



# Deep residual learning with attention mechanism for breast cancer classification

Chean Khim Toa<sup>1</sup> · Mahmoud Elsayed<sup>1</sup> · Kok Swee Sim<sup>1</sup>

Accepted: 7 August 2023 / Published online: 31 August 2023

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

## Abstract

Invasive ductal carcinoma (IDC) is a common form of breast cancer that affects women. In traditional medical practice, physicians have to manually test and classify areas which are suspected to be cancerous. However, the literature strongly indicates that the manual segmentation process performed by medical practitioners is neither time efficient nor accurate, as it relies on their subjective judgment. This paper introduces a model called residual attention neural network breast cancer classification (RANN-BCC) to help medical practitioners in the cancer diagnostic process. RANN-BCC utilizes residual neural network (ResNet) as an expert-supportive method to aid medical practitioners in cancer diagnosis. The implementation of RANN-BCC can support the classification of whole slide imaging (WSI) into non-IDC and IDC without prior information about the presence of a cancerous lesion. The classification results demonstrate that the RANN-BCC model has achieved 92.45% accuracy, 0.98 recall, 0.91 precision, and 0.94 *F*-score which has outperformed other models such as CNN, AlexNet, Residual Neural Network 34 (ResNet34), and Feed-Forward Neural Network. The developed RANN-BCC model aims to help medical experts to classify IDC and non-IDC of breast cancer by learning the feature content of medical images.

**Keywords** Breast cancer · Classification · Deep learning · Invasive ductal carcinoma (IDC) · Whole slide imaging (WSI)

## 1 Introduction

In 2021, the National Center for Health Statistics estimated that approximately 1,898,160 new cases of malignancy breast cancer would be diagnosed in the same year, with approximately 608,570 cancer death projected to occur in the United States. In Malaysia, about 1 in 19 women will be diagnosed with breast cancer. According to the World Health Organization (WHO), breast cancer became the most commonly diagnosed cancer in 2021, accounting for 12% of new cancer cases worldwide each year (Siegel and Miller 2021). Among the various forms of breast cancer, invasive ductal carcinoma (IDC) is common, representing approximately 80% of breast cancer incidence upon diagnosis (Makki 2015). Early determination is crucial during the diagnosis of breast cancer because breast cancer

survival is highly influenced by the diagnosis stage of the malignancy. Early determination enables medical experts to provide appropriate treatment to the patients, thereby reducing mortality (Youlden et al. 2012; Wang 2017). An informative diagnosis of the various cancer classification is essential to aid medical professionals in selecting legitimate treatments. Technological advancements in screening tests to identify early-stage cancer cells were recommended (Wang 2017).

Mammography is the standard screening test for detecting breast cancer, but its effectiveness is limited for patients under 40 years old and those with high-density breast tissue. It is also less sensitive to tumors smaller than 1 mm, and may not provide conclusive evidence of breast cancer (Onega et al. 2016). Another screening test is called contrast-enhanced (CE) digital mammography which can deliver higher accuracy in diagnosis compared with other screening tests in high-density breasts region cases, but its availability is limited due to the high cost and the elevated levels of radiation involved in the procedure (Lewis et al. 2017). Another method for detecting breast cancer is the

✉ Kok Swee Sim  
kssim@mmu.edu.my

<sup>1</sup> Faculty of Engineering and Technology, Multimedia University, Melaka, Malaysia

use of magnetic resonance imaging (MRI) in conjunction with mammography. MRI is a medical imaging tool that can detect small-sized tumors that are difficult to visualize with mammography. However, MRI has its drawbacks. They are high cost, low specificity, injection of contrast agent, and the chance of over-diagnosis (Hua et al. 2015). In addition, a biopsy test is considered the definitive way for a confirmative and comprehensive diagnosis of breast cancer. Invasive breast cancer detection such as invasive microscopic examination is employed to identify breast cancer in a microscopic setting. Whole slide imaging (WSI) is another commonly used imaging modality in the microscopic field for investigating breast cancer. WSI provides high-resolution histopathology images that aid in visualizing cellular features and tissue structures (Cruz-Roa et al. 2014).

Currently, many medical practitioners still rely on manual identification of invasive ductal carcinoma (IDC) in the breast. However, this approach is time-consuming and operator dependent as it involves scanning a large area to identify IDCs. Moreover, prior knowledge of the abnormality presence is required by medical practitioners for manual delineation of breast cancer mass. The discrepancy in differentiated diagnosis opinions among medical experts and radiologists requires a dual reading procedure (Yap and Yap 2016). Another approach is the semi-automatic detection and classification of breast cancer abnormalities (Sim et al. 2014; Ting et al. 2017). However, it is challenging to apply common image processing techniques to locate various types of mammograms in medical images, as malignant lesions can appear at different locations and have different intensity distributions.

Recently, the use of machine learning has shown immense potential in addressing a wide range of tasks and challenges faced by the healthcare industry. Genetic programming, a subset of machine learning, is a method for automatically generating computer programs or mathematical models to solve complex problems without the need for explicit programming by humans. The uniqueness of genetic programming is the ability to evolve programs or mathematical models, allowing it to handle a wide range of problems. Recently, D'Angelo et al. (2023) introduced the use of genetic programming to develop a classifier for diabetic foot (DF). The authors proposed an explainable genetic programming classifier (X-GPC), which aims to produce a model that can provide a human-readable explanation of the diabetic foot ulcer (DFU) diagnosis. Besides from the genetic programming, the author also mentioned about evolutionary algorithms, a class of optimization algorithms inspired by the process of natural selection and evolution in biological systems. This type of algorithm used to find the optimal or near-optimal solution

to complex problems such as the data in the biomedical field (D'Angelo and Palmieri 2020).

Deep learning, a subset of machine learning, has emerged as a groundbreaking approach that can mimic the workings of the human brain's neural networks. This technique enables an end to end learning, where the model learns all the steps between the initial input phase and the final output result. It automatically learns and extracts patterns and representations from complex medical data. One of the most significant applications of deep learning in the healthcare industry is medical imaging analysis. Traditional diagnostic methods often rely on human expertise to interpret information in the images and are subject to human error. Algorithms of deep learning, on the other hand, can automatically learn to interpret that information, enabling faster and more efficient diagnoses (Araújo et al. 2017).

