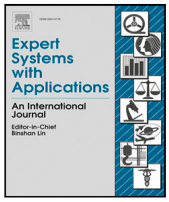
Expert Systems With Applications 249 (2024) 123569

Contents lists available at ScienceDirect

Expert Systems With Applications

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



RDTNet: A residual deformable attention based transformer network for breast cancer classification

Babita, Deepak Ranjan Nayak *

Department of Computer Science and Engineering, Malaviya National Institute of Technology, Jaipur, 302017, Rajasthan, India

ARTICLE INFO

Keywords:
Breast cancer
Convolution neural networks
Deformable convolution
Histopathological image
Vision transformer
RDTNet

ABSTRACT

Accurate and timely detection of breast cancer plays a pivotal role in reducing the mortality rate. Deep learning models, especially CNNs, have recently shown astounding performance in detecting breast cancer from histopathological images. However, their drawbacks lie in the limited capacity to capture subtle lesion information. Vision transformers (ViTs) have emerged as a promising technique due to their ability to capture global feature dependencies through self-attention. Nevertheless, applying ViTs in medical imaging is challenging due to the unavailability of large training data and their limited ability to capture local contextual information. To address these challenges, we propose a residual deformable attention-based transformer network (RDTNet) for breast cancer classification, which can capture local and global contextual details from the histopathological images. In RDTNet, we introduce a residual deformable transformer layer called RDTL after a backbone network. The RDTL comprises multi-head deformable self-attention mechanisms (MDSA) and residual connections, enabling fine-grained and category-specific lesion feature extraction. The experimental results on a benchmark dataset indicate the superiority of the RDTNet over state-of-the-art methods. Notably, our model achieves a higher image-level accuracy of 99.00%, 98.87%, 98.84%, and 97.80% and a patient-level accuracy of 96.41%, 94.82%, 93.91%, and 91.25% for 40×, 100×, 200×, and 400× magnifications, respectively. The improved performance of RDTNet can be attributed to the integration of RDTL with a backbone network.

1. Introduction

Breast cancer (BC) is one of the most lethal forms of cancer in women worldwide. According to WHO, breast cancer affects over 2.3 million individuals each year, making it the most prevalent cancer among adults. In 95% of countries, breast cancer ranks as either the primary or secondary cause of female cancer-related fatalities. Approximately 80% of breast and cervical cancer-related deaths occur in low and middle-income nations (Sung, Ferlay, & Siegel, 2021). As per the report by the International Agency for Research on Cancer (Guida et al., 2022), breast cancer constituted 25% of the roughly 4.4 million female cancer-related deaths in 2020. Such alarming statistics showcase the significance of detecting breast cancer at early stages.

Despite the use of several medical imaging modalities such as X-rays (mammograms) (Wang, Wang, Feng and Zhang, 2021), ultrasound (sonography) (Chiang, Huang, Chen, Huang, & Chang, 2018), thermography (Pramanik et al., 2018), and magnetic resonance imaging (MRI) (Wang, Sun et al., 2021), the histopathological imaging is widely adopted as the most reliable method by pathologists for identification of breast cancer with an exceptionally high degree of precision (Gurcan et al., 2009; Rubin, Strayer, Rubin, et al., 2008). It uses biopsy procedures to obtain cell and tissue samples that are further mounted on a microscope slide and are stained and examined under a microscope.

Primary Motivation: The visual inspection of histopathological images is time-consuming, and needs a high level of pathological proficiency. Furthermore, it has been frequently observed that when the same set of images is interpreted by different medical specialists, there is a substantial level of inter-observer variability. Hence, it is of utmost importance to design an automated computer-aided diagnosis (CAD) system using AI-powered cutting-edge techniques to detect breast cancer timely and accurately. This also aids in lessening the workload of pathologists.

Recent years have witnessed the significance of deep neural network based methods in a wide range of applications including designing smart systems (Połap, Jaszcz, Wawrzyniak and Zaniewicz, 2023; Połap, Srivastava and Jaszcz, 2023) and automated CAD systems. Among all deep neural network architectures, convolutional neural networks (CNNs) have been explored the most for breast cancer classification via histopathological images (Sharma, Nayak, Balabantaray, Tanveer, & Nayak, 2023). Though CNNs have shown enhanced performance due

* Corresponding author.
E-mail addresses: 2021rcp9035@mnit.ac.in (Babita), drnayak.cse@mnit.ac.in (D.R. Nayak).

https://doi.org/10.1016/j.eswa.2024.123569
Received 24 December 2023; Received in revised form 9 February 2024; Accepted 24 February 2024
Available online 5 March 2024

to their ability to derive hierarchical features from histopathological images, there is still room for performance improvement to meet real-time diagnosis requirements. In addition, it has been observed that a few CNN-based methods do not support end-to-end learning and are validated using limited training images. Further, conventional CNNs fail to capture minute lesion information and their relationships due to their limited ability to model wide-range global feature interdependencies. Recently, various approaches have been proposed to address these issues and incorporate global context into CNNs. Attention mechanisms aim to enhance the capability of CNNs by allowing them to focus on crucial regions and ignore the irrelevant ones, thus improving their potential in medical image analysis for early disease diagnosis (Das, Nayak, & Pachori, 2023; He, Li, Li, Wang, & Fu, 2020; Krishna, Suganthi, Bhavsar, Yesodharan, & Krishnamoorthy, 2023). These mechanisms have not been fully explored for breast cancer detection using histopathological images.

1.1. Innovation

Recent progress in computer vision has shown the proficiency of vision transformers (ViT) over CNN in several complex vision tasks (Han et al., 2022). This is primarily due to their potential to explore long-range feature dependencies through multi-head self-attention mechanisms (Dosovitskiy, Beyer, Kolesnikov, & Weissenborn, 2021). However, ViTs are not scalable because of high computational cost and large-scale memory requirements. Further, ViTs have shown their limitations in capturing local contextual information, which is essential along with global details for a thorough interpretation of complex lesion regions. In light of this, we propose a novel convolutional transformer network known as RDTNet for breast cancer classification by integrating a pre-trained CNN architecture and a residual deformable transformer layer. The proposed RDTNet enjoys the benefits of both CNN and ViT, allowing the model to capture local and global contextual feature dependencies.

