Automatic Calcification Morphology and Distribution Classification for Breast Mammograms With Multi-Task Graph Convolutional Neural Network

Hao Du, Melissa Min-Szu Yao, Siqi Liu, Liangyu Chen, Wing P. Chan, and Mengling Feng

***Abstract*—The morphology and distribution of microcalcifications are the most important descriptors for radiologists to diagnose breast cancer based on mammograms. However, it is very challenging and time-consuming for radiologists to characterize these descriptors manually, and there also lacks of effective and automatic solutions for this problem. We observed that the distribution and morphology descriptors are determined by the radiologists based on the spatial and visual relationships among calcifications. Thus, we hypothesize that this information can be effectively modelled by learning a relationship-aware representation using graph convolutional networks (GCNs). In this study, we propose a multi-task deep GCN method for automatic characterization of both the morphology and distribution of microcalcifications in mammograms. Our proposed method transforms morphology and distribution characterization into node and graph classification problem and learns the representations concurrently. We trained and validated the proposed method in an in-house dataset and public DDSM dataset with 195 and 583 cases,respectively. The proposed method reaches good and stable results with distribution AUC at 0.812 ± 0.043 and 0.873 ± 0.019, morphology AUC at 0.663 ± 0.016 and 0.700 ± 0.044 for both in-house and public datasets. In both datasets, our proposed method demonstrates statistically significant improvements compared to the baseline models. The performance improvements brought by our proposed multi-task mechanism can be attributed to the association between the distribution and morphology of calcifications in mammograms, which is interpretable using graphical visualizations and consistent with the definitions of descriptors in the standard BI-RADS guideline. In short, we explore, for the first time, the application of GCNs in microcalcification characterization that suggests the potential of using graph learning for more robust understanding of medical images.**

***Index Terms*—Calcification characterization, graph convolutional network, mammogram analysis.**

I. Introduction

According to Global Cancer Statistics 2020, breast cancer has overtaken lung cancer as the most common cancer around world [1]. Even so, the good news is that the 5-year survival rate for breast cancer can be as high as 90% if it is detected early before it progresses to metastatic cancer [2]. Mammography is currently the most effective tool for early detection of breast cancer, and it is widely adopted in breast cancer screening [3]. Mammography images commonly have high resolution, which enables the detection of microcalcifications (MCs) at an early stage. MC clusters are important early signs of breast cancer, accounting for approximately 50% of the diagnosed cases [4], [5]. An MC cluster contains at least 3 individual MCs where each MC is a small amount of calcium deposits in breast tissue and appears as small bright spots in mammograms [6].

Different types of MCs are associated with different probabilities of malignancy [7]. Formally, the American College of Radiology Breast Imaging Reporting and Data System (ACR BI-RADS) classifies calcifications into either the 'typically benign' or 'suspicious' category based on the morphology and distribution of calcifications [8]. Morphology describes the form of calcifications based on shape, size, brightness, roughness etc. Distribution describes how calcifications spread throughout the breast tissue. The morphology and distribution of calcifications, illustrated in Fig. 1, are the most important characteristics considered by radiologists to provide appropriate follow-up recommendations.

Manuscript received 19 January 2022; revised 31 August 2022 and 17 January 2023; accepted 20 February 2023. Date of publication 27 February 2023; date of current version 7 August 2023. This work was supported in part by the National Research Foundation Singapore through its AI Singapore Programme under Grant AISG-GC-2019-001-2A, in part by the NMRC Health Service Research under Grant MOH-000030-00, and in part by the Ministry of Education through the Higher Education Sprout Project in Taiwan under Grants DP2-109-21121-01-A-03 and DP2-110-21121-01-A-03. *(Hao Du and Melissa Min-Szu Yao contributed equally to this work.) (Corresponding author: Mengling Feng.)*

Hao Du and Mengling Feng are with the Saw Swee Hock School of Public Health and Institute of Data Science, National University of Singapore, Singapore 119260 (e-mail: duhao@u.nus.edu; ephfm@nus.edu.sg).

Melissa Min-Szu Yao and Wing P. Chan are with the Department of Radiology, Wan Fang Hospital and Department of Radiology, School of Medicine, College of Medicine, Taipei Medical University, Taipei 110, Taiwan (e-mail: manseiyiu@gmail.com; wingchan@tmu.edu.tw).

Siqi Liu is with the NUS Graduate School for Integrative Science and Engineering, National University of Singapore, Singapore 119260 (e-mail: e0272316@u.nus.edu).

Liangyu Chen is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: cly82811955@gmail.com).

Digital Object Identifier 10.1109/JBHI.2023.3249404

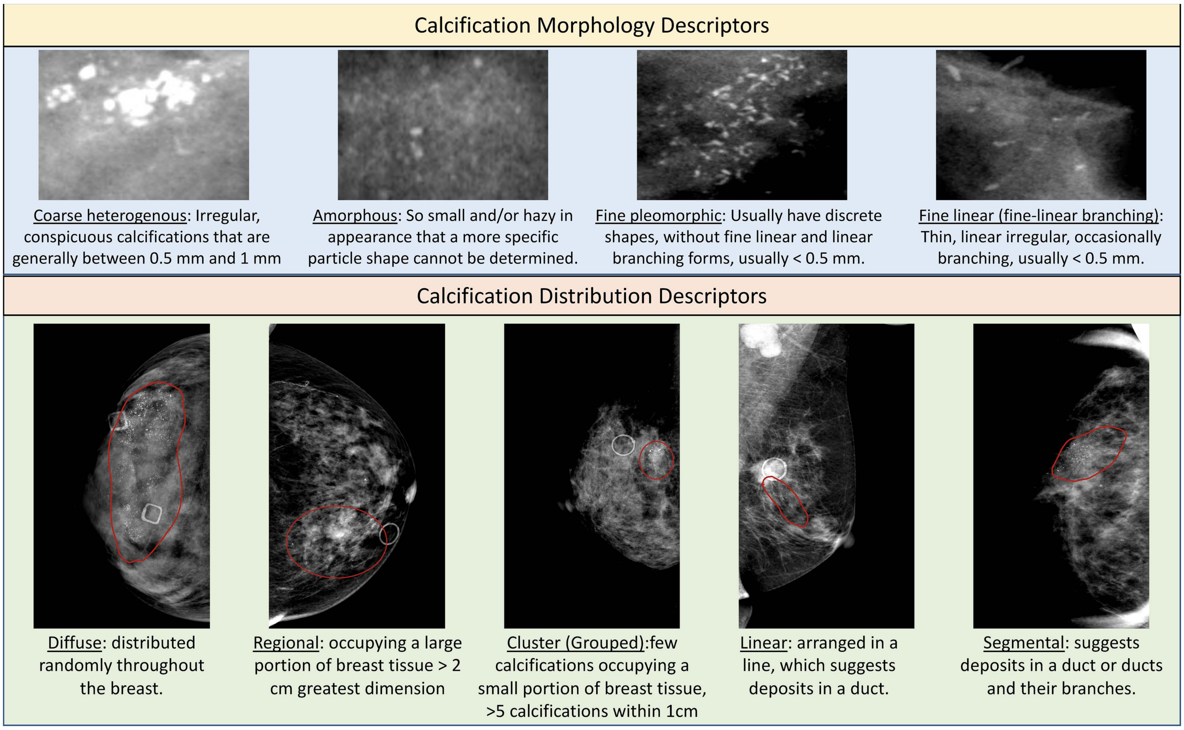
Calcification Morphology Descriptors
Coarse heterogenous: Irregular, conspicuous calcifications that are generally between 0.5 mm and 1 mm
Amorphous: So small and/or hazy in appearance that a more specific particle shape cannot be determined.
Fine pleomorphic: Usually have discrete shapes, without fine linear and linear branching forms, usually < 0.5 mm.
Fine linear (fine-linear branching): Thin, linear irregular, occasionally branching, usually < 0.5 mm.
Calcification Distribution Descriptors
Diffuse: distributed randomly throughout the breast.
Regional: occupying a large portion of breast tissue > 2 cm greatest dimension
Cluster (Grouped):few calcifications occupying a small portion of breast tissue, >5 calcifications within 1cm
Linear: arranged in a line, which suggests deposits in a duct.
Segmental: suggests deposits in a duct or ducts and their branches.

Fig. 1. Examples of morphology and distribution types. Types of suspicious morphology include coarse heterogeneous, fine pleomorphic, amorphous and fine linear (fine-linear branching). The types of distribution includes diffuse, regional, cluster(grouped), linear and segmental. All morphology and distribution descriptors are listed in the order of increasing risk of malignancy from left to right.