Hence, this paper aims to apply deep learning methods for non-IDC and IDC classification. Deep learning models are well suited for processing medical imaging due to the availability of a large number of sample images for training. The proposed model, residual attention neural network breast cancer classification (RANN-BCC), aims to assist medical practitioners in investigating medical images of breast cancer quickly and effectively. RANN-BCC utilizes a residual neural network (ResNet) as a supportive tool to classify breast cancer lesions, thereby reducing the time required for breast cancer diagnosis.

To evaluate the performance of the RANN-BCC model, a classification was conducted using a dataset of non-IDC and IDC images, and the results were compared with other deep learning models. The paper is organized as follows. A review of related work is shown in Sect. 2. The structure of the RANN-BCC model is explained in Sect. 3. The results and discussion of the RANN-BCC and other deep learning models are presented and discussed in Sect. 4. Finally, the study is summarized in Sect. 5.

## 2 Related works

### 2.1 Whole slide images

Whole slide imaging (WSI) is a technology that produces digital images by scanning and digitization of entire glass (histology) slides. WSI is considered as a digital file that is comparable to the glass slides under a microscope. WSI is increasingly being used by pathology departments, scientists, and pathologists for educational, clinical, and research activities (Hanna et al. 2020). A trained and experienced histopathologist can make accurate diagnoses of biopsy specimens based on WSI data. However, with the different dimensions of WSIs and the increasing number of cancer

cases, the analysis of WSIs will be time-consuming and even difficult if there is a lack of histopathologists (Khened et al. 2021). Figure 1 shows the typical workflow of digital pathology research, where several image analysis techniques are used to perform segmentation, detection, and classification.

In the past, most of the research methods involved histological primitives' segmentation and handcrafted feature extraction that describe the arrangement and appearance of these primitives to distinguish malignant from benign areas. Petushi et al. (2006) introduced the tissue microtexture classification to segmentate nuclei and extract two features which are spatial position and surface density of nuclei. Dundar et al. (2011) presented a computerized classification of intraductal breast lesions that can distinguish between actionable subtypes and ductal hyperplasia. Niwas et al. performed the breast lesions classification using log-Gabor complex wavelet bases which could evaluate the color texture features of the segmented nucleus. Those previous methods involved manual handcrafted features to extract the feature contents of patches divided from WSI. Those methods not only involved numerous preprocessing steps but also the classification accuracy was dependent on the accuracy of the previous step. In recent years, deep learning had provided a state-of-the-art result in various image analyses. Deep learning does not require the use of a handcrafted feature, instead, it will automatically learn the feature content of patches divided from WSI. With the rapid adoption of deep learning in imaging, the wider accessibility of WSIs now attracts the application of deep learning.

## 2.2 Deep learning in image classification

The deep learning model was useful in the development of medical research and currently received a lot of attention due to its superiors' classification of a large set of training data. These deep learning models showed outstanding capability in mimicking humans, including in the field of medical imaging (Tan et al. 2017; Ting and Sim 2017).

Among different types of deep learning models, convolutional neural network (CNN) is commonly used in classifying the image. CNN consists of several layers of neural computer connections that can greatly improve the field of computer vision with minimal systematic processing. The architecture of CNN consists of several parts such as the convolutional layer, pooling layer, and fully connected layer. A convolutional layer will learn the feature representation of the image by detecting line, edge, and other pattern forms. For computing different feature maps, several kernels will be applied to the image and get the convoluted features. Those features will then be passed to the pooling layer which is used to reduce the computational burden by decreasing the feature map resolution. After that, those features will be flattened and fed into the fully connected layer to classify them into various classes. CNN can learn a hierarchical representation of a model, from low-level to high-level functions, and extract the most important functions of a specific model (Krizhevsky et al. 2012). Since deep CNN architectures usually involve numerous layers in a neural network, with potentially millions of weight parameters to be estimated, a large number of samples are required to form the model and set the parameters. This suggests that deep learning models are suitable for handling medical imaging since a large number of medical sample images are available to perform training. Recently, the deep learning-based system was suggested by a researcher on the application such as lung cancer (Hua et al. 2015; Kumar et al. 2015), breast cancer classification (Wang et al. 2016; Ting et al. 2019), cognitive classification (Toa et al. 2021), Alzheimer's disease (AD) (Ji et al. 2019; Suk et al. 2014), and even pain quantification (Elsayed et al. 2020). Moreover, recent studies mention that deeply learned features can provide a more effective feature-learning technique for image classification as compared to handcrafted features (Toa et al. 2021; Arevalo et al. 2016). Cruz-Roa et al. provided automatic detection of IDC in WSI using CNN. The authors mentioned that the use of the deep learning method yielded a better result in the detection of IDC as compared to an approach using

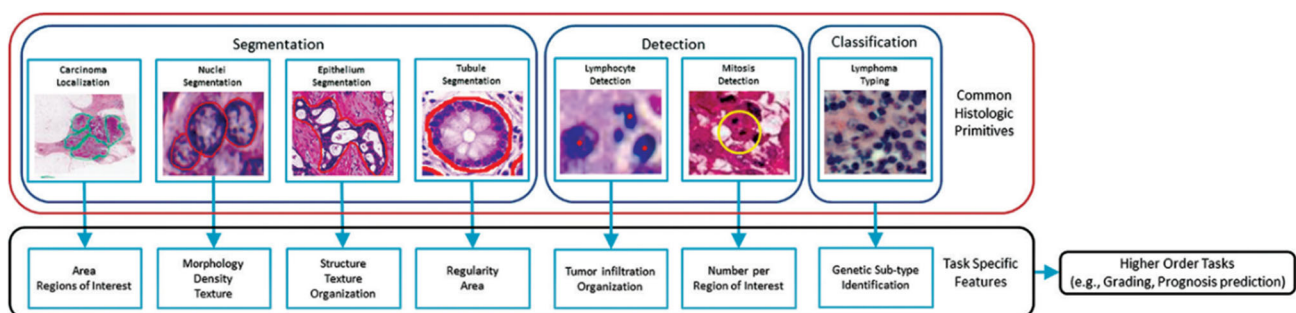


Fig. 1 The workflow of pathology research for segmentation, detection, and classification (Janowczyk and Madabhushi 2016)

handcrafted features (Cruz-Roa et al. 2014). Janowczyk and Madabhushi performed the analysis on the digital pathology image. The authors used the deep learning method to produce results superior to the handcrafted feature-based classification approach (Janowczyk and Madabhushi 2016).