1.2. Contribution

To the best of our knowledge, the proposed method is the first transformer-based method particularly designed for breast cancer classification. The primary contributions of this paper are as follows:

- We propose an end-to-end learnable residual deformable attention-based transformer network (RDTNet) for breast cancer detection in histopathological images by integrating a novel transformer module with a pre-trained CNN (backbone).
- We introduce a residual deformable transformer layer called RDTL on the top of the backbone to model both local and global feature dependencies, thereby facilitating the learning of salient features from the lesion regions. The RDTL comprises a multi-head deformable self-attention (MDSA) module and residual connections, which are mainly responsible for modeling wide-range dependencies from high-level feature maps. The deformable convolution is harnessed in the self-attention mechanism to adapt to various geometric variations of lesion regions.
- We evaluate RDTNet on a benchmark dataset and use different backbone networks to test the effectiveness of the proposed RDTL. In addition, we perform a comparative analysis with existing attention mechanisms and state-of-the-art breast cancer classification methods. The implementation of RDTNet is available at https://github.com/RCP9035/RDTNet.

The subsequent sections of the paper are organized as follows: Section 2 presents a comprehensive analysis of related works. In Section 3, the details of the dataset used are provided, while Section 4 presents the problem statement. The proposed methodology and its components are discussed in Section 5. Section 6 presents the implementation details, results, and comparisons. In Section 7, conclusions are derived along with a few recommendations for possible future work.

2. Related work

In this section, we provide an extensive overview of the relevant literature on breast cancer detection using histopathology images. In addition, the importance of ViT models and their challenges are discussed.

2.1. Deep learning-based breast cancer detection methods

Earlier the practice of feature engineering held a prominent position within the medical imaging research community. To this end, a few efforts have been made towards classifying breast cancers using histopathological images. For instance, Zhang, Zhang, Coenen, Xiao, and Lu (2014) utilized a one-class kernel-based principal component analysis (KPCA) ensemble for binary classification of breast cancer biopsy images. Wang, Hu, Li, Liu, and Zhu (2016) obtained an optimal feature set by SVM with a chain-like agent genetic algorithm (CAGA) and extracted shape-based four and color space-based 138 textural features for the classification of cell nuclei into benign and malignant classes. Spanhol, Oliveira, and Petitjean (2015) used a combination of hand-crafted feature descriptors such as GLCM, LBP, and PFTAS and a set of traditional classifiers such as SVM, 1-NN, and RF for the classification of histopathological images. They released a large dataset named BreakHis and achieved the highest accuracy of 85%. These approaches are heavily dependent on the quality of features extracted from histopathological images. Another concern in these schemes is the choice of a suitable classifier.

In recent years, deep learning methods have caught remarkable attention for the classification of histopathological images due to their ability to learn hierarchical features automatically. Spanhol, Oliveira, Petitjean, and Heutte (2016) used image patches extracted from whole slide images to train a CNN for breast cancer classification. In Bayramoglu, Kannala, and Heikkilä (2016), a customized CNN-based approach independent of magnifications of histopathology images was proposed for breast cancer classification. A single-task CNN was designed to predict malignancy, while a multi-task CNN was designed to predict both malignancy and image magnification levels. In Spanhol, Oliveira, and Cavalin (2017), Spanhol et al. addressed the problem of increased complexity and longer training time in CNN. They used a pre-trained CNN model as a feature extractor and a classifier to perform breast cancer histopathological image classification. Gupta and Bhavsar (2018) proposed a sequential model that uses multi-layered deep features extracted from DenseNet and demonstrated improved performance over a benchmark dataset. Later, Gupta and Bhavsar (2019) proposed a model by integrating multi-layer features extracted from a pre-trained ResNet model for improved breast cancer histopathology image classification. Saini and Susan (2022) developed a model named VGGIN-Net in which features were extracted from a specific block of VGG-19 architecture and fed to multiple inception modules for further refinement. The refined features were finally concatenated and classified into benign and malignant classes. Chhipa et al. (2023) proposed a self-supervised pre-training method called magnification prior contrastive similarity (MPCS), to learn efficient representations without labels on histopathological images.

Inspired by the success of attention mechanisms in CNN architectures, Xu, Liu, Hou, Liu, and Garibaldi (2019) proposed a deep hybrid attention method for the classification of breast cancer. They selected sequences of regions from raw histopathological images using a hard visual attention algorithm. Then, for each region, abnormal parts were identified using a soft attention mechanism. Wu et al. (2019) proposed two separate attention blocks along with CNN for more efficient feature learning from histopathology images. Toğaçar et al. (2020) designed a model called BreastNet, which incorporates a residual architecture with attention modules and achieved an accuracy of 98.80%. Chattopadhyay et al. (2022) developed an end-to-end dense residual dual-shuffle attention network (DRDA-Net) to extract complex lesion patterns from

Table 1
Summary of state-of-the-art approaches for histopathological breast cancer classification.

	Author	Approach	Dataset	Accuracy (%)
Traditional methods	Zhang et al. (2014)	One-class KPCA	Breast Image set	92.28
			3D OCT	92.06
			UCI Wisconsin	97.28
	Wang et al. (2016)	SVM with CAGA	BCH dataset	95.83
	Spanhol et al. (2015)	CLBP, GLCM, LBP LPQ, ORB, PFTAS	BreaKHis	83.8 (40×)
CNN-based methods	Bayramoglu et al. (2016)	Single and Multi task CNN	BreaKHis	83.08 (40×)
	Spanhol et al. (2016)	AlexNet variant	BreaKHis	89.6 (40×)
	Spanhol et al. (2017)	CaffeNet variant	BreaKHis	84.6 (40×)
	Gupta and Bhavsar (2018)	DenseNet-169 + XGboost	BreaKHis	94.71 (40×)
	Gupta and Bhavsar (2019)	ResNet + quadratic SVM	BreaKHis	96.81 (40×)
	Saini and Susan (2022)	VGG-16 with Inception blocks	BreaKHis	97.10 (40×)
			IDC Dataset	86.78
	Chhipa et al. (2023)	Efficient-net b2 and ResNet-50 as encoder with MPCS	BreaKHis	93.45 (40×)
			BACH	91.85
			BCCD	96.36
CNN with attention	Wu et al. (2019)	VGG19 with spatial and channel attention	BreaKHis	91.4 (40×)
	Toğaçar, Özkurt, Ergen, and Cömert (2020)	BreastNet	BreaKHis	98.80 (40×)
	Chattopadhyay, Dey, Singh, and Sarkar (2022)	DRDA-Net	BreaKHis	95.72 (40×)
	Ijaz et al. (2023)	VGG16 and VGG19 with CBAM	BreaKHis	94.44 (40×)
			BACH	
			PCam	
			LC25000	
	Krishna et al. (2023)	DarkNet19 + ABN	BreaKHis	98.4 (40×)

histopathological images. Ijaz et al. (2023) introduced a modality-specific CBAM-VGGNet model where they trained VGG-16 and VGG-19 models on the same domain using histopathology datasets and used them as fixed feature extractors. They incorporated the convolutional block attention module (CBAM) to enhance the ability of the model to focus on salient features. Recently, Krishna et al. (2023) proposed an interpretable decision-support model called ABN-DCN based on a backbone CNN model and attention branch network (ABN) for classification of breast cancer histopathological images. An overview of the existing approaches is summarized in Table 1.