Recently, numerous deep learning based computer-aided diagnosis (CADx) methods have been developed in medical imaging, especially mammography [9], [10], [11], [12], [13], [14], [15], [16]. Lotter et al. developed a convolutional neural network based CADx system to perform malignancy classification of mammograms and digital breast tomosynthesis [16]. The system outperformed five breast-imaging specialists in datasets from U.K., USA and China [16]. Liu et al. introduced anatomy-aware graph convolutional network into mammogram mass detection task [13]. The proposed model showed statistically significant improvements compared to the state-of-the-art performance. More specifically, for calcifications, many CADx methods have been proposed by researchers to classify calcification clusters into benign or malignant [17], [18], [19], [20], [21], [22], [23], [24], [25]. Alam et al. [17], [19] selected calcification density, distances from cluster centroids, cluster areas and calcification sizes to discriminate between benign and malignant calcification clusters. Singh et al. [18] utilized shape and texture features to determine malignancy. Although the effectiveness of these features have been proven, existing CADx methods are unable to characterize the MCs into the descriptors of morphology and distribution, as recommended by ACR BI-RADS [8]. **Automatic characterization of calcifications is important to reproduce the chain of reasoning for mammogram interpretation, leading to more accurate and robust understanding of mammograms.**

To address this challenge, we formulate the characterization of calcifications in mammograms as a multi-task classification problem and propose a graph neural network framework. Firstly, we transform the calcifications in mammography images to graphical data to represent the spatial and visual information. That is, each calcification is represented as a node, and nodes are connected according to their geometric relationships with their nearby calcifications. Following the transformation, we formulate the morphology classification into a 'node classification task' and the distribution classification into a 'graph classification' task. We propose a multi-task model with graph convolutional neural networks (GCNs) to solve both tasks. GCN is a deep learning based method that extends convolutional operations to graphical data. GCN is designed to aggregate each vertex's feature with the features from the neighboring vertices to learn relationship-aware representations for graph or node classification tasks. By employing GCNs [26], [27], [28], we incorporate both local patch features and topological structures. We developed a multi-task learning framework to automatically abstract data representations that are applicable to both the morphology and distribution classification tasks. This ensures the generalizability of the proposed model. Our main contributions are as follows:

1) We transform information of calcification in mammography images into graphical representations.
2) We propose a deep GCN based framework to model the node and graph embeddings for both morphology and distribution tasks.
3) We develop a multi-task GCN-based solution to characterize both the morphology and distribution descriptors simultaneously. We demonstrated with extensive experiments that the proposed multi-task training strategy

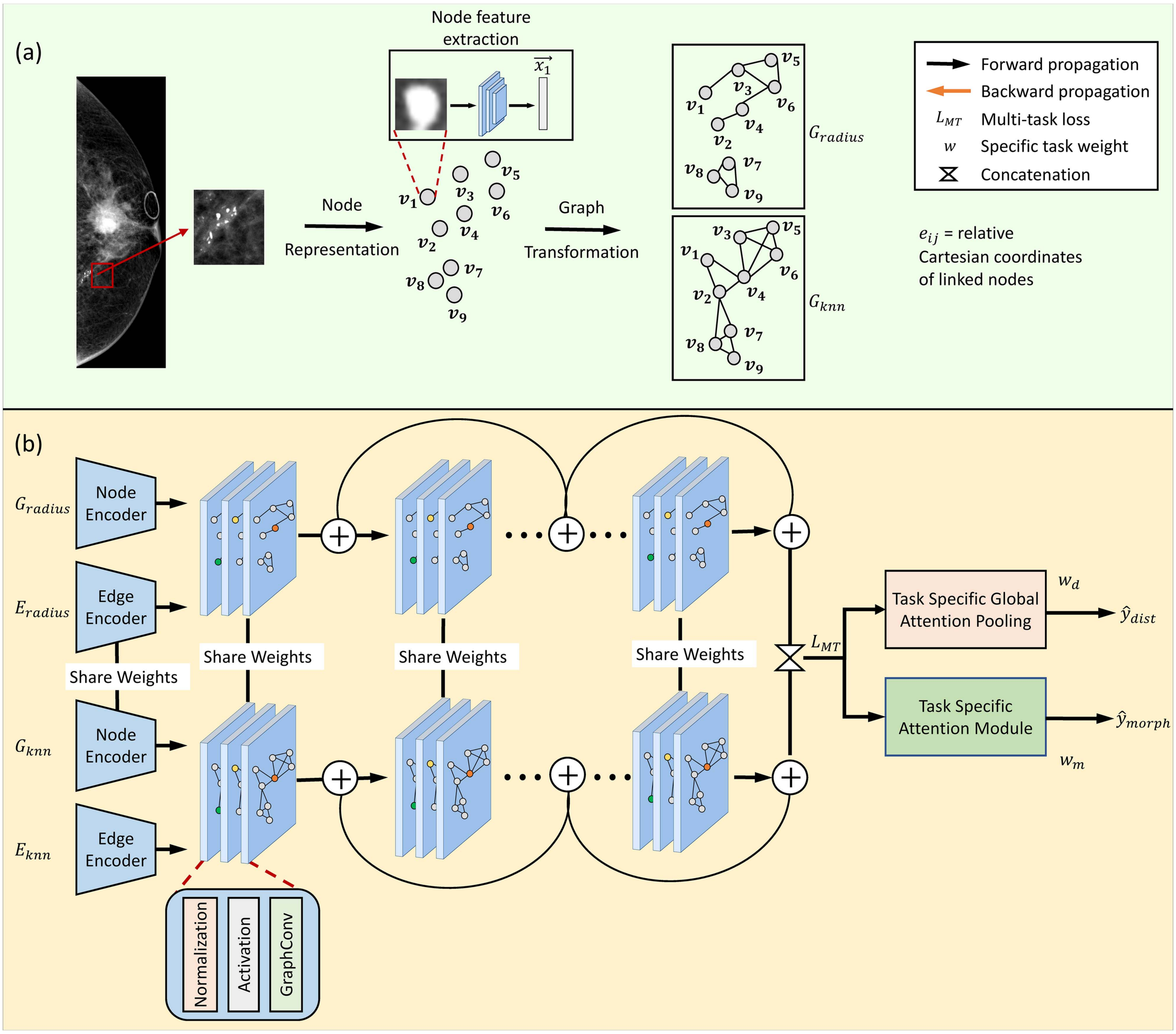

Fig. 2. Proposed framework demonstration. (a) Illustration of graph construction for calcification clusters. (b) Illustration of multi-task deep GCN with inputs from (a).

leads to better and more robust performance compared to models trained on a single task and other baseline models.

II. Methodology

A. Problem Definitions

The structure of proposed model is divided into graph construction and multi-task GCN. In the first step, we transform the calcifications in mammography images into graphical data by using a convolutional neural network (CNN) based feature extractor and graph transformation functions. Following graph construction, the proposed GCN jointly learns representations for node and graph classification with the multi-task training strategy. The end-to-end framework is illustrated in Fig. 2.

Let x^I be a mammography image, x^c be the set of calcifications in the image. A set of N mammography images $X = \{x_i\}_{i=1}^N$ where $x_i = (x_i^I, x_i^c)$ are included in our dataset. We transform image set X to graphical set $\mathcal{G}$ with $G_i \in \mathcal{G}$ and $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$ and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ are the sets of vertices and edges, respectively. e_{ij} represents an edge connecting vertices v_i and v_j if the edge $e_{ij} \in \mathcal{E}$. A vertex v and an edge e in the graph are associated with vertex features $h_v \in \mathbb{R}^D$ and edge features $h_e \in \mathbb{R}^C$ respectively, where D and C are dimensions of vertex and edge features. There are two tasks to investigate: (1) node (morphology) classification, where each vertex v has a label y_v and we aim to learn function f and representation r_v such that the vertex label could be predicted as $y_v = f(r_v)$; (2) graph (distribution) classification, where the graph has a label y_g and we aim to learn function g and representation vector r_g to predict the label of the graph as $y_g = g(r_g)$. The focus of this study is developing a multi-task GCN to effectively learn the node and graph embeddings for morphology and distribution classification of calcifications. These calcifications' locations, x^C, are annotated by radiologists. If deployed as a real-world CADx application, the proposed model should be equipped with a detection module which automatically detects calcifications. The detection module is not included in this study so as not to dilute the focus of the study. Such detection modules can be developed based on several existing studies which achieved accuracy and ROC-AUC over 95% [29], [30], [31]. The impacts of integrating a calcification detection module can be further studied in future studies.