### 3 Materials and methods

#### 3.1 Materials

Invasive ductal carcinoma (IDC) is a common subtype of breast cancer. The applied digital databases are made publicly available and were collected in a previous study (Cruz-Roa et al. 2014; Janowczyk and Madabhushi 2016). Figure 2 shows the non-invasive ductal carcinoma (non-IDC) and invasive ductal carcinoma (IDC) in whole slide imaging (WSI). The dataset consisted of 162 WSI breast cancer specimens scanned at  $40\times$ . From these WSI, 277,524 patches of size  $(50 \times 50)$  were extracted and converted into Portable Network Graphics (PNG) format with 198,738 non-IDC (0) patches and 78,786 IDC (1) patches. The filename of each patch includes the x-and y-coordinates of the cropped patch location and its category (0 or 1).

#### 3.2 Methods

To achieve the aim of identifying and classifying breast cancerous lesions, we designed a sophisticated neural network architecture named residual attention neural network breast cancer classification (RANN-BCC). It consists

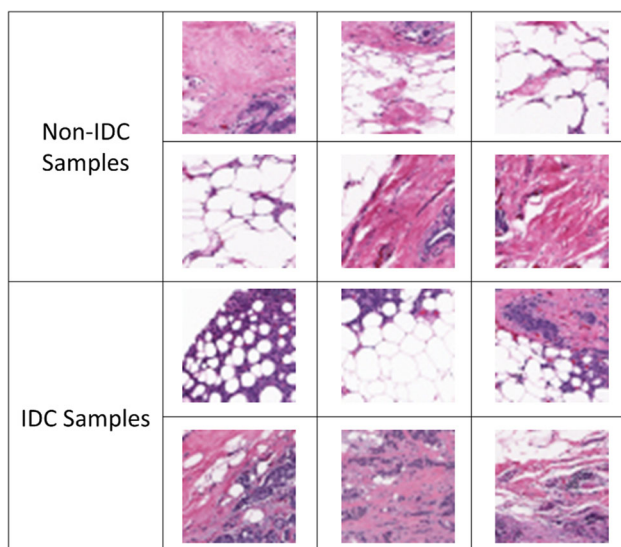


Fig. 2 Examples of non-IDC and IDC in whole slide imaging (WSI)

of six different building blocks. These six building blocks have utilized many deep learning conceptions such as residual learning, attention mechanism, convolution, and deconvolution. Figure 3 is the overall design of the architecture. The subsections below will explain each building block individually.

##### 3.2.1 Block 1: feature extractor

This block includes a residual neural network 34 (ResNet34) to map significant features of breast cancer images to feature maps (He et al. 2016). ResNet34 is an architecture that is used to solve vanishing gradient problems when constructing more layers. Figure 4 shows the ResNet34 architecture. The parameter used is shown in Table 1.

From Fig. 4, the residual connections shown between layers are important in solving many deep learning problems because they allow gradients to flow directly through the network without going through non-linear activation functions, which solves common neural network training issues such as vanishing gradients. In other words, as shown in Fig. 5, the residual connections link the previous layer output to the new layer.

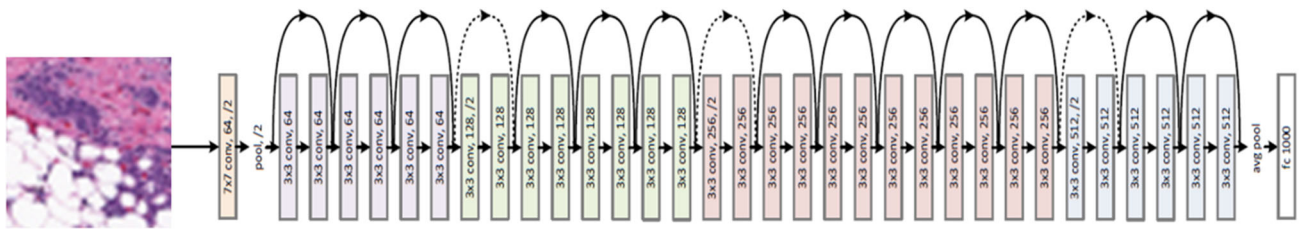
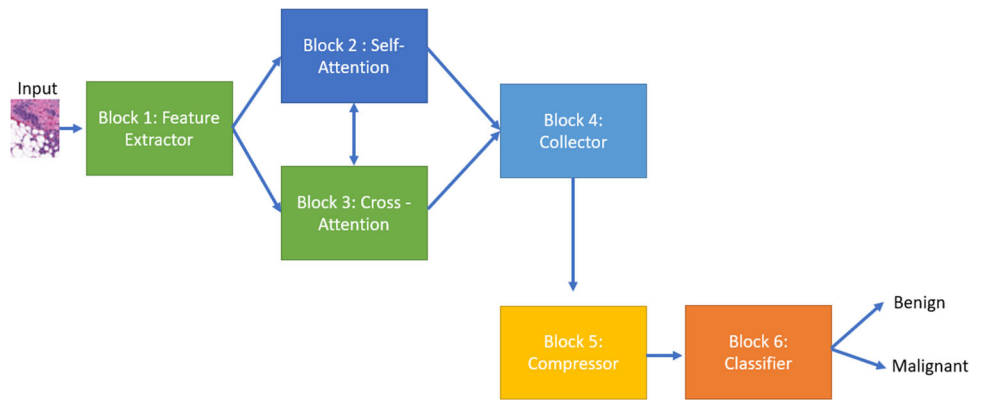
As aforementioned, each inputted image to this building block will result in the creation of 512 maps. Each of which carries some important features that would help the classifier in identifying the cancerous tumor. Figure 6 demonstrates what the feature maps could look like.

