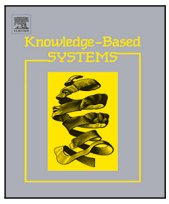
Knowledge-Based Systems 316 (2025) 113372

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



An interpretable CNN-based model for mass classification in mammography

Guobin Li [a,*], Mou Zhou [b], Yu Fu [a], Nashid Alam [a,c], Erika Denton [d], Reyer Zwiggelaar [a]

[a] Department of Computer Science, Aberystwyth University, United Kingdom
[b] School of Intelligent Technology and Engineering, Chongqing University of Science and Technology, China
[c] School of Computing and Mathematics, Manchester Metropolitan University, United Kingdom
[d] Department of Radiology, Norfolk and Norwich University Hospital, United Kingdom

ARTICLE INFO

Keywords:
Mammography
Deep learning CADx systems
Convolutional neural networks (CNNs)
Confounding information
Interpretability

ABSTRACT

Mammography is the primary screening method for lesion visualization and detecting early potentially cancerous changes in breast tissue. The application of deep learning based computer-aided diagnosis (CADx) systems to mammography mass classification poses several challenges: confounding information is learned by a deep learning model, and it can be difficult for mammographic readers to understand how and why it makes a specific decision. In this work, we present a framework for interpretable convolutional neural network-based mammographic abnormality classification. In addition to predicting whether a mass lesion is benign or malignant, our work aims to follow the reasoning processes of mammographic readers in detecting clinically relevant semantic features, such as the shape characteristics of the mass. The framework includes model training that incorporates a combination of data with original images and data with pixel-wise annotations, leading to improved performance of the model. The proposed training method based on DenseNet121 achieved an improved accuracy of $86.1 \pm 3.4\%$ compared to $67.9 \pm 3.8\%$ for the original model on mass classification. The results show the proposed method highlighted the classification-relevant parts of the image, whereas the original method highlighted healthy tissue and confounding information. An interpretable algorithm is developed that explains the model using features representing specific clinical characteristics, thereby aiding in prediction. This allows mammographic readers to verify the model's output for plausibility instead of relying on it blindly.

1. Introduction

Screening mammography is widely employed and has shown its significance, especially for invasive breast masses when they are too small to be palpable or cause symptoms [1]. The visual inspection of a mammogram typically requires the lesion to be identified as either probably benign or malignant, which is then confirmed by histology [2]. Visual inspection usually requires two mammographic readers and is time-consuming, subjective, and prone to errors [3–5]. Mammographic computer-aided diagnosis (CADx) systems have been playing an increasingly important role in assisting and improving the work of clinical experts [6].

Deep learning has increasingly been applied in mammographic imaging [7,8], with convolutional neural networks (CNNs) emerging as the most popular architecture. For mammographic mass diagnosis, CNNs directly process regions of interest (ROIs), applying consecutive convolution and pooling operations to estimate malignancy probabilities. However, the application of these technologies in actual medical scenarios faces considerable challenges.

One major challenge is confounding learning, where the model relies on incorrect information or reasoning to make a decision, even if the decision is correct. For example, in our previous work [9], we observed that the model often ignored the mass itself and instead focused on identifying calcium, which could lead to misclassifications when masses and calcifications coexist in the same image. This problem is exacerbated by the fact that there are few publicly available mammography datasets, so many models are trained on relatively few cases. The second challenge is clinical acceptance. The "black-box" nature of deep learning models can make it difficult for clinicians to trust and rely on AI-based diagnostic systems [10,11].

Building on these key observations, we propose a framework for interpretable CNN-based mammographic abnormality classification, as shown in Fig. 1. The framework consists of two main components: model training and model interpretability. Drawing inspiration from prior works on breast mass classification [12–15], the model training component aims to mitigate confounding in deep learning models (Fig. 1: Part A). Specifically, an activation map is defined to contain

* Corresponding author.
E-mail address: gul12@aber.ac.uk (G. Li).

https://doi.org/10.1016/j.knosys.2025.113372
Received 4 October 2024; Received in revised form 24 February 2025; Accepted 18 March 2025
Available online 28 March 2025

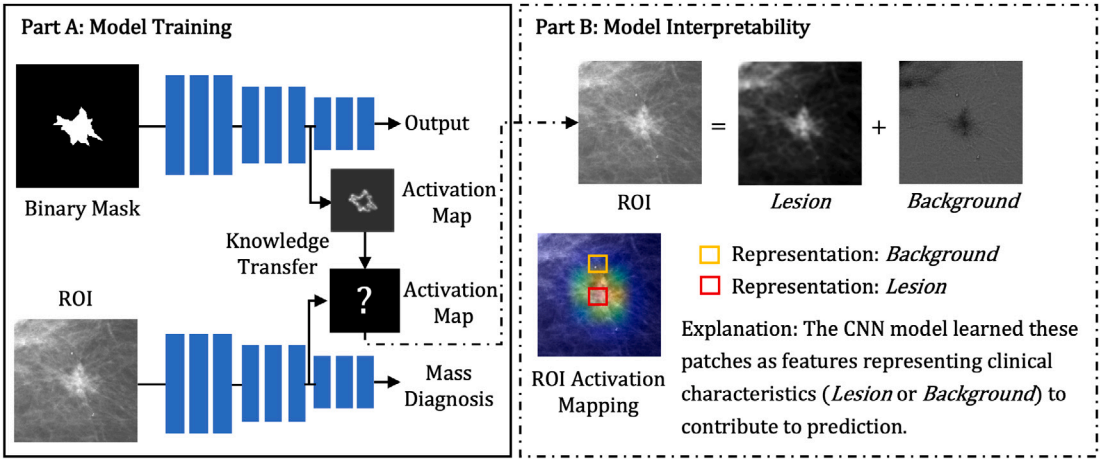


Fig. 1. Schematic representation of the proposed framework: model training and model interpretability. Part A: A CNN model uses knowledge transfer based on activation maps for mammography mass classification. Firstly, an activation map is defined to contain relevant information, especially shape knowledge from binary masks will be used to improve the final mass classification. Secondly, knowledge transfer based on activation maps transfers shape knowledge from binary masks to a model during the ROIs learning process. Part B: The interpretable algorithm was built to facilitate a clinical radiology understanding of the activation maps employed by the model.

shape knowledge from binary masks, which is used to improve the final mass classification. Additionally, knowledge transfer based on these activation maps facilitates the integration of shape knowledge into the model during the ROI learning process.

The model interpretability component introduces a novel algorithm to provide explicit reasoning for the model's decision-making process, enabling mammographic readers to verify and understand predictions (Fig. 1: Part B). For a specific mass lesion, this algorithm explains how the model identifies classification-relevant parts within the activation map, linking them to defined clinical characteristics, such as distinguishing between *Lesion* and *Background*, and how this contributes to prediction.

The first contribution of this paper is a novel methodology for integrating shape knowledge into a CNN, aiming to improve breast cancer diagnosis. (1) Unlike previous approaches that use the binary masks as input in a network, in our proposed approach, the binary maps are made available to the convolutional layers of the CNN. This introduces new activation maps at each layer, serving as a source of shape-related features from binary masks. (2) The traditional method of the integration of shape knowledge into a CNN relies on complex fusion models with multiple convolution layers. We propose a knowledge transfer mechanism utilizing a spatially-aware loss function to reduce the computational complexity of a CNN.

The second contribution of this paper is a novel algorithm for model interpretability. Traditional class activation mapping (CAM) highlights ROIs the CNN used to make a prediction. Our interpretable algorithm extends this by not only identifying relevant regions but also providing clinical radiological insights (i.e. predicted tissue type information) about the highlighted areas. This ensures a deeper understanding of how specific regions contribute to diagnosis and aligns predictions with clinical reasoning.

2. Related work

In this section, we provide an overview of the two core components of our proposed framework. First, we discuss model training, which enhances the generalization of deep CNN models. Then, we explore model interpretability to improve the understanding of the model's decision-making process.

2.1. Background on model training

2.1.1. Integration of shape knowledge

Prior to the emergence of deep learning, machine learning methods were based on hand-crafted features [16]. Researchers designed features to capture both textural and morphological characteristics, which relied on accurate binary masks [17,18] and were subsequently used with various classifiers. This is attributed to the fact that the likelihood of malignancy has been found to be highly correlated with lesion morphology, such as the shape characteristics of the mass [19]. Thus, the incorporation of the binary mask can significantly facilitate the extraction of representative features towards this end.

Earlier works [20–22] were focused on comparing CNN-based architectures to machine learning methods based on hand-crafted features. The majority of research has drawn the conclusion that shape features and CNN-based features are complementary, which leads to increased performance when they are combined. This supports the fact that CNNs, that make heavy use of the convolution operation, are better suited for extracting texture features but lack in capturing shape features.

Binary masks can be a source of shape-related characteristics to assist in a CNN pipeline for automatic diagnosis, after an in-depth analysis of the deep learning based CADx, the following trends have been identified: The first trend focuses on methodologies designed for integrating binary masks into a CNN. In particular, it concerns works that process the grey-level information of input mammogram images, incorporating the use of a binary mask. For example, Dhungel et al. [23,24] concatenated the binary mask with the ROI before feeding it to the network. Some work [25,26], fed a masked mass ROI to the network, which was produced by the element-wise multiplication of the ROI image with the binary mask. However, these methods are limited to distinguishing mass, making it challenging to capture its surrounding tissue which contains relevant information and is considered by radiologists for diagnosis [27]. Tsochatzidis et al. [28] modified the masked mass ROI, which are contour delineations acquired for masses drawn on ROI that correspond to different shape types. During experimentation, it was shown that shape and texture features had a complementary role.