The literature reveals that a few proposed methods do not support end-to-end learning. Further, the conventional CNNs used have limited capabilities in capturing and detecting a wide range of relationships in breast cancer histopathology images, resulting in lower performance. Although attention mechanisms-based CNN methods have improved the performance of CNNs, they still lack the ability to fully capture the global feature inter-dependencies, which are crucial for histopathology image analysis.

2.2. Vision Transformer (ViT)

Vision transformer (ViT) has recently received a significant attention in a wide array of computer vision tasks due to its capability to effectively handle wide range feature dependencies. It has demonstrated dramatic success in accurately capturing complex visual patterns across different tasks. Recently, some efforts have been made to explore the potential of ViT in medical image analysis tasks (Gheflati & Rivaz, 2022; Mo et al., 2023; Shamshad et al., 2023). However, ViT face challenges due to its high computational requirements and substantial data requirements, making it less scalable in medical imaging applications. They can also occasionally fall short in capturing local contextual information, which is crucial for understanding of complex lesion regions. As a result, a transition has emerged towards the development of hybrid architectures which combines both CNN and ViT (He, Yang, & Xie, 2023; Huang, Li, Xiao, Shen, & Xu, 2022). These architectures possess merits of both CNN and ViT, which facilitates in achieving state-of-the-art performance in various computer vision tasks. In addition, they offer a more interpretable framework compared to conventional CNN based methods. Considering the strengths of such hybrid architectures, we introduce a convolutional transformer network called RDTNet for breast cancer histopathological image classification. This architecture combines a pre-trained CNN model and a novel transformer layer to effectively capture both local and global contextual information from histopathological images, leading to improved classification performance.

Table 2
Description of BreakHis dataset (Spanhol et al., 2015).

Class	#Patients	Magnification				Total
		40×	100×	200×	400×	
Benign	24	625	644	623	588	2480
Malignant	58	1370	1437	1390	1232	5429
Total	82	1995	2081	2013	1820	7909

3. Dataset used

To validate the proposed RDTNet, we consider a publicly available benchmark dataset named BreaKHis, which was proposed by Spanhol et al. (2015). The digitized images in the dataset were captured from breast tissue slides by an Olympus BX-50 system microscope equipped with a relay lens at a 3.3× magnification and a Samsung digital color camera SCC-131AN.

The images were collected from 82 patients and the dataset comprised a total of 7909 microscopy biopsy images. Of which, there are 2480 benign breast tumor samples and 5429 malignant samples. These images were of dimensions 700 × 460 × 3. The dataset incorporated varying image magnifications including 40×, 100×, 200×, and 400×, with respect to objective lenses of 4×, 10×, 20×, and 40×. The pixel size in the object plane for these magnifications was 0.49 μm, 0.20 μm, 0.10 μm, and 0.05 μm, respectively. The sample images at four different magnifications for benign and malignant cases are shown in Figs. 1 and 2, respectively. Table 2 presents the distribution of images for each class across four magnifications. The more details of the dataset can be found in Spanhol et al. (2015).

4. Problem statement

Given a dataset of breast cancer histopathological images $\{(x^{(i)}, y^{(i)})\}_{i=1}^{m}$, where $x^{(i)}$ represents the ith histopathological image and $y^{(i)}$ denotes its corresponding label, indicating the presence ($y^{(i)} = 1$) or absence ($y^{(i)} = 0$) of breast cancer, and m is the total number of samples.

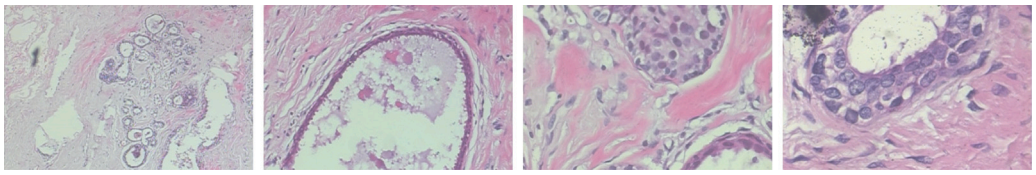


Fig. 1. Sample histopathological images of benign class at different magnifications: (a) 40× (b) 100× (c) 200× (d) 400×.

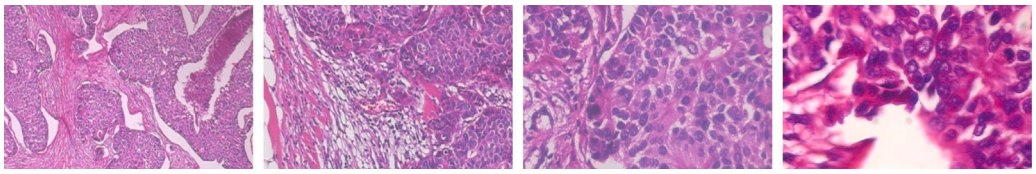


Fig. 2. Sample histopathological images of malignant class at different magnifications: (a) 40× (b) 100× (c) 200× (d) 400×.

The objective is to train an end-to-end model by learning a mapping function $f(x)$ that accurately predicts the probability of benign and malignant class in unseen histopathological images. Mathematically, the problem can be formulated as follows.

Given $\{(x^{(i)}, y^{(i)})\}_{i=1}^{m}$, where $x^{(i)} \in \mathbb{R}^{H\times W\times 3}$ and $y^{(i)} \in \{0, 1\}$, goal is to learn a model $f(x) : \mathbb{R}^{H\times W\times 3} \rightarrow [0, 1]$, to predict the probabilities of benign and malignant class such that $\hat{y} \approx y = f(x)$.