##### 3.2.2 Block 2: self-attention block

The input to this building block is the features extracted from the input image using a Residual Network (He et al. 2016), where the average pooling and classification layers (last two layers) of a ResNet34 (He et al. 2016) will be removed to obtain features of shape  $k \times k \times d$ , where  $k$  is spatial size and  $d$  is number of dimensions. We then apply an adaptive average pooling layer. We denote the generated features as  $F$ .

Self-attention was proposed by (Vaswani et al. 2017), and it was later implemented as an attention mechanism (Bahdanau et al. 2015) on its input. It is mainly utilized in our system to extract relationships from the images. The definition of the attention components and mathematical formulation will be presented here. The self-attention mechanism projects its input using three projections into a key ( $K$ ), query ( $Q$ ), and value ( $V$ ). It then performs a dot-product operation to find the similarity between the query and the key, and then generates attention weights which signify the importance of each query with all the keys. It then multiplies these attention weights with the projected

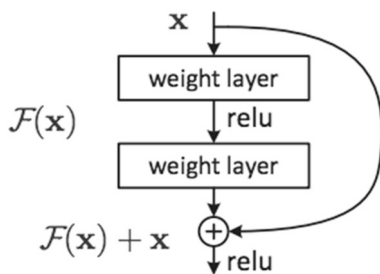
**Fig. 3** An overview of the residual attention neural network breast cancer classification (RANN-BCC) architecture



**Fig. 4** The ResNet34 architecture

**Table 1** Parameters of ResNet34

Layer name	Output size	ResNet34-layer
conv1	$350 \times 350 \times 64$	$7 \times 7, 64, \text{stride } 2, \text{padding } 3$
pool1	$175 \times 175 \times 64$	$3 \times 3, \text{max pool, stride } 2, \text{padding } 1$
conv2_x	$175 \times 175 \times 64$	$\begin{bmatrix} 3 \times 3, & 64 \\ 3 \times 3, & 64 \end{bmatrix} \times 3$
conv3_x	$88 \times 88 \times 128$	$\begin{bmatrix} 3 \times 3, & 128 \\ 3 \times 3, & 128 \end{bmatrix} \times 4$
conv4_x	$44 \times 44 \times 256$	$\begin{bmatrix} 3 \times 3, & 256 \\ 3 \times 3, & 256 \end{bmatrix} \times 6$
conv5_x	$22 \times 22 \times 512$	$\begin{bmatrix} 3 \times 3, & 512 \\ 3 \times 3, & 512 \end{bmatrix} \times 3$



**Fig. 5** Residual learning building block (He et al. 2016)

value and sums the vectors to get a representation of each query contextualized with all its important values.

$$Q = W_q \hat{Q}, K = W_K \hat{K}, V = W_v \hat{V}. \tag{1}$$

The self-attention is defined as a function of the similarity between the  $Q$  and the  $K$ , normalized with the softmax function to generate probability values that sum to one, and it is mathematically defined as shown in Eq. 2:

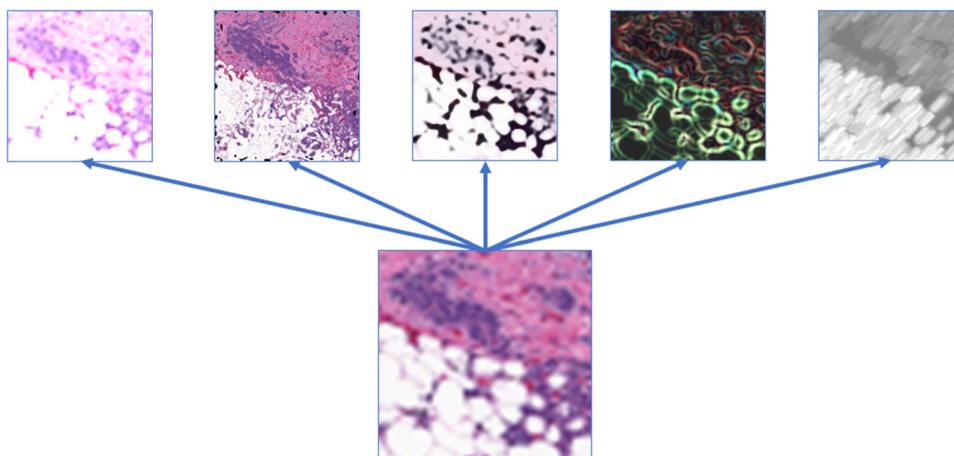
$$A = \text{Attention}(Q, K, V) = \text{softmax}(QK^T)V. \tag{2}$$

The self-attention mechanism output described in Eq. 2 is then fed to a final linear layer as shown in Eq. 3:

$$O = W_o A + b_o. \tag{3}$$

To improve the attention performance, it will be modeled as a “multi-head” and then concatenate the outputs of each head, as in Eqs. 4 and 5:

**Fig. 6** A demonstration of the feature maps created by the feature extractor building block



$$A_i = \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{4}$$

$$O = \text{Concatenate}(A_i, \dots, A_h)W_o + b_o, \tag{5}$$

where  $O$  is the output,  $h$  is the number of heads and  $d_k$  is each head dimensionality, which is computed as  $d_{model}/\text{number of heads}$ .

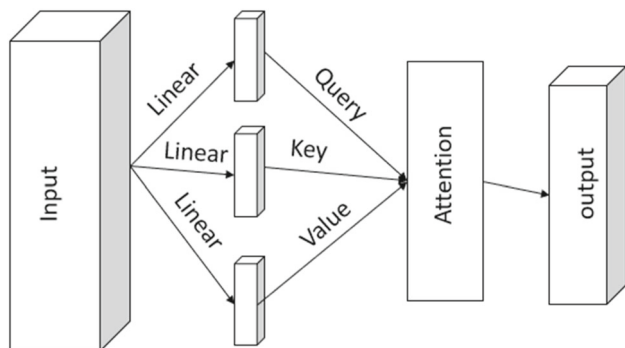
In the system, as shown in Eq. 6, the input to the self-attention block is the features extracted from the feature extractor block, denoted as  $F$ . The  $Q$ ,  $K$ , and  $V$  are projected using three separate linear layers, followed by the attention mechanism:

$$S = \text{Attention}(W_{qF}F, W_{kF}F, W_{vF}F). \tag{6}$$

It is important to mention here that due to how the self-attention mechanism operates, applying self-attention to the visual features is equivalent to exploring visual relationships between the visual elements. Figure 7 shows the architecture of the self-attention block.

