Edge detection and graph neural networks to classify mammograms: A case study with a dataset from Vietnamese patients

Article *in* Applied Soft Computing · December 2022

DOI: 10.1016/j.asoc.2022.109974

CITATIONS
3

READS
166

5 authors, including:

Linh Duong Tuan
KTH Royal Institute of Technology - Sweden
24 PUBLICATIONS **283** CITATIONS
SEE PROFILE

Phuong Nguyen
Università degli Studi dell'Aquila
79 PUBLICATIONS **1,035** CITATIONS
SEE PROFILE

Edge detection and graph neural networks to classify mammograms: A case study with a dataset from Vietnamese patients

Linh T. Duong[a,*], Cong Q. Chu[b,*], Phuong T. Nguyen[c,**], Son T. Nguyen[b], Binh Q. Tran[d]

[a] *Institute of Research and Development, Duy Tan University, Vietnam*
duongtuanlinh@duytan.edu.vn
[b] *Hanoi Oncology Hospital, Vietnam*
{congcq, sonnt}@bvubhn
[c] *Department of Information Engineering, Computer Science and Mathematics*
University of L'Aquila, 67100 - L'Aquila, Italy
phuong.nguyen@univaq.it
[d] *National Institute of Nutrition, Ministry of Health, Vietnam*
tranquangbinh@dinhduong.org.vn

Abstract

Mammograms are breast X-ray images and they are used by doctors, among other purposes, as an effective means of detecting breast cancer. Screening mammography is crucial since it allows doctors to understand better the situation and have suitable intervention. The classification of medical modalities is a prerequisite for development of computer-aided diagnosis tools in healthcare, and various techniques have been proposed to automatically classify from mammography images. Though there have been several tools developed, they have been mostly validated with data collected from Western women. Based on our initial investigations, breast anatomy in Vietnamese women differs from that of Western women, due to denser breast tissue. In this paper, to tackle the issue of detecting mammograms, we propose MammoGNN – a novel approach using the synergy between image processing techniques and graph neural networks. To validate the conceived tool, we curated a mammogram dataset from 2,351 Vietnamese women. MammoGNN obtains a maximum accuracy of 100%

*The first two authors contributed equally to the paper
**Corresponding author

on independent and shuffle test sets for both classification of BI-RADS scores and breast density types. The experimental results also demonstrate that our proposed approach outperforms different baselines. We anticipate that the proposed approach can be deployed as a non-invasive pre-screening tool to assist doctors in performing their diagnosis activities.

1. Introduction

Various studies show that breast cancer is among the most common cause of deaths worldwide, and mortality rate of the disease has increased significantly in different countries in recent years [1, 2, 3]. Given the circumstances, the early detection of the disease in women with no symptoms is crucial, as it helps doctors have timely intervention. In this respect, a mammogram is an X-ray image of the breast, and doctors use mammograms to observe early signs of breast cancer. In fact, screening mammography modality in population has been demonstrated to help reduce breast cancer mortality by 40–63% [4].

Monitoring breast cancer with the support of Medical Imaging techniques is beneficial to the diagnosis and prognosis of the disease. In recent years, deep learning has enabled greatly increased accuracies for the computer aided detection of different diseases from medical images [5, 6, 7, 8, 9]. For one specific task, screening for diabetic retinopathy, it has already led to the first FDA (United States of America Food and Drug Administration) approved system for fully automated diagnosis. Similar systems are currently being developed for many other tasks, and will soon become mature enough to be deployed in practice. Despite the benefits, such a method suffers a low accuracy [10]. Although many computer-assisted detection and diagnosis (CAD) software have been developed and in clinical use since last three decades, various studies reveal that these CAD systems contribute to only a small gain in prediction accuracy of screening mammograms [11, 12, 13, 14].

While the incidence rate of breast cancer in Vietnam is low compared to that in Western countries, it has been worryingly increasing in recent years [15]. Moreover, though there have been several automatic tools developed, most of

them have been dedicated to work on data from Western women. Based on our initial investigations, breast anatomy in Vietnamese women differs from that of Western ones, due to denser breast tissue. These differences in tissue density could increase *(i)* the chance that breast cancer may go undetected during screening; and *(ii)* the risk of getting breast cancer. This triggers the need for pre-screening facilities to conduct early diagnosis for Vietnamese women.

An important aspect is that deep learning based methods have to be trained on large patient datasets. It is a well-known limitation of these techniques that they currently have a limited ability to generalize to populations whose characteristics differ from those they were trained on, e.g., due to anatomical differences related to gender or ethnicity. This is closely related to the fact that many of these approaches are "black boxes" that neither provide insight on which factors drive their decision, nor have dedicated mechanisms to assess their own ability to classify specific images.

In this paper, we propose MammoGNN, a practical approach to the classification of mammograms, exploiting advanced image processing techniques and graph neural networks [16]. First, we employ edge detection algorithms to transform original mammograms into a suitable format. Afterwards, various graph neural networks are used to classify the input data. To evaluate MammoGNN, we curated a mammography dataset from Vietnamese women who had been screened at Hanoi Hospital Oncology. The dataset was collected and labelled by three experienced radiologists following the existing guidance [17]. The evaluation shows that the proposed framework obtains promising results both with respect to efficiency and effectiveness. Altogether, we aim at developing an app that works as a non-invasive pre-screening tool, serving doctors at large. Such a tool is of highly importance for many areas in Vietnam, where there exists a lack of well-trained specialists.

In this respect, our paper makes the following contributions:

- A practical approach to transform images into a GNN-processable format using the Prewitt algorithm, being able to dramatically reduce computational expense;

- A classification engine for mammograms exploiting three state-of-the-art graph neural network technologies, namely GCN [18], GAT [19], and GraphConv [20]. To the best of our knowledge, our work is the first attempt to deploy GNN to classify mammograms;

- A new full-field Digital Mammography dataset,[1] collected from more than 2,000 Vietnamese women and labeled by three expert radiologists, following the ACR BI-RADS® Atlas 5^{th} Edition standard [17].

The paper is organized as follows. Section 2 provides background related to mammography and graph neural networks. Afterwards, Section 3 presents in detail the proposed MammoGNN approach. In Section 4, we explain the dataset and metrics used for evaluation. The experimental results are reported and analyzed in Section 5. The section also lists the probable threats to validity of the outcomes. Section 8 reviews the related work, and finally Section 9 sketches future work and concludes the paper.

2. Background

This section introduces the background related to the classification of mammograms and graph neural networks.

2.1. Classification of mammography

Breast Imaging-Reporting and Data System (BI-RADS) is a risk assessment and quality assurance tool developed by American College of Radiology to standardize breast imaging reporting. The guideline facilitates outcome monitoring and quality assessment to mammography, ultrasound, and MRI imaging [17]. Following BI-RADS, there exists a classification of various categories which are shown in Table 1.

[1]The dataset is available upon request. Please contact the first author (duongtuanlinh@duytan.edu.vn) or the corresponding author (phuong.nguyen@univaq.it) for more detail.

Table 1: Categories of BI-RADS.

Name	Description
BI-RADS 0	Incomplete information and needs additional imaging evaluation such as mammographic views or ultrasound, and/or for mammography
BI-RADS 1	Negative symmetrical and no masses, architectural distortion, or suspicious calcification regions
BI-RADS 2	Assessed benign tumors, but with 0% probability of malignancy
BI-RADS 3	Possibly benign tumors, but less than 2% probability of malignancy. The subjects need examining in short interval follow-up around 6 months
BI-RADS 4	Suspicious for malignancy with from 2% to 94% probability of malignancy for mammography and ultrasound, these can be further divided into one of three assessment sub-categories:
BI-RADS 4A	Low suspicion for malignancy with probabilities from 2% to 9%
BI-RADS 4B	Moderate suspicion for malignancy with probabilities from 10% to 49%
BI-RADS 4C	High suspicion of malignancy with probabilities from 50% to 94%, biopsy examination should be considered
BI-RADS 5	Highly suggestive, with a probability of malignancy higher than 95%, and patients need to have appropriate treatment
BI-RADS 6	Biopsy-proven malignancy

When there are multiple findings, the BI-RADS category for an exam is assigned the highest one in the following hierarchy, from the lowest to the highest scores: 1, 2, 3, 6, 0, 4, 5. Similarly, the assignment of breast density type or composition is classified by taking into account the chance that a mass can be obscured by fibroglandular tissue following guidance of the BI-RADS edition 2013 [17]. It is categorized into one of four assessments followed by a description as shown in Table 2.

In heterogeneously dense (Type C) or extremely dense (Type D), the incidence of invasive ductal carcinoma is higher than that in other groups of breast density, likely as a combination of the amount of gland present as well as possible observation error due to the lower sensitivity of detection. Therefore, screening ultrasound and contrast-enhanced Magnetic Resonance Imaging

Table 2: Types.

Name	Description
A	Breasts are completely fatty. Thus, mammography is highly sensitive in this setting
B	There are scattered areas of fibroglandular density. The term density describes the degree of X-ray attenuation of breast tissue, but not discrete mammographic findings
C	Breasts are heterogeneously dense, which may obscure small masses. Some areas in the breasts are sufficiently dense to obscure small masses
D	Breasts are extremely dense, which lowers the sensitivity of mammography

(MRI) techniques are potentially beneficial as an adjunct to screening mammography.

2.2. Graph neural networks

Deep Learning techniques such as Convolutional Neural Networks, Recurrent Neural Networks, autoencoders, or transformers [21] are successful in dealing with hidden pattern of Euclidean geometry data, e.g., images, texts, or videos. However, there is another type of data called non-Euclidean geometry, and such data is represented as graphs with complex relationship and interdependencies between objects. For this type of data, the aforementioned deep learning techniques appear to be less applicable.

A graph is a structure to analyze the pair-wise relationship between objects and entities by means of two components, i.e., *vertices* and *edges*, $G = (V, E)$, where V is a set of *vertices* and E is the edges between *nodes*. If there is a directional dependency between nodes occur, then the edges are directed; otherwise, edges are undirected. Correspondingly, graph neural networks have been developed to deal with this type of data [22].

A graph is often represented as an adjacency matrix $\mathbf{A}$ of size $(N \times N)$, where N is the number of nodes. If each node is characterized by a set of M features, then a dimension of feature matrix X is $(N \times M)$. Graph-structured data is complex, and thus it brings a lot of challenges for existing machine

learning algorithms. Hence, GNNs are designed to perform inference on data described by graphs. By simplifying the problems into simpler representations or transform them into representations from different perspectives, GNN models enable to solve complex problems. In fact, GNNs are a good way to deal with abstraction concepts such as relationships and interactions. As how GNNs are named, they contribute to a convenient method for edge-level, nodel-level, and graph-level classification task, graph visualization and graph clustering.

