Feature learning based on connectivity estimation for unbiased mammography mass classification

Li, Guobin; Zwiggelaar, Reyer

Published in:
Computer Vision and Image Understanding

DOI:
10.1016/j.cviu.2023.103884

Publication date:
2024

Citation for published version (APA):
Li, G., & Zwiggelaar, R. (2024). Feature learning based on connectivity estimation for unbiased mammography mass classification. *Computer Vision and Image Understanding*, *238*, Article 103884. https://doi.org/10.1016/j.cviu.2023.103884

Computer Vision and Image Understanding 238 (2024) 103884

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Feature learning based on connectivity estimation for unbiased mammography mass classification

Guobin Li [*], Reyer Zwiggelaar

Department of Computer Science, Aberystwyth University, UK

ARTICLE INFO

Communicated by Samuel Kadoury

MSC:
41A05
41A10
65D05
65D17

Keywords:
Breast cancer
Texture features
Deep learned features
Interpretability

ABSTRACT

Breast cancer is the most commonly diagnosed female malignancy worldwide. Recent developments in deep convolutional neural networks have shown promising performance for breast cancer detection and classification. However, due to variations in appearance and small datasets, biased features can be learned by the networks in distinguishing malignant and benign instances. To investigate these aspects, we trained a densely connected convolutional network (DenseNet) to obtain representative features of breast tissue, selecting texture features representing different physical morphological representations as the network's inputs. Connectivity estimation, represented by a connection matrix, is proposed for feature learning. To make the network provide an unbiased prediction, we used k-nearest neighbors to find k training samples whose connection matrices are closest to the test case. When evaluated on OMI-DB we achieved improved diagnostic accuracy $73.89 \pm 2.89\%$ compared with $71.35 \pm 2.66\%$ for the initial CNN model, which showed a statistically significant difference ($p = 0.00036$). The k training samples can provide visual explanations which are useful in understanding the model predictions and failures of the model.

1. Introduction

Breast cancer, diagnosed in over two million women each year, stands as the most frequently identified non-skin cancer (Sung et al., 2021). Mammography is the primary screening method for lesion visualisation and detecting early changes in breast tissue. Radiologists analyse mammograms to identify any signs of abnormalities, such as masses, microcalcifications, or architectural distortions (Chen et al., 2014; Hamidinekoo et al., 2018a). They differentiate between benign and potentially malignant findings based on their expertise.

Computer-Aided Diagnoses (CAD) systems are designed as tools to assist radiologists in the detection and/or classification of mammographic abnormalities (Oliver et al., 2010; Alam et al., 2018). Machine learning, and in particular deep learning technologies, have been proposed to build CAD systems and have shown remarkable progress in mammography lesion classification (Jiao et al., 2018; Shen et al., 2019; Li et al., 2020; Yu et al., 2021).

Convolutional Neural Networks (CNNs) are one of the most popular deep learning technologies. Hamidinekoo et al. (2018b) have published a comprehensive review of CNNs in mammographic image processing. A CNN consists of multiple convolution layers stacked on top of each other. It is trained by feeding suitable inputs, learning hierarchical features layer by layer and then producing the final output.

One critical concern for CNNs is the lack of training data (Hamidinekoo et al., 2017). In addition, masses and calcifications both appear in mammography as common clinical signs. Feature-wise results that were obtained from CNNs, display a fixation on the calcification over other features. The latter can be used by radiologists to determine if the mass is benign or malignant. This is likely because calcifications are strongly associated with some typical breast cancers (such as ductal carcinoma in situ and invasive cancers) (Mordang et al., 2018). Thus, the model will easily disregard the mass and instead prioritise the identification of calcium, potentially resulting in the misclassification of the sample.

Some works have proposed integrating handcrafted features with CNNs to improve mass classification (Arevalo et al., 2015, 2016; Hamidinekoo et al., 2018a). However, some handcrafted features are highly correlated with each other (Zhang et al., 2021). The high dimensionality of features increases the complexity of the CNNs, yet the corresponding increase in performance is relatively limited. Inspired by Domingues et al. (2012), Dhahbi et al. (2015) and Swiderski et al. (2017), breast mass diagnosis depends on the shape and margin of mass rather than conventional hand-crafted features, and Li et al. (2021) proposed extracting shape features by using binary masks and texture features from CNNs. They integrated these two separate features to achieve improved accuracy. The binary mask was produced by an automatic mask segmentation algorithm. It is difficult to evaluate the segmentation accuracy on some public datasets without binary mask labelling.

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

https://doi.org/10.1016/j.cviu.2023.103884
Received 25 April 2023; Received in revised form 23 October 2023; Accepted 9 November 2023
Available online 11 November 2023

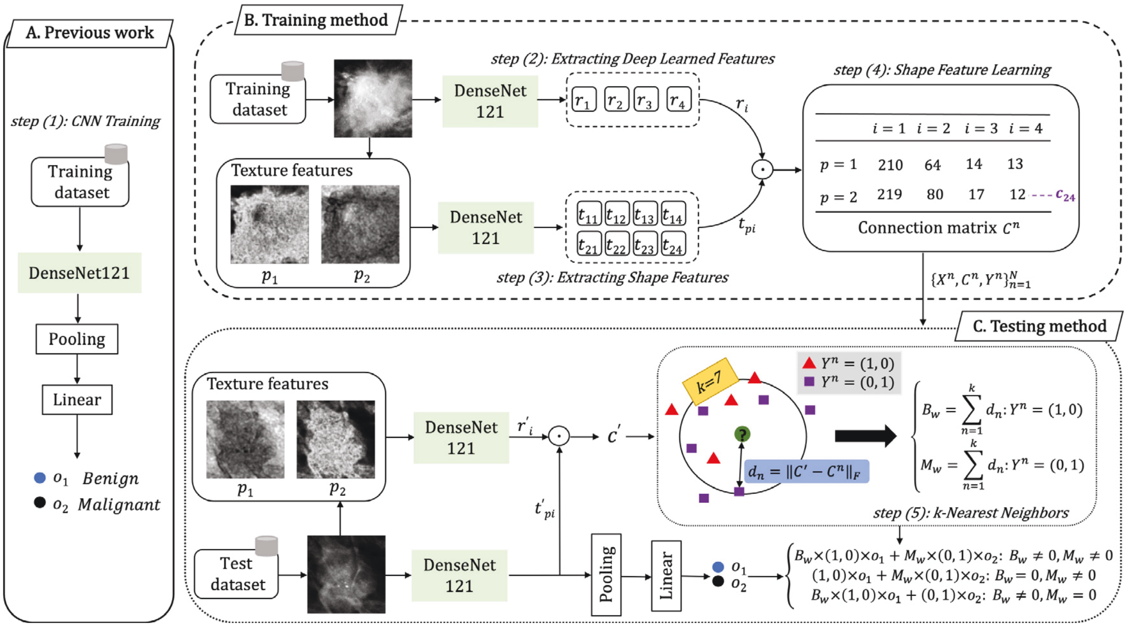

A. Previous work
step (1): CNN Training
Training dataset
DenseNet121
Pooling
Linear
o_1 Benign
o_2 Malignant
B. Training method
Training dataset
step (2): Extracting Deep Learned Features
DenseNet 121
r_1 r_2 r_3 r_4
r_i
Texture features
p_1 p_2
DenseNet 121
t_{11} t_{12} t_{13} t_{14}
t_{21} t_{22} t_{23} t_{24}
t_{pi}
step (3): Extracting Shape Features
step (4): Shape Feature Learning
$i=1$ $i=2$ $i=3$ $i=4$
$p=1$ 210 64 14 13
$p=2$ 219 80 17 12 c_{24}
Connection matrix C^n
$\{X^n, C^n, Y^n\}_{n=1}^N$
C. Testing method
Texture features
p_1 p_2
DenseNet 121
r'_i
C'
t'_{pi}
Test dataset
DenseNet 121
Pooling
Linear
o_1
o_2
k=7
$Y^n = (1,0)$
$Y^n = (0,1)$
$d_n = \|C' - C^n\|_F$
$B_w = \sum_{n=1}^{k} d_n : Y^n = (1,0)$
$M_w = \sum_{n=1}^{k} d_n : Y^n = (0,1)$
step (5): k-Nearest Neighbors
$B_w \times (1,0) \times o_1 + M_w \times (0,1) \times o_2$: $B_w \neq 0, M_w \neq 0$
$(1,0) \times o_1 + M_w \times (0,1) \times o_2$: $B_w = 0, M_w \neq 0$
$B_w \times (1,0) \times o_1 + (0,1) \times o_2$: $B_w \neq 0, M_w = 0$

Fig. 1. The architecture of our proposed method. Step (1): CNN training. We train DenseNet121 for benign or malignant breast lesion classification. The trained DenseNet121 is used in subsequent steps. Step (2): deep learned features are extracted from the trained DenseNet121 using the training dataset. Step (3): shape features are extracted from the trained DenseNet121 using the texture-images. Step (4): A connection matrix is built to provide feature learning. For a test case, we use step (5): k-Nearest Neighbors to find k training samples to support the trained DenseNet121's diagnosis.