### 3.2.3 Block 3: cross-attention block

The only difference between this block and the self-attention block is that the  $Q$  is a projection of the model’s input,



**Fig. 7** Architecture of self-attention block

and the  $K$  and  $V$  are projections of different features. Here, the input will query the other features rather than querying itself. In our system, the query is the output ( $O$ ) of the self-attention layer, and the keys and values are the features extracted from the feature extractor building block. Then, the first cross-attention layer output will be fed as  $K$  and  $V$  to a second cross-attention layer, where the  $Q$  are features extracted from the first CNN. The purpose of adding this block is to cross-reference and confirm the weights that project the importance of the features resulting from performing the self-attention on the output of the feature extractor building block.

Note that block 1 and block 2 include a residual (He et al. 2016) and layer normalization (Ba et al. 2016) layer at the output. At the end of each block, a positional-wise feed-forward network composed of two linear layers with a ReLU activation function in between is included to add non-linearity to our network, also with residual and layer normalization layers at the output. The input to the second layer self-attention is the output of the first layer cross-attention and the input to the second layer cross-attention is the output of the second layer self-attention, and so on.

### 3.2.4 Block 4: collector

This block and the next were partly inspired by squeeze-and-excitation networks (SENet) (Hu et al. 2020) which was originally designed for image recognition. The collector-building block was mainly added to our system to filter the feature maps before going into the classification stage. Adding this block into our classification system, it provides us with an effective and learnable approach to replace image processing filtering techniques. It is important to note that SENets modifies the equal weighting of the feature maps by adding a content-aware mechanism that adaptively weights each channel. This is different from what CNN does which is to weight all the feature maps

equally. Figure 8 shows the inner architecture of the collector and the compressor building blocks. The only reason this block is separated here from the next is to emphasize the different two objectives namely filtering and dimension reduction.

### 3.2.5 Block 5: compressor

This building block is mainly added to our classification system to reduce the dimensionality with maintaining the important features that have been extracted in the previous building blocks. This dimension reduction step is planned to enhance the efficiency and accuracy of the classifier building block. Figure 8 shows the architecture of blocks 4 and 5 combined. They can be considered as one block, the only reason we divided them into two here is to highlight the two objectives that they both are designed to filter the feature maps and reduce their dimensions before entering the classifier.

### 3.2.6 Block 6: classifier

As aforementioned, our system consists of six building blocks where block 4 and block 5 can be combined. The output of the compressor building block (block 5) is then fed to the classifier building block. This output is then run through a classification layer with two output classes: (0) non-IDC and (1) IDC. We implement the cross-entropy loss to optimize our network. The cross-entropy loss is given as shown in Eq. 7:

$$CE = -\frac{1}{n} \sum_{j=1}^n \sum_{i=1}^c y_i \log \hat{y}_i, \tag{7}$$

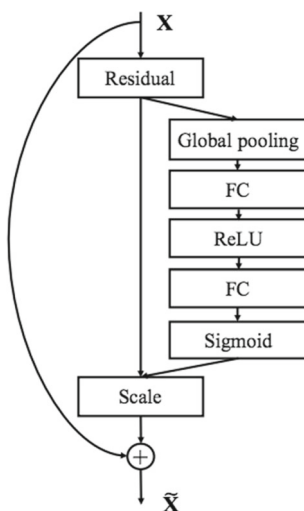


Fig. 8 Combined architecture of the collector and the compressor building block

where  $y_i$  is the class label which is either 0 or 1,  $\hat{y}_i$  is the predicted probability of the class,  $c$  is the number of classes (2 in our case) and finally  $n$  is the number of samples in the batch. The complete network is optimized with the Adam optimizer (Kingma and Ba 2015) with a batch size of 15. We set an initial learning rate of  $2e-4$  and it is then reduced by a factor of 0.8 every 3 epochs. The model is trained for 25 epochs with early stopping, which is a state-of-the-art approach for monitoring the training model performance and stopping training once the model performance begins to degrade. The first layer is an adaptive average pooling, followed by a convolutional layer, and finally, a sigmoid is applied to facilitate the classification process. Figure 9 shows the architecture of the classifier building block.

## 4 Results and discussion

The experiment results from our proposed residual attention neural network breast cancer classification (RANN-BCC) model are provided. The results will be compared with existing methods used in the classification of non-invasive ductal carcinoma (non-IDC) and invasive ductal carcinoma (IDC). The first method is the convolutional neural network (CNN). Cruz-Roa et al. proposed the use of CNN to perform the automatic detection of IDC (Cruz-Roa et al. 2014). The model adopts 3 layers of CNN architecture which employs 16 feature maps for the first layer, 32 feature maps for the second layer, and 7200 features flattened for a fully connected layer. A kernel size of  $8 \times 8$  was used in the convolutional layer and  $2 \times 2$  was used in the pooling layer. The second method is the AlexNet network used by Janowczyk and Madabhushi on digital