We call u and v as two nodes in a graph, and x_u and x_v their corresponding feature vectors. The encoder function $Enc(u)$ and $Enc(v)$ convert the feature vectors to z_u and z_v.

$$h_v^0 = X_v(feature\_vector) \tag{1}$$

$$h_v^k = \sigma(W_k \sum \frac{h_v^{k-1}}{|N(v)|} + B_k h_v^{k-1}) \tag{2}$$

where $W_k \sum \frac{h_v^{k-1}}{|N(v)|}$ is the average of all the neighbors of node v, while $B_k h_v^{k-1}$ is the previous layer embedding of node v augmented with B_k – a trainable weight matrix and it represents a self-loop activation for node v.

In the k^{th} encoder layer, given a target node i, the importance of a neighboring node $j \in \Gamma$ is computed as:

$$e_{(k)}^{ij} = \sigma({a_s^{(k)}}^T W^{(k)} h_i^{k-1} + {a_r^{(k)}}^T W^{(k)} h_j^{k-1}) \tag{3}$$

where $W^{(k)}$ represents the trainable parameters of the k^{th} encoder layer. The importance coefficients of a node comparable is normalized by the softmax function:

$$\alpha_{ij}^{(k)} = \frac{exp(e_{ij}^{(k)})}{\sum_{l \in N_i} exp(e_{ij}^{(k)})} \tag{4}$$

where N_i represents the neighboring nodes of node i. Equation 4 normalizes the sum of the importance scores of all neighboring nodes to 1. The final representation of node i output by the k^{th} encoder layer is computed using the following formula:

$$h_i^{(k)} = \sum_{j \in N_i} \sigma(\alpha_{ij}^{(k)} W^{(k)} h_j^{(k-1)}) \tag{5}$$

By *link-level prediction*, it is necessary to represent the relationship among nodes in graph and predict if there is a connection between two entities. Meanwhile in *node-level classification*, the task is to understand node embedding for every node in a given graph by looking at the labels of their neighbors. With *graph-level classification*, the entire graph needs to be classified into suitable categories. This is similar to image classification, yet the target substitutes for the graph domain. Graph classification can be applied in a wide range of both academic and industrial problems, e.g., classifying a molecular structure into a meaningful category.

Graph clustering deals with the grouping of data represented in the form of graphs. There are two different forms of clustering performed on graph-structured data. The nodes of the graph can be clustered into groups of densely connected regions based on either edge weights or edge distance by seeking vertex clustering. And the graph clustering form analyze the graphs as the objects to be clustered and clusters these objects based on similarity.

Using popular CNNs and or Transformers, models can recognize objects in images and videos, obtaining promising accuracy [23, 24, 25, 26]. In the scope of this paper, we employ different GNN techniques following the graph classification paradigm to the detection of mammograms. The next section introduces our proposed approach in detail.

3. MammoGNN: A practical solution to classification of mammography

This section introduces the proposed approach to the classification of BI-RADS, based on image processing and three Graph Neural Network techniques, including GCN [18], GAT [19] and GraphConv [20]. The preprocessed and

datasets and the code used for this research are made online available to facilitate future research.[2]

3.1. Data transformation

To provide inputs for GNN, it is necessary to transform the input images into a computable format. All original images with the DICOM (Digital Imaging and Communications in Medicine) format are extracted metadata, converted to lighter format as Joint Photographic Experts Group (JPEG) and then resized to the size of $512 \times 512 \times 1$.

An edge represents a local change of intensity in an image which implies that a local minina or maxima will appear for the change of intensity of the edge region. In this study, we make use of the Prewitt algorithm [27] and its variant [28], which are spatial filters that can detect contours, to extract edged images. Following convolution of each 3×3 sub-matrix by filters on both the horizontal and vertical axes, gradient for each sub-matrix has been validated. Due to conversion of all grayscale images, each pixel can be changed in a edge if the gradient magnitude exceeds the gray value of 128 [28].

The edge maps are then used to build proper graphs for each dataset. Afterward, the Prewitt algorithm [27, 28] is applied to preprocess on a resized image, which is then converted into graph data by following the steps below:

- Every pixel being grayscale intensity value equal to or higher than 128 is counted as a node or a graph point. This means that nodes only locate on the important edges of the edge transformed image. A node feature contains the intensity of the relative pixel;
- An edge connects between to vertexes which map neighboring pixels in the original images;
- A graph is constructed from a single image. This implies that all the vertexes and edges as well are formed from a relevant image the property

[2] https://github.com/linhduongtuan/GraphMammoNet

of the same graph. Attributes of the vertex knowing as grayscale intensity values are normalized graph-wise. Normalization operator is performed by subtracting the mean of all attributes under every graph from the original value of preprocessed image and then dividing it by the standard deviation. While nodes are connected together via edges in a preprocessed image instead of the entire image, so graph datasets are composed from lower dimensional data. FFDM data is classified following BI-RADS, and breast density type consists of contrast regions and shape, transformed edges can be different, this leads to complexion of every represent graph. Indeed, such morphology of each graph is able useful for classification using Graph Neural Networks. Finally, thanks to PyG [29], its build-in functions help us construct five types of datasets including *Adjacency*, *Node attribute*, *Nodel label*, *Graph indicator*, and *Graph label*, representing the graph data of all the FFDMs.

3.2. A GNN-based classification engine

In this section, we describe the classification engine which is built on top of three graph neural networks (GNNs), namely GATConv [19], GCNConv [18] and GraphConv [20] presented as follows.

- Graph attention networks (GATs) [19] are applied to operate on graph-structured data, leveraging masked self-attentional layers to address the limitations of prior methods based on graph convolutions or their approximations. By stacking layers in which nodes are able to attend over their neighborhoods' features, specifying different weights to different nodes in a neighborhood can be performed, without resorting to costly matrix operation, or knowing the graph structure in advance. So several key challenges of spectral-based graph neural networks simultaneously are tackled, and making model-based GATs readily applicable to inductive as well as transductive problems.

- Another approach for graph-structure data, an efficient variant of convolutional neural networks GCNConv [18] has been introduced to learn

directly on graphs. The convolutional architecture enables to operate a localized first-order approximation of spectral graph convolutions. The algorithm scales linearly in the number of graph edges and learns hidden layer representations that encode both local graph structure and features of nodes.

- GraphConv [20] consists of a layer architecture originated from a theoretical point of view and is related to the 1-dimensional Weisfeiler-Leman graph isomorphism heuristic (1-WL) [30]. A generalization of 1-GNNs, so-called k-GNNs, has been developed based on the k-WL. Its layer can take higher-order graph structures at multiple stack into statement. These higher-order structures play a crucial role in the characterization of social networks, molecule graphs and image recognition as well.

3.3. Network architecture

One of the most distinctive characteristics of a deep learning architecture is the application of neural networks with several layers. Surprisingly, most Graph Neural Network are quite shallow with just only several layers. To explain the circumstance of Graph Neural Network, over-smoothing and bottleneck are observed. The former is the phenomenon of the node features being likely to converge to the same vector and get nearly indistinguishable as the result of using many GCNConv layers [31, 32, 33]. The latter is resulted from information from exponentially many neighbors in fixed-size vectors [34]. Therefore, we propose Graph Neural Network with only three layers, and we construct three GNN models based on backbones of either GATConv [19], or GCNConv [18], or GraphConv [20]. Input and output channels at the first GNN layer are equal to the number of node features of relevant input data and 256. The next two input and output channels of the other GNN layers are set as 256. Following two first GNN layers, there are activation function of ReLu [35] and Dropout [36] of 0.5, whereas the last GNN layer is connected to ReLu activation and a linear layer with input of 256 and output as numbers of categories. In detail, the configuration of our proposed models is shown in Fig. 1 and in Table **??**.

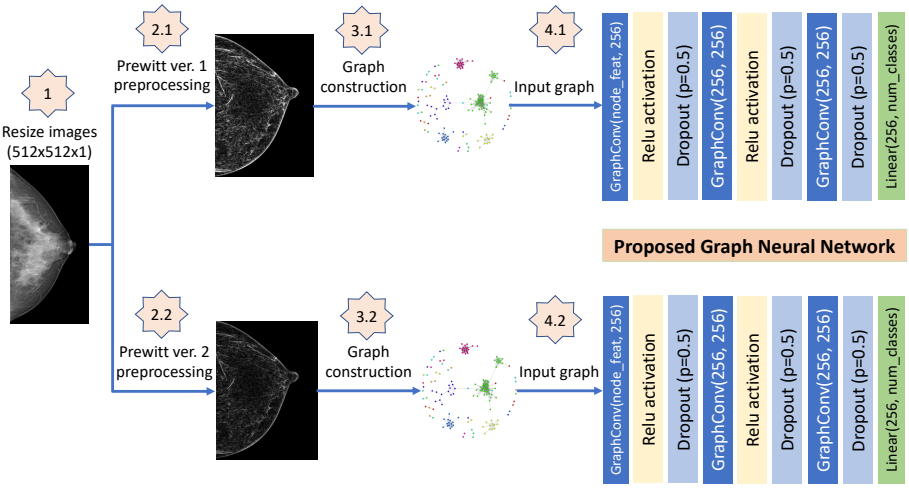


Figure 1: Architecture of the proposed MammoGNN model.

Table 3: Configuration of our proposed MammoGNN.

Stage	Layer Fuction Name	Parameter Numbers
1	GAT/GCN/GraphConv	Numbers of node features
2	ReLU activation	Not applicable
3	Dropout	0.5
4	GraphConv	256 -> 256
5	ReLU activation	Not applicable
6	Dropout	0.5
7	GAT/GCN/GraphConv	256 -> 256
8	Dropout	0.5
9	Linear	256 -> number classes

4. Evaluation materials and methods

This section explains the datasets and settings used to evaluate our proposed approach. First, the research questions to study the systems' characteristics are listed in Section 4.1. Afterward, Section 4.2 gives an overview of the datasets that we curated from individuals screened at Hanoi Oncology Hospital (Vietnam). The experimental settings are described in Section 4.3, while the metrics utilized to measure the prediction performance are explained in Section 4.4.

4.1. Research Questions

Through a series of experiments on the collected datasets, we study our proposed approach by answering the following research questions:

- **RQ**$_1$: *Which edge detection technique contributes to a more timing efficient and effective classification?* We perform prediction with the graph data produced by using two versions of the Prewitt edge detection algorithm, aiming to identify the version that brings the best prediction accuracy as well as timing efficiency on the curated datasets.

- **RQ**$_2$: *Which Graph Neural Network model brings the best recommendation performance?* Three different Graph Neural Network models, namely GATConv [19], GCNConv [18] and GraphConv [20] have been used as the classification engine, this aims to find the most effective one, achieving the best prediction performance.

- **RQ**$_3$: *How does MammoGNN compare with the considered baselines?* We are interested in understanding if our proposed approach outperforms a state-of-the-art baseline built on top of a conventional Convolutional Neural Network for medical modalities classification [1].