The second trend focuses on methodologies designed for integrating shape features into a CNN. In particular, it concerns works that may include textual features extracted from ROIs, shape features extracted from binary masks, and then integrate these two separate features in the feature fusion block. For example, Yan et al. [29] automatically extracted textural and shape features from ROIs and binary masks separately and then fused them based on machine learning algorithms. Carneiro et al. [30] initially trained a separate CNN model for ROIs and binary masks. Then, using the features learned from ROIs and binary masks to train a final CNN classifier that estimated the patient's risk of developing breast cancer. The poor performance was likely caused by the multi-stage process training. To construct a CNN architecture trained in an end-to-end fashion, Li et al. [31] proposed a

fully automatic dual-path CNN architecture that used two subnetworks to process binary masks and ROIs separately. At a late stage in the final CNN classifier, the information from the two paths was fused before reaching a final decision. While achieving competitive results on publicly available mammography datasets, this approach is computationally expensive and relies on multiple CNNs, bypassing the inherent self-learning capabilities of a single model.

The proposed method incorporates knowledge transfer, as described in Section 2.1.2. Knowledge transfer aims to improve the training of one network by relying on knowledge transferred from another network. We propose CNN learning for binary masks to extract shape features, and then transfer shape knowledge into a CNN for ROI learning, instead of a CNN-based fusion model for integrating shape knowledge into a CNN, with the goal of improving breast cancer diagnosis and reducing the model's computational complexity.

2.1.2. Knowledge transfer

Attention transfer is a knowledge transfer method that transfers attention from one network to another network with the goal of improving the performance of the latter. In attention transfer, given the spatial attention maps of a teacher network, the goal is to train a student network that will not only make correct predictions but will also have attention maps that are similar to those of the teacher. In general, one can place transfer losses with respect to attention maps computed across several layers. For instance, some studies [15,32] have aligned activation maps between models in the spatial domain. A new loss function is introduced, incorporating spatial distance into the standard cross-entropy loss to minimize point-wise and/or pair-wise discrepancies from activation maps between models.

Given the spatial attention maps of a network from binary masks, which can be a source of shape-related characteristics. To train a new network for ROIs that will not only make correct predictions but will also have attention maps that are similar to those from the binary masks, aiming to steer the attention of the network toward the shape characteristics of the mass.

2.2. Background on model interpretability

2.2.1. CAM explanation

As deep learning models are increasingly applied in breast cancer diagnostic assistance systems, this raises questions about the ability to understand and interpret its decision-making process. Once a deep learning model has been trained, the output of its layer refers to a representation of the training data along with the corresponding labels. The output is hierarchical and evolves through the layers of a CNN, enabling the model to recognize patterns and make predictions. In computer vision, the output of the data learned by each layer of the CNNs has been commonly used to understand the decision-making process.

Arevalo et al. [21] visualized the first layer output equivalent to local kernels that work as filters over the image and found they showed a set of edges in different orientations. Because the output of layers is often high-dimensional, it can be defined with respect to various layers of the network so that they are able to capture both low-, mid-, and high-level representation information. Methods based on class activation maps (CAM) [33], converted the high-dimensional output from the ResNet into a low-dimensional spatial map, which highlighted the area that the model focused on during its decision-making process. However, although these traditional methods have already been used in medical imaging, they are considered post-hoc analyses because they only focus on the image area framed by the model and cannot explain how the model predicts [34].

Hand-crafted features have recently emerged as a promising approach for enhancing model interpretability. Hand-crafted features are extracted from medical images using predefined mathematical equations [35]. These engineered equations have been designed to capture different characteristics of images [36]. For example, the second-order features are based on matrices including grey-level co-occurrence matrix (GLCM) to extract texture information [35]. In the past few years, the analysis between CNN-based features and hand-crafted features has been established [37–40]. This traditional analysis consists of three steps: CNN-based feature extraction, hand-crafted feature extraction and correlation calculation. Several methods are used to quantify their relationships, including the Pearson correlation coefficient, Euclidean distance, and Earth mover's distance. However, these approaches primarily measure straight-line distances, neglecting the spatial relationships within the features.

The proposed interpretable algorithm incorporates hand-crafted features, as described in Section 2.2.2. To capture spatial relationships and provide a more robust measure of similarity between texture and CNN-based features, we introduce a new representation of these features in a low-dimensional space. This representation enables the evaluation of their relationship using Euclidean distance. Our algorithm relies on CAM mechanisms to highlight parts of an input image, explains that it considers these parts of images similar to the texture features, and provides a score for the probability of the specific diagnosis for this image.

2.2.2. GLCM texture features

Within a selected ROI, there are several subregions showing different texture statistics, for example, *variance* emphasizes the region inside the mass [40]. In this study, we use GLCM to obtain the *variance* texture feature. An element of the GLCM, $p(m,n)$, is defined as the joint probability that grey-level m and grey-level n occur together.

$$variance = \sum_{m,n=0}^{L-1} (m-\mu)^2 \cdot p(m,n) \tag{1}$$

$$\mu = \frac{1}{L^2} \sum_{m,n=0}^{L-1} m \cdot p(m,n) \tag{2}$$

where L is defined as the number of grey levels in ROIs. The feature *variance* is a two-dimensional (2D) image that keeps the same size as the ROI.

3. Experimental setup

3.1. Mammography datasets

We evaluated the proposed framework using publicly available mammographic datasets: CBIS-DDSM [2] and Breast Cancer Digital Repository (BCDR) [41].

- The CBIS-DDSM is a modernized subset of Digital Database for Screening Mammography (DDSM) [42], in which 3568 digitized mammograms are formatted in DICOM format (1696 masses and 1872 calcifications). On top of that, the CBIS-DDSM has already been partitioned into the training (1318 masses) and test set (378 masses). For our study, we utilized the full set of mass lesions, where all cases are classified as benign or malignant. With respect to class balance, both training and test sets have the same ratio of the two classes, which was $1:1.07$.
- The BCDR provides annotated patient cases of breast cancer including mammography images and lesion outline annotations provided by radiologists, pre-computed image-based descriptors as well as related clinical data. This repository is continuously being enriched and currently, contains cases of 1734 patients with mammography and ultrasound images. It consists of two different types of sub-repositories: (1) a Film Mammography-based Repository (BCDR-FM) and (2) a Full Field Digital Mammography-based Repository (BCDR-DM). Both FM and DM repositories are divided into several sub-datasets including a different number of cases, which form

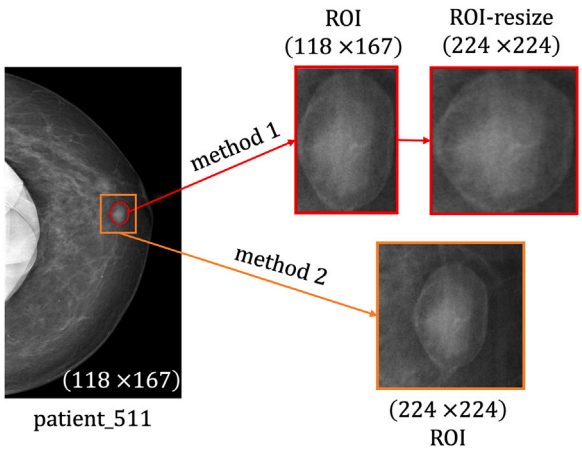


Fig. 2. Example mass from the BCDR database. Two different patch contexts from the training images were compared without implementing any variations in the image intensity. Method 1: Bounding box of the annotated mass: (original ROI). Method 2: mass-centred ROI includes regions $\max(224, width) \times \max(height, 224)$: (mass-centred ROI).

a common ground for a fair comparison between various CADx systems for mammographic cancer analysis.

The present work takes subsets from both FM and DM sub-datasets, containing the following: (1) BCDR-D01 contains 141 segmentations, all of which represent suspicious masses and are used in this study. Among them, 85 are benign, and 56 are malignant. (2) BCDR-F01 contains 362 segmentations, with mass lesions occurring in 231 segmented images. To create a balanced dataset, we randomly selected a subset of 133 mass lesions, comprising 59 benign and 74 malignant cases. In summary, there are a total of 274 mammographic lesions in the BCDR dataset, 144 lesions are benign and 130 are malignant.

Regarding the data division for the evaluation of the proposed training method based on a CNN, five-fold cross-validation was employed to create independent training and test sets. The training datasets were further split 80 : 20 to create independent sets. The average accuracy on five-fold cross-validation demonstrates the classification performance representing the proposed method.

3.2. Datasets preprocessing

In terms of the ROIs selection, masses are centre cropped. The design of the size of CNN inputs is 224 × 224. The used approach is to crop the contextual rectangular region with proportional padding, so that mass-centred ROI includes regions $\max(224, width) \times \max(height, 224)$ instead of the bounding box. The selected ROIs are then all resized into an identical dimension 224 × 224 by bicubic interpolation. Accordingly, the ground truth binary masks are cropped and resized but with the nearest neighbour interpolation. Two different patch contexts from the training images were compared without implementing any variations in the image intensity, as shown in Fig. 2.