5. Methodology

In this section, we provide a detailed description of our proposed RDTNet model and its basic components. We first discuss the overview of the RDTNet architecture to perform breast cancer classification and then describe each component of it.

5.1. Overview

As shown in Fig. 3, our proposed RDTNet is an end-to-end architecture which is composed of three key components: a backbone network, a residual deformable transformer layer (RDTL) and a classifier.

At first the histopathological image undergoes an initial passage through a backbone CNN network, resulting in the extraction of hierarchical feature maps F. Subsequently, these feature maps are fed into the proposed RDTL to model wide-range feature dependencies. The RDTL introduces a multi-head deformable self-attention (MDSA) module which leverages the extraction of salient and category-specific features from the crucial regions of the histopathological images. In addition, it utilizes skip connections for better feature flow within the network. The resultant feature are finally fed to a classification layer to classify the image as benign or malignant. Each component of RDTNet is discussed in detail in the following subsections.

5.2. Backbone network

In our proposed RDTNet, we used the pre-trained Xception (Chollet, 2017) as the backbone network for high-level feature extraction. Xception is a deep convolution network which was pre-trained on ImageNet dataset. It employs depthwise separable convolutions which are more efficient than classical convolutions in terms of computational complexity. The input to the backbone is a histopathological image and it generates a set of high-level feature maps $F \in \mathbb{R}^{H\times W\times C}$, where H, W, and C represent the height, width, and number of channels, respectively. It is worth noting that the high level feature representations F are derived from the last convolution layer of the backbone network.

5.3. Residual Deformable Transformer Layer (RDTL)

While the feature F accommodates the local contextual details from input images, it does not contain global contextual information which are crucial in case of histopathological image analysis. Therefore, we introduced RDTL to capture both local and global feature relationships within the feature maps. This ultimately facilitates the model to learn salient patterns from the lesion regions. The RDTL consists of a few major components such as layer normalization, MDSA, 1×1 convolution, and residual connections. The purpose of introducing MDSA module in RDTL is to extract long-range feature dependencies using deformable self-attention, resulting in a refined feature map F_{df}.

The RDTL takes a feature map $F \in \mathbb{R}^{H\times W\times C}$ as input, which is generated from the backbone network. Then, it utilizes layer normalization to normalize every feature of the activations, which facilitates stabilizing and speeding up the learning process (Ba, Kiros, & Hinton, 2016). The normalized feature map F_n is then fed to the MDSA module to obtain an attention feature map $F_{df} \in \mathbb{R}^{H\times W\times C}$, which can be mathematically expressed as follows:

$$F_n = LayerNorm(F) \tag{1}$$

$$F_{df} = \text{MDSA}(F_n) \tag{2}$$

To enhance feature propagation throughout the RDTL, residual connections have been introduced. We obtain F' after adding the attention feature map with F using residual connection. The feature map F' is further fed to layer normalization, followed by a 1×1 convolution for a better feature representation, resulting in a final feature map F_{res} using a residual connection.

$$F' = F + F_{df} \tag{3}$$

$$F_{res} = F' + Conv2D(LayerNorm(F')) \tag{4}$$

where $Conv2D$ represents the convolution operation with kernel size 1×1 followed by ReLU activation.

5.3.1. Multi-head Deformable Self-Attention (MDSA)

In histopathological images, lesions carry unique patterns with random spatial distributions. Hence, it is of utmost importance to extract these characteristics from different lesions for better decision-making. Self-attention (SA) is the core component in ViT models, which helps to model wide-range global feature relationships and allows the model to focus on critical areas of the input. However, the SA utilizes the

