image Processing*, 6144: 1877–1885. https://doi.org/10.1117/12.657023

28. Shao Z, Bian H, Chen Y, Wang Y, Zhang J, Ji X, et al. (2021) Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Adv Neural Inf Process Syst* 34: 2136–2147.

29. Singh NK, Raza K, (2021) Medical image generation using generative adversarial networks: A review. *Health Inf Comput Perspect Healthcare* 2021: 77–96. https://doi.org/10.1007/978-981-15-9735-0_5

30. Ho J, Jain A, Abbeel P, (2020) Denoising diffusion probabilistic models. *Adv Neural Inf Process Syst* 33: 6840–6851.

31. Li T, Sahu AK, Zaheer M, Sanjabi M, Talwalkar A, Smith V, (2020) Federated optimization in heterogeneous networks, In: *Proceedings of Machine Learning and Systems*, 2: 429–450.

32. Khan HN, Shahid AR, Raza B, Dar AH, Alquhayz H, (2019) Multi-view feature fusion based four views model for mammogram classification using convolutional neural network. *IEEE Access* 7: 165724–165733. https://doi.org/10.1109/ACCESS.2019.2953318

33. Zhao Y, Li M, Lai L, Suda N, Civin D, Chandra V, (2018) Federated learning with non-iid data, preprint, arXiv:1806.00582. https://doi.org/10.48550/arXiv.1806.00582

AIMS Press