The first method is the bounding box of the annotated mass (original ROI). The second method is $\max(224, width) \times \max(height, 224)$, and it is this mass-centred ROI we used in this project (mass-centred ROI). Whether the width or height of the bounding box is smaller than 224, these contextual ROIs can decrease the structural distortion led by ROIs resize, potentially changing it from an **oval** to a **round** appearance, as illustrated in Fig. 2. Additionally, in the process of ROI selection, it is essential to include not only the mass abnormality itself but also the surrounding tissue. This approach captures relevant contextual information, as radiologists often consider the mass and its neighbourhood when making a diagnosis. Including this surrounding area has been shown to significantly improve classification performance, as noted in the previous study [27].

Table 1
Performance comparison of DenseNet model with different augmentation configurations during fine-tuning on the OMI-DB.

Reference	Methodology	Accuracy
Ortega-Martorell et al. [46]	Original ROI	54.4%
Hamidinekoo et al. [27]	Double bounding box	57.4%
Li et al. [31]	Double bounding box, flip	60.9%
Shen et al. [4]	Original ROI with an overlapping ratio	53.6%
Proposed method	Crop, flip, rotate	63.6%

3.3. Model architecture

DenseNet is a strongly performing deep learning model for binary classification of breast tumours [43]. We utilize DenseNet121 [44] for benign or malignant mammography mass classification. The classifier layer is modified since the original average pooling and fully connected layers are developed to classify 1000 categories instead of two. Specifically, the average pooling layer and fully connected layer were replaced with a simplified architecture consisting of a fully connected layer and a softmax activation function for two output classes, enabling the model to effectively distinguish between benign and malignant cases.

Transfer learning based on pre-trained weights using the ImageNet [45] dataset represents primitive features that tend to be preserved across different tasks, whereas the classification layer with randomly initialized weights represents higher-order features that are more related to specific tasks and require further training. In addition, as shown in Table 1, the diagnosis performance of the DenseNet121 model with different augmentation configurations was evaluated during fine-tuning on the OMI-DB dataset. All the listed algorithms come from related works and have been implemented by us for comparison.

Several prior studies have explored ROI extraction strategies for mammographic analysis. For example, Ortega-Martorell et al. [46] used the bounding box of the annotated mass (original ROI). Hamidinekoo et al. [27] proposed that the ROI was extracted with a size equal to double the square bounding box of the abnormality. Li et al. [31] introduced mass-centred ROI, which includes regions two times the size of the bounding box and then augmented with horizontal and vertical flips. Shen et al. [4] extracted the original ROI with a minimum overlapping ratio of 0.9. In alignment with these methodologies, the current study follows the ROI extraction and resizing approach detailed in Section 3.2. To avoid overfitting, the selected ROIs are augmented with horizontal and vertical flips and random rotation (with augmentation probability of 50% for each instance) after data division. Notably, this data preprocessing pipeline has contributed to improved model generalization. Finally, the DenseNet121 model is trained utilizing the proposed data preprocessing strategy, ensuring robustness in feature learning and facilitating its deployment for downstream diagnostic analysis.

4. Methodology

4.1. Model training

The proposed training method can be divided into two branches (see Fig. 3). The first branch provides activation map extraction. The second branch provides knowledge transfer. The activation maps from the binary mask are transferred to the CNN model when it is trained on breast tissue to provide mass classification.

4.1.1. Activation map

Activation map definition. Our approach contrasts directly with previous works that rely on (a) gradient-based activation mapping using no class discrimination and (b) class discrimination to highlight the parts of an input image, whilst our approach is an inherently interpretable learning process based on how the network explores the image itself.

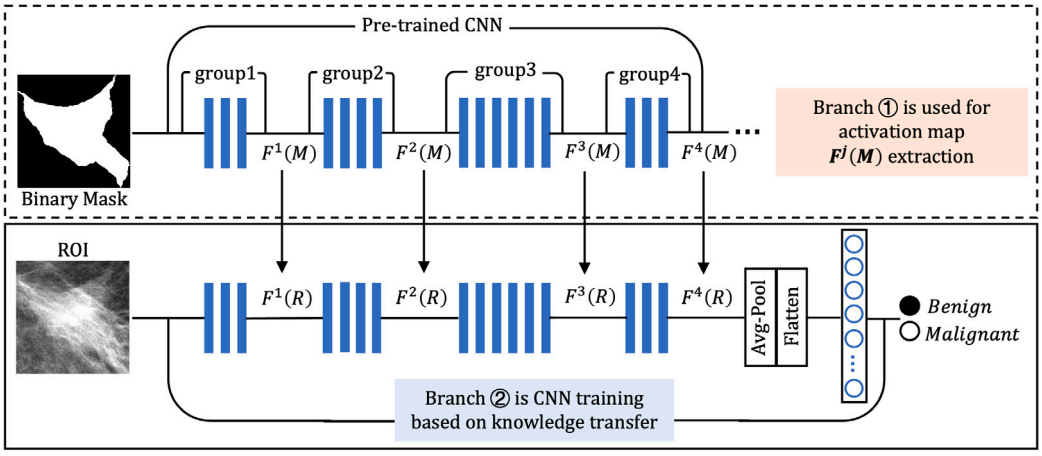


Fig. 3. The architecture of the proposed method. This is divided into two branches, the first branch defines activation map $F^j(M)$ to capture representation knowledge from binary masks. It can be directly used in a pre-trained model and will not participate in the training process in the second branch. The second branch provides knowledge transfer, the knowledge from binary mask activation is transferred to the CNN model when it is trained on ROIs.

Considering a CNN layer and its corresponding activation tensor $A \in R^{H\times W\times C}$, which consists of C feature planes with spatial dimensions $H \times W$. To define a spatial activation mapping function, we take as input the 3D tensor A and output a spatial activation map, i.e. the absolute value of a hidden neuron activation can be used as an indication of the importance of that neuron with respect to the specific input. More specifically, in this work, we will consider the sum of absolute values raised to the power of 2:

$$F = \frac{\sum_{i=1}^{C} |A_i|^2}{C} \tag{3}$$

where a spatial activation map $F \in R^{H\times W}$ by computing statistics of these values across the channel dimension. The activation map is defined only using neuron activation when the network is evaluated on a given input. It can be directly used in a pre-trained model and will not participate in the training process in the second branch.

Activation map visualization. To further illustrate what kind of information from this function can be used for improving the performance of CNN architectures, activation maps are extracted to compare the information change over different datasets and layers. As shown in Fig. 4, we visualized activation maps of two inputs with differences in mass classification (benign vs malignant): binary masks (81.8% test accuracy), and ROIs (63.6% test accuracy).

Furthermore, activation maps focus on different parts for different layers in the networks. In the low- and mid-level layers, neuron activation levels are higher for the most discriminative regions. In the high-level layer, they reflect the whole object. For example, low and mid-level activation maps of binary masks will have higher activations around the shape/boundary of the masses and high-level activation will correspond to the whole mass. However, binary masks are limited to distinguishing between foreground and background, making it challenging to capture the nuanced features of tumours accurately. ROIs exhibit diverse characteristics in terms of shape, texture, and density and are commonly used by mammographic readers in breast mass diagnosis. Activation maps of an ROI are abstract, as the poor performance indicates. The confusion matrix achieves unbalanced results, which indicates the model tends to be more accurate for malignant diagnosis while performing poorly for benign cases.

This ideal also follows the reasoning process of mammographic readers in detecting clinically relevant semantic features in each lesion, such as using the shape characteristics of the mass and subsequently estimating the patient's risk of developing breast cancer based on BI-RADS classification.

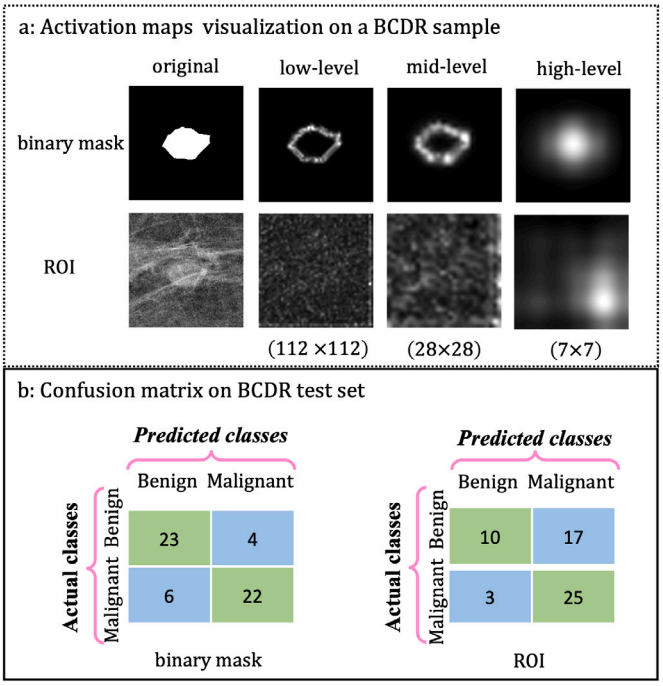


Fig. 4. (a) Activation maps, at three levels, for the DenseNet121 network on one BCDR example. (b) The resulting confusion matrices on BCDR test set: binary masks (81.8% test accuracy), ROIs (63.6% test accuracy).