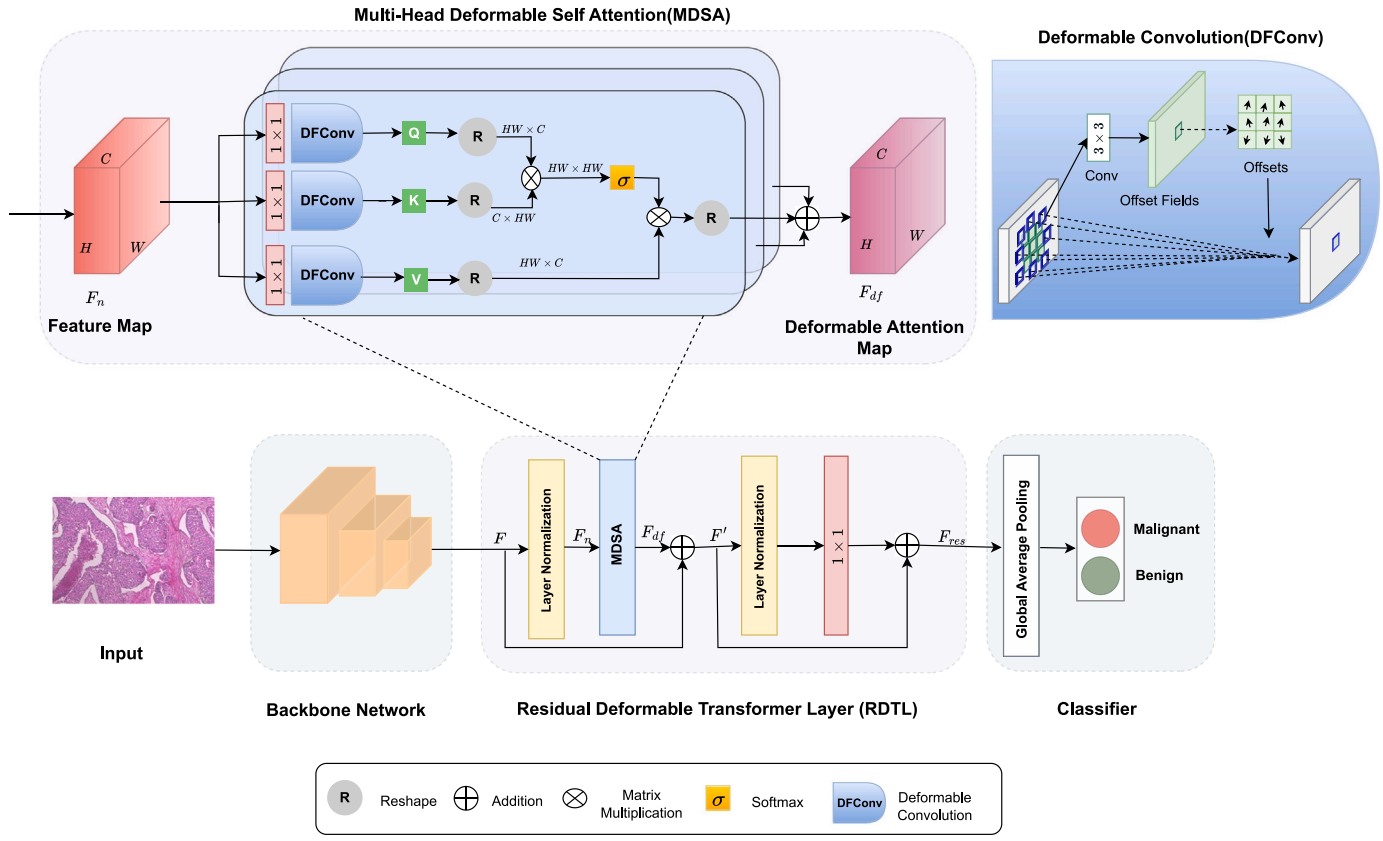


Fig. 3. The overall framework of the proposed RDTNet for breast cancer classification. It consists of a backbone network, a RDTL, and a classification layer.

standard convolution of the fixed kernel size of 1×1, thereby extracting features of fixed receptive field. It is essential to take into account the need for diverse receptive fields while extracting features from histopathological images to capture lesions of various shapes and sizes.

Therefore, we applied a dynamic filtering method known as *deformable convolution* ($DFConv$) across the spatial dimensions in SA to generate query, key, and value feature tensors, and hence it is named deformable self-attention (DSA) in this study. Fig. 3 depicts the structure of the proposed DSA. Deformable convolution (Dai et al., 2017) provides the benefit of an adjustable receptive field that changes depending on the lesion's scale and, therefore, has the capability to adapt to geometrical variations of the lesions. In addition, it is learnable. In *deformable convolution*, the normal grid $I = \{(-1,-1),(-1,0),\ldots,(0,1),(1,1)\}$ over the feature map is augmented with offset $\{\Delta d_n | n = 1,\ldots,D\}$, where $D = |I|$. Therefore, for every location d_0 in the output feature map, $DFConv$ is computed as

$$DFConv(d_0) = \sum_{d_n \epsilon I} w(d_n).x(d_0 + d_n + \Delta d_n) \tag{5}$$

As shown in Fig. 3, the feature map $F_n \in \mathbb{R}^{H \times W \times C}$ is derived from layer normalization and is fed as input to MDSA. Like in traditional SA (Das & Nayak, 2023; Zhang, Goodfellow, Metaxas, & Odena, 2019), we applied convolution operations with kernel size 1×1 over F_n to obtain a representative feature maps. Further, different from traditional SA, *deformable convolution* is applied on each output to obtain query, key, and value feature tensors as follows.

$$Q(F_n) = DFConv(Conv2D(F_n)) \tag{6}$$

$$K(F_n) = DFConv(Conv2D(F_n)) \tag{7}$$

$$V(F_n) = DFConv(Conv2D(F_n)) \tag{8}$$

Here, $Q(F_n)$, $K(F_n)$ and $V(F_n)$ are the query, key and value projections of feature map F_n, $Conv2D$ is the convolutional operation with kernel size 1×1 followed by ReLU activation and $DFConv$ represents the *deformable convolution* with kernel size 3×3.

We reshape the spatial dimension of $Q(F_n)$, $K(F_n)$, and $V(F_n)$ to HW and produce tensors of size $HW \times C$, where every row indicates a feature vector corresponding to a specific spatial position. To capture the spatial similarities of these feature vectors, we perform matrix multiplication of $K(F_n)$ and $Q(F_n)$ and subsequently apply softmax activation to generate a spatial attention map $A_{sp} \in \mathbb{R}^{HW \times HW}$ as

$$A_{sp} = \sigma(Q(F_n) \otimes K(F_n)^T) \tag{9}$$

where $\sigma(.)$ denotes the softmax activation, $\otimes$ indicates the matrix multiplication, and $K(F_n)^T$ denotes the transpose of the key tensor. Further, the attention map A_{sp} is multiplied with $V(F_n)$ and the product is reshaped back to $H \times W \times C$ by spatial dimension expansion, thus obtaining the deformable attention feature map $F_{df} \in \mathbb{R}^{H \times W \times C}$ which can be mathematically defined as

$$F_{df} = A_{sp} \otimes V(F_n) \tag{10}$$

To explore rich spatial contextual information from the input feature maps, we use three heads in the proposed MDSA which has been decided empirically. Finally, we perform the element wise addition of the three deformable attention feature maps to obtain final attention feature map $F_{df} \in \mathbb{R}^{H \times W \times C}$. So, we can breakdown the Eq. (2) as

$$MDSA(F_n) = F^1_{df} \oplus F^2_{df} \oplus F^3_{df} \tag{11}$$

where, $\oplus$ represents the element wise addition and F^h_{df} indicates the output obtained at head $h \in \{1,2,3\}$. With the introduction of MDSA, the model could able to extract adequate spatial feature relationships from the input feature map. In a nutshell, the proposed transformer layer over the backbone network aids to extract the global and local

feature dependencies using multi-head self mechanism and deformable convolutions. The pseudo-code for the RDTL is provided in Algorithm 1.

Algorithm 1: Pseudo-code for RDTL

Input: F : Feature map obtained from backbone network
h : Number of heads
k : Kernel size
Output: F_{res} : Output feature map of RDTL

$F_n := LayerNorm(F)$
$i := 0$
/* This is MDSA module */
for i *to* h **do**
  $Q_i := DFConv_{k=3}(Conv2D_{k=1}(F_n))$
  $K_i := DFConv_{k=3}(Conv2D_{k=1}(F_n))$
  $V_i := DFConv_{k=3}(Conv2D_{k=1}(F_n))$
  $A^i_{sp} := \sigma(Q_i \otimes K_i^T)$
    // $\otimes$ and σ denote matrix multiplication and softmax activation
  $f^i_{df} := A^i_{sp} \otimes V_i$ // Deformable attention map of i^{th} head
$F_{df} := \sum_{i=1}^{h} f^i_{df}$
$F' := F + F_{df}$
$F_{res} := F' + Conv2D_{k=1}(LayerNorm(F'))$

The RDTL builds on a basic block called MDSA module that consists of three heads. A head takes as input $F_n \in \mathbb{R}^{H\times W\times C}$ and applies three projections to obtain queries (Q), keys (K), and values (V). The module computes the product between queries and keys, and applies softmax to produce attention scores, where $Q \in \mathbb{R}^{HW\times C}$ and $K \in \mathbb{R}^{HW\times C}$. The matrix product QK^T leads to $O(N^2C)$ complexity, where $N = H * W$.