Another critical concern for CNNs is the lack of explainability. As CNNs become more complex, it becomes more complicated to understand their decision-making process. Common approaches such as Saliency Maps and GradCAM have been commonly used to understand which parts of an image a CNN used to make a prediction (Simonyan et al., 2013; Selvaraju et al., 2017). A new output-level, gradient-based CNN explanation approach, High-Resolution Class Activation Mapping (HiResCAM) (Draelos and Carin, 2020) is designed as an improvement on GradCAM. However, recent studies (Mueller et al., 2019; Kenny et al., 2021) have reported that the evaluation of these techniques has lagged behind CNNs development, in some cases providing explanations that radiologists do not understand or find too complex. In recent years, some workers (Papernot and McDaniel, 2018; Ortega-Martorell et al., 2022) have proposed the use of k-nearest neighbors (kNN) training, to explain the prediction on the test dataset. This could help not only to understand the CNNs but even identify k-nearest training samples to assist in determining subsequent treatment.

Based on these two observations (i.e. lack of data and lack of explainability), we are motivated to construct a breast cancer diagnosis model: in order to solve the biased learning by integrating shape features and simultaneously providing visual explanations of the model predictions. To achieve this, our methods as illustrated in Fig. 1 contain the following novel aspects: (a) the combination of deep learned, shape and texture features for the classification of mammographic abnormalities; (b) building connectivity estimation for feature learning; and (c) using k-nearest neighbors to identify similar cases.

The desired properties of the proposed model: (1) ensure that the model uses accurate information, such as the shape of masses to do predictions; (2) propose a visualisation of the decision-making process in the model.

2. Methods

2.1. Dataset

In this study, we used the Optimam Medical Image Database (OMI-DB) (Halling-Brown et al., 2020), which contains NHS Breast Screening Programme (NHSBSP) images from multiple breast screening centres across the UK and has been created to serve as a large repository of de-identified medical images to support research involving medical imaging. The database contains digital mammograms in a standard DICOM format. Both craniocaudal (CC) and mediolateral oblique (MLO) views are available for most cases. Each view is treated as a separate image in this study. It also contains pixel-level annotations for the regions of interest (ROI) and their pathology, including the BI-RADS (Martin et al., 2006) assessment. BI-RADS scores range from B0 to B6, however in this study, only B2 and B5 cases are used. B2 is regarded as benign, and B5 is regarded as malignant. It further labels each ROI as containing calcifications or masses.

We used a dataset containing 1870 breast lesions, 914 lesions are benign and 956 are malignant. This dataset is created by sampling image patches from ROIs and resizing them to 224 × 224 using interpolation. Data augmentation is used such that each training image is randomly flipped and rotated, to generate a total of 4488 images.

2.2. Related work

2.2.1. CNN training

Hamidinekoo et al. (2018c) have published a comparative study on various types of CNNs for binary classification of breast tumours. DenseNet is a strongly performing deep learning model because it directly connects from any layer to all subsequent layers to receive the feature maps of all preceding layers. This is done for concatenating features while disregarding redundant feature maps during training. Since the total number of our dataset is limited, we have selected DenseNet121 (Huang et al., 2017), which represents limited parameters in our proposed approach.

This model can be divided into the first convolution layer, four dense blocks, three transition layers and the classification layer. The initial convolution layer incorporates a 7 × 7 convolution and a 3 × 3 max pooling on the input images. For dense blocks, dense blocks 1, 2, 3 and 4 have 6, 12, 24 and 16 bottleneck layers implemented by a 1 × 1 convolution before the 3 × 3 convolution layer, which is helpful in disregarding redundant feature maps. Between the dense blocks, there

Table 1
Details of the DenseNet121 structure. Conv, max, average and FC represent the convolution, max pooling, average pooling and fully-connected layers, respectively.

Layer	DenseNet121	Output Size
Convolution	7 × 7 conv, stride 2	112 × 112
Pooling	3 × 3 max, stride 2	56 × 56
Dense block (1)	$\begin{bmatrix}1\times 1\text{conv}\\ 3\times 3\text{conv}\end{bmatrix}\times 6$	56 × 56
Transition layer (1)	$1\times 1\text{conv}$ $2\times 2\text{averge, stride}2$	28 × 28
Dense block (2)	$\begin{bmatrix}1\times 1\text{conv}\\ 3\times 3\text{conv}\end{bmatrix}\times 6$	28 × 28
Transition layer (2)	$1\times 1\text{conv}$ $2\times 2\text{averge, stride}2$	14 × 14
Dense Block (3)	$\begin{bmatrix}1\times 1\text{conv}\\ 3\times 3\text{conv}\end{bmatrix}\times 6$	14 × 14
Transition layer (3)	$1\times 1\text{conv}$ $2\times 2\text{averge, stride}2$	7 × 7
Dense block (4)	$\begin{bmatrix}1\times 1\text{conv}\\ 3\times 3\text{conv}\end{bmatrix}\times 6$	7 × 7
Classification	$7\times 7\text{max}$ 2FC, softmax	2 × 1

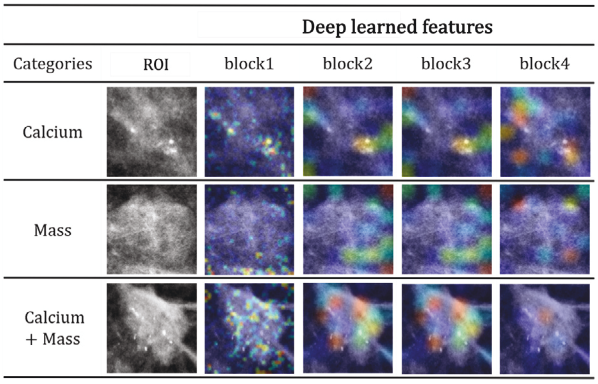


Fig. 2. Deep learned feature extraction. Breast lesions containing only calcium, only mass and calcium + mass. Block 1, 2, 3 and 4 represent the deep learned features corresponding to the breast lesions. They are rescaled to be within [0, 1] and a low cutoff of 0.06 is used to remove background noise. To gain a better understanding of these deep learned features and hence the model behaviour, they are resized to 224 × 224 using interpolation and viewed in colour, and the original ROI is added as a background. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

are transition layers, which are composed of batch normalisation, 1 × 1 convolution, and 2 × 2 average pooling operations to reduce the feature maps. The details of the various layers are shown in Table 1.

DenseNet121 is trained to diagnose whether each ROI is benign or malignant. The classification layer is modified since the original average pooling and fully-connected layers are developed to classify 1,000 categories instead of two. We replace the average pooling with a max pooling layer which is more sensitive to outliers, followed by a dense layer with a ReLu activation of 256 neurons and a dropout layer with a 0.2 rate, and then the final dense layer, a sigmoid activation with only two classes.