4.1.2. Activation map-based knowledge transfer

In knowledge transfer, given the spatial activation maps of binary masks, the goal is to train a CNN on ROIs that will not only make correct predictions but will also have activation maps that are similar to those of the binary masks. We can define the distance between binary mask activation and ROI activation as:

$$d^j = \left\| \frac{vec(F^j(M))}{\|vec(F^j(M))\|_2} - \frac{vec(F^j(R))}{\|vec(F^j(R))\|_2} \right\|_2 \tag{4}$$

where $F^j(M)$ and $F^j(R)$ are, respectively, the jth CNN layer's activation maps corresponding to the binary mask and ROI input, vec means the activation maps in vectorized form, and $\|...\|_2$ indicates the l_2 normalization. As can be seen, we make use of l_2-normalized activation maps, i.e. each vectorized activation map with: $\frac{vec(F)}{\|vec(F)\|_2}$. Then, we want

the ROI activation to be similar to the binary mask activation during the ROI training process, we use transfer loss with respect to distance computed for different inputs [15], i.e.:

$$\ell_{AT} = \ell(\omega, x) + \frac{\beta}{2}\sum_{j\in\Omega} d^j \tag{5}$$

where ω denotes the trainable weights and $\ell(\omega, x)$ denotes a standard cross-entropy loss function of a CNN model. Ω denotes the indices of the CNN's corresponding trainable layers to which we want to transfer activation maps. When combined, the model loss function adds an extra distance function, which can be easily computed during forward propagation, and needs minimization.

Knowledge transfer can also be used between different models. For instance, to train the DenseNet121 model on ROIs, the first branch's activation map extracted from the binary mask could be chosen from a different model and will not participate in the training process. In this project, we assume that activation maps are extracted from the same model, but, if needed, knowledge transfer can be incorporated with different models.

4.2. Model interpretability

For the interpretability of an activation map, the current methods proposed by [37–40,47] have found similar semantic information between deep learned features and hand-crafted features, especially second-order GLCM texture features. The use of such hand-crafted features has caused a loss of spatial information as texture features need dimensionality reduction, e.g. the mean of each texture feature.

Compared with the variations in appearance of the ROIs, texture features provide a quantitative measure of the diversity or uniformity of textures within ROIs. We have investigated how deep learning based networks perform on texture features in [9] and found some clinical representations such as that the shape of a mass can be learned from texture features. Unlike the previous method [9], which provided texture features to improve the performance of the models. In this paper, we aim to investigate the behaviour of networks for the texture feature learning process to explain how the networks perform so well. In our study, Fig. 5, the proposed algorithm contains three steps.

Step 1: ROI preprocessing. By analysing the texture feature *variance*, we gain valuable insights into the internal structure of the ROI, enhancing our understanding and characterization of ROI content. We regard the *variance* feature as a clinical *Lesion* characteristic of an ROI. The defined *Lesion/ variance* increases the contrast between mass and no-mass area in the breast image and contains less information compared to the ROI. The remaining information is excluded by the *Lesion* from the ROI and is defined as *Background*.

Step 2: Patch replacement. Unlike our previous work [9], which directly used texture features as inputs sent to the model to investigate the deep learned features. The second step is a small patch replacement based on a projection window. The projection windows remain consistent with the areas in the ROI, which are projected by the activation map F^4 (7×7 in the case of the last activation map of DenseNet121). The size of the projection window should be (32×32). The projection window in the ROI will be replaced with *Lesion* and *Background* as inputs sent to the model to investigate the projection value in the activation map. We found that replacing one patch as well as keeping the remaining ROI unchanged can increase the sensitivity of the model to the replaced patch.

For the given activation map $F^4 = \left\{F^4_{p,q}\right\}^{(7,7)}_{(p,q)=(1,1)}$, the inputs of the CNN model can then be noted as $\left\{ROI, ROI_{L(p,q)}, ROI_{B(p,q)}\right\}^{(7,7)}_{(p,q)=(1,1)}$, where $ROI_{L(p,q)}$ represents pth row, qth column patch in the ROI will be replaced with *Lesion*, and $ROI_{B(p,q)}$ represents pth, row qth column patch in the ROI will be replaced with *Background*.

Step 3: Representation understanding. The patch replacement allows the model to consider how relevant the projection value $F^4_{p,q}$ in the activation map is present with the two replaced patches (*Lesion* and *Background*), so that the distance between convolutional activation tensors can be interpreted as how strong a defined clinical representation is present in the replaced patch of the input image.

As illustrated in Fig. 5, given the images $\left\{ROI, ROI_{L(1,3)}, ROI_{B(1,3)}\right\}$ as inputs separately sent to the CNN model to receive their corresponding activation tensors $\left\{A, A^L, A^B\right\}$, by computing the average l_2 distance between all 1×1 patches of convolutional activation tensors A and A^L:

$$\begin{aligned} dis^L_{1,3} &= \frac{1}{7\times7}\sum_{n=(1,1)}^{(7,7)}\left\|A_n - A^L_n\right\|_2 \\ dis^B_{1,3} &= \frac{1}{7\times7}\sum_{n=(1,1)}^{(7,7)}\left\|A_n - A^B_n\right\|_2 \end{aligned} \tag{6}$$

where $n \in \{(1,1),\dots,(1,W),(2,1),\dots,(7,7)\}$ indexes the 1×1 patches of the 7×7 convolutional activation tensors A and A^L.

If the patch $A_{(1,3)}$ is close to the patch $A^L_{(1,3)}$, the distance $dis^L_{1,3}$ between those convolutional activation tensors will be small — consequently, convolutional activation tensor A will have a patch $A_{(1,3)}$ that represents *Lesion* as well as the activation map will have a value $F^4_{1,3}$ that represents *Lesion*. On the other hand, if the patch $A_{(1,3)}$ is close to the patch $A^B_{(1,3)}$, the activation map will have a value $F^4_{1,3}$ that represents *Background*. We used the average l_2 distance because the surrounding pixels near the replaced patch will also be impacted during the convolutional operation.

Since each value in the activation map F^4 has a distance score $(dis^L_{p,q}, dis^B_{p,q})$, the distance score can then identify how relevant the value $F^4_{p,q}$ in the activation map is present with defined clinical *Lesion* and *Background* during the model diagnosis process.

5. Results and analysis

5.1. Model training

5.1.1. Classification performance

In general, activation maps can be extracted on output activations of each group of DenseNet121, as is shown in Fig. 3. Increasing the number of trainable weights in the model can potentially lead to overfitting, especially when it is trained on limited data. Exploring which layers tend to fine-tune information in the model is fundamental to verifying the proposed knowledge transfer. The distance $d_j | j \in 1,2,3,4$ between ROI activation and binary mask activation through different groups are compared separately. Results from the BCDR training dataset can be found in Fig. 6.

In Fig. 6, for visualization purposes, we use $(\log d^j + 10) | j \in 1,2,3,4$. In the pre-trained DenseNet121, convolution layers are based on pre-trained weights using the ImageNet dataset. The initial layers such as group 1, 2, and 3 represent primitive features that tend to be preserved across different tasks, so ROIs activation maps are similar to binary masks. Group 4 represents high-order features that are more related to specific tasks and require further training. There is a clear difference for the activation maps in group 4. Training all layers will expand the level of computational complexity and less information transfer in the lower layers. We train DenseNet121 with only one transfer loss in group 4 to test the performance.

Table 2 shows the diagnosis performance of the proposed training method. Note that the listed accuracy is obtained by averaging accuracy on five-fold cross-validation, and the confusion matrix is the sum of the five experiments.

We first examine the fine-tuned DenseNet121 with the classifier layer plus group 4 to learn features that are more related to mass classification and obtain $51.5 \pm 9.3\%$ accuracy. The poor performance and biased confusion matrix are likely caused by differences between