4.2. Data collection

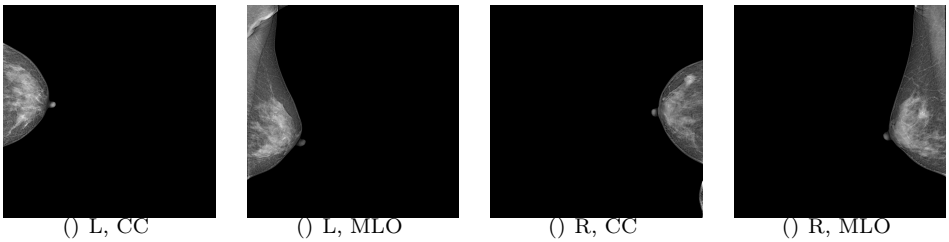
We collected a full-field digital mammography dataset from 2,351 women who were examined by radiologists at Hanoi Hospital Oncology (Vietnam) from December 2020 to June 2021. The data curation was done by strictly following ethic regulations,[3] and private information of any individuals in this study has been properly anonymized. The process of labelling data was followed the ACR BI-RADS® Atlas 5^{th} Edition [17] and performed by three doctors, who have more than 15-year experience in oncology radiology. For each individual, four images were collected and manually classified. If there is any disagreement

[3] Among others, we adhere to the Decree of medical ethics, Decision Nr 2088/BYT-QD of Vietnam Ministry of Health and the regulations of Hanoi Hospital Oncology.

Table 4: A summary of original BI-RADS and Type dataset.

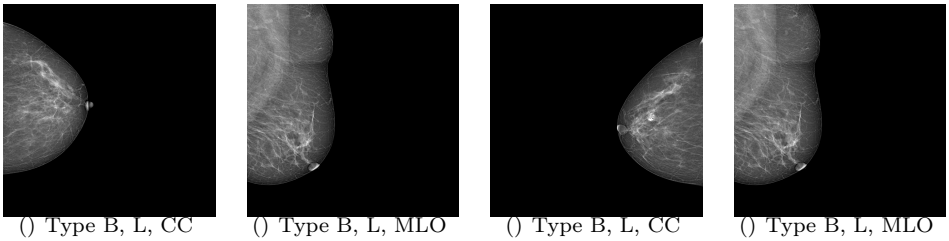
No.	Dataset	Category	# of individuals	# of images
1	BI-RADS	0	8	37
2	BI-RADS	1	1,016	4,176
3	BI-RADS	2	781	3,201
4	BI-RADS	3	199	817
5	BI-RADS	4A	208	869
5	BI-RADS	4B	63	266
7	BI-RADS	4C	26	107
8	BI-RADS	5	14	60
9	Type	A	61	258
10	Type	B	409	1,698
11	Type	C	1,666	6,874
12	Type	D	164	668

over the four mammograms, then two radiologists are involved in discussing and reaching a consensus on the final classification. In the extreme case, if no consensus is reached, then the third radiologist is asked to mediate an agreement. Finally, we obtained an FFDM dataset, including BI-RADS scores and density types. A summary of the collected dataset is shown in Table 4. Some examples of the collected images are shown in Fig. 2 and Fig. 3. **L** and **R** indicate left and right breasts, respectively. Bilateral craniocaudal (CC) and mediolateral oblique (MLO) views are standard views, which comprise routine screening mammography.



() L, CC () L, MLO () R, CC () R, MLO

Figure 2: Mammography examples of BI-RADS scores.



() Type B, L, CC () Type B, L, MLO () Type B, L, CC () Type B, L, MLO

Figure 3: Mammography examples of breast density types.

4.3. Experimental settings

For all the experiments, we split the dataset into three independent parts, i.e., 70% for training, 15% for validation and 15% for testing. We trained with 100 epochs for each model on these graph datasets, using the Adam optimizer with a learning rate of 10^{-3}. We adopted a recent implementation on Python version 3.8, PyG version 2.0.2 [29] with built-in functions such as GAT [19], GCN [18] and GraphConv [20] backbones and PyTorch version 1.10 with CUDA version 11.4.[4] The following server is used to run the experiments: Intel® Xeon® CPU E5-2680v4 @ 2.50GHz × 14 cores, 96GiB RAM, NVIDIA GeForce RTX 3090, Operating System Ubuntu 20.04.4 LTS. The number of batch sizes is set to fit to 24Gb VRAM of RTX 3090.

4.4. Evaluation Metrics

Given a set of mammogram images, there is a set labels, i.e., $G = (G_1, G_2, .., G_N)$. We compare it with the set of predicted labels, i.e., $C = (C_1, C_2, .., C_N)$ to calculate *Accuracy*, *Precision*, *Recall*, and F_1 *score*, defined as follows.

Accuracy: The metric is measured as the ratio of number of correct prediction to the total number items.

$$accuracy = \frac{\sum_i^N match_i}{\sum_i^N |G_i|} \times 100\% \tag{6}$$

[4] https://pytorch.org

Precision and Recall: *Precision* evaluates the number correctly predicted instances, meanwhile *recall* gauges the ability to find all relevant instances in the dataset, and they are computed using the following formula.

$$precision_i = \frac{match_i}{|C_i|} \quad (7) \qquad recall_i = \frac{match_i}{|G_i|} \quad (8)$$

F_1 score (F-Measure): F_1-score is computed as the harmonic mean of Precision and Recall:

$$F_1 = \frac{2 \cdot precision_i \cdot recall_i}{precision_i + recall_i} \quad (9)$$

In the next section, we analyze the experimental results by answering the research questions in Section 4.1.

5. Experimental results

We report and analyze the results obtained by conducting a series of experiments on the given dataset in Section 5.1, Section 5.2, and Section 5.3. Afterward, Section 5.4 discusses the probable threats that may adversely impact on the validity of our findings.

5.1. $\boldsymbol{RQ_1}$*:* Which edge detection technique contributes to a more timing efficient and effective classification?

In this research question, we are interested in understanding the efficiency and effectiveness as follows.

5.1.1. Efficiency

Fig. 4 and Fig. 5 show examples of mammograms transformed with the Prewitt algorithms. The images imply that using the Prewitt v1 filter to transform original to edge images brings more noisy edge images than compared to using the Prewitt v2 filter. In this respect, we hypothesize that preprocessing data with the Prewitt v1 filter consumes much more time compared to that when running with filter Prewitt v2.

To validate the hypothesis, we compute and show in Table 5 information related to the graph produced from the original dataset using the edge detection

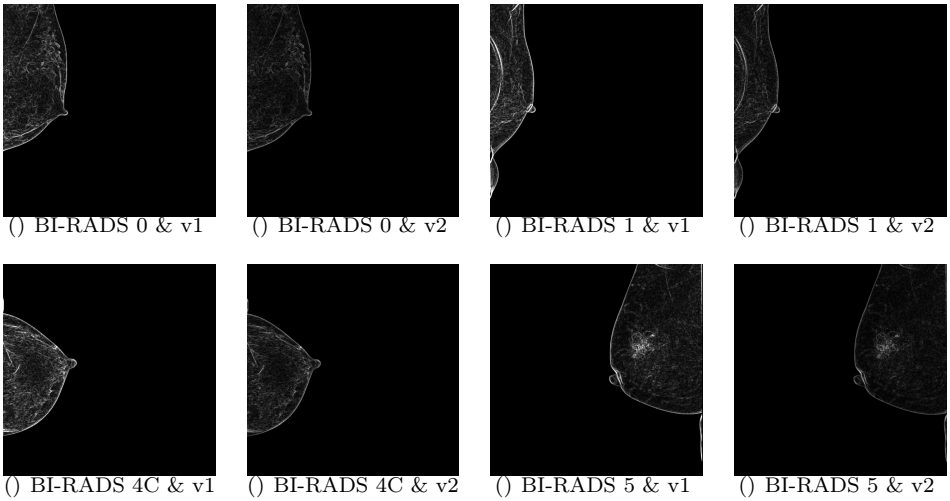


Figure 4: Mammograms of BI-RADS scores preprocessed using the Prewitt algorithms.

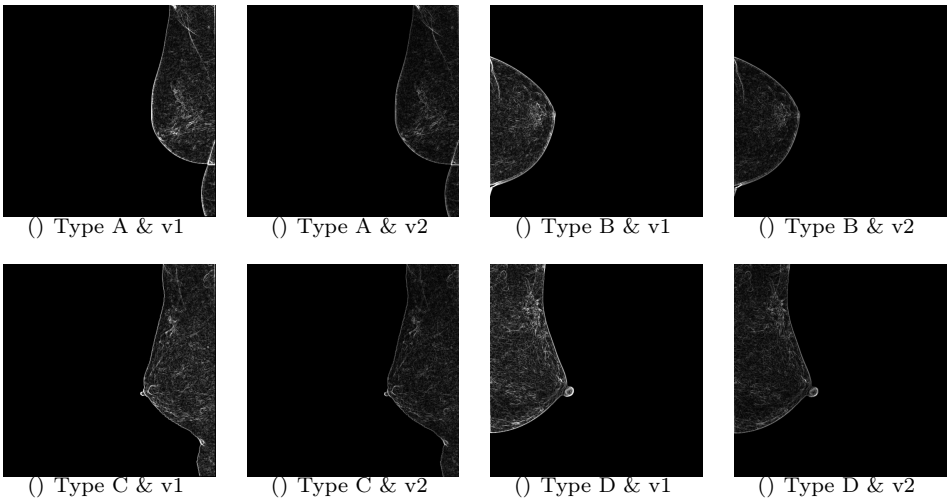


Figure 5: Mammograms of breast density types preprocessed using Prewitt's algorithms.

algorithms. From the table, we see that compared to Prewitt v1, Prewitt v2 is more efficient with respect to both timing and storage. In particular, from the BI-RADS dataset, the number of nodes obtained by Prewitt v1 is $1,685$ while the corresponding number by Prewitt v2 is 302. Similarly, by the Type dataset, using Prewitt v1 and Prewitt 2, we get 3,321 and 785 nodes, respectively. Correspondingly, the number of edges also dramatically decreases when changing from Prewitt v1 to Prewitt v2.

Table 5: Summary of preprocessed data.

Dataset	BI-RADS		Type	
Filter	Prewitt v1	Prewitt v2	Prewitt v1	Prewitt v2
# of graphs	9,633	9,633	9,488	9,488
# of features	9	9	5	5
# of classes	8	8	4	4
# of nodes	1,685	302	3,321	785
# of edges	5,778	844	11,742	2,328
Degree	3.43	2.79	3.54	2.97
Dataset size (MB)	8,039	1,743	6,607	1,343
Preprocessed data size (MB)	1,600	345	1,300	278
Processing time (minutes)	360	240	360	240
Training time (minutes)	7	3	6	2

Concerning the dataset size, there is a dramatic decrease when moving from Prewitt 1 to Prewitt v2. For instance, on the BI-RADS dataset, we need more than 8,000 MB to store the data produced by Prewitt v1 while we need only 1,743 MB to store the same data produced by Prewitt 2. Similarly also by the Type images, using Prewitt v2 to transform the data always yields a small file size, compared to using Prewitt v1. Overall, Table 5 suggests that using Prewitt v2 brings a substantial efficiency in storage and processing time. Images preprocessed by Prewitt version 2 are less noisy.