We use transfer learning based on pre-trained weights using the ImageNet (Russakovsky et al., 2015) dataset representing primitive features that tend to be preserved across different tasks, whereas the classification layer with randomly initialised weights represents higher-order features that are more related to specific tasks and require further training. Subsequently, DenseNet121 is fine-tuned on a training and validation dataset of OMI-DB. The trained Densent121 will be used for downstream analysis.

2.2.2. Extracting deep learned features

Heatmaps have been widely applied to protect patients from biased CNN results. The most recent method, HiResCAM (Draelos and Carin, 2020) faithfully represents the locations within the images that a CNN has used to make a decision, even if these locations are outside the object of interest. In this method, define s_m as a DenseNet's raw score for class m before a sigmoid function is applied to provide predicted probabilities. We first compute the gradient of s_m with respect to a collection of feature maps $F = \{F_j\}_{j=1}^{N}$ produced by a convolutional layer, where j means the depth of feature maps. For a 2D input, this gradient is 3-dimensional, $[width, height, N]$, matching the shape of the collection of feature maps. After computing the gradient, the heatmap is produced by element-wise multiplying the gradient and the feature maps before summing over the feature dimension:

$$r_m^{HiResCAM} = \sum_{j=1}^{N} \frac{\partial s_m}{\partial F_j} \odot F_j \tag{1}$$

where $r_m^{HiResCAM}$ is referred to as the deep learned features, and $\odot$ represents a Hadamard product. N indicates the depth of the feature maps.

To investigate the unbiasedness of DenseNet121, we extract low-level, mid-level and high-level knowledge of the model during training. For an input patch $X \in R^{224\times 224}$, deep learned features that are extracted from the pooling layer of four dense blocks. Hence, dense block 1, 2, 3, and 4 have gradients, which sizes are [28 × 28 × 128], [14 × 14 × 256], [7 × 7 × 512] and [7 × 7 × 1024], respectively. The deep learned features r_i come from four dense blocks that have sizes of 28 × 28, 14 × 14, 7 × 7 and 7 × 7, respectively. Fig. 2shows deep learned features extracted from three categories of breast lesions. We are only interested in the features that have positive values for positive influence on the target class which saves computational costs. Negative values are likely to belong to other categories in the image.

As shown in Fig. 2, breast lesions contain only calcium, only mass and calcium + mass. Block 1, 2, 3 and 4 represent the deep learned features corresponding to the breast lesions. To gain a better understanding of these deep learned features, they are resized and viewed in colour, and the original ROI is added as a background. The first row shows the identified areas that are close to calcium. The second row shows a typical breast lesion which only contains a mass. Here, block 1 and block 2 seem to catch the boundary information, and block 3 and block 4 identify background and light patches of the breast lesion which look like calcium. The last row shows a typical breast lesion which not only contains a mass but also contains calcium, the model prefers to locate calcium and it is difficult to discern other information.

2.2.3. Similarity between deep learned features and GLCM texture-images

Lao et al. (2017), Chowdhury et al. (2021), Zhang et al. (2021), Li et al. (2022) have demonstrated that the deep learned features have similar information as hand-crafted features in many areas of medical image analysis. To gain a better understanding of deep learned features for binary classification of breast lesions, we use texture features to investigate the similarity between them.

Within a selected ROI, there are several subregions showing different texture statistics, for example, *homogeneity* emphasises the surrounding tissue, *dissimilarity* emphasises the transition region between the mass and the surrounding tissue, and *auto-correlation* emphasises the region inside the mass. In this study, we use Grey Level Co-occurrence Matrix (GLCM) (Löfstedt et al., 2019) to obtain three GLCM texture-images, namely, *homogeneity*, *dissimilarity*, *auto-correlation*. An element of the GLCM, $p(i,j)$, is defined as the joint probability that grey-level i and grey-level j occur together. Texture-images were defined as follows.

Homogeneity:

$$\sum_{i=0}^{L-1}\sum_{j=0}^{L-1} \frac{1}{1+(i-j)^2} \cdot p(i,j) \tag{2}$$

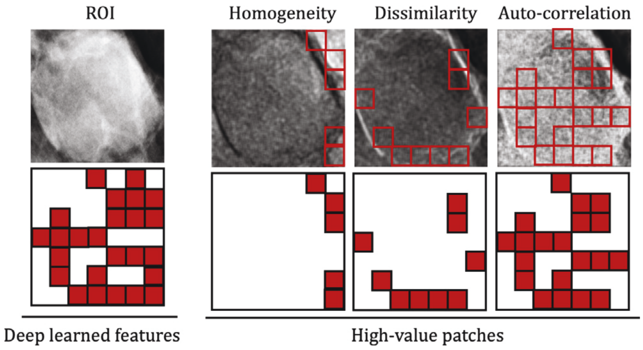


Fig. 3. The correlation between deep learned features and texture-images. This breast lesion contains a mass. The deep learned features are extracted from the third dense block and resized to 224 × 224 (e.g. represented as a 7 × 7 matrix). Positive values are marked as red patches to analysis in deep learned features. *Homogeneity* contains 5 red patches as deep learned features. *Dissimilarity* contains 9 red patches and *Auto-correlation* contains 20 red patches as deep learned features. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Dissimilarity:

$$\sum_{i=0}^{L-1}\sum_{j=0}^{L-1}|i-j|\cdot p(i,j) \tag{3}$$

Auto-correlation:

$$\sum_{i=0}^{L-1}\sum_{j=0}^{L-1}(i\cdot j)\cdot p(i,j) \tag{4}$$

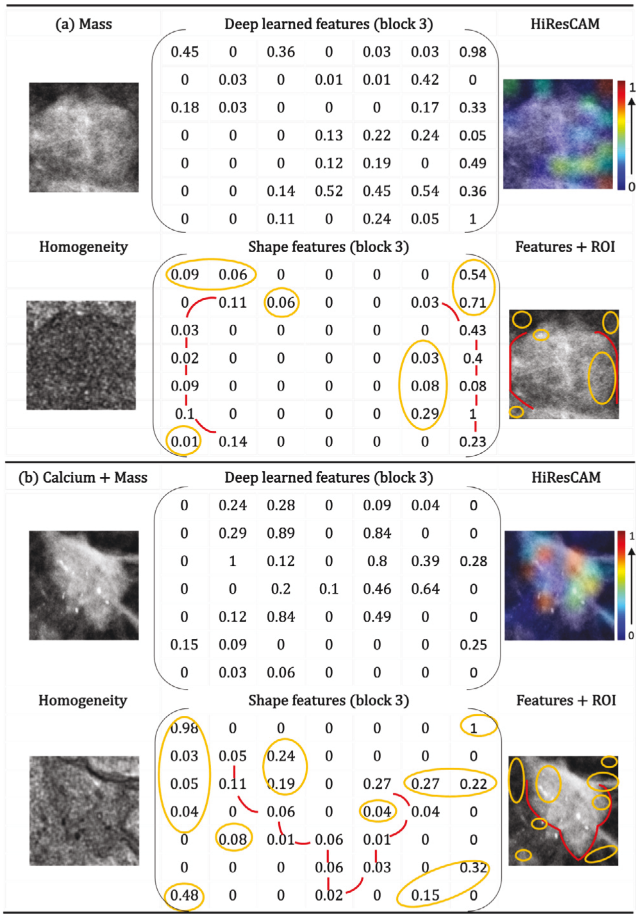
where L is defined as the number of grey-levels in a ROI. Each texture-image keeps the same size as ROI.

Fig. 3 shows an example case to compare its deep learned features and texture-images. The deep learned features are extracted from the third dense block and resized to 224 × 224 (e.g. represented as a 7 × 7 matrix). We marked features with positive values as red patches. If one deep learned feature has a positive value in a texture-image at the same position, we keep the red patch. If one deep learned feature corresponds to a low value in a texture-image, the red patch is removed. In high-value patches, there are five red patches in *homogeneity*. To gain a better intuition about texture-images, they are added with corresponding high-value patches and marked as red squares.