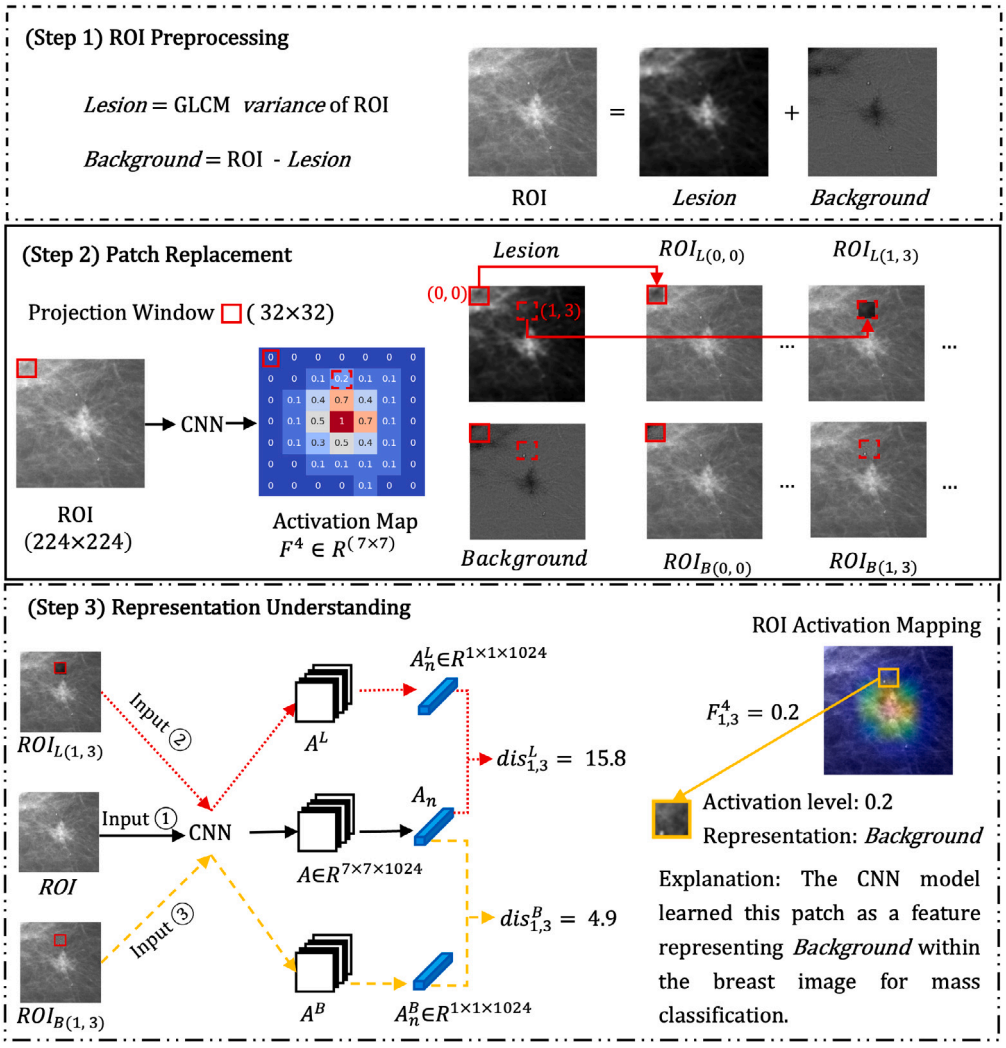


Fig. 5. The architecture of the proposed interpretable algorithm. It can be divided into three steps, in the first, we defined *Lesion* and *Background* in the ROI preprocessing. The second step is a small patch replacement based on a projection window. The projection windows remain consistent with the areas in the ROI, which are projected by the activation map F^4 (7×7 in the case of the last activation map of DenseNet121). The size of the projection window should be (32×32). The projection window in the ROI will be replaced with clinical *Lesion* and *Background* as inputs sent to the DenseNet121 model. The third step calculates the average l_2 distance between convolutional activation tensors of the ROI and replaced ROIs. The distance score can then identify how relevant the value in the activation map is, with defined clinical *Lesion* and *Background* during the model diagnosis process.

Table 2
Mass classification performance in DenseNet121 (fine-tune vs knowledge transfer).

Dataset	Pre-training	Methods	Trainable layers	Accuracy	Confusion matrix
BCDR	ImageNet	Fine-tune	group 4 + classifier layer	$51.5 \pm 9.3\%$	$\begin{pmatrix} 53 & 91 \\ 42 & 88 \end{pmatrix}$
		Knowledge transfer		$68.2 \pm 5.5\%$	$\begin{pmatrix} 90 & 54 \\ 33 & 97 \end{pmatrix}$
BCDR	DDSM	Fine-tune	group 4 + classifier layer	$67.9 \pm 3.8\%$	$\begin{pmatrix} 94 & 50 \\ 38 & 92 \end{pmatrix}$
		Knowledge Transfer		$86.1 \pm 3.4\%$	$\begin{pmatrix} 121 & 23 \\ 15 & 115 \end{pmatrix}$

natural images and mammographic images [9]. The confusion matrix reveals a tendency of the model to favour malignant diagnoses. The proposed method trained the model with the same layers and achieved an improved accuracy of $68.2 \pm 5.5\%$ compared with the classical fine-tuning method. The confusion matrix achieves more balanced results. In addition, when we pre-train the model on DDSM, the accuracies are further improved to $67.9 \pm 3.8\%$ and $86.1 \pm 3.4\%$ for the fine-tuning method and knowledge transfer. In all combinations, knowledge transfer achieves improved accuracy and unbiased learning compared with the no knowledge transfer models. It should be noted that in both cases the improvements are statistically significant (with $p = 0.0082$ and $p = 0.0069$).

5.1.2. Visualization

We visualize activation maps of the DenseNet121 model (see Fig. 7), including knowledge transfer and no-knowledge transfer (classical fine-tuned method) on ROI classification. To evaluate the effectiveness of the proposed method, activation maps are rescaled to be within [0, 1]

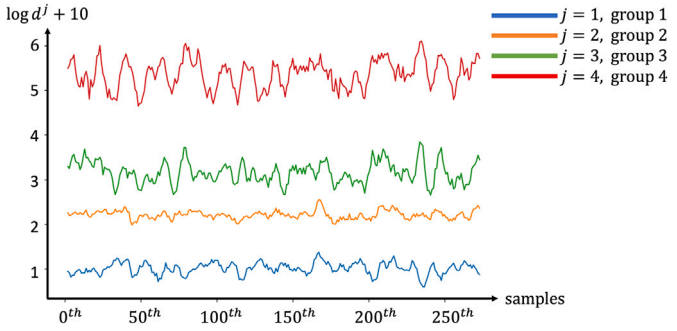


Fig. 6. The distance distribution between ROI activation and binary mask activation for different groups. For each sample, the distance increases for subsequent groups/layers.

and resized to 224 × 224 using interpolation. ROI activation mapping adds the original ROI as a background and shows the activation map as a colour overlay. The binary masks have been given to accurately localize masses within the breast tissue. The comparison between them is shown in Fig. 7.

A mass tends to have dense tissues that appear white on mammograms. Microcalcification and salt/pepper noise appear in a mammogram as bright dots and can impact the mass diagnosis of the CNN model. For example, Figs. 7a and 7b both show feature-wise results (activation maps) that were obtained from the fine-tuning method, displaying a fixation on the microcalcification over masses. The suspicious areas are marked with red circles in the ROIs. This is likely because calcifications are strongly associated with some typical breast cancers (such as ductal carcinoma in situ and invasive cancers) [48]. Thus, the model will easily disregard the mass and instead prioritize the identification of bright dots, potentially resulting in the misclassification of the mass diagnosis.

Another factor influencing the mass identification of the model is dense breast tissue, which also appears as increased density on mammograms due to its fibrous and glandular composition. For instance, in Fig. 7c, the expanded ROI may encompass dense breast tissue, resembling potential masses. Additionally, an abnormality may be concealed within these dense areas, as demonstrated in Fig. 7d. The healthy tissue is marked with red circles in the ROIs. The activation levels of the fine-tuned method highlight the whole ROI, and more irrelevant information including dense breast tissue is used for mass classification, probably causing the poor performance as shown in Table 2.

To address these challenges, we propose knowledge transfer, leveraging binary masks to aid the model in accurately localizing masses within the breast tissue while also preserving inherent capacity of the CNN model for self-learning to acquire new representations for the masses. In the high-level layer of DenseNet121, the model accurately localized the mass within the breast tissue and learned the features that represent the mass.

5.2. Model interpretability

5.2.1. Visualization

Existing interpretable techniques [49,50] include localization as in Fig. 8a, which shows a CAM that highlights which parts of the ROI are used for decision making, but there is no explanation of the CAM. A correlation heatmap [37,39] was proposed (see in Fig. 8(b) to find similar hand-craft features to explain the abstract deep learned features, but there is no explanation of what parts of the training set these associations are learned from. Many recently published interpretable algorithms [9,51], as shown in Fig. 8c, have explained that the model considers some characteristics of a test sample similar to training samples it has seen before, and provides a score for the confidence of the specific diagnosis for this test sample.

As shown in Fig. 7, the ROI activation mapping can highlight which parts of the ROI are used for decision making. The activation levels of knowledge transfer are more likely to highlight the mass and less likely to highlight the surrounding healthy tissue. The surrounding tissue may contain microcalcification, noise artefacts or normal dense tissue in ROIs. They would also be activated as non-negative values in activation maps. So it is important to provide an explanation of what attributes of the activation the model considered for classification decisions. Fig. 8d shows an example of the proposed algorithm: the activation map localizes relevant regions, and associates the relevant region with specific clinical characteristics (*Lesion* or *Background*) to provide a representation to support understanding for the decision-making process.

5.2.2. Clinical use

Our interpretable algorithm provides a second opinion, rather than a full decision, aimed at improved overall human–machine collaboration. Using the proposed framework with some examples (benign and malignant) is shown in Fig. 9. For example, Fig. 9a shows a breast image containing noise or microcalcifications. The proposed interpretable approach can give an explanation that the tissue including noise and microcalcifications are learned as features representing *Background* by the model to contribute to the prediction. The model clearly captured the high-level representation *Lesion* located within the mass areas.

Fig. 9b shows a breast image containing dense tissue, which makes it hard to see the boundary of the masses. The proposed method indicates what kind of distribution of the mass the model learned as features representing *Lesion* to make a decision. We can also find that the representation *Lesion* has a higher activation level compared with the *Background* in the activation map, which has more impact on the prediction. This validates that the proposed method can effectively explain why the proposed training method achieved improved accuracy.