5.4. Classifier

In classification layer, we use a global average pooling (GAP) layer followed by a fully connected layer with two nodes to classify the breast cancer into benign or malignant class.

6. Experiments and results

In this section, we present the experimental settings with implementation details and the performance metric used to evaluate the proposed and existing methods. In addition, we provide the results and comparisons with different attention mechanisms as well as state-of-the-art breast cancer classification techniques. Further, we present the results of the ablative experiments carried out to test the influence of each component of the proposed transformer layer.

6.1. Experimental setup and performance metrics

We employed the BreaKHis dataset to evaluate the proposed RDTNet, which accounts for varying magnification levels, necessitating the training of images magnified at 40×, 100×, 200×, and 400× separately. All input images were uniformly resized to 299 × 299 for input dimensional compatibility with ImageNet pre-trained Xception architecture. We used images from 58 patients (70%) chosen at random for training and the remaining 24 patients (30%) for testing as adopted in Spanhol et al. (2015). We augmented the training images using rescaling, rotation, vertical flip, horizontal flip, height shift, width shift, and zooming, which mitigates the issue of overfitting. We optimized our model using Adam optimizer and used categorical cross-entropy loss function during training.

To select the best hyperparameters, we conducted a series of experiments by varying their values and effect of a few notable settings on the performance of our model is shown in Table 3. It can be seen that our model under setting 'S-3' achieved better performance and hence, we adopted this setting for our study. It is noteworthy that these experiments were performed for 40× magnification. Specifically, the batch size (BS) was set to 16 and the initial learning rate (L_r) was set to 0.0001 with decayed by a factor (F_{decay}) of 0.8 and weight decay (W_{decay}) 0.00001. We performed warm up training for 2 epochs before training our model for 70 epochs. A similar training strategy was adopted across different magnification factors such as 100×, 200×, and 400×. All the experiments were conducted on a workstation with NVIDIA Tesla V-100 GPU and 128 GB memory. The models were implemented using Keras and TensorFlow deep learning framework.

Table 3
Effect of different train settings for hyperparameters selection.

Hyperparameters	S-1	S-2	S-3	S-4	S-5	S-6
L_r	0.001	0.001	0.0001	0.0001	0.00001	0.00001
BS	8	8	16	16	32	64
Epochs	50	50	70	70	100	100
Input_size	299	399	299	299	399	399
W_{decay}	0.001	0.001	0.00001	0.001	0.0001	0.0001
F_{decay}	0.5	0.6	0.8	0.8	0.9	0.9
Accuracy (%)	97.49	98.83	99.0	98.46	98.34	98.71

To evaluate the performance of the proposed model as well as existing methods, we use various commonly adopted performance measures such as patient level accuracy and image level accuracy which are discussed as follows.

The patient-level accuracy is denoted as A_{PL} and is defined as

$$A_{\text{PL}} = \frac{\sum_{i=1}^{N} \text{PA}_i}{N} \tag{12}$$

where N is the total number of patients available in the test set and PA_i is the score of ith patient which is computed as

$$\text{PA} = \frac{N_{cor}}{N_P} \tag{13}$$

where N_{cor} indicates the total number of correctly classified images and N_P denotes the total number of images of a patient P.

The image level accuracy is denoted as A_{IL} and is stated as

$$A_{\text{IL}} = \frac{N_{tcp}}{N_{TI}} \tag{14}$$

where, N_{tcp} indicates the total number of correctly classified images and N_{TI} indicates the total number of images in the test set. The patient information is not taken into account while computing A_{IL}.

6.2. Results

6.2.1. Comparison with backbone models

To verify the effectiveness of Xception (Chollet, 2017) as a backbone network in the proposed RDTNet, we compared it with a few contemporary pre-trained CNN such as MobileNet (Howard et al., 2017), DenseNet121, DenseNet169 (Huang, Liu, Van Der Maaten, & Weinberger, 2017), and InceptionV3 (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016). All these models were pre-trained on ImageNet dataset and were later fine-tuned on BreaKHis dataset. The results in terms of image level accuracy and patient level accuracy are shown in Tables 4 and 5, respectively. It can be seen from Tables 4 and 5 that our proposed RDTL improves the overall performance when compared with the backbone models across all magnifications. The reason of superior performance can be attributed to the inclusion of MDSA in the RDTL. Also, it can be observed that Xception with RDTL achieved an higher image level accuracy of 99.00%, 98.87%, 98.84%, and 97.80% for 40×, 100×, 200×, and 400×, respectively. Further, a higher performance in terms of patient level accuracy was obtained by Xception compared to other pre-trained models.

Table 4
Image level accuracy (in %) of the proposed model by adopting different backbone networks.

Model		Magnification factor			
		40×	100×	200×	400×
MobileNet (Howard et al., 2017)	Backbone	97.99	96.58	96.65	95.39
	Backbone + RDTL	98.49	98.02	96.68	96.70
DenseNet121 (Huang et al., 2017)	Backbone	97.35	97.75	97.63	96.25
	Backbone + RDTL	98.59	97.83	98.17	97.18
DenseNet169 (Huang et al., 2017)	Backbone	98.24	97.92	98.27	97.32
	Backbone + RDTL	98.83	98.02	98.45	97.45
InceptionV3 (Szegedy et al., 2016)	Backbone	98.06	97.89	98.34	96.81
	Backbone + RDTL	98.66	97.92	98.55	97.32
Xception (Chollet, 2017)	Backbone	98.53	97.95	98.67	97.61
	RDTNet	99.00	98.87	98.84	97.80

Table 5
Patient level accuracy (in %) of the proposed model by adopting different backbone networks.