To be specific, *auto-correlation* not only contains 11 patches as deep learned features but also has 9 patches that are not included in *dissimilarity*. Considering these two texture-images are highly correlated with each other, the number of selected texture-images will affect the complexity of the model. We choose *homogeneity* and *auto-correlation* to improve the mass classification in the biased model.

2.3. Proposed methods

The proposed method for a graphical overview see Fig. 1 can be divided into five steps: step (1): CNN training. We use OMI-DB to train DenseNet121 for benign or malignant breast lesion classification; step (2): Extracting deep learned features. HiResCAM is utilised to produce deep learned features and represents the locations within the ROI that CNN has used to make a decision; step (3): Extracting shape features. Selecting textural ROIs representing different physical representations as inputs used by the trained CNN and extracting shape features; step (4): Feature learning. We build a connection matrix to learn shape features for each training sample; step (5): k-Nearest Neighbors. Once all of the connection matrices for the training set are calculated, we use k-nearest neighbors to find k training samples whose connection matrices are closest to the test case and use the neighbours to further improve the CNN's diagnosis.



Deep learned features (block 3)

0.45	0	0.36	0	0.03	0.03	0.98
0	0.03	0	0.01	0.01	0.42	0
0.18	0.03	0	0	0	0.17	0.33
0	0	0	0.13	0.22	0.24	0.05
0	0	0	0.12	0.19	0	0.49
0	0	0.14	0.52	0.45	0.54	0.36
0	0	0.11	0	0.24	0.05	1

Shape features (block 3)

0.09	0.06	0	0	0	0	0.54
0	0.11	0.06	0	0	0.03	0.71
0.03	0	0	0	0	0	0.43
0.02	0	0	0	0	0.03	0.4
0.09	0	0	0	0	0.08	0.08
0.1	0	0	0	0	0.29	1
0.01	0.14	0	0	0	0	0.23

Deep learned features (block 3)

0	0.24	0.28	0	0.09	0.04	0
0	0.29	0.89	0	0.84	0	0
0	1	0.12	0	0.8	0.39	0.28
0	0	0.2	0.1	0.46	0.64	0
0	0.12	0.84	0	0.49	0	0
0.15	0.09	0	0	0	0	0.25
0	0.03	0.06	0	0	0	0

Shape features (block 3)

0.98	0	0	0	0	0	1
0.03	0.05	0.24	0	0	0	0
0.05	0.11	0.19	0	0.27	0.27	0.22
0.04	0	0.06	0	0.04	0.04	0
0	0.08	0.01	0.06	0.01	0	0
0	0	0	0.06	0.03	0	0.32
0.48	0	0	0.02	0	0.15	0

Fig. 4. Visualisation of deep learned features. Breast lesions containing only mass (a) and calcium + mass (b). Deep learned features corresponding to (a) and (b) that are extracted from the third dense block. These are rescaled to be within [0, 1] and resized to 224 × 224 using interpolation. HiResCAM adds the original ROI as a background and shows deep learned features in colour. To gain a better intuition about deep learned features based on the *homogeneity* texture-image, they are shown in two colours. The red lines represent the shape of mass in the original ROI. The yellow ellipses represent calcium, background or other factors used by the trained DenseNet121 to make predictions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Step (1): CNN training and step (2): Extracting Deep learned features are implemented in Section 2.2. Steps (3)-(5) will work on top of the previous steps and be introduced in this section.

2.3.1. Extracting shape features

The trained DenseNet121 is employed for classifying benign or malignant lesions using texture-images. The first input image is *auto-correlation* to reduce the calcium's emphasis in the lesion. The second input image is *homogeneity* because we find the model considers not only the lesion itself but also its neighbourhood which contains relevant information, which is considered by radiologists for diagnosis (Hamidinekoo et al., 2017). Then deep learned features are extracted from texture-images and ROIs to compare the differences between them as shown in Fig. 4.

Similar to Fig. 2, we use the same breast lesions containing only mass (a) and calcium + mass (b) to compare the deep learned features from the original ROI and *homogeneity* texture-image. The deep learned features are extracted from the third dense block (e.g. represented as a 7 × 7 matrix). The distribution of features from the ROI is scattered. The distribution of features based on *homogeneity* is more concentrated, which we manually divided into two groups. The first group used red lines to connect them correspond to the shape of the

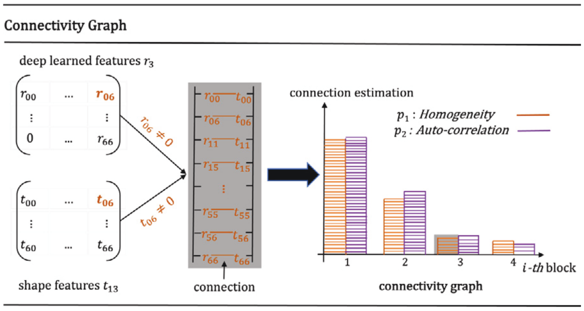
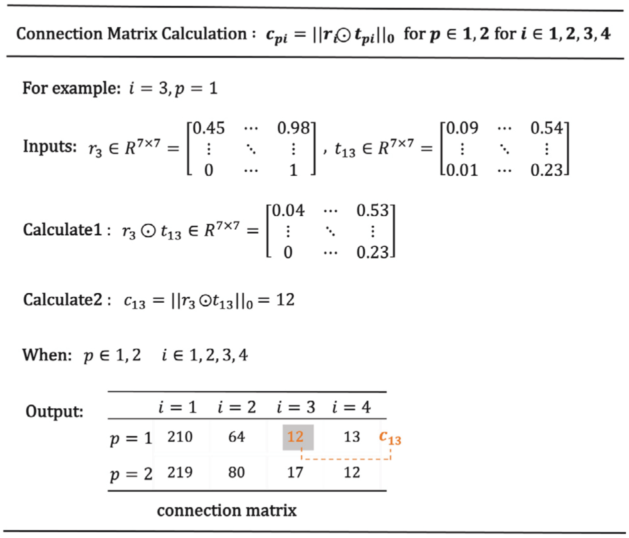


Fig. 5. Connectivity graph of example deep learned features r_3 and shape features t_{13} in Fig. 4(a). If a positive deep learned feature has a positive value for a shape feature at the same position, we indicate a connection between them. So there are 12 connections between r_3 and t_{13} in Fig. 4(a). In the connectivity graph, the x-axis represents ith block, and the y-axis indicates the connection between features r_i and t_{pi}. Each texture-image has been assigned a different colour. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



Connection Matrix Calculation : $c_{pi} = ||r_i \odot t_{pi}||_0$ **for** $p \in 1, 2$ **for** $i \in 1, 2, 3, 4$

For example: $i = 3, p = 1$

Inputs: $r_3 \in R^{7\times7} = \begin{bmatrix} 0.45 & \cdots & 0.98 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}$, $t_{13} \in R^{7\times7} = \begin{bmatrix} 0.09 & \cdots & 0.54 \\ \vdots & \ddots & \vdots \\ 0.01 & \cdots & 0.23 \end{bmatrix}$

Calculate1 : $r_3 \odot t_{13} \in R^{7\times7} = \begin{bmatrix} 0.04 & \cdots & 0.53 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0.23 \end{bmatrix}$

Calculate2 : $c_{13} = ||r_3 \odot t_{13}||_0 = 12$

When: $p \in 1, 2$ $i \in 1, 2, 3, 4$

Output:

	$i = 1$	$i = 2$	$i = 3$	$i = 4$	
$p = 1$	210	64	12	13	c_{13}
$p = 2$	219	80	17	12	

connection matrix

Fig. 6. Connection matrix calculation of the example deep learned features and shape features in Fig. 4(a). The connection matrix of size 2 × 4. Here, $p = 1$ indicates texture *homogeneity* and $p = 2$ indicates *auto-correlation*, and i means ith dense block.

mass in the ROI. The second group used yellow corresponding to the background, calcium and other characteristics in the ROI. In summary, shape-related features are learned by texture-images, which are easily ignored through the next layers in a biased model.