Table 6: Number of parameters, size, and training time.

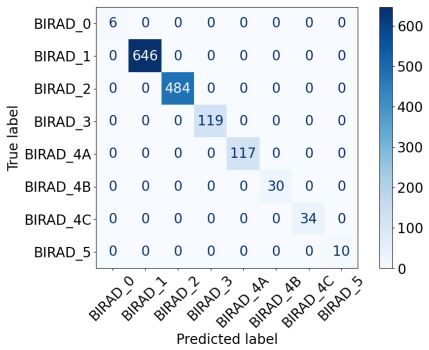
No.	Backbone	# Params	Size (MB)	Training time (minutes)	
				Prewitt v1	Prewitt v2
1	GAT	2,628	1.2	less than 6	less than 2.5
2	GCN	2,436	1.5	less than 7	less than 3
3	GraphConv	4,644	2.2	less than 5	approximate 2

Referring to Table 6, we also see that Prewitt v2 is more training efficient, i.e., it contributes to much less time required for a model to learn from the input data. Time for edge detection and graph preparation using Prewitt v2 is approximately 4 hours on the dedicated platform with one CPU Xeon E5 2680V4 and 64 GB RAM, while processing with Prewitt v1 requires around 6

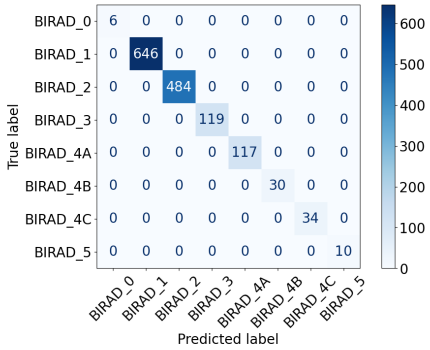
hours. Training time for dataset processed with Prewitt v2 is also much shorter than that with Prewitt version 1.

5.1.2. Effectiveness

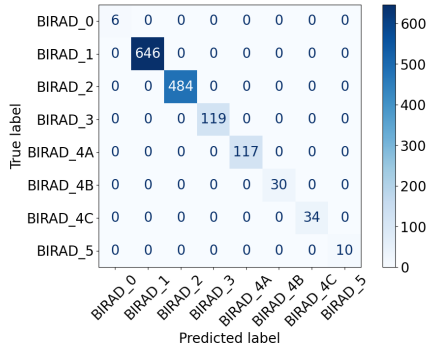
Next, we analyze the prediction performance by using data processed with the two algorithms. Fig. 6 and Fig. 7 show the prediction performance results on the the test sets for both datasets. The proposed models combined with two preprocessing data techniques work very well on both classification tasks for BI-RADS and breast density type datasets.



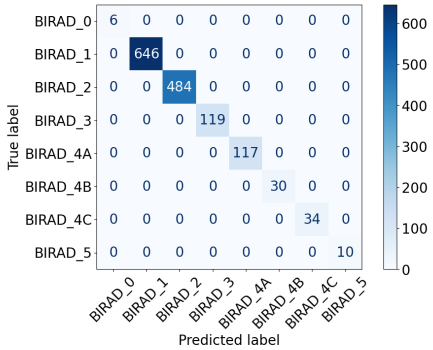
() Prewitt v1 & GAT



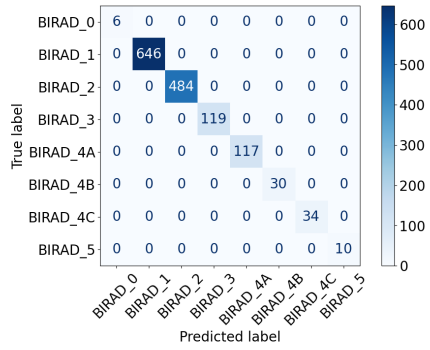
() Prewitt v1 & GCN



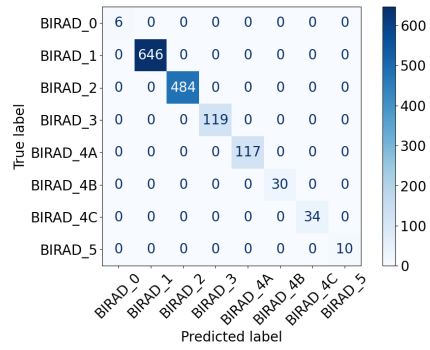
() Prewitt v1 & GraphConv



() Prewitt v2 & GAT

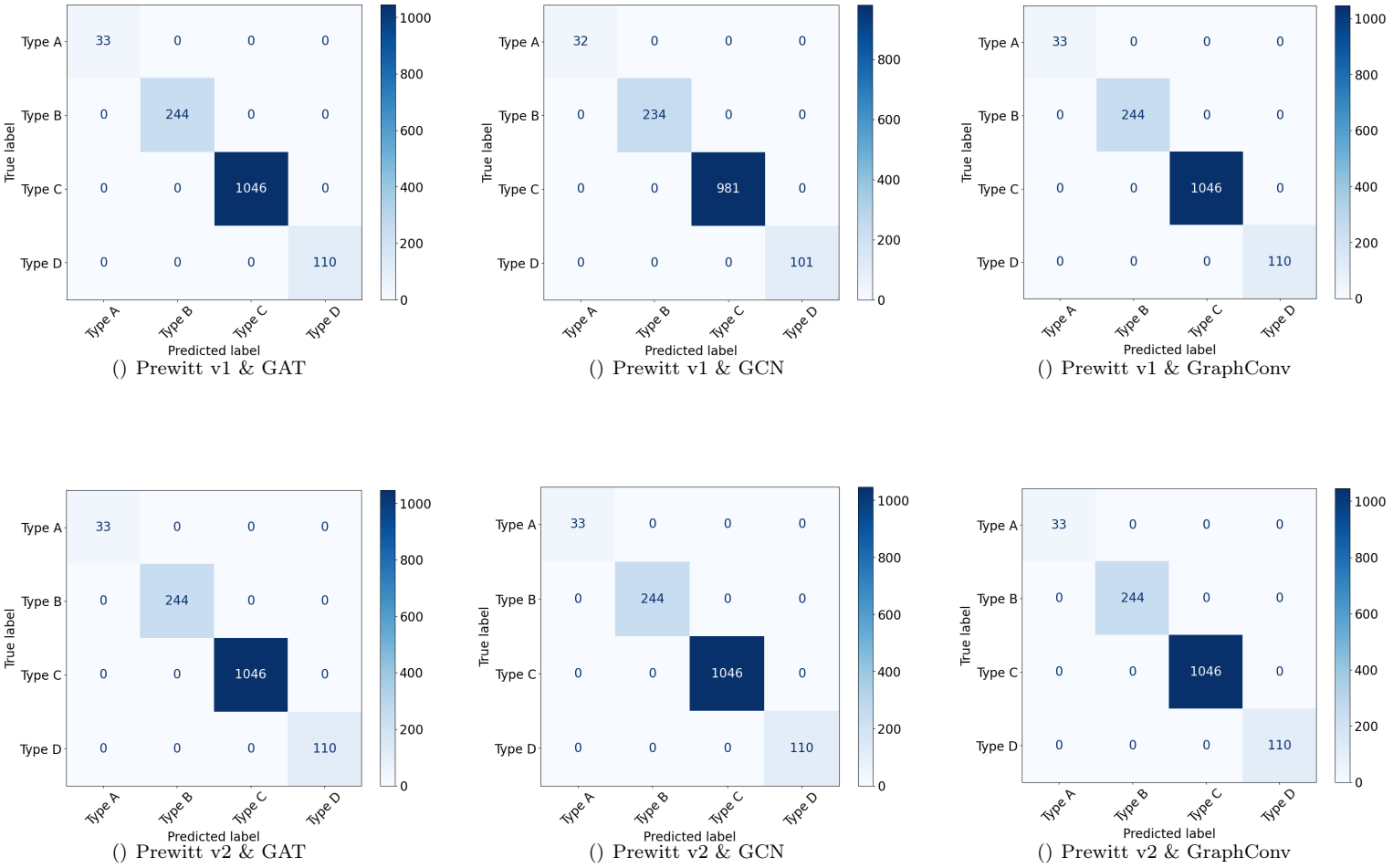


() Prewitt v2 & GCN



() Prewitt v2 & GraphConv

Figure 6: Confusion matrices for the BI-RADS dataset.



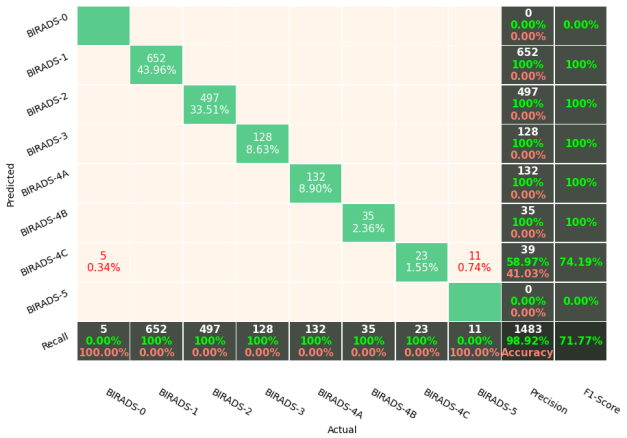
() Prewitt v1 & GAT () Prewitt v1 & GCN () Prewitt v1 & GraphConv

() Prewitt v2 & GAT () Prewitt v2 & GCN () Prewitt v2 & GraphConv

Figure 7: Confusion matrices for breast density types.

It is evident that MammoGNN returns a perfect match for all the categories. In particular, the confusion matrices for the BI-RADS dataset are shown in Fig. 6, and we can see that all the images are correctly classified. For instance, the BI-RADS_1 category with 646 images is the largest one, and all the mammograms are properly sorted to their real category.

The overall prediction performance for the BI-RADS recognition task is absolutely correct, which raises question about the process of our proposed models design. Indeed, we have been adjusting scale ratios of Depth, Width, and Resolution very carefully to achieve the best metrics results. We can show you our results if we reduce the width of our proposed models. In this circumstance, using the width of 32 instead of 256 the obtained metrics on the test set is 98.92% as shown in Figure 8. Thus, the proposed model can not overcome the imbalanced data, especially category BI-RADS-5 will be predicted to BI-RADS-4C class.