B. Graph Construction

Graph construction is demonstrated in part (a) of Fig. 2. For each mammography image with calcifications (x_i^I, x_i^c), we define a set of patches as $P = \{p_1, p_2, \ldots, p_n\}$, where p represents an image patch that locates at the center of a calcification with dimension $M \times M$. We extract high level features from patches P with a convolutional neural network (CNN) as a feature extractor. We concatenate extracted features with the normalized coordinates of the patches to form the node feature h_v. The edge features h_e are defined as relative Cartesian coordinates of linked nodes. Following node and edge feature extraction, we construct two types of graphs based on the spatial connectivity relationship between calcifications:

1) K-nearest neighbor (KNN) graph G_{knn}: Creates edges if the nodes are within the k nearest neighbors. KNN graphs have been widely adopted in point cloud classification and segmentation [32], [33], [34], image classification [35], etc. However, it may cause information loss from disconnected neighbors in dense calcification clusters or introduce noise when the node is an outlier from the calcification cluster.
2) Radius graph G_{radius}: Creates edges based on node positions to all other nodes within a given distance. The radius graph solves the limitations introduced by the KNN graph described above. However, it is affected by a constant distance threshold which may cause information loss for vertices beyond the threshold.

The process of graph construction is shown in Algorithm 1.

Algorithm 1: Calcification Characterization Multi-graph Construction Algorithm.

```
Input: A mammography image $x^I$ and a set of
       calcifications with coordinates of the
       calcifications' centers
       $x^C = \{(cx_1, cy_2), (cx_2, cy_2), \ldots, (cx_n, cy_n)\}$;
       $n$ is the number of calcifications
Result: Adjacency Matrix for KNN graph$A^{KNN}$,
        Adjacency Matrix for radius graph $A^{radius}$
        Vertex Feature Matrix $H_V$, Edge Feature
        Matrix $H_E$
Function ConstructGraph($x^I, x^C, n$):
    for $i \leftarrow 1$ to $n$ do
        Segment patch $p_i$ at $(cx_i, cy_i)$ from $X^I$ ;
        $H_V \longleftarrow CNN\_Feature\_Extractor(p_i)$;
        if $j \in NearestNeighbor(k)$ then
            $A_{ij}^{KNN}$ = 1;
            $H_E^{KNN}$ = relative Cartesian coordinates;
        end
        $d_{ij} = \sqrt{(cx_i^2 - cx_j^2) + (cy_i^2 - cy_j^2)}$ ;
        if $d_{ij} < r$ then
            $A_{ij}^{radius} = 1$ ;
            $H_E^{radius}$ = relative Cartesian coordinates;
        end
    end
```

C. Deep Graph Convolutional Network

The constructed multi-graph inputs are then fed into deep GCN, as illustrated in Fig. 2(b). The weights of the proposed GCN are shared across multi-graph inputs. The design of weight sharing targets to learn the common features that can describe the characteristics of both graphs. Following [28] and [36], we used GCN blocks with Normalization → ReLU → GraphConv → Addition and GENeralized Aggregation Networks (GENconv) as GraphConv backbone. In GENconv, the message construction function $p^{(l)}$ is defined to apply on vertex feature $h_v^{(l)}$, neighbor vertex's feature $h_u^{(l)}$ and edge feature $h_{e_{vu}}^{(l)}$ to construct the message to propagate. $p^{(l)}$ is defined as:

$$
\begin{aligned}
m_{vu}^{(l)} &= \rho^{(l)}\left(h_v^{(l)}, h_u^{(l)}, h_{e_{vu}}^{(l)}\right) \\
&= \mathrm{ReLU}\left(h_u^{(l)} + \mathbb{1}(h_{e_{vu}}^{(l)}) \cdot h_{e_{vu}}^{(l)}\right) + \epsilon, u \in \mathcal{N}(v) \qquad (1)
\end{aligned}
$$

where the ReLU$(\cdot)$ represents the rectified linear unit activation function [37], $\mathbb{1}(\cdot)$ is an indicator function which equals to 1 when edge features exist otherwise 0, and ϵ is a small positive constant. SoftMaxAgg$_\beta$ is then used as the message aggregation function and defined as:

$$
\begin{aligned}
m_v^{(l)} &= SoftMaxAgg_\beta(\cdot) \\
&= \sum_{u \in \mathcal{N}(v)} \frac{\exp(\beta m_{vu}^{(l)})}{\sum_{i \in \mathcal{N}(v)} \exp(\beta m_{vu}^{(l)})} \cdot m_{vu}^{(l)}, \qquad (2)
\end{aligned}
$$

where $\mathcal{N}(v)$ is the set of neighbors of vertex v and β is a hyper-parameter which controls the aggregation function. Message normalization *MsgNorm* is then introduced to address the over-smoothing and gradient vanishing problem in training deep GCNs. MsgNorm normalizes the features of the aggregated message $m_v^{(l)}$ by combining them with other features during the vertex update phase. Suppose MsgNorm is applied to a multi-layer perceptron (MLP) vertex update function MLP$(h_v^{(l)} + m_v^{(l)})$, the vertex update function becomes as follows:

$$
\begin{aligned}
h_v^{l+1} &= \phi^{l(l)}(h_v^{(l)}, m_v^{(l)}) \\
&= \mathrm{MLP}\left(h_v^{(l)} + s \cdot \|h_v^{(l)}\|_2 \cdot \frac{m_v^{(l)}}{\|m_v^{(l)}\|_2}\right) \qquad (3)
\end{aligned}
$$

where s is a learnable scaling factor. The aggregated message $m_v^{(l)}$ is first normalized by its ℓ_2 norm and then scaled by the ℓ_2

norm of $h_v^{(l)}$ by a factor of s. The scaling factor s is set to be a learnable scalar with an initialized value of 1.

D. Multi-Task Learning

In this study, the proposed multi-task GCN is trained to jointly perform morphology and distribution classification. In general, the model was trained by a multi-task loss $L_{MT} = w_m L_m + w_d L_d$ where $w_m L_m$ and $w_d L_d$ are weighted cross-entropy loss for morphology and distribution classification, respectively. In ACR BI-RADS guideline, morphology and distribution of calcifications are equally important. Therefore, we introduced GradNorm [38] to learn both tasks at an equal pace. To explain GradNorm in the proposed method, we define the necessary quantities as below:

- W: The subset of the full network weights $W \subset \mathcal{W}$. The weights of the last shared layer is generally chosen as W.
- $G_W^{(i)}(t) = \|\nabla_W w_i(t) L_i(t)\|_2$: The L_2 norm of the gradient over the weighted loss $w_i(t)L_i(t)$ for task i with respect to W, at training step t.
- $\overline{G}_W(t) = E_{task}[G_i^W(t)]$: The average value of gradient norms over all tasks for training step t.
- $\tilde{L}_i(t) = \frac{L_i(t)}{L_i(0)}$: The loss ratio as the inverse training rate of task i at step t;
- $r_i(t) = \frac{\tilde{L}_i(t)}{E_{task}[\tilde{L}_i(t)]}$: The relative inverse training rate of task i at step t.

In order to balance the gradient magnitudes $G_W^{(i)}$ for each task, the mean gradient norm across all tasks $\overline{G}_W$ is set as the common scale target. The relative inverse training rate of task i, $r_i(t)$, is used to balance the learning pace of all tasks. The target gradient norm for task i is:

$$G_W^{(i)}(t) \rightarrow \overline{G}_W(t) \times [r_i(t)]^{\alpha}, \tag{4}$$

where α controls the strength of the restoring force which pulls tasks back to a common training rate. A higher value of α indicates a higher strength to enforce training rates to be balanced.