So far, deep learned features and shape features have been extracted from the ROIs and texture-images, it is natural to integrate these two separated features in the feature fusion block. As a result, a great deal of the spatial information inherent in shape features is lost. As the number and distribution of shape features are different in each case, we propose a connection matrix to do shape feature leaning.

2.3.2. *Shape feature learning*

Inspired by papers by Strange et al. (2014) and Chen et al. (2014), the connectivity estimation was proposed for microcalcification cluster classification. It was shown that the number and distribution of microcalcifications can be used for classification purposes.

In our study, we used the number and distribution of shape features learned from texture-images. We build the connectivity between deep learned features and shape features. To do so, we define r_i as deep learned features of the ith dense block of the ROIs, and t_{pi} as shape features of the ith dense block of the pth texture-image. Here, $p \in 1, 2$ represents the texture-images *homogeneity* and *auto-correlation*, respectively. $i \in 1, 2, 3, 4$ represents ith dense block. Features are extracted from the four dense blocks of the trained DenseNet121 to increase the robustness of the proposed method. The resulting connectivity graph corresponding to the r_3 and t_{13} in Fig. 4(a) is shown in Fig. 5 .

After generating a connectivity graph of each sample, a connection matrix inspired by Pan et al. (2021) can be extracted to capture the connectivity properties of shape features. The connection matrix will constitute the feature space for the classification of breast lesions. Here, we use $C = \{c_{pi}\}$ to represent a connection matrix. c_{pi} denotes the number of connection between features r_i and t_{pi}. The connection matrix is defined as follows:

$$c_{pi} = \| r_i \odot t_{pi} \|_0 \qquad for\, p \in 1, 2 \qquad for\, i \in 1, 2, 3, 4 \tag{5}$$

where $\odot$ represents a Hadamard product, $\|*\|_0$ means the number of non-zero coordinates. Assuming the example in Fig. 4(a), the connection matrix size is 2 × 4, representing 2 texture-images and 4 blocks as shown in Fig. 6.

2.3.3. *k-nearest neighbors*

The k-nearest neighbors (kNN), which has been successfully applied in previous studies (Papernot and McDaniel, 2018; Ortega-Martorell et al., 2022), can be used to increase the model confidence estimated by the distance between the test case and the model's training samples. For the proposed method, kNN is utilised to integrate shape features into the classifier layer to solve the biased learning problem. To achieve this, we use k-nearest neighbors to find k training samples whose connection matrices are closest to a test case and use the neighbours to further improve the CNN's diagnosis.

Using N mammogram ROIs as the training samples can be indicated as $\{X^n, C^n, Y^n\}_{n=1}^{N}$, where $X^n \in R^{224\times224}$ represents the nth ROI, $C^n \in R^{2\times4}$ represents the connection matrix corresponding to the nth ROI, and $Y^n = (1, 0)$ or $Y^n = (0, 1)$ is the corresponding label indicating a benign or malignant case.

The test case follows similar steps as the training method (see Fig. 1). Firstly, deep learned features are extracted from the test case and a connection matrix $C' = \{c'_{pi}\}$ is calculated. Then, we use k-nearest neighbors based on the Euclidean distance to find the k training samples whose connection matrices are closest to this test case. The distance between the test case and k-nearest training samples is defined as $\{d_n | n \in 1, 2, \dots, k\}$:

$$d_n = \|C' - C^n\|_F \qquad for\, n \in 1, 2, \dots, k \tag{6}$$

where $\|*\|_F$ means the Frobenius Norm. Given the test case's prediction $O(o_1, o_2)$ from the trained DenseNet121 and its k-nearest training samples, we regard the k-nearest training samples as confidence to confirm the classification of the test case.

Definition 1. Confidence indicates how likely the prediction is to be correct according to k-nearest training samples, The confidence of the test case with each label is defined as (B_w, M_w):

$$\begin{cases} B_w = \sum_{n=1}^{k} d_n \,:\, Y^n = (1, 0) \\ M_w = \sum_{n=1}^{k} d_n \,:\, Y^n = (0, 1) \end{cases} \tag{7}$$

where B_w is the distance accumulation of nearest training samples whose labels are benign, and M_w is the distance accumulation of nearest neighbours whose labels are malignant. When B_w is high, there is stronger support for the benign label assigned to the test case in the training samples. Instead, when B_w is low, it means no training samples supporting the benign label.

Definition 2. We use B_w to multiply o_1 to calibrate the prediction of the trained DenseNet121. The same as M_w, we use M_w to multiply o_2

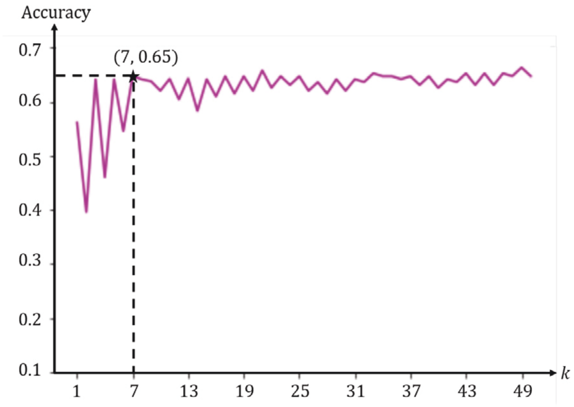


Fig. 7. Test error of kNN classifier for our connection matrix.

to calibrate the prediction of the trained DenseNet121. Considering B_w or M_w is able to be zero, the calibration of the trained DenseNet121 $O(o'_1, o'_2)$ is defined as:

$$\begin{cases} B_w \times (1,0) \times o_1 + M_w \times (0,1) \times o_2 \;:\; B_w \neq 0, M_w \neq 0 \\ (1,0) \times o_1 + M_w \times (0,1) \times o_2 \;:\; B_w = 0, M_w \neq 0 \\ B_w \times (1,0) \times o_1 + (0,1) \times o_2 \;:\; B_w \neq 0, M_w = 0 \end{cases} \tag{8}$$

if B_x or M_x is zero, we keep the corresponding trained DenseNet121's prediction. The predicted label for the test case is the one assigned the largest empirical o'-value, i.e. $\text{argmax}O(o'_1, o'_2)$.

3. Results and analysis

3.1. Implementation details

In the proposed model, the classification performance depends on the number of neighbours. In order to find the best performing value of k, we have compared results obtained with the kNN classifier to achieve the final benign or malignant classification. Fig. 7 shows the performance on the test dataset when k is in the 1–50 range. We find the accuracy curve is smoother after $k = 7$. Considering a higher value k will increase the computational complexity of the calibration, so, we choose $k = 7$ to assess the performance of the proposed model (see Fig. 7).

3.2. Comparison with state-of-the-art

Additionally, we have compared the proposed model with two related state-of-the-art interpretable deep learning models which provide modified CNNs to improve their performance. The results are listed in Table 2, where the performance of these compared methods are obtained from the results presented in their papers. All the listed algorithms achieved interpretation of test cases using k-Nearest Neighbors.

Papernot and McDaniel (2018) has achieved high performance on the MNIST data. MNIST is a dataset of simple handwritten digits, it will be easier for the model to do classification.

Ortega-Martorell et al. (2022) tried to improve the breast mass classification problem. However, they did not achieve improved accuracy compared with the initial CNN model.

The proposed model not only learns features that represent data from classical DenseNet121 but also learns shape features from texture-images, and combines these features to do the final classification to achieve improved accuracy.