() Prewitt v1 & GraphConv with width of 32 hidden channels

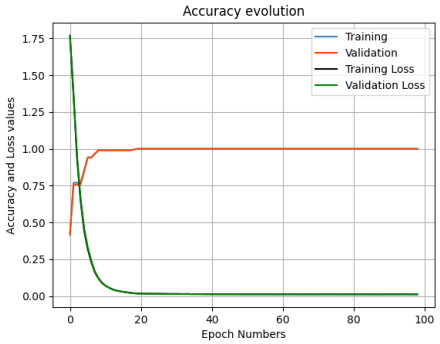
Figure 8: Accuracy for the classification of BI-RADS scores using a model with width of 32 hidden channels.

Concerning the Type dataset, we also encounter a similar outcome, i.e., all the testing mammograms are correctly classified into their real categories, resulting in a perfect match for all the settings. For example, by the largest category, i.e., Type C with 1,046 images, MammoGNN correctly classifies all

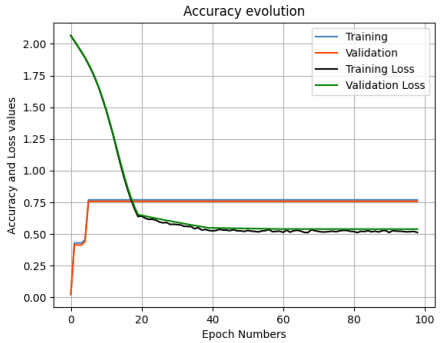
of them. We conclude that by both datasets, our proposed approach obtains a maximum prediction performance by all the experimental configurations with respect to the choice of neural networks and image processing techniques.

Finally, we also monitor accuracies and losses on training and validation sets during the training phases. As shown in Fig. 9 and Fig. 10, all training procedures converge early, i.e., after a couple of epochs, and accuracy values for both training and validation sets come very close together. Notably, the pattern is almost identical for both edge detection techniques, i.e., Prewitt v1 and Prewitt v2. This essentially means that there is neither overfitting nor underfitting on training phases or test steps. Looking at the curves, we can see the models based on GraphConv converge fastest than those that use GATConv or GCNConv.

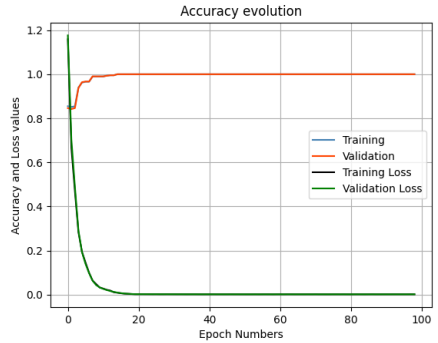
Answer to RQ$_1$. By all the experimental configurations, though both edge detection algorithms contribute to a maximum prediction accuracy, Prewitt version 2 is more timing efficient, compared to Prewitt version 1.



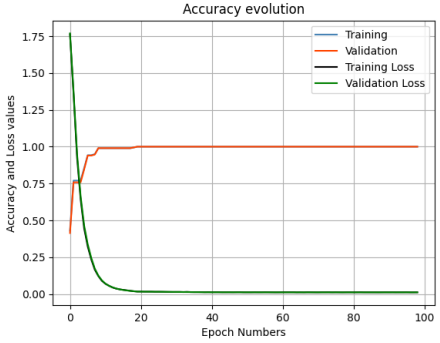
() Prewitt v1 & GAT



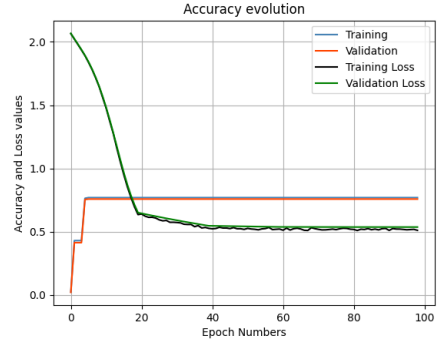
() Prewitt v1 & GCN



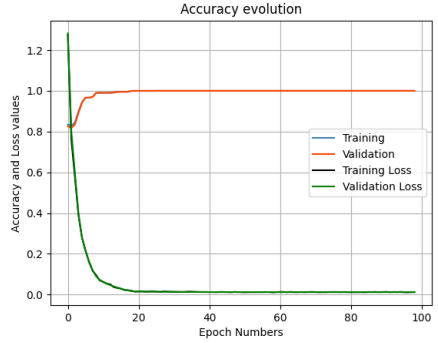
() Prewitt v1 & GraphConv



() Prewitt v2 & GAT

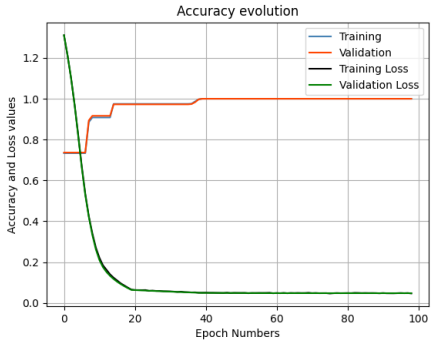


() Prewitt v2 & GCN

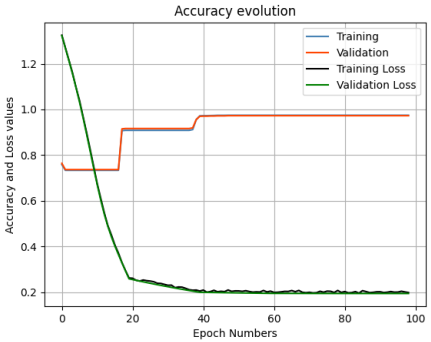


() Prewitt v2 & GraphConv

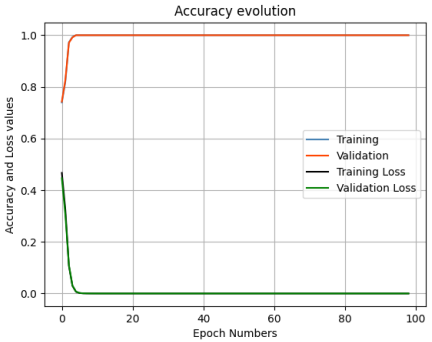
Figure 9: Accuracy and loss curves for the classification of BI-RADS scores.



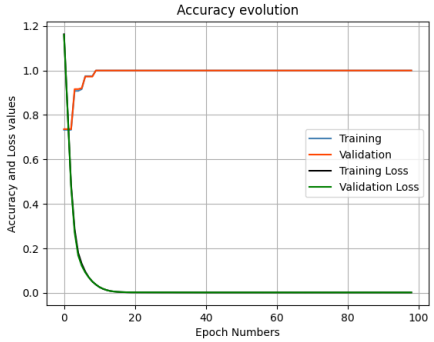
() Prewitt v1 & GAT



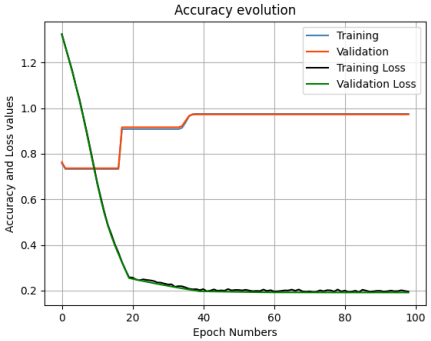
() Prewitt v1 & GCN



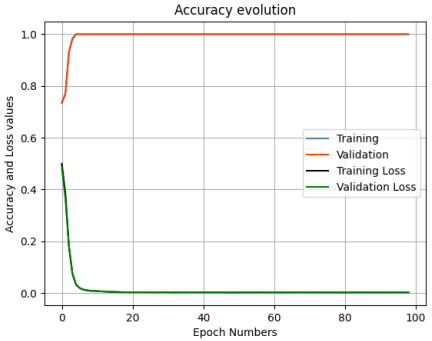
() Prewitt v1 & GraphConv



() Prewitt v2 & GAT



() Prewitt v2 & GCN



() Prewitt v2 & GraphConv

Figure 10: Accuracy and loss curves for the classification of breast density types.

5.2. ***RQ***$_2$*:* Which Graph Neural Network model brings the best recommendation performance?

In the BI-RADS score classification task, we can see in Fig. 9 that among others, using a model with the GraphConv layers combined with the Prewitt v1 filter converges faster and smoother. Though the other models using GATConv or GCNConv layers are slower at the first epochs, they still saturate from the 60^th epoch. The prediction performance of all the models is measured using confusion matrices in Fig. 6, and an accuracy of 100% on the independent test set is obtained.

For the classification of breast density types, the results in Fig. 10 demonstrate that using either GATConv with both types of the filters or GraphConv with the Prewitt v2 filter converges very quickly after a few epoch of training phases. Similar to classifying BI-RADS, models built with GCNConv bring later convergence during training stages. However, our proposed models still achieve perfect performance on the independent test set (Fig. 7). Moreover, using GATConv layers requires much more VRAM memory compared to the other architectures, reducing the number of batch sizes and increasing time to train such models-based GATConv layers.

As shown in Table 6, there are differences among the experimental configurations for GATConv, GCNConv, and GraphConv. The *# Params* column corresponds to the number of parameters needed to store the weights and biases of the network. Our proposed models are considerably compact and contain less than $5,000$ parameters. Correspondingly, the *File size* column indicates the size of file to store the parameters (in MB). They all are also very light size and less than 3 MB. Together, we assume that the proposed models are suitable for running on edge devices.

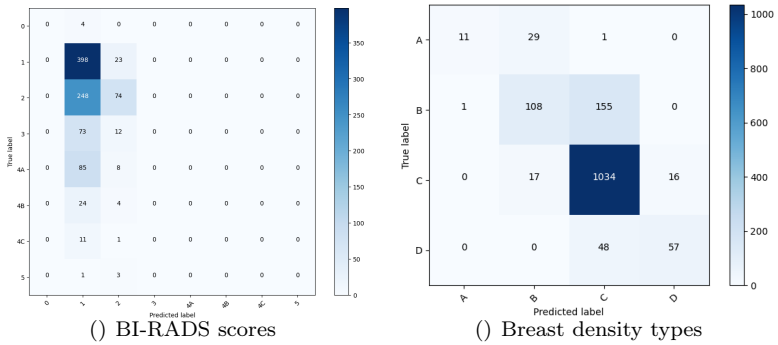
It is worth noting that our proposed methods provide not only perfect metric values, but also overcome for the heavily class imbalanced datasets such as BI-RADS score 0, BI-RADS score 5, breast density type **A**, and breast density type **D**. Intuitively, this suggests that Graph Neural Network maps nodes to the same region only if these vertexes have identical sub-trees with indistin-

guishable features on the corresponding nodes. This phenomenon implies that its aggregation scheme has to be related to an injection or one-to-one mapping, so-called injective. Therefore, injective multiple sets functions can be represented by a Graph Neural Network's aggregation.

Answer to RQ$_2$. On the given dataset, GraphConv outperforms GATConv, GCNConv with respect to both prediction performance and timing efficiency.

5.3. ***RQ***$_3$*:* How does MammoGNN compare with the considered baselines?