Equation (4) provides the target gradient norms for task i. At each training step t, we update the loss weights $w_i(t)$ to bring gradient norms close to the target for task i. L_1 loss between the actual gradient norms and the target at each time step for each task is introduced as L_{grad} and we summed L_{grad} across both morphology and distribution classification tasks.

$$L_{grad}(t; w_i(t)) = \sum_i |G_W^{(i)}(t) - \overline{G}_W(t) \times [r_i(t)]^{\alpha}|_1 \tag{5}$$

III. Experimental Results and Discussion

A. Datasets

1) TMU Dataset: We collected a full field digital mammogram dataset for this study from the Wan Fang Hospital, Taipei Medical University (TMU), from June 2010 and October 2018. The dataset contains 387 mammography images from 200 patients who were classified as ACR BI-RADS category 4 and 5 with documented calcifications from the original radiological reports. All cases were confirmed breast cancers from biopsy tests. Descriptors of morphology and distribution were annotated by a senior radiological technologist and carefully reviewed by a panel with two senior radiologists in a joint meeting. Our clinical annotators are breast imaging experts to ensure the reliability of the ground truths in the annotation process. The radiological technologist is a senior radiographical technologist with 15 years of experience in mammogram reading. The review panel consists of the professor in the Department of Radiology, Taipei Medical University, and chief of breast imaging in Wang Fang Hospital with 32 and 20 years of experience, respectively. To assess the impact of inter-observer variability on the ground truths, we evaluated the agreement between the radiological technologist and one of the senior radiologists in the review panel using the Cohen's kappa [39]. The inter-observer kappa values are 0.978 and 0.992 on annotating distribution and morphology descriptors, respectively. The kappa results indicate a high degree of agreement between annotators, thus inter-observer variability has little impact on obtained ground truths (kappa 0.81–1.00: almost perfect agreement [40]).

The study was jointly approved by NUS Institutional Review Board (NUS-IRB) (Approval No. 2019/00159) and Joint Institutional Review Board of Taipei Medical University (TMU-JIRB) (Approval No. N202006039). We excluded 5 cases with no biopsy confirmation, malignant phyllodes tumor or low image qualities. The basic characteristics of the final cohort is shown in online Appendix Table 1 [41].

2) CBIS-DDSM Dataset: We validated our proposed method on the CBIS-DDSM (Curated Breast Imaging Subset of DDSM) dataset. CBIS-DDSM [42] is an updated and standardized version of the Digital Database for Screening Mammography (DDSM) dataset [43]. The DDSM dataset is a publicly available database of 2,620 scanned film mammography studies. The cases were annotated with region of interests (ROIs) for calcifications and masses, and BI-RADS descriptors for calcification morphology, calcification distribution, mass shape, mass margin and breast density. Following the same inclusion criteria as the TMU dataset, we included cases which were classified as ACR BI-RADS category 4 and 5 with annotated calcifications. We excluded cases which contained calcifications with more than one morphology type, because CBIS-DDSM does not provide separate ROI annotations for multiple morphology descriptors. The number of such cases is relatively small ($<10\%$). As a result, we extracted 583 mammography images from CBIS-DDSM for this study.

B. Experiments and Results

1) Implementation Details: The experiments were implemented with PyTorch framework and Pytorch Geometric package [44], [45]. The dimension of calcification patches was set at 14×14 (1.32 mm $\times$ 1.32 mm), as the size of calcifications are generally less than 14 pixels in mammograms [46]. Hyper-parameters were selected through grid search over potential parameters. Empirically, the hidden size in proposed network was set at 128, α was set at 1.5, k was set for KNN at 4 and distance threshold for radius graph was set at 112. Initial learning rate was set at 10^{-3} and decayed by $\frac{1}{10}$ every 10 epochs. The models were trained by Adam optimizer on an Ubuntu server with 4 NVIDIA V100 GPU cards for 100 epochs. The models were trained and validated independently in TMU and DDSM dataset in 5-fold cross validation manner. The splitting of the folds was performed on the patient level such that there was no overlapping of mammograms from same patients between training and testing folds. No statistically significant differences were found between training and testing folds in all demographic variables (details in online Appendix Table 2-6) [41]. We also conducted ablation studies to evaluate the model's performance after removing each proposed module in order to understand the proposed module's contribution to the overall model. To ensure reproducibility, our implementations of the experiments are publicly available on GitHub.[1]

2) Performance Comparison: To the best of our knowledge, there is no state-of-the-art models to characterize morphology and distribution of calcifications in mammography images. In order to establish baselines for comparison, we employed multiple popular CNN and GCN models that have been widely and successfully applied in medical imaging as baseline models. For CNN baseline models, we regarded both the distribution and morphology classification tasks as a multi-classification problem. Distribution baseline models take mammography images X^I as input to predict the types of distribution. For morphology baseline models, the set of patches P defined in Section II-B is used as inputs. Each patch p is located at the center of a calcification and considered as an independent input to baseline models. Similar to vertices in constructed calcification graphs, each patch is associated with a morphology label. The baseline models classify the patch set into morphology categories. For GCN baseline models, we evaluate the models' performance for graph and node classification tasks separately. The employed baseline models include:

1) ResNet [47]: ResNet has been one of the most successful and popular network architectures in computer vision field since proposed in 2015. Residual blocks with skip connections were proposed to solve the problem of gradient vanishing in training deep neural networks. ResNet and its variants have been successfully adopted in many applications such as medical image classification, segmentation, synthesis etc [48], [49], [50], [51]. In this study, we used ResNet-50 as a baseline comparison.
2) DenseNet [52]: In DenseNet architectures, layers are densely connected with each other to reuse features and preserve global state. DenseNet demonstrated excellent results on small benchmark datasets such as Cifar-10 and Cifar-100 [53], [54]. We used DenseNet-121 as one of the baseline models.
3) MobileNet [55]: MobileNet was proposed primarily for mobile and embedded devices. MobileNet uses depthwise separable convolutions to achieve high accuracy with low latency. Many researchers have demonstrated the effectiveness of MobileNet architectures in various medical applications such as diabetic foot ulcer and lung disease detection [56], [57]. We used MobileNetV2 [58] in this study.
4) EfficientNet [59]: EfficientNets is proposed using network architecture search, which performs compound scaling in depth, width, and resolution. EfficientNets achieved the state-of-the-art performance in various benchmark datasets with significantly reduced parameters compared to other models. We adopted EfficientNet-B0 in experiments of this study.
5) GCN (vanilla) [27]: GCN is proposed to generalize convolution operations to non-Euclidean graphical data. GCN has been successfully applied in medical tasks such as COVID-19 classification [60], drug discovery [61] and brain fMRI analysis [62].
6) Graph attention network (GAT) [63]: GAT is one of the most successful variant of the vanilla GCN. GAT introduced masked self-attention into graph convolution operations to apply weights to information propagation from neighbouring vertices. GAT has demonstrated its potential in medical tasks such as Alzheimer's disease analysis [64], identification of bipolar disorder [65] and medical image enhancements [66].

We addressed several image quality issues to ensure the fair comparison with baseline models. The CBIS-DDSM dataset was collected from scanned analog films. As a result, the image quality is much poorer compared to digital mammograms. We applied the preprocessing techniques including CLAHE enhancement and lesion segmentation [67]. For the TMU dataset, a small amount of collected mammography images were overexposed ($<10\%$), showing bright and white areas in breasts. This overexposure problem actually does not affect the performance of the proposed model because the overexposed areas do not overlap with calcifications and the proposed model takes calcification patches as inputs, however, the overexposure may affect the performance of baseline models because the baseline models take the entire mammography images as inputs. To help the baseline model overcome this issue, we preprocessed the images by removing the overexposed regions from the affected mammograms.

In our experiments, both distribution and morphology classification tasks are formulated as multi-class classification tasks. Following the standard medical guideline BI-RADS fifth edition [8], the number of classes are 5 and 4 for distribution and morphology descriptors, respectively. The examples of the descriptors are shown in Introduction. We used the multi-class AUC as primary evaluation metrics [68]. AUC was evaluated at the node and graph level for morphology and distribution classification, respectively. In addition, we evaluated precision, recall, F1-score and accuracy for comparative purposes. All performance metrics were evaluated with weighted average method across multiple classes [69].

95% confidence intervals (95% CI) and statistical tests were used for performance comparison. Confidence intervals were computed with 1000 bootstraps [70]. Randomized permutation tests were used to test for statistically significant differences [71]. To overcome multiple comparison, the significance level was adjusted to 0.008 using Bonferroni correction [72].

[1] [Online]. Available: https://github.com/DuHao10086/multi_task_calcification.