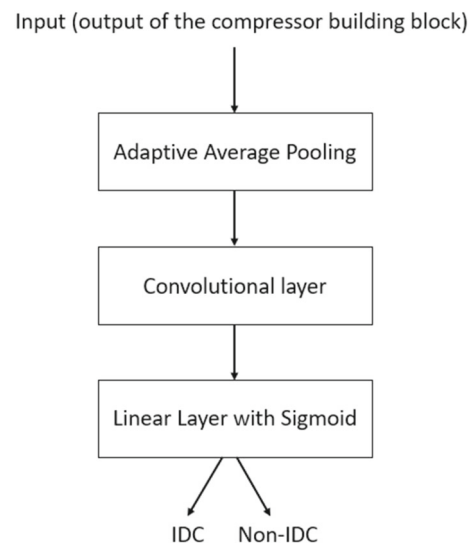


Fig. 9 The architecture of the classification building block

pathology image classification (Janowczyk and Madabhushi 2016). The AlexNet model consists of 3 convolutional layers and 1 fully connected layer. The 1st and 2nd convolutional layers consist of 32 feature maps, the 3rd convolutional layer consists of 64 feature maps, and the fully connected layer consists of 1024 flattened features. A kernel size of  $5 \times 5$  was used in the convolutional layer and  $3 \times 3$  was used in the pooling layer. Moreover, to make the result more significant, other baseline models such as feed-forward neural network and ResNet34 will be compared with our model. Feed-forward neural network is a type of artificial neural network where features were performed in a single direction, starting from input nodes, moving through the hidden nodes, and towards output nodes. This neural network consists of 4 layers with 2500 input dimensions, 100 hidden dimensions, and 2 output dimensions. The residual neural network 34 (ResNet34) model is an architecture that has 34 layers deep. The model introduced the use of the residual network to solve the problem of vanishing gradient when constructing more layers. ResNet34 model consists of 6 layers with 64 features maps in 1st and 2nd layers, 128 features maps in 3rd layers, 256 features maps in 4th layers, 512 features maps in 5th layers, and 25,088 flattened features in the fully connected layer.

All the deep learning models will be compared using 4 classification metrics which are accuracy, recall, precision, and *F*-score as shown in Eqs. 8–11.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \times 100\%, \quad (8)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (9)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (10)$$

$$F\text{-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (11)$$

where TP is a true positive in which the model correctly predicts the IDC class, TN is a true negative in which the model correctly predicts the non-IDC class, FN is a false negative in which the model incorrectly predicts the actual IDC class, and FP is false positive in which the model incorrectly predicts the non-IDC class. Tables 2, 3, 4, 5

show the result of classification metrics for deep learning models.

For the model accuracy in classification, as shown in Table 2, the RANN-BCC model can obtain the highest accuracy of 92.45%. It is then followed by AlexNet (90.28%), CNN (89.56%), ResNet34 (79.49%), and feed-forward neural network (71.18%). The accuracy of Resnet34 is lower than that of CNN and AlexNet. Through the introduction of other mechanisms, such as self-attention, cross-attention, collector, and compressor combined with ResNet34, the RANN-BCC model is designed, and its accuracy can achieve 92.45%. This shows that by introducing the use of other mechanisms, we have improved the accuracy from 79.4 to 92.45%, an increment of 13.05%.

In the recall metric, as shown in Table 3, all models have high accuracy within an error margin of 0.05. This indicates that all models have a low rate of incorrect predictions of the actual IDC class. For the precision metric, RANN-BCC has achieved the highest value of 0.91, followed by CNN and AlexNet with a value of 0.87, ResNet34 with a value of 0.76, and feed-forward neural network with a value of 0.71. It shows that RANN-BCC has a lower rate of incorrect predictions of actual non-IDC class as compared to other models. Apart from that, the feed-forward neural network has the lowest value of precision, indicating that the model has the highest rate of incorrect prediction of actual non-IDC class. For ResNet34, although it has the highest recall of 1, it has a lower precision of 0.76, indicating a low rate of incorrect predictions for actual IDC class but a high rate of incorrect predictions for actual non-IDC class. Thus, the model is biased toward the actual IDC class.

As for the *F*-score, as shown in Table 5, it is used to calculate the harmonic mean between precision and recall. Since RANN-BCC has high precision and recall rate, it is undoubtedly having the highest value of 0.94, followed by CNN and AlexNet with the same values of 0.92, ResNet34 with a value of 0.86, and feed-forward neural network with a value of 0.81. The RANN-BCC with the highest *F*-score value indicates that the model has low false positives and low false negatives. Based on the result of classification metrics for all deep learning models, RANN-BCC shows the best performance since the model is able to achieve

**Table 2** Result of our model compared to other models in terms of accuracy

Model	Accuracy (%)
Convolutional neural network (CNN) (Cruz-Roa et al. 2014)	89.56
AlexNet (Janowczyk and Madabhushi 2016)	90.28
Feed-forward neural network	71.18
Residual neural network 34 (ResNet34)	79.49
Residual attention neural network breast cancer classification (RANN-BCC)	92.45



**Table 3** Result of our model compared to other models in terms of recall

Model	Recall
Convolutional neural network (CNN) (Cruz-Roa et al. 2014)	0.98
AlexNet (Janowczyk and Madabhushi 2016)	0.98
Feed-forward neural network	0.95
Residual neural network 34 (ResNet34)	1
Residual attention neural network breast cancer classification (RANN-BCC)	0.98

**Table 4** Result of our model compared to other models in terms of precision

Model	Precision
Convolutional neural network (CNN) (Cruz-Roa et al. 2014)	0.87
AlexNet (Janowczyk and Madabhushi 2016)	0.87
Feed-forward neural network	0.71
Residual neural network 34 (ResNet34)	0.76
Residual attention neural network breast cancer classification (RANN-BCC)	0.91

**Table 5** Result of our model compared to other models in terms of *F*-score

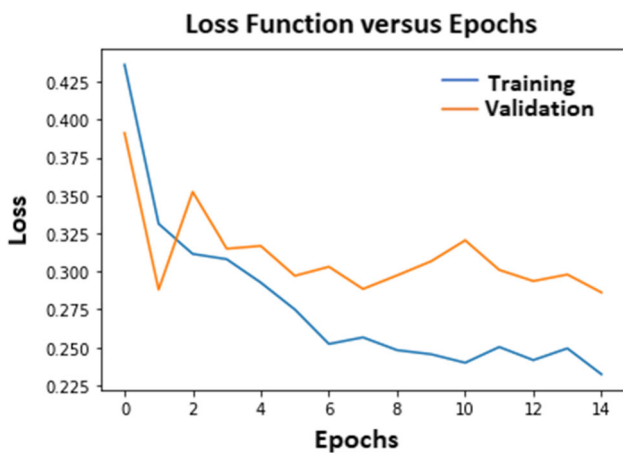
Model	<i>F</i> -score
Convolutional neural network (CNN) (Cruz-Roa et al. 2014)	0.92
AlexNet (Janowczyk and Madabhushi 2016)	0.92
Feed-forward neural network	0.81
Residual neural network 34 (ResNet34)	0.86
Residual attention neural network breast cancer classification (RANN-BCC)	0.94

higher accuracy, recall, precision, and *F*-score when classifying the IDC and non-IDC class of breast cancer.