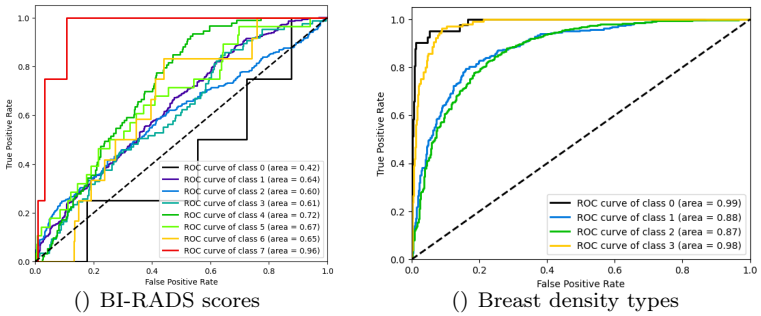
Using the dataset, we compare our approach with a set of pre-trained backbone models, including a family of ResNet [37, 38], EfficientNet v1 [39], EfficientNet v2 [40], and DeiT [41]. We also combined CNN or Transformer backbones with various augmentation techniques and resolution of images. All the experiments were performed using the TIMM codebase [42] with various settings of hyperparameters for the training phase. The time required to train with image size of $512 \times 512 \times 1$ with batch sizes of 110 and mixed precision of 32-bit using ResNet26 [37] was around 48 hours.



() BI-RADS scores () Breast density types

Figure 11: Confusion matrices for BI-RADS scores and breast composition using ResNet26.

Following the standard procedure, subjects diagnosed BI-RADS score and breast density types are examined their FFDMs taking two breast with two views of each breast. Addition, any human body anatomy is imperfect symmetric, that is why radiologists must use their perception get reasoning diagnosis. It is challenging for machine perception using traditional approaches.



() BI-RADS scores () Breast density types

Figure 12: ROC curves for BI-RADS scores and breast composition using ResNet26.

Table 7: Accuracy, Precision, Recall and F1-score for BI-RADS classification using ResNet26D.

	BI-RADS scores							
	0	1	2	3	4A	4B	4C	5
Precision	0.000	0.472	0.592	0.000	0.000	0.000	0.000	0.000
Recall	0.000	0.945	0.230	0.000	0.000	0.000	0.000	0.000
$\mathbf{F}_1$	0.000	0.629	0.3311	0.000	0.000	0.000	0.000	0.000
Accuracy	**0.487**							

We assume that Graph Neural Network representation graph with reasoning inference is the key to solve such problems.

In contrast to the BI-RADS classification task, breast composition classification tasks work much better using architectures of CNNs or Transformers. We encounter negative results as shown in Fig. 11(a), Fig. 12(a) and Table 7. As shown in Fig. 12(b) and Table 8, the ResNet26 model gains an accuracy of 81.92%, but it does not overcome the phenomenon of imbalanced data. Eventually, our proposed models outperform performance with respect to accuracy, heavily imbalanced data and computational cost.

Answer to RQ$_3$. MammoGNN considerably outperforms the considered baselines in all the test configurations by different quality indicators.

5.4. Threats to Validity

This section explains the probable threats to internal, external validity of our work as follows.

Table 8: Accuracy, Precision, Recall and F_1-score for breast density type classification with ResNet26D.

	Breast density type			
	A	B	C	D
Precision	0.917	0.701	0.835	0.782
Recall	0.268	0.409	0.969	0.54
$\mathbf{F_1}$	0.415	0.517	0.897	0.61
Accuracy	**0.819**			

Internal validity. These are internal factors that might impact on the evaluation. A possible threat is the comparison with the baselines. We minimize the threat by running experiments using the original implementations, as well as testing all the related tools on our collected dataset, and comparing them with the same set of metrics.

External validity. The threat to *external validity* of the approach concerns the generalizability of our approach, i.e., if it is valid outside the scope of this paper. Such a threat is minimized by evaluating the systems using a real dataset collected on our own.

Construct validity. This is related to the experimental settings used in our work, concerning the simulated configurations to evaluate the approach. We performed the evaluation by means of a training set and a test set, which may not represent a real-world usage. To target a fair comparison, we made use of the same configurations to compare the systems.

6. Results of the external dataset

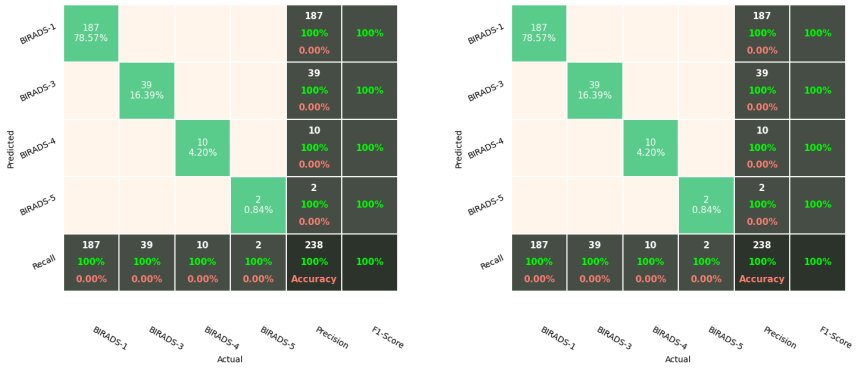
To be sure for our proposed models obtain consistent and concrete resutls, and generalizability, we try to evaluate our experiment procedues on another full-field digital mammography dataset. During reviewing literature throughgoutly, we curate FFDM dataset with BI-RADS labelled from Alsolami *et al.* [**?**]. We found a database from Sheikh Mohammed Hussein Al-Amoudi Center of Excellence in Breast Cancer at King Abdulaziz University (Jeddah, Saudi Arabia). The dataset contains 1416 cases; all cases include images with two types

of views (CC and MLO) for both breasts (right and left), resulting in a total of 5,662 mammogram images. The dataset was classified into 1 to 5 categories in accordance with BI-RADS [43]. Unfortunately, this dataset downloaded from Kaggle [5] has only 4 classes including BI-RADS-1 (1.865 images), BI-RADS-3 (387 images), BI-RADS-4 (102 images), and BI-RADS-5 (24 images). It is a little bit different from their description on the paper such as lack of category BI-RADS-3. And then we tried to apply our proposed approach to get results for an external validation as follows. The original images of the dataset were detected edges using Prewitt v1 and v2, and then the edge images were converted to Graph-structured data. Later on, we splitted the entire dataset into training/validation/testing with a ratio of 80%/10%/10%. Hereby, we show the training/validation curves and prediction performance of our proposed model on the test set using edge detection algorithms Prewitt v1 and v2 combined with GraphConv layer for MammoGNN. Training phages for the dataset are also converge very fast, which achieves at maximum epochs of 10. The prediction performance on the test set from King Abdulaziz University Mammogram Dataset is also absolutely correct, which is the same as our present results on the manuscript. Hence, we concretely believe that our proposed models work well and consistently on heterogeneous datasets (the results in detail in Fig. 13).

7. Discussion

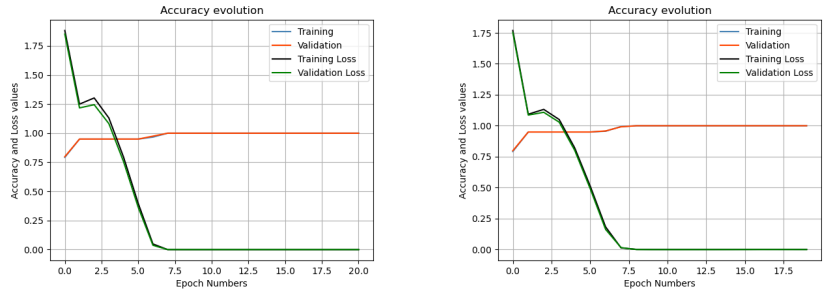
In this study, we introduce our own FFDM datasets, which are collected at Hanoi Hospital Oncology and labelled by three independent experts. To the best of knowledge, these datasets are the first publicly available one, examined on Vietnamese women. In the context of this study, we only take consideration into medical imaging and the standard BI-RADS guidance for classifying both BI-RADS scores and breast density types from FFDMs. Following the guidance, doctors must use FFDMs from both breasts and two views (CC and

[5] https://9h.fit/OHptMT



() Prewitt v1 & GraphConv

() Prewitt v2 & GraphConv



() Training curve for Prewitt v1

() Training curve for Prewitt v2

Figure 13: Confusion matrices and training curves for BI-RADS score from the external dataset.

MLO) to give final conclusion. In addition, human body is imperfectly bilaterally symmetric. Thus, it is challenging for any conventional approach such as CNNs and Transformers to detect. In this respect, Graph Neural Network models have the traits to deal with the detection of FFDMs. To ensure the superiority of our proposed models, we have also compared their performance against some pre-trained models such as the ResNet family [37], EfficientNet version 1 family [39], EfficientNet version 2 family [40], and DeitT family [41]. Indeed, the experimental results show that MammoGNN outperforms the baselines.

There are many techniques to detect edges of an image such as Prewitt [27], Sobel [44], Canny [45], to name a few. Such these filters can operate typical algorithms that may produce different levels of noise of a new image from the original images. Moreover, using distinct edge detection algorithms also affect the ability to be successful in generating graph data. For examples,

we failed to construct graphs for the BI-RADS dataset using Canny method [45], but succeeded when using the Sobel technique [44]. Following our own experience during surveillance of the preprocessing data phase on many other publicly available datasets, each dataset has its own typical character, and it is more suitable with certain filter algorithms. Among others, the Prewitt method brings the best computational cost for both preprocessing and training data because it produces less noisy images. Eventually, edge detection plays a crucial role in not only in generating proper graph data, but also in balancing prediction performance and cost in terms of both preprocessing data stage and training phase.

8. Related Work

In our work, we deal with the recognition of mammograms, thus in this section we analyze the most notable studies, paying our attention to the datasets used, the graph representation, and accuracy during the training phase and the final prediction.

In recent, Alsolami *et al.* [43] publised an open mammogram dataset, named King Abdulaziz University Breast Cancer Mammogram Dataset (KAU-BCMD) version 1. The dataset was collected from the Sheikh Mohammed Hussein Al-Amoudi Center of Excellence in Breast Cancer at King Abdulaziz University. It contains 1,416 cases. Each subject has two views for both sides of breasts, totally resulting in 5,662 images based on the BI-RADS criteria. Three independent radiologists annotated, reviewed, and labeled every category for the dataset. This for Saudi wome contains standard imaging modalities for breast cancer with BI-RADS scores. However, the authors have not proposed any machine learning to recognize feature maps of the data yet.