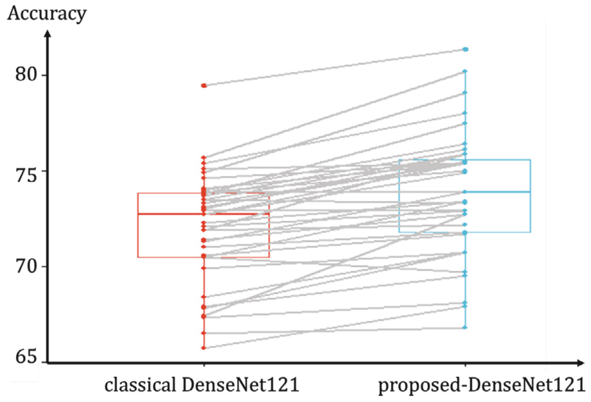


Fig. 8. Accuracy distribution analysis for classical DenseNet121 and proposed-DenseNet121, showing a box plot indicating a paired comparison.

3.3. The signification of shape feature learning

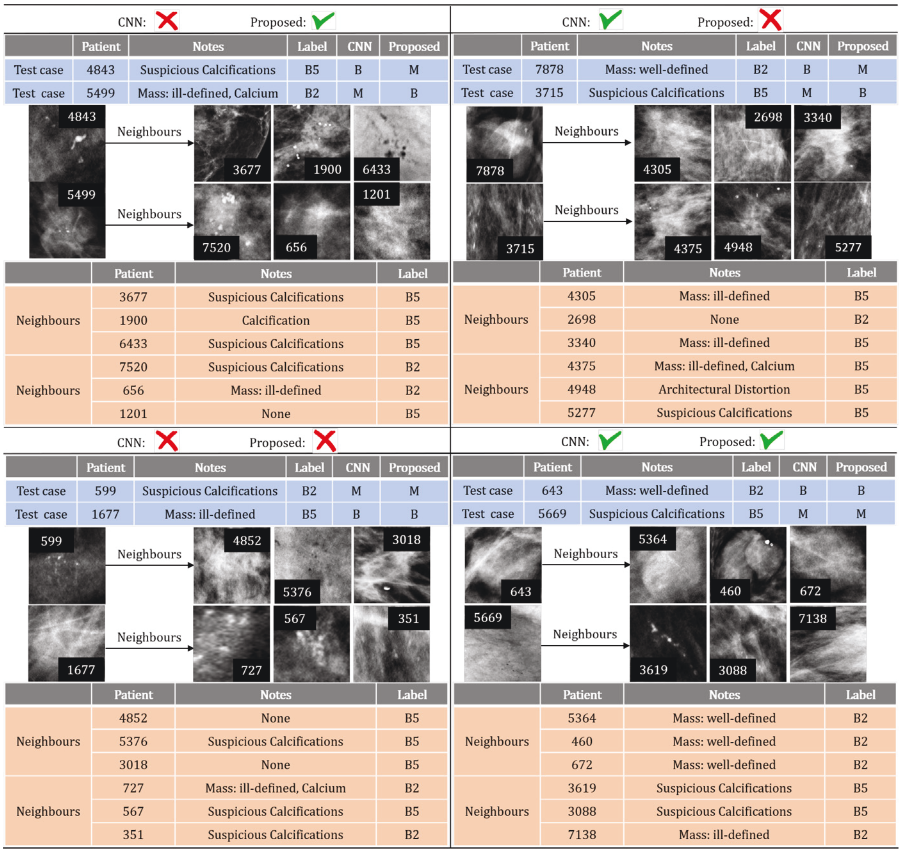
Ten-fold cross-validation was employed to assess the effectiveness of the proposed model. The average accuracy on cross validation, demonstrating the classification performance representing the proposed model. We first compare the proposed model with the classical DenseNet121. Average accuracy on cross validation in each model was estimated as 71.35 ± 2.66% for classical DenseNet121, 73.89 ± 2.89% for the proposed-DenseNet121. To further characterise performance, a t-test is used to compare the accuracy of the classical DenseNet121 and the proposed-DenseNet121, which showed a statistically significant difference ($p = 0.00036$), with further details provided in the box-plot shown in Fig. 8.

We are interested in studying the importance of the shape feature in the learning process. We replaced DenseNet121 with several well-known CNN backbone architectures, e.g. VGG16 (Simonyan and Zisserman, 2014), Resnet50 (He et al., 2016). The classification comparison between classical CNNs and the proposed-CNNs has been listed in Table 3. It can be observed that all the proposed-CNNs achieve improved accuracy compared with the classical CNNs, demonstrating the advantage of integrating shape feature learning into CNNs.

3.4. k-Nearest neighbors visualisation of the test cases

Similar to the study by Ortega-Martorell et al. (2022), which proposed that the k-nearest training samples can be utilised as explanations of the predictions. Fig. 9 shows a number of examples where the proposed-DenseNet121 and classical DenseNet121 are compared and showing (in)correct classification. It also provides three nearest training samples for the selected test cases. The neighbours can provide insight into which characteristics are used for diagnosis and why the cases were classified.

For example, (1) for both cases 4843 and 5499 – two cases that were misclassified by DenseNet121 but correctly classified by the proposed-DenseNet121 – the associated metadata shows that our method uses factors not limited to calcification, such as the shape of the mass. Especially, patient 656 has the same shape mass; (2) For four misclassified cases in our proposed-DenseNet121 (patient 7878, 3715, 599, 1677), our method is based on the trained DenseNet121 and uses positive deep learned features learned from the ROI as a benchmark to build connection matrix. Some of the less useful characteristics in the trained DenseNet121 will be retained in our model and these might be the cause of failure cases. However these misclassified cases could be used to understand the decision-making process for failure cases. For example, for case 7878: the associated metadata indicates that the trained DenseNet121 may have used incorrect information to make a decision, even if the decision is correct, and for case 1677:



CNN: ✗ Proposed: ✓

	Patient	Notes	Label	CNN	Proposed
Test case	4843	Suspicious Calcifications	B5	B	M
Test case	5499	Mass: ill-defined, Calcium	B2	M	B

	Patient	Notes	Label
Neighbours	3677	Suspicious Calcifications	B5
	1900	Calcification	B5
	6433	Suspicious Calcifications	B5
Neighbours	7520	Suspicious Calcifications	B2
	656	Mass: ill-defined	B2
	1201	None	B5

CNN: ✓ Proposed: ✗

	Patient	Notes	Label	CNN	Proposed
Test case	7878	Mass: well-defined	B2	B	M
Test case	3715	Suspicious Calcifications	B5	M	B

	Patient	Notes	Label
Neighbours	4305	Mass: ill-defined	B5
	2698	None	B2
	3340	Mass: ill-defined	B5
Neighbours	4375	Mass: ill-defined, Calcium	B5
	4948	Architectural Distortion	B5
	5277	Suspicious Calcifications	B5

CNN: ✗ Proposed: ✗

	Patient	Notes	Label	CNN	Proposed
Test case	599	Suspicious Calcifications	B2	M	M
Test case	1677	Mass: ill-defined	B5	B	B

	Patient	Notes	Label
Neighbours	4852	None	B5
	5376	Suspicious Calcifications	B5
	3018	None	B5
Neighbours	727	Mass: ill-defined, Calcium	B2
	567	Suspicious Calcifications	B5
	351	Suspicious Calcifications	B2

CNN: ✓ Proposed: ✓

	Patient	Notes	Label	CNN	Proposed
Test case	643	Mass: well-defined	B2	B	B
Test case	5669	Suspicious Calcifications	B5	M	M

	Patient	Notes	Label
Neighbours	5364	Mass: well-defined	B2
	460	Mass: well-defined	B2
	672	Mass: well-defined	B2
Neighbours	3619	Suspicious Calcifications	B5
	3088	Suspicious Calcifications	B5
	7138	Mass: ill-defined	B2

Fig. 9. Example cases to indicate the performance of the developed approach. Each type contains two test cases with different labels. Test cases are represented with added notes marked with a blue background. The three nearest neighbouring cases are visualised with added notes marked with an orange background. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Comparison of proposed model with two recent studies.