Model		Magnification factor			
		40×	100×	200×	400×
MobileNet (Howard et al., 2017)	Backbone	90.05	91.75	90.46	87.85
	Backbone + RDTL	93.34	92.32	92.39	89.49
DenseNet121 (Huang et al., 2017)	Backbone	90.67	91.21	91.49	88.78
	Backbone + RDTL	92.32	93.09	91.72	89.60
DenseNet169 (Huang et al., 2017)	Backbone	91.42	91.34	91.78	86.38
	Backbone + RDTL	95.05	91.53	92.39	89.69
InceptionV3 (Szegedy et al., 2016)	Backbone	90.54	90.32	92.13	86.18
	Backbone + RDTL	91.13	91.88	93.24	89.88
Xception (Chollet, 2017)	Backbone	92.22	91.16	91.52	86.71
	RDTNet	96.41	94.82	93.91	91.25

6.2.2. Comparison with existing attention approaches

To demonstrate the potential of RDTL, we compared it with the state-of-the art attention mechanisms such as squeeze-and-excitation (SE) (Hu, Shen, & Sun, 2018), bottleneck attention module(BAM) (Park, Woo, Lee, & Kweon, 2018), convolutional block attention module (CBAM) (Woo, Park, Lee, & Kweon, 2018), category attention block (CAB) (He et al., 2020) and self-attention (Dosovitskiy et al., 2021) and the results are shown in Table 6. It can be observed that the proposed RDTL yields superior performance in the context of image level and patient level accuracy across all magnifications. Further, a comparable performance was obtained with CBAM and SE. It is worth noting that these attention modules have been placed after the last convolutional layer of the backbone network. For this experiment, we considered the best performing pre-trained model as the backbone.

6.2.3. Comparison with existing approaches

We performed a comparative analysis of our proposed RDTNet against existing breast cancer detection methods on the BreaKHis dataset as shown in Table 7. It is evident that our approach achieved a higher classification performance compared to existing methods at both image and patient levels. A few studies (Gupta & Bhavsar, 2017; Song et al., 2018; Spanhol et al., 2015) employed conventional methods of feature engineering and classification, and obtained a lower performance. While other approaches (Bayramoglu et al., 2016; Spanhol et al., 2017, 2016) utilized pre-trained convolutional neural networks (CNNs) for deep feature extraction and applied various classifiers for improved breast cancer classification. However, these methods do not support end-to-end learning. Though the methods reported in Benhammou, Tabik, Achchab, and Herrera (2018) and Saini and Susan (2022) support end-to-end training, they yielded a comparatively lower classification performance than the proposed approach. The attention-based networks proposed in Chattopadhyay et al. (2022), Krishna et al. (2023), Toğaçar et al. (2020), and Wang, Wang, Li et al. (2021) obtained a comparable performance. While the self-supervised pre-training method, MPCS-RP (Chhipa et al., 2023), obtained lower performance than the proposed approach. In summary, the proposed RDTNet is end-to-end learnable and it eliminates the need for traditional feature engineering. Moreover, it leads to a better performance compared to attention mechanisms based networks as well as other existing CNN-based models. This can be mainly attributed to the introduction of MDSA mechanism in RDTL and its proficiency in extracting complex lesion patterns from histopathological images. It is worth mentioning here that the reported values of the existing methods have been taken from their original papers. Moreover, several methods have not reported both image and patient level accuracy. The RDTNet has 35.5 million trainable parameters and it requires approximately 2.4×10^3 seconds of run time for training. During inference, our RDTNet processes each image in 0.025 ms. It is evident from the table that a significant number of approaches exhibit sub-optimal performance at 400× magnification. This can be attributed to the likelihood of incomplete tissue structures in an image patch magnified by 400×, leading to misclassification in a few cases. In future, the learning ability of the proposed model can be enhanced to handle these cases.

6.2.4. Generalization ability of RDTNet

In order to test the generalization capability of RDTNet in comparison to other state-of-the-art CNN methods, we evaluated its performance using an external dataset known as UCSB (Drelie Gelasca, Obara, Fedorov, Kvilekval, & Manjunath, 2009). The UCSB dataset comprises 58 images stained with H&E (26 malignant and 32 benign) and is used for histopathological breast cancer classification. It is noteworthy that the UCSB dataset in this experiment is merely adopted for testing. Table 8 depicts the results of trained RDTNet on the UCSB dataset. It can be observed that our model obtained an accuracy of 97.83% on the UCSB dataset and outperformed state-of-the-art CNN-based models, demonstrating its superiority and better generalization capability on unseen data. This is attributable to learning salient features by the RDTNet during training.

6.2.5. Ablation study

To quantify the influence of the individual components of our proposed transformer layer, we conducted a set of ablative experiments and the results are reported in Table 9. All these experiments were carried out by adopting Xception as backbone. We mainly measured the impact of residual connection (RC) and deformable convolution over key, query and value tensors on the classification performance. It can be observed that the removal of the RC in the RDTL resulted in a reduced classification performance. While the inclusion of RC improved the image-level accuracy by 5.28%, 7.41%, 9.67%, and 4.12% for 40×, 100×, 200× and 400×, respectively. At the patient level, the corresponding improvements are 5.33%, 4.74%, 5.99%, and 7.07% for 40×, 100×, 200× and 400×, respectively. Also, a slight improvement was observed when the deformable convolution was applied over one of the three tensors: query (Q), key (K), and value (V). However, the proposed model achieves the highest classification performance when deformable convolution was employed over all three tensors along with a RC. Specifically, it obtained a higher performance of 99.00%, 98.87%, 98.84%, and 97.80% in terms of image level accuracy for 40×, 100×, 200×, and 400×, respectively, while it yielded a patient-level accuracy of 96.41%, 94.82%, 93.91%, and 91.25%, respectively. This demonstrates the effectiveness of utilizing deformable convolution in the standard SA mechanism.

6.2.6. Interpretation using GradCAM visualization maps

To verify the interpretability of our RDTNet, we employed GradCAM (Selvaraju et al., 2017) technique that produces heat maps, revealing where the model directs its focus during decision-making. Figs. 4 and 5 depict the GradCAM visualization maps obtained by the backbone and proposed RDTNet for benign and malignant samples,

Table 6
Performance comparison with popular attention mechanisms by adopting Xception as backbone.