From a clinical point of view, this interpretable algorithm aims to provide mammographic readers with the means not to simply trust the model but to check its output for plausibility.

1. The developed approach provides an explanation of the attributes of noise or microcalcifications that the model considered for classification decisions. If a model simply relied upon confounders such as noise or microcalcification to do mass classification, the model would fail to generalize.
2. The developed approach identifies the obscured aspect of masses where the margins are hidden by surrounding fibroglandular tissue. It will give a visualization of an interpretable activation mapping containing the portion of the margin that the model learned for decision-making. This visualization aids in understanding how the model distinguishes the mass from the surrounding tissue and can be verified by mammographic readers to assess the accuracy of the model's diagnoses.

From a methodological perspective, this approach allows for the comparison of the distribution of representations learned by various models for decision-making. By applying it across different models, researchers can evaluate the performance and sensitivity of each model on the same dataset.

6. Limitation and future work

We achieved improved performance of CNN models for mass diagnosis and explanation of what attributes of the activation map are used for classification, yet there are several limitations for further investigation. Firstly, although we have adopted pretraining, there is still significant room for improvement. In order to retrain the CNN models and further mitigate the overfitting problem, several modifications can be explored, including exploiting large amounts of unannotated data, by pre-training the model with pretext tasks in a self-supervised manner [52,53]. Secondly, at the moment, binary masks directly come from

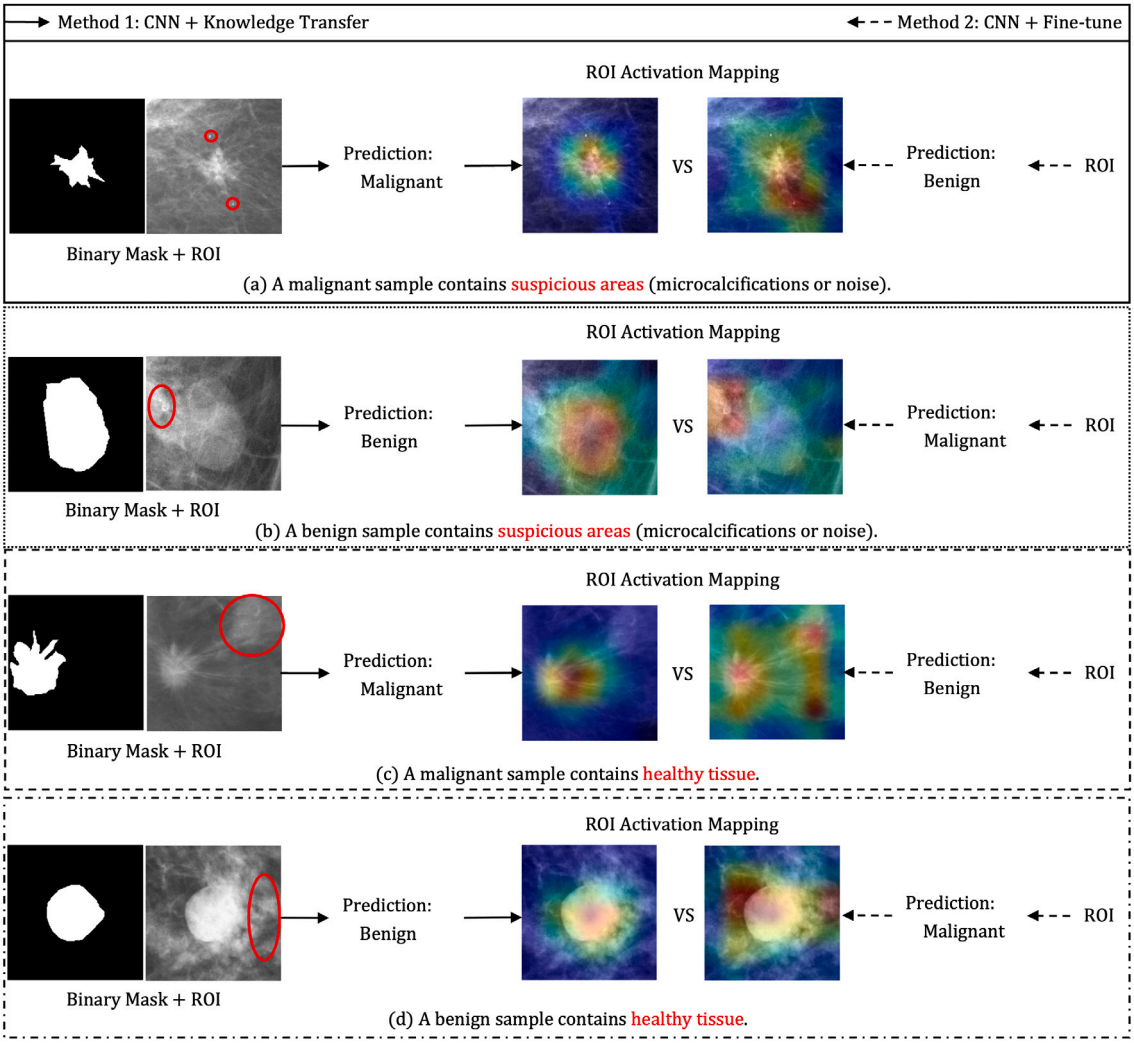


Fig. 7. The distribution of activation maps between knowledge transfer and no-knowledge transfer (classical fine-tuned) methods on ROI classification. Compared with the ROI activation mapping based on the no-knowledge transfer method, the activation levels based on the knowledge transfer method are more accurately localized masses at the same position as binary masks.

pixel-wise annotations, which are very difficult to obtain in practice. Enlightened by recent explorations of automatic segmentation algorithms [24,54], various deep learning approaches can be investigated to improve the usability of this project. Thirdly, the definition of features directly utilizes traditional texture features, which are used to divide the ROI into mass area and no-mass area. Inspired by [55], filter Integration obtains de-noising algorithms that can be incorporated so that an additional suggestion of mass area recognition can be provided in image pre-processing to improve the models' interpretability.

7. Conclusion

A framework for interpretable CNN-based mammographic image analysis is proposed for classifying mammographic masses, utilizing knowledge transfer from binary masks. An activation map is defined to contain valuable information, particularly shape knowledge from binary masks, which is used to enhance the final mass classification. The CNN model achieved a mapping from relative ROIs to class labels, and the model also has activation maps that are similar to those of the binary masks. The proposed method based on DenseNet121 achieved improved accuracy of $86.1 \pm 3.4\%$ compared to $67.9 \pm 3.8\%$ for the original model on mass classification. The results show that knowledge transfer achieved improved accuracy and unbiased learning compared to without knowledge transfer. The proposed interpretable algorithm defined clinical characteristics. The projection window in the ROI will be replaced with defined characteristics as inputs sent to the model to compare the distance between activation tensors of the ROI and replaced ROIs. The results show that it is able to highlight parts of the image in ROI activation mapping, and give an explanation of what attributes are used for classification. This provides mammographic readers with the means not to simply trust the model but to check its output for plausibility.

CRediT authorship contribution statement

Guobin Li: Writing – original draft, Methodology, Investigation, Conceptualization. **Mou Zhou:** Writing – review & editing. **Yu Fu:** Methodology. **Nashid Alam:** Data curation. **Erika Denton:** Writing – review & editing. **Reyer Zwiggelaar:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