	Datasets	Accuracy		Method	
		CNNs	modified-CNNs	Explanation of test cases	Improved accuracy
Papernot and McDaniel (2018).	MNIST	99.2%	99.1%	✓	×
Ortega-Martorell et al. (2022).	CBIS-DDSM	72%	72%	✓	×
Proposed-DenseNet121.	OMI-DB	71.35%	73.89%	✓	✓

Table 3
Breast mass classification of classical CNNs and proposed-CNNs on OMI-DB.

CNNs	Accuracy	k	p_value
DenseNet121	71.35 ± 2.66%	7	0.00036
proposed-DenseNet121	73.89 ± 2.89%		
VGG16	70.3 ± 2.56%	7	0.00026
proposed-VGG16	72.9 ± 2.31%		
Resnet50	67.3 ± 3.54%	7	0.0007
proposed-Resnet50	71.1 ± 3.39%		

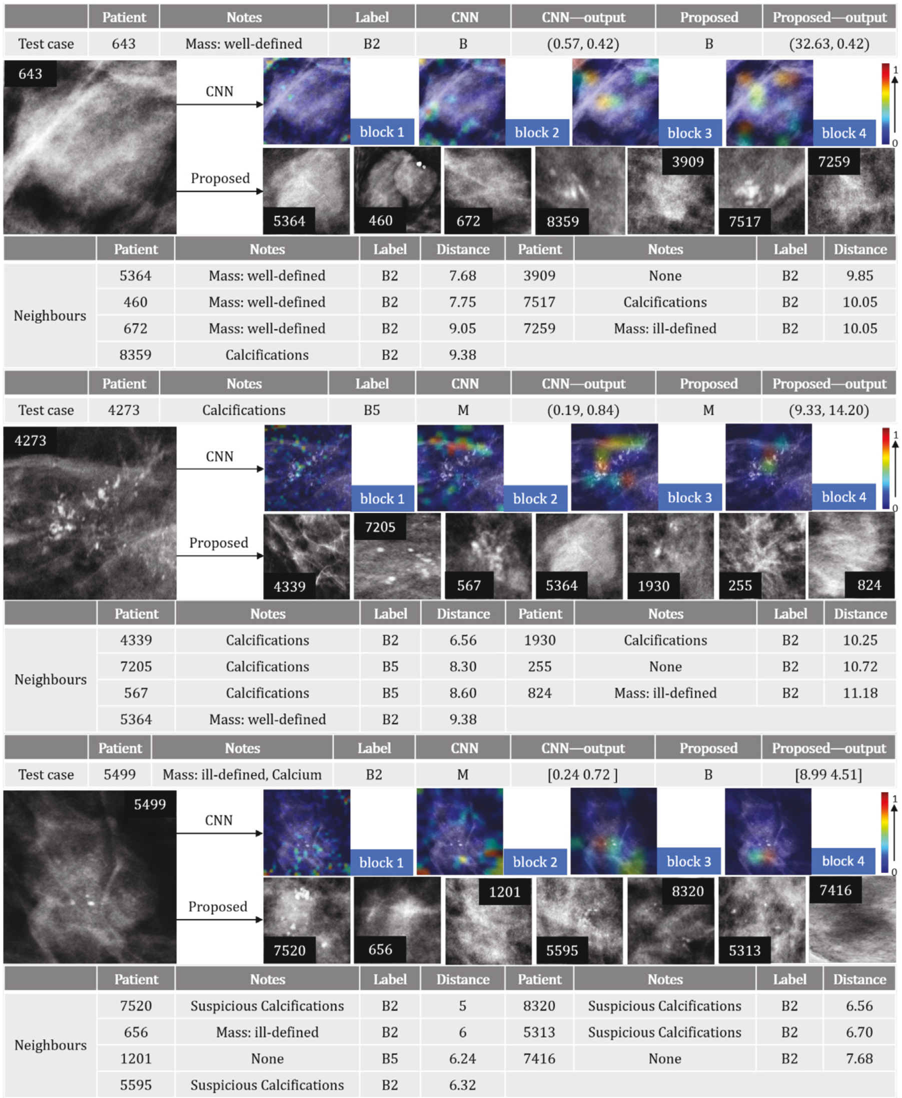
the associated metadata provides case 727 which has the same shape as mass resulting in the trained DenseNet121 making the incorrect prediction; (3) If the trained DenseNet121 uses the right characteristics to make correct predictions, cases 643 and 5669 show our model can retain the same characteristics resulting in correct predictions.

Fig. 10 compares classical DenseNet121 with the proposed-DenseNet121 for three breast lesions. To gain a better understanding of the DenseNet121 behaviour, we extracted deep learned features from four dense blocks and viewed them in colour. The seven nearest training samples of each test case are also provided to compare with the test case. Test case 643 is a breast lesion containing only a mass. The DenseNet121 makes a correct classification, however, the highlighted areas are distributed over the different blocks. The associated metadata provided by our model finds four training samples that have a mass and two samples that have calcifications. Our model not only learns features that represent a mass but also retains some features learning that characterise calcium. In case 4273, the associated metadata indicates there are some features representing a mass or other factors also activated by the DenseNet121 which are illustrated by two neighbouring cases: 5364 and 824. Test case 5499 is a breast lesion containing a mass and calcium. There are some features representing a mass which are also activated by the DenseNet121, but they have less effect on the prediction. The associated metadata shows our model used the features representing calcium and used case 656 as confidence to confirm the prediction of the DenseNet121.

4. Discussion

The focus of this work is to construct a breast cancer diagnosis model able to (1) overcome biased learning when the model is learned on a limited dataset; (2) understand the characteristics of a patient and facilitate the decision-making process in breast cancer diagnosis.

To illustrate how it works, we conduct a comparative analysis between our proposed method and some recent studies. For biased learning, Li et al. (2021) proposed extracting shape features by using



	Patient	Notes	Label	CNN	CNN—output	Proposed	Proposed—output
Test case	643	Mass: well-defined	B2	B	(0.57, 0.42)	B	(32.63, 0.42)

	Patient	Notes	Label	Distance	Patient	Notes	Label	Distance
Neighbours	5364	Mass: well-defined	B2	7.68	3909	None	B2	9.85
	460	Mass: well-defined	B2	7.75	7517	Calcifications	B2	10.05
	672	Mass: well-defined	B2	9.05	7259	Mass: ill-defined	B2	10.05
	8359	Calcifications	B2	9.38				

	Patient	Notes	Label	CNN	CNN—output	Proposed	Proposed—output
Test case	4273	Calcifications	B5	M	(0.19, 0.84)	M	(9.33, 14.20)

	Patient	Notes	Label	Distance	Patient	Notes	Label	Distance
Neighbours	4339	Calcifications	B2	6.56	1930	Calcifications	B2	10.25
	7205	Calcifications	B5	8.30	255	None	B2	10.72
	567	Calcifications	B5	8.60	824	Mass: ill-defined	B2	11.18
	5364	Mass: well-defined	B2	9.38				

	Patient	Notes	Label	CNN	CNN—output	Proposed	Proposed—output
Test case	5499	Mass: ill-defined, Calcium	B2	M	[0.24 0.72]	B	[8.99 4.51]

	Patient	Notes	Label	Distance	Patient	Notes	Label	Distance
Neighbours	7520	Suspicious Calcifications	B2	5	8320	Suspicious Calcifications	B2	6.56
	656	Mass: ill-defined	B2	6	5313	Suspicious Calcifications	B2	6.70
	1201	None	B5	6.24	7416	None	B2	7.68
	5595	Suspicious Calcifications	B2	6.32				

Fig. 10. Comparison of DenseNet121 predictions and proposed-DenseNet121 on three test cases. Deep learned features are extracted from four dense blocks and viewed in colour. Seven nearest training samples of each test case are visualised. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

binary masks and texture features from CNNs. They integrated these two separate features to achieve improved accuracy. The binary mask was produced by an automatic mask segmentation algorithm. However, some public datasets such as OMI-DB do not have binary mask labelling. It is difficult to evaluate the shape features learned from the binary mask. Furthermore, there is no assessment of whether CNNs use shape features to do diagnosis.