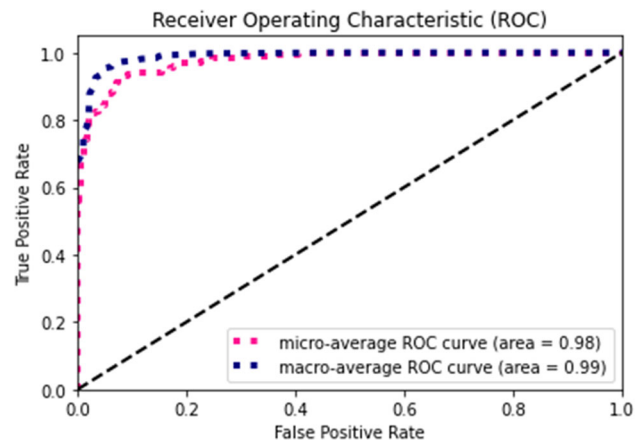
To show that RANN-BCC has a good generalization capability, we have plotted the curve for the loss function and receiver operating characteristic (ROC). The loss function is a method to evaluate how well the model performs on the dataset. Figure 10 shows the plotted graph for the validation loss and training loss of the RANN-BCC model. From the graph, we can see that the training line (blue) and the validation line (orange) are close to each other in exponential decay. This shows that the model has

good generalization capability and it is not overfitting to the breast cancer dataset.

The next one is the ROC, which is a useful method to measure how well the model can distinguish between the IDC class and the non-IDC class. The area under the curve (AUC) is a measurement tool used to measure the area underneath the ROC curve with a score from 0 to 1. The higher the AUC score, the better the model is at predicting the IDC class and non-IDC class. Figure 11 shows the plotted curve for the ROC curve of the RANN-BCC model. We can see that there are 2 types of curves which are



**Fig. 10** Loss graph of training and validation process



**Fig. 11** Receiver operating characteristic (ROC) curve

micro-average and macro-average. Micro-average is a summation of the TP, FP, and FN of the model, while macro-average takes the average of the precision and recall of the model. From the curve, we can see that the AUC score of the micro-average and the macro-average is equal to 0.98 and 0.99, respectively, which is approximately 1. This indicates that the RANN-BCC model has a good generalization capability to distinguish between the IDC class and the non-IDC class.

## 5 Conclusion

In this paper, we introduced and designed the residual attention neural network breast cancer classification (RANN-BCC) model to classify the given breast cancer dataset into invasive ductal carcinoma (IDC) and non-invasive ductal carcinoma (non-IDC). We demonstrated that our model had outperformed other deep learning models and showed the significance of each block of the RANN-BCC model. We found that the accuracy could be improved from 79.49 to 92.45% through the implementation of Residual Neural Network 34 (ResNet34) integrated with self-attention, cross-attention, collector, and compressor. We believe this integrative developed deep learning approach will not only help medical practitioners to classify IDC and non-IDC of breast cancer by learning the feature content of medical images but also will contribute to the field of computer-aided diagnostics by inspiring more similar and effective deep learning approaches.

**Funding** This research received no external funding.

**Data availability** The data can be obtained from the following link: <https://www.kaggle.com/paultimothymooney/breast-histopathology-images>.

**Code availability** The source code of the current work is available from the corresponding author on reasonable request.

## Declarations

**Conflict of interest** The authors of this work declare no conflict of interest.

**Consent to participate** Not applicable.

## References

Araújo T, Aresta G, Castro E, Rouco J, Aguiar P, Eloy C, Polónia A, Campilho A (2017) Classification of breast cancer histology images using Convolutional Neural Networks. *PLoS ONE* 12(6):e0177544. <https://doi.org/10.1371/journal.pone.0177544>

Arevalo J, González FA, Ramos-Pollán R, Oliveira JL, Guevara Lopez MA (2016) Representation learning for mammography mass lesion classification with convolutional neural networks. *Comput Methods Programs Biomed* 127:248–257. <https://doi.org/10.1016/j.cmpb.2015.12.014>

Ba JL, Kiros JR, Hinton GE (2016) Layer Normalization. <http://arxiv.org/abs/1607.06450>

Bahdanau D, Cho KH, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, pp. 1–15

Cruz-Roa A, Basavanthally A, González F, Gilmore H, Feldman M, Ganesan S, Shih N, Tomaszewski J, Madabhushi A (2014) Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. *Med Imaging Digit Pathol* 9041(216):904103. <https://doi.org/10.1117/12.2043872>

D'Angelo G, Palmieri F (2020) Discovering genomic patterns in SARS-CoV-2 variants. *Int J Intell Syst* 35(11):1680–1698. <https://doi.org/10.1002/int.22268>

D'Angelo G, Della-Morte D, Pastore D, Donadel G, De Stefano A, Palmieri F (2023) Identifying patterns in multiple biomarkers to diagnose diabetic foot using an explainable genetic programming-based approach. *Future Gener Comput Syst* 140:138–150. <https://doi.org/10.1016/j.future.2022.10.019>

Dundar MM, Badve S, Bilgin G, Raykar V, Jain R, Sertel O, Gurcan MN (2011) Computerized classification of intraductal breast lesions using histopathological images. *IEEE Trans Biomed Eng* 58(7):1977–1984. <https://doi.org/10.1109/TBME.2011.2110648>

Elsayed M, Sim KS, Tan SC (2020) A novel approach to objectively quantify the subjective perception of pain through electroencephalogram signal analysis. *IEEE Access* 8:199920–199930. <https://doi.org/10.1109/access.2020.3032153>