A new Optimal Multi-Level Thresholding-based Segmentation with DL enabled Capsule Network (OMLTS-DLCN) breast cancer diagnosis model utilizing digital mammograms was proposed by Kavitha *et al.* [46]. The OMLTS-DLCN model requires an Adaptive Fuzzy based median filtering (AFF) technique as a pre-processing phase to eliminate the noise that occurs in the FFDM

images. In addition, Optimal Kapur's based Multilevel Thresholding with Shell Game Optimization (SGO) algorithm (OKMT-SGO) was used for breast cancer segmentation. Besides, the proposed model includes a CapsNet based feature extractor and Back-Propagation Neural Network (BPNN) classification machine was ultilized to detect the occurence of breast cancer. The diagnostic outcomes of the presented OMLTS-DLCN technique were investigated employing benchmark Mini-MIAS and DDSM datasets. The prediction performance of the OMLTS-DLCN model achieved higher accuracy of 98.50% and 97.55% on the Mini-MIAS dataset and DDSM dataset, respectively.

Chakravarthy *et al.* [47] presented a new customized method of integrating the concept of deep learning with the extreme learning machine (ELM), which was optimized using a simple crow-search algorithm (ICS-ELM). The algorithm focused on detecting the input mammograms as either normal or abnormal. The digital mammograms for this work were Curated Breast Imaging Subset of DDSM (CBIS-DDSM), Mammographic Image Analysis Society (MIAS), and INbreast datasets. The work used ResNet-18 based deep extracted features with proposed Improved Crow-Search Optimized Extreme Learning Machine (ICS-ELM) algorithm. The proposed pipeline was compared with the existing Support Vector Machines (RBF kernel), ELM, particle swarm optimization (PSO) optimized ELM, and crow-search optimized ELM. The maximum overall classification accuracy was achieved for the proposed method with 97.19% for DDSM, 98.14% for MIAS and 98.27% for INbreast datasets, respectively.

Leman *et al.* [48] collected a huge database consisting of 119,139 bilateral screening mammograms in 57,617 consecutive and multi-demographic cases screened at 5 hospitals from September 18, 2017, to February 1, 2021. Their proposed deep learning model, Tyrer-Cuzick, and National Cancer Institute Breast Cancer Risk Assessment Tool (NCI BCRAT) risk models were measured as regards performance metrics and area under the receiver operating characteristic curves (AUC). Concerning cancers detected per thousand patients screened, the deep learning model, Tyrer-Cuzick, and NCI BCRAT showed increased risk of 8.6 (95% confidence interval [CI]=7.9-9.4), 4.4 (95% CI=3.9-4.9),

and 3.8 (95% CI=3.3-4.3), respectively. Similarly, AUCs of the deep learning model, Tyrer-Cuzick, and NCI BCRAT were 0.68 (95% CI=0.66-0.70), 0.57 (95% CI=0.54-0.60), 0.57 (95% CI=0.54-0.60), respectively. The statistic values were significant difference.

Wu *et al.* [1] performed experiments with convolutional neural networks on a dataset of 229,426 digital screening mammogramphy examinations with 1,001,093 images from 141,473 unique patients screened between 2010 and 2017. In the whole dataset, 8,842 lesions from 8,080 images with diagnosis were confirmed by biopsy, which reveals that a single breast can contain multiple lesions of differing types [49]. Afterward, the authors extracted information from image patches of 256 $\times$ 256 pixels as one of the four classes: "malignant", "benign", "outside" and "negative". They trained a deep convolutional neural network (DCNN) to classify the dataset following a ratio of training (80%), validation (10%) and testing (10%) sets. The best result was an AUC of 0.8 on a test set containing 464 benign and 136 malignant lesions.

Zhang *et al.* [50] combined a CNN architecture with a Graph Convolutional Neural Network (GCN) to classify six mammographic abnormal breast types using the mini-MIAS dataset [51]. First, the authors used a CNN pipeline to extract individual image-level features; then, they fed to a GCN to assess a relation-aware representation. These features were fused via a dot product and a linear projection with trainable weights. Their proposed model gains high prediction performance with a sensitivity of 96.20 $\pm$ 2.90%, a specificity of 96.00 $\pm$ 2.31% and an accuracy of 96.10 $\pm$ 1.60% after ten folds of cross validation split.

By using full field digital mammogram (FFDM) images from the INbreast dataset [52], Du *et al.* [53] proposed the zoom-in mechanism, which mimics a behavior of radiologists during examining medical images by zooming into region of interests (ROIs) for a close-up examination, with a hierarchical graph-based model to detect abnormal lesions. First, the authors employed a pre-trained CNN trained on lesion patches to extract features. Afterward, these features were used as input for a Graph Attention model to classify whether to mag-

nify or not into the next level to predict a benign or malignant mammogram. Their proposed model achieves an AUC of 0.943 in binary mammography classification into normal/benign breast or malignant breast cancer. Similar to mini-MIAS, the INbreast dataset is relatively small, and this indeed poses a threat to the external validity of the approach.

Ragab *et al.* [54] introduced a new computer-aided detection system for classifying benign and malignant mass tumors in mammograms using their designed convolutional neural networks. Two datasets DDSM and CBIS-DDSM were extracted from $9,368$ ROIs images. The approach obtained a maximum accuracy of 87.2% and AUC of 0.94. Agarwal *et al.* [55] investigated the performance of VGG16, ResNet50, and InceptionV3 for classifying mass and non-mass breast regions for the CBIS-DDSM and INBreast database. A dataset with $1,592$ and 112 images from CBIS-DDSM and INbreast has been populated to use as the evaluation data. In CBIS-DDSM, a total of $81,766$ images with a half of positive and the other of negative images are extracted to ROIs patch. The highest accuracy obtained by using InceptionV3 is 98%.

Zhao *et al.* [56] employed AlexNet and SVM for classifying benign and malignant tumors in a total of 115 ROIs images, which are extracted from MIAS. The authors also used rotation and flipping augmentation to generate to $4,600$ patches as a way to enrich the original dataset. The experimental results show that AlexNet achieves a classification accuracy of 97.57%. Dhungel *et al.* [57] demonstrated that the performance of a modified ResNet can be comparable to human perception in binary classification task. In the study, the INbreast dataset with FFDM of CC and MLO mammograms views was used to validate the proposed approach, and the highest AUC is 0.80. Levy and Jain [58] made use of AlexNet and GoogLeNet as the classification engine and built a CNN for classifying $1,820$ cropped mammograms consisting of mass tumors. The approach obtained the best accuracy of 92.9% using GoogLeNet with strong augmentation pre-processing. Compared to Levy and Jain [58], our proposed classification engine performs better even in a larger dataset.

A recent study [59] achieved the best ROC of 96% for 2-category classification on a subset DDSM by using pre-trained ResNet50. The dataset consists of $1,592$ images with benign or malignant mass and $2,340$ normal mammograms. Wang *et al.* [60] investigated the performance of AI techniques based on three publicly available databases namely DSSM, INbreast, MIAS, and one private dataset UKy. In the study, the authors studied FFDM mammograms for 2-class classification task. Their models AlexNet, VGG16, and ResNet50 reported AUC values between 0.88 and 0.95 on internal validation sets. However, the authors demonstrated inconsistency of performance models with AUC under 0.65 on external validation sets MIAS, INbreast, and UKy.

All the aforementioned studies have been tested with a variety of datasets, and obtained practical results. Meanwhile, in the present work, we combined different strategies for data transformation and recognition. We also demonstrated that our proposed approach is both effective and efficient, obtaining a high prediction accuracy and ROC. In this respect, we suppose that the approach is feasible in practice, and ready to be deployed to support doctors in their diagnosing tasks.

9. Conclusions and Future Work

In this paper, we proposed a workable solution to the recognition of BI-RADS scores and breast density types. Our proposed models have been conceptualized based on two core structures: detection of edge images and graph neural networks. The edges of an image are generated in the preprocessing stage using variants of the Prewitt algorithms [27, 28]. Therefore, the computational expense in our study is optimized compared to the typical CNN or ViT based models. Our approach works effectively on an FFDM dataset collected by our own. The prediction performance reaches 100%, and the model allows to us overcome the problems, such as overfitting and category imbalance, which usually occur on medical imaging datasets. Graph Neural Network is rather compact as it is small in size, can be run on conventional CPU, enabling administrators to deploy the framework on lightweight devices such as smartphones

and embedded on computers with general configurations. We plan to deploy the models in the form of a mobile app. This can come in handy for many remote regions in developing countries like Vietnam, where there is a lack of well-trained oncology radiologists.

References

[1] N. Wu, Z. Huang, Y. Shen, J. Park, J. Phang, T. Makino, S. Gene Kim, K. Cho, L. Heacock, L. Moy, et al., Reducing false-positive biopsies using deep neural networks that utilize both local and global image context of screening mammograms, Journal of Digital Imaging (2021) 1–10.

[2] D. T. T. Toan, D. T. Son, L. X. Hung, L. N. Minh, D. L. Mai, L. N. Hoat, Knowledge, attitude, and practice regarding breast cancer early detection among women in a mountainous area in northern vietnam, Cancer Control 26 (1) (2019) 1073274819863777, pMID: 31331185. arXiv:https://doi.org/10.1177/1073274819863777, doi:10.1177/1073274819863777. URL https://doi.org/10.1177/1073274819863777

[3] P. D. Y. Trieu, C. Mello-Thoms, P. Brennan, Female breast cancer in vietnam: a comparison across asian specific regions, Cancer biology & medicine 12 (2015) 238–45. doi:10.7497/j.issn.2095-3941.2015.0034.

[4] L. Tabár, B. Vitak, H.-H. T. Chen, M.-F. Yen, S. W. Duffy, R. A. Smith, Beyond randomized controlled trials: organized mammographic screening substantially reduces breast carcinoma mortality, Cancer: Interdisciplinary International Journal of the American Cancer Society 91 (9) (2001) 1724–1731.

[5] Y. Wu, Q. Zhang, Y. Hu, K. Sun-Woo, X. Zhang, H. Zhu, L. jie, S. Li, Novel binary logistic regression model based on feature transformation of xgboost for type 2 diabetes mellitus prediction in healthcare systems, Future Generation Computer Systems 129 (2022) 1–12.

doi:https://doi.org/10.1016/j.future.2021.11.003.
URL https://www.sciencedirect.com/science/article/pii/S0167739X21004325

[6] U. Ahmed, G. Srivastava, U. Yun, J. C.-W. Lin, Eandc: An explainable attention network based deep adaptive clustering model for mental health treatment, Future Generation Computer Systems 130 (2022) 106–113. doi:https://doi.org/10.1016/j.future.2021.12.008.
URL https://www.sciencedirect.com/science/article/pii/S0167739X21004891

[7] G. Manogaran, C. Thota, D. Lopez, V. Vijayakumar, K. M. Abbas, R. Sundarsekar, Big Data Knowledge System in Healthcare, Springer International Publishing, Cham, 2017, pp. 133–157. doi:10.1007/978-3-319-49736-5_7.
URL https://doi.org/10.1007/978-3-319-49736-5_7

[8] N. Muralidharan, S. Gupta, M. R. Prusty, R. K. Tripathy, Detection of covid19 from x-ray images using multiscale deep convolutional neural network, Applied Soft Computing 119 (2022) 108610. doi:https://doi.org/10.1016/j.asoc.2022.108610.
URL https://www.sciencedirect.com/science/article/pii/S1568494622001119

[9] S. Chakraborty, K. Mali, A radiological image analysis framework for early screening of the covid-19 infection: A computer vision-based approach, Applied Soft Computing 119 (2022) 108528. doi:https://doi.org/10.1016/j.asoc.2022.108528.
URL https://www.sciencedirect.com/science/article/pii/S1568494622000667

[10] C. D. Lehman, R. F. Arao, B. L. Sprague, J. M. Lee, D. S. Buist, K. Kerlikowske, L. M. Henderson, T. Onega, A. N. Tosteson, G. H. Rauscher, et al., National performance benchmarks for modern screening digital

mammography: update from the breast cancer surveillance consortium, Radiology 283 (1) (2017) 49–58.