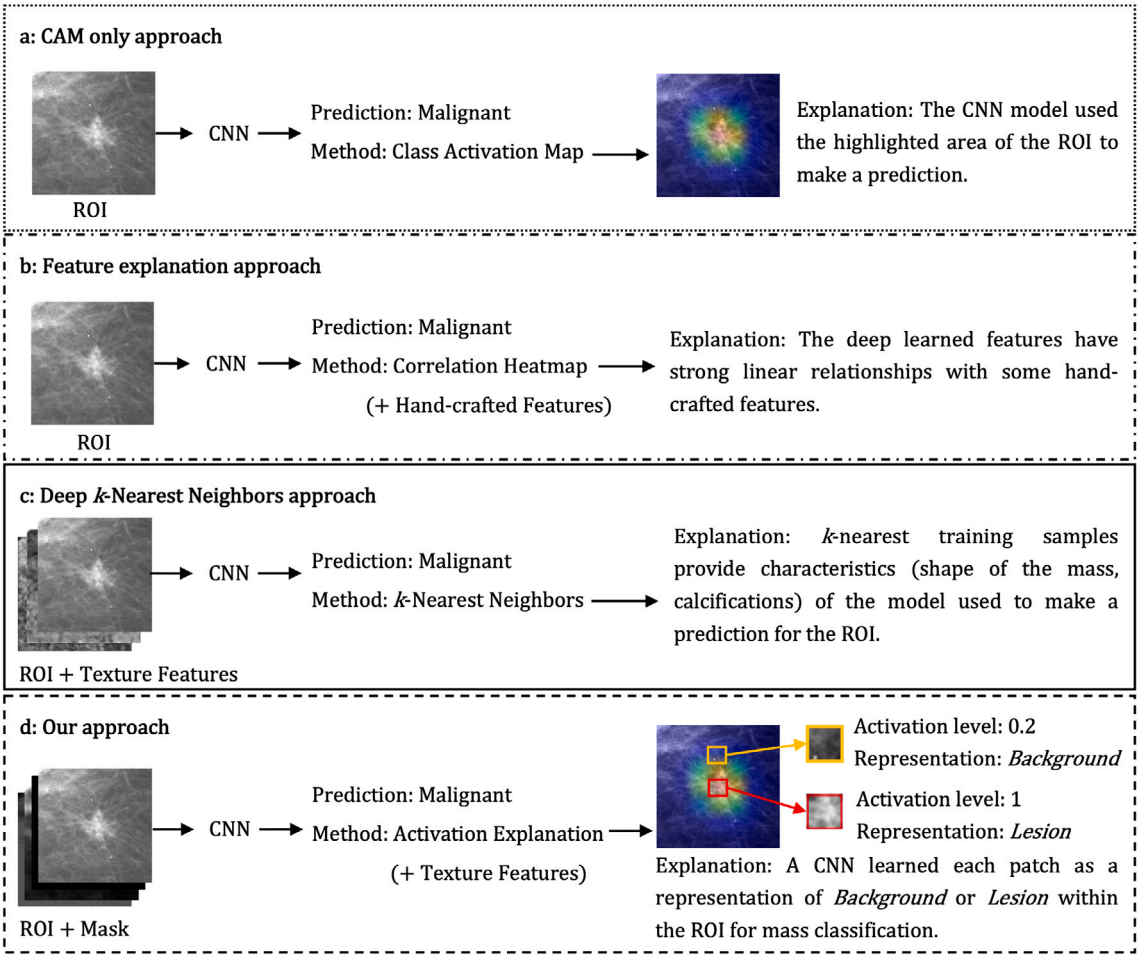


Fig. 8. (a) An interpretable approach [49,50] might point out which parts of the ROI are used for decision marking, but there is no explanation of these parts. (b) An explainable approach [37,39] for the abstract features used Pearson correlation coefficients of hand-crafted features and deep learned features to find the linear associations that exist between them. (c) Another interpretable approach [9,51] has proposed the use of k-Nearest Neighbours to find k similarly training samples and which characteristics explain the prediction on the test sample. (d) Our framework that highlights parts of the ROI, identifies the highlighted parts as two defined clinical characteristics to explain what attributions of the parts are used for classification.

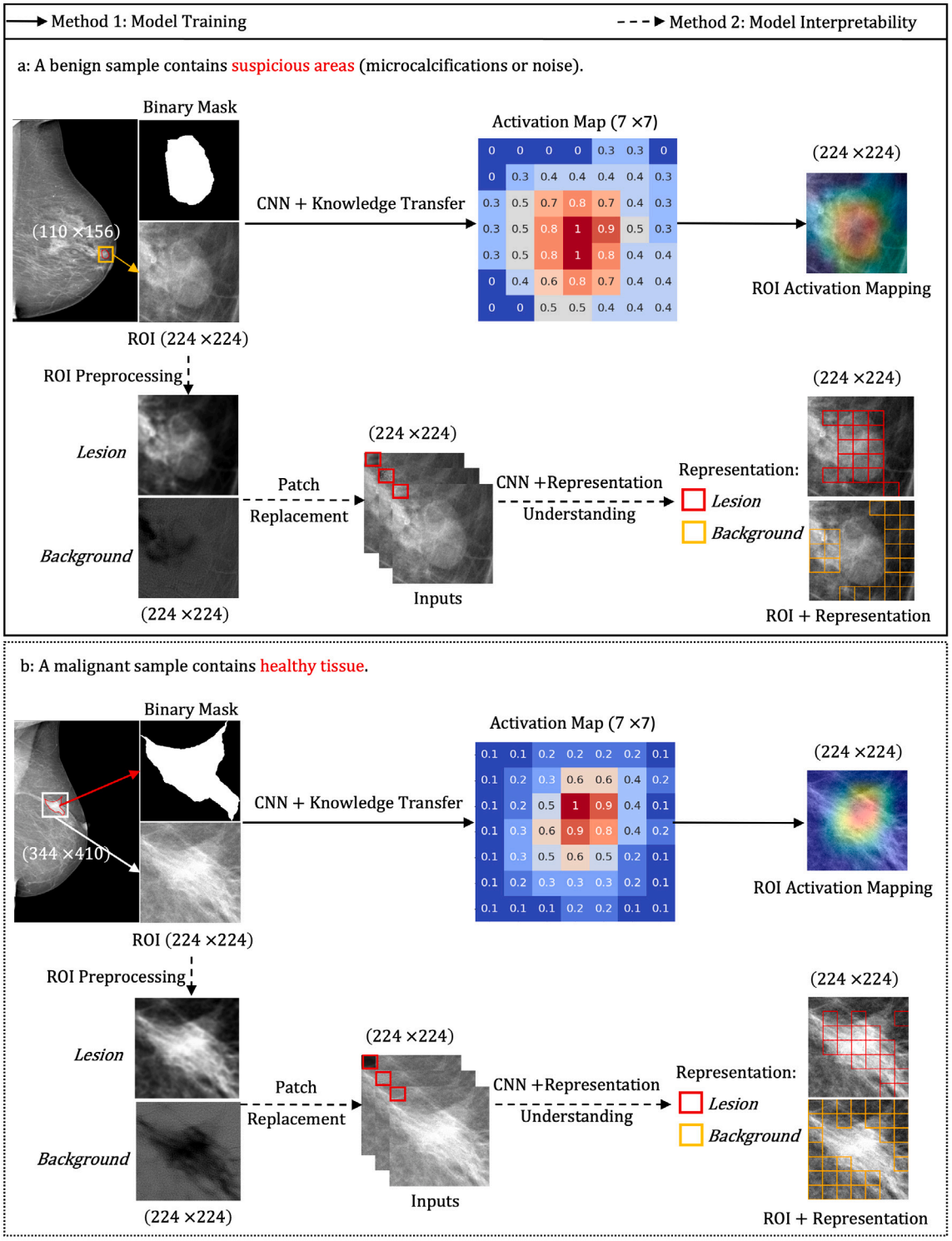


Fig. 9. Visualization of the proposed framework including model training and model interpretability.

References

[1] P. Mc Leod, B. Verma, Variable hidden neuron ensemble for mass classification in digital mammograms, IEEE Comput. Intell. Mag. 8 (1) (2013) 68–76.

[2] R.S. Lee, F. Gimenez, A. Hoogi, K.K. Miyake, M. Gorovoy, D.L. Rubin, A curated mammography data set for use in computer-aided detection and diagnosis research, Sci. Data 4 (1) (2017) 1–9.

[3] A. Oliver, J. Freixenet, J. Marti, E. Perez, J. Pont, E.R. Denton, R. Zwiggelaar, A review of automatic mass detection and segmentation in mammographic images, Med. Image Anal. 14 (2) (2010) 87–110.

[4] L. Shen, L.R. Margolies, J.H. Rothstein, E. Fluder, R. McBride, W. Sieh, Deep learning to improve breast cancer detection on screening mammography, Sci. Rep. 9 (1) (2019) 12495.

[5] B. Swiderski, S. Osowski, J. Kurek, M. Kruk, I. Lugowska, P. Rutkowski, W. Barhoumi, Novel methods of image description and ensemble of classifiers in application to mammogram analysis, Expert Syst. Appl. 81 (2017) 67–78.

[6] Y. Jiang, Computer-aided diagnosis of breast cancer in mammography: Evidence and potential, Technol. Cancer Res. Treat. 1 (3) (2002) 211–216.

[7] A. Hamidinekoo, E. Denton, A. Rampun, K. Honnor, R. Zwiggelaar, Deep learning in mammography and breast histology, an overview and future trends, Med. Image Anal. 47 (2018) 45–67.

[8] Y. LeCun, K. Kavukcuoglu, C. Farabet, Convolutional networks and applications in vision, in: Proceedings of 2010 IEEE International Symposium on Circuits and Systems, IEEE, 2010, pp. 253–256.

[9] G. Li, R. Zwiggelaar, Feature learning based on connectivity estimation for unbiased mammography mass classification, Comput. Vis. Image Underst. 238 (2024) 103884.

[10] B. Edwards, FDA guidance on clinical decision support: Peering inside the black box of algorithmic intelligence, 2017, ChilmarkResearch https://www.chilmarkresearch.com/fda-guidance-clinical-decision-support.

[11] S. Soffer, A. Ben-Cohen, O. Shimon, M.M. Amitai, H. Greenspan, E. Klang, Convolutional neural networks for radiologic images: A radiologist's guide, Radiology 290 (3) (2019) 590–606.

[12] G. Carneiro, J. Nascimento, A.P. Bradley, Automated analysis of unregistered multi-view mammograms with deep learning, IEEE Trans. Med. Imaging 36 (11) (2017) 2355–2365.

[13] T. Kooi, B. van Ginneken, N. Karssemeijer, A. den Heeten, Discriminating solitary cysts from soft tissue lesions in mammography using a pretrained deep convolutional neural network, Med. Phys. 44 (3) (2017) 1017–1027.

[14] H. Li, D. Chen, W.H. Nailon, M.E. Davies, D.I. Laurenson, Signed laplacian deep learning with adversarial augmentation for improved mammography diagnosis, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22, Springer, 2019, pp. 486–494.

[15] S. Zagoruyko, N. Komodakis, Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer, 2016, arXiv preprint arXiv:1612.03928.