TABLE I
THE PERFORMANCE COMPARISON ON TMU DATASET BETWEEN BASELINE MODELS AND PROPOSED MODEL ON DISTRIBUTION AND MORPHOLOGY CLASSIFICATION

Type	Methods	Distribution				
		ROC-AUC±std [95% CI]	**Precision±std [95% CI]**	**Recall±std [95% CI]**	**F1-score±std [95% CI]**	**Accuracy±std [95% CI] (%)**
Baselines	ResNet	0.660±0.018 [0.645, 0.674]	0.535±0.042 [0.501, 0.569]	0.550±0.033 [0.521, 0.578]	0.463±0.032 [0.433, 0.488]	55.026±3.266 [52.152, 57.805]
	DenseNet	0.574±0.050 [0.531, 0.618]	0.465±0.070 [0.408, 0.522]	0.530±0.035 [0.503, 0.561]	0.452±0.040 [0.422, 0.487]	52.985±3.518 [50.367, 56.156]
	MobileNet	**0.688±0.026 [0.667, 0.709]**	**0.553±0.021 [0.535, 0.569]**	**0.568±0.031 [0.543, 0.596]**	**0.484±0.036 [0.453, 0.515]**	**56.809±3.124 [54.308, 59.669]**
	EfficientNet	0.674±0.041 [0.642, 0.712]	0.550±0.099 [0.435, 0.574]	0.540±0.029 [0.516, 0.567]	0.464±0.043 [0.428, 0.501]	53.967±2.953 [51.565, 56.732]
	GCN	0.623±0.051 [0.580, 0.665]	0.395±0.099 [0.314, 0.489]	0.509±0.014 [0.497, 0.520]	0.389±0.034 [0.361, 0.416]	50.880±1.362 [49.731, 52.013]
	GAT	0.625±0.053 [0.582, 0.673]	0.425±0.113 [0.329, 0.521]	0.511±0.021 [0.495, 0.529]	0.393±0.037 [0.362, 0.422]	51.130±2.056 [49.487, 52.902]
Proposed	**Multi-task, multi-graph, 8-layer GCN**	**0.812±0.043 [0.774, 0.846]**	**0.630±0.055 [0.581, 0.679]**	**0.635±0.036 [0.598, 0.661]**	**0.597±0.026 [0.574, 0.619]**	**63.467±3.573 [59.754, 66.073]**

(a) Classification performance on distribution descriptors

Type	Methods	Morphology				
		ROC-AUC±std [95% CI]	**Precision±std [95% CI]**	**Recall±std [95% CI]**	**F1-score±std [95% CI]**	**Accuracy±std [95% CI] (%)**
Baselines	ResNet	0.555±0.004 [0.552, 0.559]	0.361±0.048 [0.319, 0.401]	0.573±0.058 [0.569, 0.579]	0.445±0.050 [0.440, 0.456]	57.340±5.777 [56.860, 57.911]
	DenseNet	**0.594±0.017 [0.580, 0.609]**	**0.380±0.086 [0.304, 0.452]**	**0.578±0.060 [0.573, 0.584]**	0.449±0.056 [0.447, 0.461]	**57.840±5.978 [57.251, 58.390]**
	MobileNet	0.571±0.012 [0.560, 0.580]	0.361±0.075 [0.294, 0.421]	0.574±0.058 [0.569, 0.579]	0.452±0.057 [0.448, 0.460]	57.413±5.850 [56.873, 57.861]
	EfficientNet	0.588±0.017 [0.573, 0.602]	0.367±0.046 [0.322, 0.404]	0.577±0.060 [0.572, 0.583]	0.453±0.058 [0.449, 0.463]	57.751±5.949 [57.193, 58.329]
	GCN	0.559±0.018 [0.542, 0.572]	0.336±0.062 [0.284, 0.386]	0.575±0.056 [0.570, 0.581]	0.447±0.042 [0.457, 0.470]	57.467±5.630 [57.001, 58.100]
	GAT	0.555±0.030 [0.530, 0.581]	0.344±0.048 [0.302, 0.386]	0.576±0.058 [0.571, 0.583]	**0.460±0.049 [0.465, 0.479]**	57.656±5.793 [57.123, 58.306]
Proposed	**Multi-task, multi-graph, 8-layer GCN**	**0.663±0.016 [0.648, 0.676]**	**0.471±0.043 [0.466, 0.526]**	**0.596±0.049 [0.591, 0.601]**	**0.518±0.034 [0.517, 0.529]**	**59.661±4.958 [59.128, 60.169]**

(b) Classification performance on morphology descriptors

† **Statistical tests were performed between baseline and proposed models on all evaluation metrics.** $p < 0.0001$ **for all comparisons, which are below the adjusted significance level (0.008). All comparisons were statistically significant.**
Best results in baseline models are represented in red bold. Results from the proposed method are highlighted using black bold.

3) Results: As Tables I and II shows, our proposed model demonstrated leading performance across both tasks in two datasets. For the classification task on distribution, compared with the baseline models, ResNet, DenseNet, MobileNet, EfficientNet, vanilla GCN and GAT, our proposed model demonstrated a mean ROC-AUC improvement of 0.152, 0.238, 0.124, 0.138, 0.189 and 0.187 in the TMU dataset, respectively. In addition, our proposed model achieved mean improvements of 0.077, 0.067, 0.113 and 6.658 on precision, recall, F1-score and accuracy respectively, compared to best results in baseline models. For the classification task on morphology, the improvements of ROC-AUC, precision, recall, F1-score and accuracy were 0.069, 0.091, 0.018, 0.058 and 1.821 in the TMU dataset, respectively, compared to best results in baseline

TABLE II
THE ROC-AUC COMPARISON ON CBIS-DDSM DATASET BETWEEN BASELINE MODELS AND PROPOSED MODEL ON DISTRIBUTION AND MORPHOLOGY CLASSIFICATION

Type	Methods	Distribution ROC-AUC±std [95% CI]	Morphology ROC-AUC±std [95%CI]
Baselines	ResNet	0.632±0.039 [0.599, 0.666]	0.603±0.020 [0.588, 0.621]
	DenseNet	0.668±0.049 [0.626, 0.710]	0.611±0.032 [0.578, 0.636]
	MobileNet	0.660±0.035 [0.632, 0.694]	0.599±0.027 [0.578, 0.624]
	EfficientNet	**0.672±0.030 [0.648, 0.700]**	**0.613±0.014 [0.602, 0.627]**
	GCN	0.605±0.038 [0.570, 0.635]	0.580±0.041 [0.547, 0.616]
	GAT	0.622±0.039 [0.584, 0.651]	0.552±0.022 [0.531, 0.570]
Proposed	**Multi-task, multi-graph, 8-layer GCN**	**0.873±0.019 [0.859, 0.891]**	**0.700±0.044 [0.661, 0.735]**

† **Statistical tests were performed between baseline and proposed models on ROC-AUC in both tasks.** $p < 0.0001$ **for all comparisons, which are below the adjusted significance level (0.008). All comparisons were statistically significant.**
Best results in baseline models are represented in red bold. Results from the proposed method are highlighted using black bold.

models [Table I]. Similarly, the proposed model outperformed all baseline models in the mean ROC-AUC in the CBIS-DDSM dataset, with maximum improvements of 0.268 and 0.148 in distribution and morphology classification tasks, respectively [Table II]. ROC-AUC results of each type of distribution and morphology descriptors are shown in Appendix Table 7 and 8. [41] The improvements on distribution classification task can be attributed to the design of GCN which captures the geometrical relationships between calcifications, thereby improving the ability to distinguish distribution types. For morphology, the improvements can be attributed to the message propagation from neighboring vertices with the same morphology type. Calcifications with the same morphology tend to locate in a nearby region or cluster. Therefore, the feature propagation from neighbors enhances the proposed model to distinguish morphology.

C. Ablation Study

1) Ablation Experiments for Multi-Task Network: We separately trained task-specific models by removing the distribution or morphology branch respectively [Tables III and IV]. Although statistical significance was not reached due to the limited sample size, the multi-task model outperformed the task-specific models in both tasks. For the distribution classification task, the multi-task architecture demonstrated 0.009 and 0.022 higher ROC-AUC than the task-specific architecture on the TMU and CBIS-DDSM datasets, respectively. For the morphology classification task, the proposed multi-task model achieved mean ROC-AUC improvements of 0.020 and 0.062 compared to the task-specific architectures in the two datasets. The improvements can be attributed to the fact that distribution and morphology are associated and jointly affect the radiologists' decision-making on malignancy diagnosis. For example, in ductal carcinoma in situ and invasive ductal carcinoma, fine linear or linear branching calcifications often have a segmental ductal distribution [73]. Fine pleomorphic and linear branching calcifications in a segmental distribution are highly suspicious for malignancy [73]. The design of the multi-task network learns the shared representation from the morphology and distributed labels, thus achieving improvements on both tasks.

2) Ablation Experiments for Depth of Deep GCNs: To investigate the effectiveness of depths of Deep GCN, we compared with different number of graph convolutional layers in the proposed network. The experiment results showed that relative larger number of GCN layers improves the performance, though no statistical significance was found due to the limited study sample size. In the TMU dataset, when the number of GCN layers increase from 2 to 8 layers, the mean ROC-AUC of distribution and morphology classification tasks increased by 0.009 and 0.016, respectively. When the number of GCN layers was further increased to 16 layers, the performance of the two tasks dropped by 0.011 and 0.028, respectively. A similar trend was also observed in the experiments on the CBIS-DDSM dataset.