Hanna MG, Parwani A, Sirintrapun SJ (2020) Whole slide imaging: technology and applications. *Adv Anat Pathol* 27(4):251–259. <https://doi.org/10.1097/PAP.0000000000000273>

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. *IEEE Conf Comput vis Pattern Recognit (CVPR)* 2016:770–778. <https://doi.org/10.1109/CVPR.2016.90>

Hu J, Shen L, Albanie S, Sun G, Wu E (2020) Squeeze-and-excitation networks. *IEEE Trans Pattern Anal Mach Intell* 42(8):2011–2023. <https://doi.org/10.1109/TPAMI.2019.2913372>

Hua KL, Hsu CH, Hidayati SC, Cheng WH, Chen YJ (2015a) Computer-aided classification of lung nodules on computed tomography images via deep learning technique. *Oncotargets Ther* 8:2015–2022. <https://doi.org/10.2147/OTT.S80733>

Janowczyk A, Madabhushi A (2016) Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. *J Pathol Inform* 7(1):29. <https://doi.org/10.4103/2153-3539.186902>

Ji H, Liu Z, Yan WQ, Klette R (2019) Early diagnosis of alzheimer's disease using deep learning. In: Proceedings of the 2nd International Conference on Control and Computer Vision, pp. 87–91. <https://doi.org/10.1145/3341016.3341024>

Khened M, Kori A, Rajkumar H, Krishnamurthi G, Srinivasan B (2021) A generalized deep learning framework for whole-slide image segmentation and analysis. *Sci Rep* 11(1):11579. <https://doi.org/10.1038/s41598-021-90444-8>

Kingma DP, Ba JL (2015) Adam: a method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, pp. 1–15

Krizhevsky A, Sutskever I, Hinton G (2012) ImageNet Classification with Deep Convolutional Neural Networks. *Neural Information Processing Systems*. <https://doi.org/10.1145/3065386>

Kumar D, Wong A, Clausi DA (2015) Lung nodule classification using deep features in CT images. In: Proceedings -2015 12th

- Conference on Computer and Robot Vision, CRV 2015, pp. 133–138. <https://doi.org/10.1109/CRV.2015.25>
- Lewis TC, Pizzitola VJ, Giurescu ME, Eversman WG, Lorans R, Robinson KA, Patel BK (2017) Contrast-enhanced digital mammography: a single-institution experience of the first 208 cases. *Breast J* 23(1):67–76. <https://doi.org/10.1111/tbj.12681>
- Makki J (2015) Diversity of breast carcinoma: histological subtypes and clinical relevance. *Clin Med Insights Pathol* 8:23–31. <https://doi.org/10.4137/CPath.S31563>
- Onega T, Goldman LE, Walker RL, Miglioretti DL, Buist DS, Taplin S, Geller BM, Hill DA, Smith-Bindman R (2016) Facility mammography volume in relation to breast cancer screening outcomes. *J Med Screen* 23(1):31–37. <https://doi.org/10.1177/0969141315595254>
- Petushi S, Garcia FU, Haber MM, Katsinis C, Tozeren A (2006) Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer. *BMC Med Imaging* 6:14. <https://doi.org/10.1186/1471-2342-6-14>
- Siegel RL, Miller KD (2021) Cancer statistics, 2021. *CA Cancer J Clin* 71(1):7–33. <https://doi.org/10.3322/caac.21654>
- Sim KS, Chia FK, Nia ME, Tso CP, Chong AK, Abbas SF, Chong SS (2014) Breast cancer detection from MR images through an auto-probing discrete Fourier transform system. *Comput Biol Med* 49:46–59. <https://doi.org/10.1016/j.compbiomed.2014.03.003>
- Suk H-I, Lee S-W, Shen D, Initiative ADN (2014) Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *Neuroimage* 101:569–582. <https://doi.org/10.1016/j.neuroimage.2014.06.077>
- Tan YJ, Sim KS, Ting FF (2017) Breast cancer detection using convolutional neural networks for mammogram imaging system. In: 2017 International Conference on Robotics, Automation and Sciences (ICORAS), pp. 1–5. <https://doi.org/10.1109/ICORAS.2017.8308076>
- Ting FF, Sim KS (2017) Self-regulated multilayer perceptron neural network for breast cancer classification. In: 2017 International Conference on Robotics, Automation and Sciences (ICORAS), pp. 1–5. <https://doi.org/10.1109/ICORAS.2017.8308074>
- Ting FF, Sim KS, Chong SS (2017) Auto-probing breast cancer mass segmentation for early detection. In: 2017 International Conference on Robotics, Automation and Sciences (ICORAS), pp. 1–5. <https://doi.org/10.1109/ICORAS.2017.8308077>
- Ting FF, Tan YJ, Sim KS (2019) Convolutional neural network improvement for breast cancer classification. *Expert Syst Appl* 120:103–115. <https://doi.org/10.1016/j.eswa.2018.11.008>
- Toa CK, Sim KS, Tan SC (2021) Electroencephalogram-based attention level classification using convolution attention memory neural network. *IEEE Access* 9:58870–58881. <https://doi.org/10.1109/access.2021.3072731>
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: *Advances in Neural Information Processing Systems*, 2017-Decem(Nips), pp. 5999–6009
- Wang L (2017) Early diagnosis of breast cancer. *Sensors (switzerland)* 17(7):1572. <https://doi.org/10.3390/s17071572>
- Wang D, Khosla A, Gargeya R, Irshad H, Beck A (2016) Deep learning for identifying metastatic breast cancer. *ArXiv*, abs/1606.0
- Yap MH, Yap CH (2016) Breast ultrasound lesions classification: a performance evaluation between manual delineation and computer segmentation. *Proc SPIE*. <https://doi.org/10.1117/12.2208797>
- Youlden DR, Cramb SM, Dunn NAM, Muller JM, Pyke CM, Baade PD (2012) The descriptive epidemiology of female breast cancer: an international comparison of screening, incidence, survival and mortality. *Cancer Epidemiol* 36(3):237–248. <https://doi.org/10.1016/j.canep.2012.02.007>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.