[11] E. B. Cole, Z. Zhang, H. S. Marques, R. Edward Hendrick, M. J. Yaffe, E. D. Pisano, Impact of computer-aided detection systems on radiologist accuracy with digital mammography, American Journal of Roentgenology 203 (4) (2014) 909–916.

[12] M. Elter, A. Horsch, Cadx of mammographic masses and clustered microcalcifications: a review, Medical physics 36 (6Part1) (2009) 2052–2068.

[13] J. J. Fenton, S. H. Taplin, P. A. Carney, L. Abraham, E. A. Sickles, C. D'Orsi, E. A. Berns, G. Cutter, R. E. Hendrick, W. E. Barlow, et al., Influence of computer-aided detection on performance of screening mammography, New England Journal of Medicine 356 (14) (2007) 1399–1409.

[14] C. D. Lehman, R. D. Wellman, D. S. Buist, K. Kerlikowske, A. N. Tosteson, D. L. Miglioretti, Diagnostic accuracy of digital screening mammography with and without computer-aided detection, JAMA internal medicine 175 (11) (2015) 1828–1837.

[15] P. D. Y. Trieu, C. Mello-Thoms, P. C. Brennan, Female breast cancer in vietnam: a comparison across asian specific regions 12 (3) 238–245, publisher: Chinese Anti-Cancer Association. arXiv:26487968, doi:10.7497/j.issn.2095-3941.2015.0034.
URL https://pubmed.ncbi.nlm.nih.gov/26487968

[16] V. A. Horta, I. Tiddi, S. Little, A. Mileo, Extracting knowledge from deep neural networks through graph analysis, Future Generation Computer Systems 120 (2021) 109–118. doi:https://doi.org/10.1016/j.future.2021.02.009.
URL https://www.sciencedirect.com/science/article/pii/S0167739X21000613

[17] A. C. of Radiology, C. J. D'Orsi, et al., ACR BI-RADS atlas: breast imaging reporting and data system; mammography, ultrasound, magnetic

resonance imaging, follow-up and outcome monitoring, data dictionary, ACR, American College of Radiology, 2013.

[18] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, CoRR abs/1609.02907 (2016). arXiv:1609.02907.
URL http://arxiv.org/abs/1609.02907

[19] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio, Graph attention networks (2018). arXiv:1710.10903.

[20] C. Morris, M. Ritzert, M. Fey, W. L. Hamilton, J. E. Lenssen, G. Rattan, M. Grohe, Weisfeiler and leman go neural: Higher-order graph neural networks, CoRR abs/1810.02244 (2018). arXiv:1810.02244.
URL http://arxiv.org/abs/1810.02244

[21] L. T. Duong, N. H. Le, T. B. Tran, V. M. Ngo, P. T. Nguyen, Detection of tuberculosis from chest x-ray images: Boosting the performance with vision transformer and transfer learning, Expert Systems with Applications 184 (2021) 115519.

[22] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, P. S. Yu, A comprehensive survey on graph neural networks, IEEE Transactions on Neural Networks and Learning Systems 32 (1) (2021) 4–24. doi:10.1109/TNNLS.2020.2978386.

[23] Z. Wang, T. Chen, J. Ren, W. Yu, H. Cheng, L. Lin, Deep reasoning with knowledge graph for social relationship understanding, arXiv preprint arXiv:1807.00504 (2018).

[24] V. Garcia, J. Bruna, Few-shot learning with graph neural networks, arXiv preprint arXiv:1711.04043 (2017).

[25] X. Wang, Y. Ye, A. Gupta, Zero-shot recognition via semantic embeddings and knowledge graphs, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 6857–6866. doi:10.1109/CVPR.2018.00717.

[26] K. Marino, R. Salakhutdinov, A. Gupta, The more you know: Using knowledge graphs for image classification, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 20–28. doi:10.1109/CVPR.2017.10.

[27] J. M. S. Prewitt, Interactive decision-making for picture processing, in: 1977 IEEE Conference on Decision and Control including the 16th Symposium on Adaptive Processes and A Special Symposium on Fuzzy Set Theory and Applications, 1977, pp. 373–379. doi:10.1109/CDC.1977.271598.

[28] P. Saha, D. Mukherjee, P. K. Singh, A. Ahmadian, M. Ferrara, R. Sarkar, Graphcovidnet: A graph neural network based model for detecting covid-19 from ct scans and x-rays of chest, Scientific Reports 11 (1) (2021) 1–16.

[29] M. Fey, J. E. Lenssen, Fast graph representation learning with PyTorch Geometric, in: ICLR Workshop on Representation Learning on Graphs and Manifolds, 2019.

[30] A. Leman, B. Weisfeiler, A reduction of a graph to a canonical form and an algebra arising during this reduction, Nauchno-Technicheskaya Informatsiya 2 (9) (1968) 12–16.

[31] K. Oono, T. Suzuki, Graph neural networks exponentially lose expressive power for node classification (2021). arXiv:1905.10947.

[32] Q. Li, Z. Han, X.-M. Wu, Deeper insights into graph convolutional networks for semi-supervised learning, in: Thirty-Second AAAI conference on artificial intelligence, 2018.

[33] H. NT, T. Maehara, Revisiting graph neural networks: All we have is low-pass filters (2019). arXiv:1905.09550.

[34] U. Alon, E. Yahav, On the bottleneck of graph neural networks and its practical implications (2021). arXiv:2006.05205.

[35] A. F. Agarap, Deep learning using rectified linear units (relu), arXiv preprint arXiv:1803.08375 (2018).

[36] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, The journal of machine learning research 15 (1) (2014) 1929–1958.

[37] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition (2015). arXiv:1512.03385.

[38] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, W. Sieh, Deep learning to improve breast cancer detection on screening mammography 9 (1) 12495. doi:10.1038/s41598-019-48995-4.
URL https://doi.org/10.1038/s41598-019-48995-4

[39] M. Tan, Q. V. Le, Efficientnet: Rethinking model scaling for convolutional neural networks (2020). arXiv:1905.11946.

[40] M. Tan, Q. V. Le, Efficientnetv2: Smaller models and faster training (2021). arXiv:2104.00298.

[41] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention (2021). arXiv:2012.12877.

[42] R. Wightman, Pytorch image models, https://github.com/rwightman/pytorch-image-models (2019). doi:10.5281/zenodo.4414861.

[43] A. S. Alsolami, W. Shalash, W. Alsaggaf, S. Ashoor, H. Refaat, M. Elmogy, King abdulaziz university breast cancer mammogram dataset (kau-bcmd), Data 6 (11) (2021) 111.

[44] N. Kanopoulos, N. Vasanthavada, R. L. Baker, Design of an image edge detection filter using the sobel operator, IEEE Journal of solid-state circuits 23 (2) (1988) 358–367.

[45] J. Canny, A computational approach to edge detection, IEEE Transactions on pattern analysis and machine intelligence (6) (1986) 679–698.

[46] T. Kavitha, P. P. Mathai, C. Karthikeyan, M. Ashok, R. Kohar, J. Avanija, S. Neelakandan, Deep learning based capsule neural network model for breast cancer diagnosis using mammogram images, Interdisciplinary Sciences: Computational Life Sciences 14 (1) (2022) 113–129.

[47] S. S. Chakravarthy, H. Rajaguru, Automatic detection and classification of mammograms using improved extreme learning machine with deep learning, IRBM 43 (1) (2022) 49–61.

[48] C. D. Lehman, S. Mercaldo, L. R. Lamb, T. A. King, L. W. Ellisen, M. Specht, R. M. Tamimi, Deep learning vs traditional breast cancer risk models to support risk-based mammography screening, JNCI: Journal of the National Cancer Institute (2022).

[49] P. Xi, C. Shu, R. Goubran, Abnormality detection in mammography using deep convolutional neural networks, in: 2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA), IEEE, 2018, pp. 1–6.

[50] Y.-D. Zhang, S. C. Satapathy, D. S. Guttery, J. M. Górriz, S.-H. Wang, Improved breast cancer classification through combining graph convolutional network and convolutional neural network, Information Processing & Management 58 (2) (2021) 102439.

[51] P. SUCKLING J, The mammographic image analysis society digital mammogram database, Digital Mammo (1994) 375–386.

[52] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, J. S. Cardoso, Inbreast: toward a full-field digital mammographic database, Academic radiology 19 (2) (2012) 236–248.

[53] H. Du, J. Feng, M. Feng, Zoom in to where it matters: a hierarchical graph based model for mammogram analysis, arXiv preprint arXiv:1912.07517 (2019).

[54] D. A. Ragab, M. Sharkas, S. Marshall, J. Ren, Breast cancer detection using deep convolutional neural networks and support vector machines, PeerJ 7 (2019) e6201.

[55] R. Agarwal, O. Diaz, X. Lladó, M. H. Yap, R. Martí, Automatic mass detection in mammograms using deep convolutional neural networks, Journal of Medical Imaging 6 (3) (2019) 031409.

[56] X. Zhao, X. Wang, H. Wang, Classification of benign and malignant breast mass in digital mammograms with convolutional neural networks, in: Proceedings of the 2Nd International Symposium on Image Computing and Digital Medicine, 2018, pp. 47–50.

[57] N. Dhungel, G. Carneiro, A. P. Bradley, Fully automated classification of mammograms using deep residual neural networks, in: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), IEEE, 2017, pp. 310–314.

[58] D. Lévy, A. Jain, Breast mass classification from mammograms using deep convolutional neural networks, arXiv preprint arXiv:1612.00542 (2016).

[59] W. E. Fathy, A. S. Ghoneim, A deep learning approach for breast cancer mass detection, International Journal of Advanced Computer Science and Applications 10 (1) (2019). doi:10.14569/IJACSA.2019.0100123.
URL http://dx.doi.org/10.14569/IJACSA.2019.0100123

[60] X. Wang, G. Liang, Y. Zhang, H. Blanton, Z. Bessinger, N. Jacobs, Inconsistent performance of deep learning models on mammogram classification, Journal of the American College of Radiology (2020).