In GCNs, single layer of GCN considers nearest neighbor while networks with multiple GCN layers perform message propagation and fusion from multi-hop neighbors. As mentioned, calcifications with same morphology locate in a nearby region or cluster and distribution considers how calcifications spread over the breast. To a certain extent, when the depth of GCN increases, message propagation from more hops of neighbors enhance the network's ability in classifying nodes and graphs. However, when the network depth increases further, the message propagation from further nodes may be harmful for morphology classification because the further nodes may not have the same type of morphology. Deeper GCN in this study may also suffer from over-smoothing and gradient vanishing problems, which could be further investigated in future studies.

3) Ablation Experiments for Multi-Graph Fusion: To investigate the effectiveness of multi-graph fusion, we compared with multi-task model with single radius or KNN graph as input to GCN. The experiment results showed that the multi-graph fusion improves the robustness of the model. The improvements were statistically significant compared to the GCN model with the KNN graph, while the improvements are not statistically significant compared to the GCN model with the radius graph. Comparing with single graph GCN models, the proposed multi-graph model achieved maximum ROC-AUC improvements of 0.096 and 0.078 for the distribution classification task, 0.069 and 0.109 for the morphology classification task in two datasets. As mentioned in Section II-B, individual graph has limitations in either morphology or distribution classification task. The design of multi-graph fusion enhances the model's ability to

TABLE III
ABLATION STUDY: THE PERFORMANCE COMPARISON ON TMU DATASET BETWEEN THE ABLATION MODELS AND THE PROPOSED MODEL ON DISTRIBUTION AND MORPHOLOGY CLASSIFICATION

Type	Methods	Distribution				
		ROC-AUC ±std [95% CI]	**Precision ±std [95% CI]**	**Recall ±std [95% CI]**	**F1-score ±std [95% CI]**	**Accuracy±std [95% CI] (%)**
Ablation study	Task-specific (Dis.)	0.803±0.042 [0.762, 0.836]	0.628±0.020 [0.607, 0.645]	0.630±0.033 [0.601, 0.657]	0.583±0.029 [0.556, 0.606]	62.957±3.293 [60.095, 65.641]
	Single-graph (Rad.)	0.788±0.034 [0.758, 0.818]	0.615±0.031 [0.591, 0.647]	0.624±0.028 [0.602, 0.651]	0.581±0.014 [0.570, 0.594]	62.448±2.828 [60.206, 65.084]
	Single-graph (KNN)	*0.716±0.054 [0.668, 0.760]	*0.494±0.030 [0.469, 0.520]	*0.581±0.014 [0.569, 0.593]	*0.510±0.012 [0.500, 0.519]	*58.054±1.369 [56.916, 59.265]
	2-layer GCN	0.803±0.041 [0.763, 0.839]	0.597±0.048 [0.566, 0.644]	0.630±0.035 [0.593, 0.658]	0.580±0.027 [0.557, 0.602]	62.954±3.520 [59.248, 65.803]
	4-layer GCN	0.806±0.050 [0.761, 0.848]	0.593±0.047 [0.555, 0.637]	0.632±0.031 [0.603, 0.656]	0.578±0.034 [0.551, 0.609]	63.201±3.064 [60.264, 65.533]
	16-layer GCN	0.801±0.047 [0.761, 0.843]	0.593±0.041 [0.561, 0.629]	0.632±0.036 [0.602, 0.662]	0.578±0.034 [0.550, 0.610]	63.207±3.575 [60.186, 66.228]
Proposed	**Multi-task, multi-graph, 8-layer GCN**	**0.812±0.043 [0.774, 0.846]**	**0.630±0.055 [0.581, 0.679]**	**0.635±0.036 [0.598, 0.661]**	**0.597±0.026 [0.574, 0.619]**	**63.467±3.573 [59.754, 66.073]**

(a) Classification performance on distribution descriptors

Type	Methods	Morphology				
		ROC-AUC±std [95% CI]	**Precision ±std [95% CI]**	**Recall ±std [95% CI]**	**F1-score ±std [95% CI]**	**Accuracy±std [95% CI] (%)**
Ablation study	Task-specific (Mor.)	0.643±0.016 [0.628, 0.657]	0.452±0.073 [0.394, 0.522]	0.593±0.052 [0.587, 0.599]	0.496±0.057 [0.444, 0.543]	59.347±5.206 [58.760, 59.913]
	Single-graph (Rad.)	0.648±0.016 [0.633, 0.663]	0.469±0.057 [0.424, 0.515]	0.587±0.057 [0.537, 0.637]	0.507±0.038 [0.475, 0.540]	58.663±5.679 [53.648, 63.677]
	Single-graph (KNN)	*0.594±0.013 [0.582, 0.605]	0.458±0.048 [0.415, 0.495]	0.590±0.054 [0.540, 0.636]	0.475±0.060 [0.421, 0.525]	59.016±5.453 [54.023, 63.557]
	2-layer GCN	0.647±0.016 [0.631, 0.659]	0.381±0.102 [0.300, 0.469]	0.576±0.061 [0.523, 0.630]	0.472±0.063 [0.420, 0.530]	57.592±6.114 [52.245, 63.030]
	4-layer GCN	0.624±0.033 [0.596, 0.653]	0.395±0.112 [0.295, 0.487]	0.574±0.060 [0.522, 0.622]	0.469±0.067 [0.411, 0.526]	57.424±5.965 [52.165, 62.243]
	16-layer GCN	0.635±0.024 [0.614, 0.655]	0.422±0.058 [0.369, 0.472]	0.594±0.051 [0.551, 0.637]	0.506±0.053 [0.460, 0.552]	59.449±5.111 [55.153, 63.746]
Proposed	**Multi-task, multi-graph, 8-layer GCN**	**0.663±0.016 [0.648, 0.676]**	**0.471±0.043 [0.466, 0.526]**	**0.596±0.049 [0.591, 0.601]**	**0.518±0.034 [0.517, 0.529]**	**59.661±4.958 [59.128, 60.169]**

(b) Classification performance on morphology descriptors

† **Statistical tests were performed between baseline and proposed models on all evaluation metrics of both tasks. Statistically significant comparisons are highlighted with *.** $p < 0.0001$ **for these comparisons, which are below the adjusted significance level (0.008).**
Results from the proposed method are highlighted using black bold. Mor.=morphology, dis.=distribution, rad.=radius.

learn representations from two graphs, thereby improving on both classification tasks.

D. Discussion

In this study, we proposed a multi-task GCN model to jointly classify morphology and distribution descriptors of calcifications in mammography images. The proposed model demonstrated improved performance compared to multiple representative baseline models. The improvements were statistically significant across two datasets, suggesting the model has the potential to generalize well across different demographics and image qualities. Compared to the recent application of GCN on mammograms by Liu et al. [13], our study is focused on the classification of morphology and distribution descriptors of calcifications, rather than mass detection in mammograms. In addition, the proposed model was designed to model the morphology and distribution descriptors simultaneously, which is, to the best of our knowledge, the first application of multi-task mechanism on the characterization of calcifications in mammograms.