[16] C. Varela, S. Timp, N. Karssemeijer, Use of border information in the classification of mammographic masses, Phys. Med. Biol. 51 (2) (2006) 425.

[17] M.L. Giger, N. Karssemeijer, J.A. Schnabel, Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer, Annu. Rev. Biomed. Eng. 15 (1) (2013) 327–357.

[18] A. Jalalian, S.B. Mashohor, H.R. Mahmud, M.I.B. Saripan, A.R.B. Ramli, B. Karasfi, Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: A review, Clin. Imaging 37 (3) (2013) 420–426.

[19] W.L. Bi, A. Hosny, M.B. Schabath, M.L. Giger, N.J. Birkbak, A. Mehrtash, T. Allison, O. Arnaout, C. Abbosh, I.F. Dunn, et al., Artificial intelligence in cancer imaging: Clinical challenges and applications, CA: Cancer J. Clin. 69 (2) (2019) 127–157.

[20] G. Carneiro, J. Nascimento, A.P. Bradley, Unregistered multiview mammogram analysis with pre-trained deep learning models, in: Medical Image Computing and Computer-Assisted Intervention, MICCAI 2015, Springer International Publishing, Cham, 2015, pp. 652–660.

[21] J. Arevalo, F.A. González, R. Ramos-Pollán, J.L. Oliveira, M.A.G. Lopez, Representation learning for mammography mass lesion classification with convolutional neural networks, Comput. Methods Programs Biomed. 127 (2016) 248–257.

[22] T. Kooi, G. Litjens, B. van Ginneken, A. Gubern-Mérida, C.I. Sánchez, R. Mann, A. den Heeten, N. Karssemeijer, Large scale deep learning for computer aided detection of mammographic lesions, Med. Image Anal. 35 (2017) 303–312.

[23] N. Dhungel, G. Carneiro, A.P. Bradley, The automated learning of deep features for breast mass classification from mammograms, in: Medical Image Computing and Computer-Assisted Intervention, MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19, Springer, 2016, pp. 106–114.

[24] N. Dhungel, G. Carneiro, A.P. Bradley, A deep learning approach for the analysis of masses in mammograms with minimal user intervention, Med. Image Anal. 37 (2017) 114–128.

[25] M.A. Al-Antari, M.A. Al-Masni, M.-T. Choi, S.-M. Han, T.-S. Kim, A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification, Int. J. Med. Informatics 117 (2018) 44–54.

[26] A. Baccouche, B. Garcia-Zapirain, A.S. Elmaghraby, An integrated framework for breast mass classification and diagnosis using stacked ensemble of residual neural networks, Sci. Rep. 12 (1) (2022) 12259.

[27] A. Hamidinekoo, Z. Suhail, T. Qaiser, R. Zwiggelaar, Investigating the effect of various augmentations on the input data fed to a convolutional neural network for the task of mammographic mass classification, in: Medical Image Understanding and Analysis: 21st Annual Conference, MIUA 2017, Edinburgh, UK, July 11–13, 2017, Proceedings 21, Springer, 2017, pp. 398–409.

[28] L. Tsochatzidis, K. Zagoris, N. Arikidis, A. Karahaliou, L. Costaridou, I. Pratikakis, Computer-aided diagnosis of mammographic masses based on a supervised content-based image retrieval approach, Pattern Recognit. 71 (2017) 106–117.

[29] F. Yan, H. Huang, W. Pedrycz, K. Hirota, Automated breast cancer detection in mammography using ensemble classifier and feature weighting algorithms, Expert Syst. Appl. 227 (2023) 120282.

[30] G. Carneiro, J. Nascimento, A.P. Bradley, Unregistered multiview mammogram analysis with pre-trained deep learning models, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 652–660.

[31] H. Li, D. Chen, W.H. Nailon, M.E. Davies, D.I. Laurenson, Dual convolutional neural networks for breast mass segmentation and diagnosis in mammography, IEEE Trans. Med. Imaging 41 (1) (2021) 3–13.

[32] C. Shu, Y. Liu, J. Gao, Z. Yan, C. Shen, Channel-wise knowledge distillation for dense prediction, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 5311–5320.

[33] 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.

[34] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. 1 (5) (2019) 206–215.

[35] P. Lambin, R.T. Leijenaar, T.M. Deist, J. Peerlings, E.E. De Jong, J. Van Timmeren, S. Sanduleanu, R.T. Larue, A.J. Even, A. Jochems, et al., Radiomics: The bridge between medical imaging and personalized medicine, Nat. Rev. Clin. Oncol. 14 (12) (2017) 749–762.

[36] J.J. Van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R.G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, H.J. Aerts, Computational radiomics system to decode the radiographic phenotype, Cancer Res. 77 (21) (2017) e104–e107.

[37] T. Chowdhury, A.R. Bajwa, T. Chakraborti, J. Rittscher, U. Pal, Exploring the correlation between deep learned and clinical features in melanoma detection, in: Medical Image Understanding and Analysis: 25th Annual Conference, MIUA 2021, Oxford, United Kingdom, July 12–14, 2021, Proceedings 25, Springer, 2021, pp. 3–17.

[38] J. Lao, Y. Chen, Z.-C. Li, Q. Li, J. Zhang, J. Liu, G. Zhai, A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme, Sci. Rep. 7 (1) (2017) 10353.

[39] Y. Zhang, E.M. Lobo-Mueller, P. Karanicolas, S. Gallinger, M.A. Haider, F. Khalvati, Improving prognostic performance in resectable pancreatic ductal adenocarcinoma using radiomics and deep learning features fusion in CT images, Sci. Rep. 11 (1) (2021) 1378.

[40] G. Li, C. Thomas, R. Zwiggelaar, Comparison of deep learned and texture features in mammographic mass classification, in: 16th International Workshop on Breast Imaging, IWBI2022, Vol. 12286, SPIE, 2022, pp. 153–159.

[41] M.G. Lopez, N. Posada, D.C. Moura, R.R. Pollán, J.M.F. Valiente, C.S. Ortega, M. Solar, G. Diaz-Herrero, I. Ramos, J.P. Loureiro, T.C. Fernandes, B.F. Araújo, BCDR: A breast cancer digital repository, in: 15th International Conference on Experimental Mechanics, Vol. 1215, 2012, pp. 113–120.

[42] M. Heath, K. Bowyer, D. Kopans, P. Kegelmeyer Jr., R. Moore, K. Chang, S. Munishkumaran, Current status of the digital database for screening mammography, in: Digital Mammography, Nijmegen, 1998, Springer, 1998, pp. 457–460.

[43] A. Hamidinekoo, Z. Suhail, E. Denton, R. Zwiggelaar, Comparing the performance of various deep networks for binary classification of breast tumours, in: 14th International Workshop on Breast Imaging, IWBI 2018, Vol. 10718, SPIE, 2018, pp. 36–43.

[44] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[45] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, F. Li, Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (2015) 211–252.

[46] S. Ortega-Martorell, P. Riley, I. Olier, R.G. Raidou, R. Casana-Eslava, M. Rea, L. Shen, P.J. Lisboa, C. Palmieri, Breast cancer patient characterisation and visualisation using deep learning and fisher information networks, Sci. Rep. 12 (1) (2022) 14004.

[47] A. Hamidinekoo, Z.C. Dagdia, Z. Suhail, R. Zwiggelaar, Distributed rough set based feature selection approach to analyse deep and hand-crafted features for mammography mass classification, in: 2018 IEEE International Conference on Big Data, Big Data, IEEE, 2018, pp. 2423–2432.

[48] J. Mordang, A. Gubern-Mérida, A. Bria, F. Tortorella, R. Mann, M. Broeders, G. Den Heeten, N. Karssemeijer, The importance of early detection of calcifications associated with breast cancer in screening, Breast Cancer Res. Treat. 167 (2018) 451–458.

[49] R.R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, D. Batra, Grad-CAM: Why did you say that?, 2016, arXiv preprint arXiv:1611.07450.

[50] R.L. Draelos, L. Carin, Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks, 2020, arXiv preprint arXiv:2011.08891.

[51] N. Papernot, P. McDaniel, Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning, 2018, arXiv preprint arXiv:1803.04765.

[52] C. Abbet, I. Zlobec, B. Bozorgtabar, J.-P. Thiran, Divide-and-rule: Self-supervised learning for survival analysis in colorectal cancer, in: Medical Image Computing and Computer Assisted Intervention, MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23, Springer, 2020, pp. 480–489.

[53] X. Xie, J. Chen, Y. Li, L. Shen, K. Ma, Y. Zheng, Instance-aware self-supervised learning for nuclei segmentation, in: Medical Image Computing and Computer Assisted Intervention, MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23, Springer, 2020, pp. 341–350.

[54] H. Li, D. Chen, W.H. Nailon, M.E. Davies, D. Laurenson, Improved breast mass segmentation in mammograms with conditional residual u-net, in: Image Analysis for Moving Organ, Breast, and Thoracic Images: Third International Workshop, RAMBO 2018, Fourth International Workshop, BIA 2018, and First International Workshop, TIA 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16 and 20, 2018, Proceedings 3, Springer, 2018, pp. 81–89.

[55] D. Devakumari, V. Punithavathi, Noise removal in breast cancer using hybrid denoising filter for mammogram images, in: Computational Vision and Bio-Inspired Computing, ICCVBIC 2019, Springer, 2020, pp. 109–119.