For interpretability, Papernot and McDaniel (2018) estimated the distance between deep learned features extracted from a test case and training samples. This estimation was used to find k-nearest training samples whose labels matched the prediction on the test case and then used training samples to explain the prediction. However, the computation is high because the distance is calculated in high-dimensional deep learned features. Ortega-Martorell et al. (2022) extracted deep learned features that come from one classifier layer and did a dimensional reduction to transform high dimensional features into a two-dimensional space (created using the training samples). Mapping a test case to the two-dimensional space, where it would provide a prediction based on its neighbouring training samples. The results demonstrated a good visualisation of the decision-making process for CNNs, but it did not achieve improved accuracy compared with the initial CNNs.

Returning to the desired properties of the proposed model: (1) Accurate information, such as the shape of mass is used for model diagnosis. Fig. 4 shows deep learned features across the different layers of the trained DenseNet, which provide an indication of the factors considered by the model for diagnosis; (2) Less computation: HiResCAM is utilised to convert each high-dimensional deep learned feature into a two-dimensional matrix. This property is beneficial to the k-nearest neighbors approach in high dimensions; (3) Improved accuracy is achieved. Table 3 compares the accuracy of CNNs and the proposed-CNNs. We not only learn features of data by CNNs but also learn the shape features, and combine them to do the final classification; (4) Decision-making understanding is achieved by finding k-nearest training samples supporting the prediction. Fig. 9 shows that the proposed model can produce a visualisation of the decision-making process in the model. For example, the associated information that comes from the nearest training samples allows us to understand how the model classification process works, such as the shape or type of lesions and the BI-RADS scores.

As shown by the performance of the proposed model, there is signification room for improvement. This is likely because of the variable morphology of masses and microcalcifications, it will be difficult for CNNs to do benign or malignant classification for a lesion. Inspired by Hamidinekoo et al. (2018a) and Ortega-Martorell et al. (2022), mass and calcium are both common and important symbols in mammography for breast cancer detection and have individual successful applications in CNNs, future work could consider dual-path CNNs to do mass classification and calcium classification separately, and then combine them to improve the model's performance.

5. Conclusion

In this study, we first investigate the unbiasedness of a CNN for breast cancer diagnosis. Shape features are learned by texture-images, which are easily ignored in CNN. In order to add these features to the underlying classifier of the CNN, we propose a connection matrix for shape feature learning and use k-nearest neighbors to find k-nearest training samples whose connection matrices are closest to the test case. These neighbours are regarded as confidence to confirm the classification of the test case. When evaluated on OMI-DB we achieved improved diagnostic accuracy $73.89 \pm 2.89\%$ compared with $71.35 \pm 2.66\%$ for the initial CNN model, which showed a statistically significant difference ($p = 0.00036$). We also show a visualisation of k-nearest training samples to better understand the characteristics of a test patient and facilitate failure analysis.

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.

Data availability

The authors do not have permission to share data.

Acknowledgements

The images and data used in this publication are derived from the OPTIMAM imaging database (Halling-Brown et al., 2020), we would like to acknowledge the OPTIMAM project team and staff at the Royal Surrey NHS Foundation Trust who developed the OPTIMAM database, and Cancer Research UK who funded the creation and maintenance of the database.

All authors approved the version of the manuscript to be published.

References

Alam, N., Oliver, A., Denton, E.R., Zwiggelaar, R., 2018. Automatic segmentation of microcalcification clusters. In: Medical Image Understanding and Analysis: 22nd Conference, MIUA 2018, Southampton, UK, July 9-11, 2018, Proceedings 22. Springer, pp. 251–261.

Arevalo, J., González, F.A., Ramos-Pollán, R., Oliveira, J.L., Lopez, M.A.G., 2015. Convolutional neural networks for mammography mass lesion classification. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. EMBC, IEEE, pp. 797–800.

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

Chen, Z., Strange, H., Oliver, A., Denton, E.R., Boggis, C., Zwiggelaar, R., 2014. Topological modeling and classification of mammographic microcalcification clusters. IEEE Trans. Biomed. Eng. 62 (4), 1203–1214.

Chowdhury, T., Bajwa, A.R., Chakraborti, T., Rittscher, J., Pal, U., 2021. 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, pp. 3–17.

Dhahbi, S., Barhoumi, W., Zagrouba, E., 2015. Breast cancer diagnosis in digitized mammograms using curvelet moments. Comput. Biol. Med. 64, 79–90.

Domingues, I., Sales, E., Cardoso, J., Pereira, W., 2012. Inbreast-database masses characterization. XXIII CBEB.

Draelos, R.L., Carin, L., 2020. Use HiResCam instead of Grad-Cam for faithful explanations of convolutional neural networks. arXiv e-prints, arXiv–2011.

Halling-Brown, M.D., Warren, L.M., Ward, D., Lewis, E., Mackenzie, A., Wallis, M.G., Wilkinson, L.S., Given-Wilson, R.M., McAvinchey, R., Young, K.C., 2020. Optimam mammography image database: a large-scale resource of mammography images and clinical data. Radiol. Artif. Intell. 3 (1), e200103.

Hamidinekoo, A., Dagdia, Z.C., Suhail, Z., Zwiggelaar, R., 2018a. 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, pp. 2423–2432.

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

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

Hamidinekoo, A., Suhail, Z., Qaiser, T., Zwiggelaar, R., 2017. 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, pp. 398–409.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

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

Jiao, Z., Gao, X., Wang, Y., Li, J., 2018. A parasitic metric learning net for breast mass classification based on mammography. Pattern Recognit. 75, 292–301.

Kenny, E.M., Ford, C., Quinn, M., Keane, M.T., 2021. Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. Artificial Intelligence 294, 103459.

Lao, J., Chen, Y., Li, Z.-C., Li, Q., Zhang, J., Liu, J., Zhai, G., 2017. A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. Sci. Rep. 7 (1), 1–8.

Li, H., Chen, D., Nailon, W.H., Davies, M.E., Laurenson, D., 2020. COIN: Contrastive identifier network for breast mass diagnosis in mammography. arXiv preprint arXiv:2012.14690.

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

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

Löfstedt, T., Brynolfsson, P., Asklund, T., Nyholm, T., Garpebring, A., 2019. Gray-level invariant haralick texture features. PLoS One 14 (2), e0212110.

Martin, K.E., Helvie, M.A., Zhou, C., Roubidoux, M.A., Bailey, J.E., Paramagul, C., Blane, C.E., Klein, K.A., Sonnad, S.S., Chan, H.-P., 2006. Mammographic density measured with quantitative computer-aided method: comparison with radiologists' estimates and BI-RADS categories. Radiology 240 (3), 656–665.

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

Mueller, S.T., Hoffman, R.R., Clancey, W., Emrey, A., Klein, G., 2019. Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. arXiv preprint arXiv:1902.01876.

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

Ortega-Martorell, S., Riley, P., Olier, I., Raidou, R.G., Casana-Eslava, R., Rea, M., Shen, L., Lisboa, P.J., Palmieri, C., 2022. Breast cancer patient characterisation and visualisation using deep learning and Fisher information networks. Sci. Rep. 12 (1), 1–14.

Pan, J., Qian, Y., Li, F., Guo, Q., 2021. Image deep clustering based on local-topology embedding. Pattern Recognit. Lett. 151, 88–94.

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

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115 (3), 211–252.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626.

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

Simonyan, K., Vedaldi, A., Zisserman, A., 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Strange, H., Chen, Z., Denton, E.R., Zwiggelaar, R., 2014. Modelling mammographic microcalcification clusters using persistent mereotopology. Pattern Recognit. Lett. 47, 157–163.

Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F., 2021. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: Cancer J. Clin. 71 (3), 209–249. http://dx.doi.org/10.3322/caac.21660.

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

Yu, H., Yang, L.T., Zhang, Q., Armstrong, D., Deen, M.J., 2021. Convolutional neural networks for medical image analysis: state-of-the-art, comparisons, improvement and perspectives. Neurocomputing 444, 92–110.

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