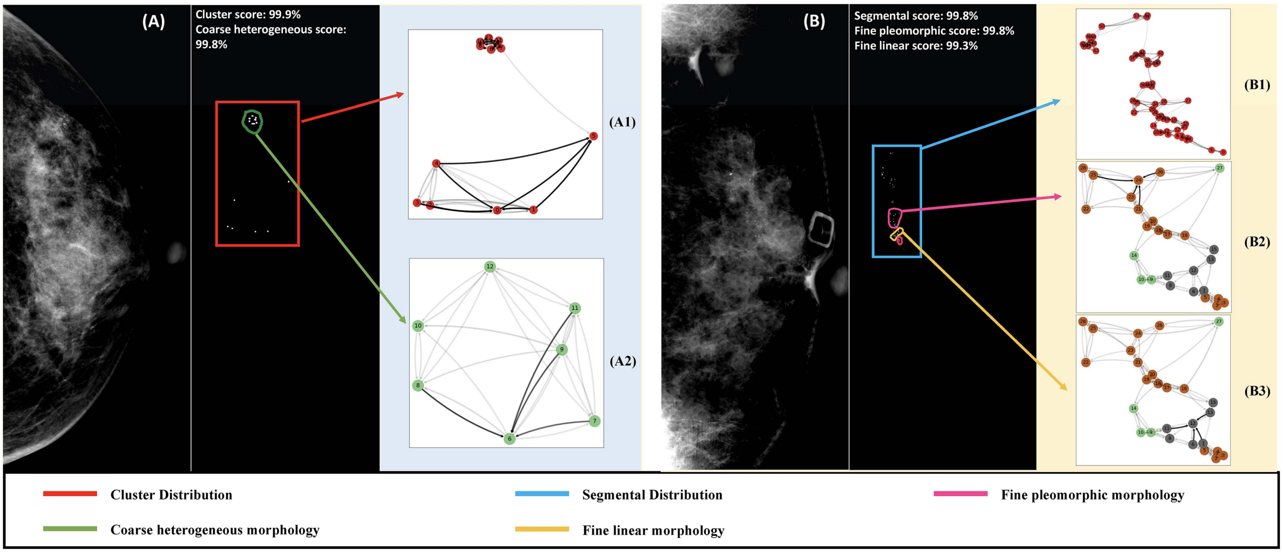

Fig. 3. Two case studies (A,B) generated from GNNexplainer for making graph prediction and node prediction. For each case study, the mammography image and radiological annotation are located at left hand side. Figure (A1) and (B1) identifies the important edges for graph prediction of case (A) and (B), respectively. Figure (A2) identifies the important subset of graph and edges for node prediction of node 6. Figure (B2) and (B3) demonstrate the important sub-graph and sub-edges for node prediction on node 24 and 12, respectively. In Figure (A2), (B2) and (B3), solid edges indicate significant contributions to node prediction. A full-size high-resolution version of the figure can be found online [41].

TABLE IV
The ROC-AUC Comparison on CBIS-DDSM Dataset Between Ablation Study Models and Proposed Model on Distribution and Morphology Classification

Type	Methods	Distribution ROC-AUC±std [95% CI]	Morphology ROC-AUC±std [95%CI]
Ablation study	Task-specific (Dist.)	0.851±0.019 [0.835, 0.867]	—
	Task-specific (Mor.)	—	0.638±0.057 [0.586, 0.685]
	Single-graph (Rad.)	0.872±0.025 [0.852, 0.893]	0.655±0.050 [0.614, 0.698]
	Single-graph (KNN)	*0.795±0.026 [0.770, 0.815]	*0.591±0.028 [0.569, 0.616]
	2-layer GCN	0.866±0.024 [0.848, 0.888]	0.675±0.041 [0.640, 0.709]
	4-layer GCN	0.869±0.019 [0.856, 0.888]	0.679±0.033 [0.646, 0.703]
	16-layer GCN	0.867±0.021 [0.849, 0.887]	0.675±0.054 [0.632, 0.724]
Proposed	**Multi-task, multi-graph, 8-layer GCN**	**0.873±0.019 [0.859, 0.891]**	**0.700±0.044 [0.661, 0.735]**

† Statistical tests were performed between baseline and proposed models on all evaluation metrics of both tasks. Statistically significant comparisons are highlighted with *. $p < 0.0001$ for these comparisons, which are below the adjusted significance level (0.008).
Results from the proposed method are highlighted using black bold.
Mor.=morphology, dis.=distribution, rad.=radius.

As discussed in Experiments (Section III-B) and Ablation Study (Section III-C), the findings in experiment results can be explained with clinical guidelines that characterize distribution and morphology descriptors of calcifications. We further introduced GNNexplainer [74], to enhance the interpretability of the proposed model and to support our findings. GNNexplainer generates explanations by identifying the subgraphs of the computational graphs and node feature subsets that have the greatest impacts on the GNN's predictions. Two case studies are shown in Fig. 3 with original mammography image, radiological annotations and explanation graphs generated from GNNexplainer. In case study (A), the calcifications are distributed in cluster distribution and the cluster marked in the green outline is identified as coarse heterogeneous morphology. GNNexplainer highlights the edges with crucial roles in node and graph prediction. For graph prediction, more edges between calcifications in the cluster are displayed, indicating that these nodes and edges are more influential in classifying the graph as cluster distribution. For node classification, GNNexplainer generated a crucial subgraph for node 6 which contains nodes from the calcification cluster and highlights the edges between calcifications in this cluster. In case study (B), crucial edges are connected across the calcifications nodes and form the segmental distribution (Figure B1). In addition, there are two kinds of morphology in case study (B): fine pleomorphic and fine linear. As shown in Fig. 3 (B2), the classification of node 24 is based on feature propagation from neighboring nodes 21, 23, 25 and 26, which are all with the same morphology. The classification of node 12 with fine linear morphology is explained in Fig. 3 (B3). The information propagation from neighboring nodes with fine linear morphology plays a crucial role in the node classification. The results from GNNexplainer supported our interpretation of the experiment results. With the development of interpretation tools on graph networks, we believe that more explanation and insights could be achieved in future research.

Moreover, we only included malignant cases with ACR BI-RADS 4 and 5 in this study. This inclusion criteria is based on the consideration that the classification of distribution and morphology descriptors are more important in malignant cases for patient care. Therefore, the effectiveness of the proposed method on benign cases has not been assessed in this study. The extraction and annotation of benign cases will continue to enrich the calcification dataset for future studies.

IV. Conclusion

We proposed a multi-task GCN model to tackle the challenging problem of characterization of calcifications morphology and distribution in mammography images, which is a essential task for any effective computerized assisted detection tools for mammography. Through experiments, we demonstrated that our proposed model outperformed the baseline and also the single-task models.

References

[1] H. Sunget al., "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer J. Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[2] V. Cokkinides, J. Albano, A. Samuels, M. Ward, and J. Thum, *American Cancer Society: Cancer Facts and Figures*. Atlanta GA, USA: Amer. Cancer Soc., 2005.

[3] S. Misra, N. L. Solomon, F. L. Moffat, and L. G. Koniaris, "Screening criteria for breast cancer," *Adv. Surg.*, vol. 44, no. 1, pp. 87–100, 2010.

[4] M. Scimeca, E. Giannini, C. Antonacci, C. A. Pistolese, L. G. Spagnoli, and E. Bonanno, "Microcalcifications in breast cancer: An active phenomenon mediated by epithelial cells with mesenchymal characteristics," *BMC Cancer*, vol. 14, no. 1, pp. 1–10, 2014.

[5] R. Baker, K. Rogers, N. Shepherd, and N. Stone, "New relationships between breast microcalcifications and cancer," *Brit. J. Cancer*, vol. 103, no. 7, pp. 1034–1039, 2010.

[6] Y. Ma, P. C. Tay, R. D. Adams, and J. Z. Zhang, "A novel shape feature to classify microcalcifications," in *Proc. IEEE Int. Conf. Image Process.*, 2010, pp. 2265–2268.

[7] W. F. Dahnert, *Radiology Review Manual*. Philadelphia, PA, USA:Lippincott Williams & Wilkins, 2017.

[8] 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*. Reston, VA, USA: Amer. College Radiol., 2013.

[9] O. Erkaymaz, M. Ozer, and M. Perc, "Performance of small-world feedforward neural networks for the diagnosis of diabetes," *Appl. Math. Comput.*, vol. 311, pp. 22–28, 2017.

[10] Z. Gao et al., "Complex networks and deep learning for EEG signal analysis," *Cogn. Neurodyn.*, vol. 15, no. 3, pp. 369–388, 2021.

[11] M. Heenaye-Mamode Khan et al., "Multi-class classification of breast cancer abnormalities using deep convolutional neural network (CNN)," *PLoS One*, vol. 16, no. 8, 2021, Art. no. e0256500.

[12] H. Li et al., "Application of deep learning in the detection of breast lesions with four different breast densities," *Cancer Med.*, vol. 10, no. 14, pp. 4994–5000, 2021.

[13] Y. Liu, F. Zhang, C. Chen, S. Wang, Y. Wang, and Y. Yu, "Act like a radiologist: Towards reliable multi-view correspondence reasoning for mammogram mass detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 5947–5961, Oct. 2022.

[14] Z. Li et al., "Domain generalization for mammography detection via multi-style and multi-view contrastive learning," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2021, pp. 98–108.

[15] Z. Cao et al., "Supervised contrastive pre-training formammographic triage screening models," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2021, pp. 129–139.

[16] W. Lotteret al., "Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach," *Nature Med.*, vol. 27, pp. 244–249, 2021.

[17] N. Alam, E. RE Denton, and R. Zwiggelaar, "Classification of microcalcification clusters in digital mammograms using a stack generalization based classifier," *J. Imag.*, vol. 5, no. 9, 2019, Art. no. 76.

[18] B. Singh and M. Kaur, "An approach for classification of malignant and benign microcalcification clusters," *Sādhanā*, vol. 43, no. 3, pp. 1–18, 2018.

[19] N. Alam and R. Zwiggelaar, "Automatic classification of clustered microcalcifications in digitized mammogram using ensemble learning," *Proc. SPIE*, vol. 10718, 2018, Art. no. 1071816.

[20] A. Papadopoulos, D. I. Fotiadis, and A. Likas, "Characterization of clustered microcalcifications in digitized mammograms using neural networks and support vector machines," *Artif. Intell. Med.*, vol. 34, no. 2, pp. 141–150, 2005.

[21] Z. Chen, H. Strange, A. Oliver, E. R. Denton, C. Boggis, and R. Zwiggelaar, "Topological modeling and classification of mammographic microcalcification clusters," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 4, pp. 1203–1214, Apr. 2015.

[22] A. Oliver et al., "Automatic microcalcification and cluster detection for digital and digitised mammograms," *Knowl.-Based Syst.*, vol. 28, pp. 68–75, 2012.

[23] M. Ciecholewski, "Microcalcification segmentation from mammograms: A morphological approach," *J. Digit. Imag.*, vol. 30, no. 2, pp. 172–184, 2017.

[24] H. Strange, Z. Chen, E. R. Denton, and R. Zwiggelaar, "Modelling mammographic microcalcification clusters using persistent mereotopology," *Pattern Recognit. Lett.*, vol. 47, pp. 157–163, 2014.

[25] Y.-Z. Shao et al., "Characterizing the clustered microcalcifications on mammograms to predict the pathological classification and grading: A mathematical modeling approach," *J. Digit. Imag.*, vol. 24, no. 5, pp. 764–771, 2011.

[26] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Proc. 30th Int. Conf. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3844–3852.

[27] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.

[28] G. Li, M. Muller, A. Thabet, and B. Ghanem, "DeepGCNs: Can GCNs go as deep as CNNs?," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9267–9276.

[29] P. Shi, J. Zhong, A. Rampun, and H. Wang, "A hierarchical pipeline for breast boundary segmentation and calcification detection in mammograms," *Comput. Biol. Med.*, vol. 96, pp. 178–188, 2018.

[30] Y. Guo et al., "A new method of detecting micro-calcification clusters in mammograms using contourlet transform and non-linking simplified PCNN," *Comput. Methods Prog. Biomed.*, vol. 130, pp. 31–45, 2016.

[31] J. G. Melekoodappattu and P. S. Subbian, "A hybridized elm for automatic micro calcification detection in mammogram images based on multi-scale features," *J. Med. Syst.*, vol. 43, no. 7, pp. 1–12, 2019.

[32] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph CNN for learning on point clouds," *ACM Trans. Graph.*, vol. 38, no. 5, pp. 1–12, 2019.

[33] L. Landrieu and M. Simonovsky, "Large-scale point cloud semantic segmentation with superpoint graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4558–4567.

[34] G. Te, W. Hu, A. Zheng, and Z. Guo, "RGCNN: Regularized graph CNN for point cloud segmentation," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 746–754.

[35] F. Monti, D. Boscaini, J. Masci, E. Rodola, J. Svoboda, and M. M. Bronstein, "Geometric deep learning on graphs and manifolds using mixture model CNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5115–5124.

[36] G. Li, C. Xiong, A. Thabet, and B. Ghanem, "DeepERGCN: All you need to train deeper GCNs," 2020, *arXiv:2006.07739*.

[37] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[38] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich, "Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 794–803.

[39] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.

[40] M. L. McHugh, "Interrater reliability: The Kappa statistic," *Biochemia Medica*, vol. 22, no. 3, pp. 276–282, 2012.

[41] H. Du, M. M.-S. Yao, S. Liu, L. Chen, W. P. Chan, and M. Feng, "Apendix tables for automatic calcification morphology and distribution classification for breast mammograms with multi-task graph convolutional neural network," 2023, Accessed: Feb. 21. [Online]. Available: https://github.com/DuHao10086/multi_task_calcification

[42] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Sci. Data*, vol. 4, no. 1, pp. 1–9, 2017.

[43] M. Heath et al., "Current status of the digital database for screening mammography," in *Digital Mammography*. Berlin, Germany:Springer, 1998, pp. 457–460.

[44] A. Paszkeet al., "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 8026–8037.

[45] M. Fey and J. E. Lenssen, "Fast graph representation learning with pytorch geometric," 2019, *arXiv:1903.02428*.

[46] F. Zhang et al., "Cascaded generative and discriminative learning for microcalcification detection in breast mammograms," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12578–12586.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[48] Z. Gu et al., "CE-Net: Context encoder network for 2D medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2281–2292, Oct. 2019.

[49] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: Achievements and challenges," *J. Digit. Imag.*, vol. 32, no. 4, pp. 582–596, 2019.

[50] D. Nie et al., "Medical image synthesis with deep convolutional adversarial networks," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 12, pp. 2720–2730, Dec. 2018.

[51] Y. Xu et al., "Deep learning predicts lung cancer treatment response from serial medical imaging," *Clin. Cancer Res.*, vol. 25, no. 11, pp. 3266–3275, 2019.

[52] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[53] K. Zhang, Y. Guo, X. Wang, J. Yuan, and Q. Ding, "Multiple feature reweight densenet for image classification," *IEEE Access*, vol. 7, pp. 9872–9880, 2019.

[54] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, 2019.

[55] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[56] R. Ito, S. Iwano, and S. Naganawa, "A review on the use of artificial intelligence for medical imaging of the lungs of patients with coronavirus disease 2019," *Diagn. Interventional Radiol.*, vol. 26, no. 5, 2020, Art. no. 443.

[57] M. Goyal, N. D. Reeves, S. Rajbhandari, and M. H. Yap, "Robust methods for real-time diabetic foot ulcer detection and localization on mobile devices," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 4, pp. 1730–1741, Jul. 2019.

[58] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[59] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[60] S.-H. Wang, V. V. Govindaraj, J. M. Górriz, X. Zhang, and Y.-D. Zhang, "COVID-19 classification by FGCNet with deep feature fusion from graph convolutional network and convolutional neural network," *Inf. Fusion*, vol. 67, pp. 208–229, 2021.

[61] T. Gaudeletet al., "Utilizing graph machine learning within drug discovery and development," *Brief. Bioinf.*, vol. 22, no. 6, 2021, Art. no. bbab159.

[62] X. Li et al., "BrainGNN: Interpretable brain graph neural network for fMRI analysis," *Med. Image Anal.*, vol. 74, 2021, Art. no. 102233.

[63] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," 2017, *arXiv:1710.10903*.

[64] X. Li et al., "Pooling regularized graph neural network for fMRI biomarker analysis," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2020, pp. 625–635.

[65] H. Yang et al., "Interpretable multimodality embedding of cerebral cortex using attention graph network for identifying bipolar disorder," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2019, pp. 799–807.

[66] X. Hu et al., "Feedback graph attention convolutional network for medical image enhancement," 2020, *arXiv:2006.13863*.

[67] J. Dabass, S. Arora, R. Vig, and M. Hanmandlu, "Mammogram image enhancement using entropy and CLAHE based intuitionistic fuzzy method," in *Proc. 6th Int. Conf. Signal Process. Integr. Netw.*, 2019, pp. 24–29.

[68] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[69] C. Ferri, J. Hernández-Orallo, and R. Modroiu, "An experimental comparison of performance measures for classification," *Pattern Recognit. Lett.*, vol. 30, no. 1, pp. 27–38, 2009.

[70] T. J. DiCiccio and B. Efron, "Bootstrap confidence intervals," *Stat. Sci.*, vol. 11, no. 3, pp. 189–228, 1996.

[71] M. Ojala and G. C. Garriga, "Permutation tests for studying classifier performance," *J. Mach. Learn. Res.*, vol. 11, no. 6, pp. 1833–1863, 2010.

[72] H. Abdiet al., "Bonferroni and Šidák corrections for multiple comparisons," *Encyclopedia Meas. Statist.*, vol. 3, pp. 103–107, 2007.

[73] P.-H. Chen, E. T. Ghosh, P. J. Slanetz, and R. L. Eisenberg, "Segmental breast calcifications," *Amer. J. Roentgenol.*, vol. 199, no. 5, pp. W532–W542, 2012.

[74] R. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "Gnnexplainer: Generating explanations for graph neural networks," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 9240–9251.