Accuracy	Attention mechanism	Magnification factor			
		40×	100×	200×	400×
A_{IL}	SE (Hu et al., 2018)	98.16	97.85	98.01	97.21
	BAM (Park et al., 2018)	98.79	97.76	97.25	95.79
	CBAM (Woo et al., 2018)	98.83	97.63	98.0	97.45
	CAB (He et al., 2020)	97.25	97.66	98.17	97.35
	Self-attention (Dosovitskiy et al., 2021)	98.66	97.04	98.01	96.98
	RDTL	**99.00**	**98.87**	**98.84**	**97.80**
A_{PL}	SE (Hu et al., 2018)	95.39	90.84	92.90	89.21
	BAM (Park et al., 2018)	92.81	93.09	92.04	88.56
	CBAM (Woo et al., 2018)	94.53	92.57	93.58	88.91
	CAB (He et al., 2020)	95.73	92.40	92.93	89.60
	Self-attention (Dosovitskiy et al., 2021)	93.68	90.50	92.56	87.15
	RDTL	**96.41**	**94.82**	**93.91**	**91.25**

Table 7
Comparison with state-of-the-art breast cancer classification methods using histopathological images. "–" indicates that the corresponding results are not reported.

Approach	Year	Accuracy	Magnification factor			
			40×	100×	200×	400×
PFTAS + QDA (Spanhol et al., 2015)	2016	A_{PL}	83.8	82.1	84.2	82.0
		A_{IL}	–	–	–	–
Alex Net (Spanhol et al., 2016)	2016	A_{PL}	88.6	84.5	83.3	81.7
		A_{IL}	89.6	85.0	82.8	80.2
Single & Multi Task CNN (Bayramoglu et al., 2016)	2016	A_{PL}	83.08	83.17	84.63	82.1
		A_{IL}	–	–	–	–
Classifier ensembling (Gupta & Bhavsar, 2017)	2017	A_{PL}	87.2	88.22	88.99	85.82
		A_{IL}	–	–	–	–
CaffeNet + LR (Spanhol et al., 2017)	2017	A_{PL}	84.0	83.9	86.3	82.1
		A_{IL}	84.6	84.8	84.2	81.6
InceptionV3 + TL (Benhammou et al., 2018)	2018	A_{PL}	91.5	85.1	86.8	82.9
		A_{IL}	90.2	85.6	86.1	82.5
FV+CSE (Song et al., 2018)	2018	A_{PL}	88.5	90.8	89.2	89.2
		A_{IL}	87.5	88.6	85.5	85.0
Deep active learning (Qi et al., 2019)	2019	A_{PL}	91.26	93.10	92.84	92.30
		A_{IL}	89.29	90.95	91.61	90.36
BreastNet (Toğaçar et al., 2020)	2020	A_{PL}	–	–	–	–
		A_{IL}	97.99	97.84	98.51	95.88
FE-BKCapsNet (Wang, Wang, Li et al., 2021)	2021	A_{PL}	–	–	–	–
		A_{IL}	92.71	94.52	94.03	93.54
DRDA-Net (Chattopadhyay et al., 2022)	2022	A_{PL}	–	–	–	–
		A_{IL}	95.72	94.41	97.43	96.84
MPCS-RP (Chhipa et al., 2023)	2023	A_{PL}	93.26	93.57	92.23	89.57
		A_{IL}	93.45	93.38	92.28	89.81
VGGIN-Net (Saini & Susan, 2022)	2023	A_{PL}	–	–	–	–
		A_{IL}	97.10	96.67	97.16	93.68
ABN-DCN (Krishna et al., 2023)	2023	A_{PL}	–	–	–	–
		A_{IL}	98.4	98.6	98.7	97.8
Ours (RDTNet)		A_{PL}	96.41	94.82	93.91	91.25
		A_{IL}	99.00	98.87	98.84	97.80

Table 8
Performance of RDTNet with UCSB dataset.

Model	Accuracy (%)
MobileNet (Howard et al., 2017)	88.88
DenseNet121 (Huang et al., 2017)	91.27
InceptionV3 (Szegedy et al., 2016)	93.66
AlexNet-BC (Liu et al., 2022)	96.10
RDTNet	97.83

respectively. It can be observed that the RDTNet accurately indicate the lesion regions and ignores the irrelevant regions within the images. While the backbone model in a few cases neglects crucial regions and focuses on unimportant regions. To facilitate the understanding of the interpretation process, potential lesion regions are marked by red arrows in the original images.

7. Conclusion

In this paper, we proposed a novel convolutional transformer called residual deformable attention based transformer network (RDTNet) for improved breast cancer classification. We introduced a residual deformable transformer layer after a pre-trained CNN network with the primary goal of extracting both local and global contextual details from histopathological images. The multi-head deformable self-attention (MDSA) module is the core component in RDTL and is introduced to build wide-range global feature dependencies. Extensive experiments and comparisons on a benchmark dataset indicate the

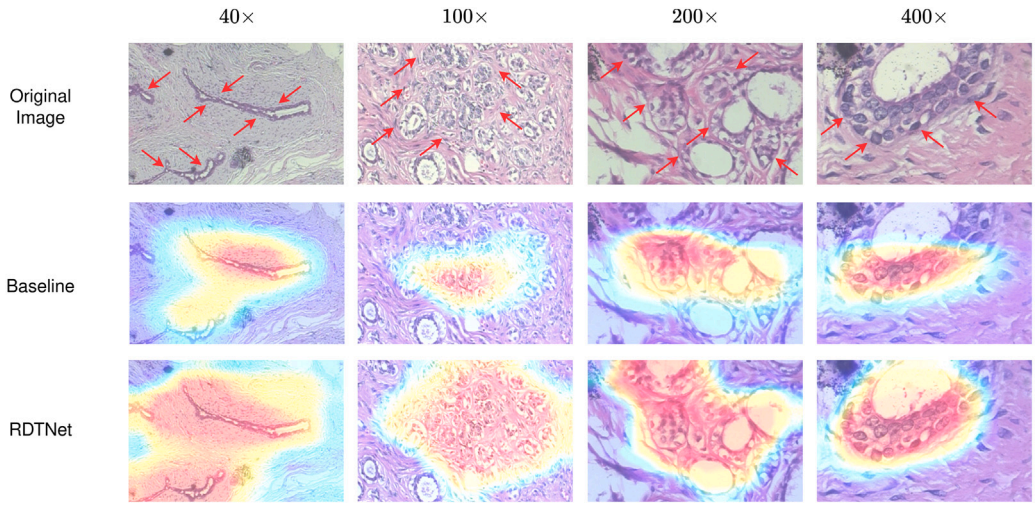


Fig. 4. GradCAM visualization maps of the backbone network and RDTNet for benign samples at different magnifications.

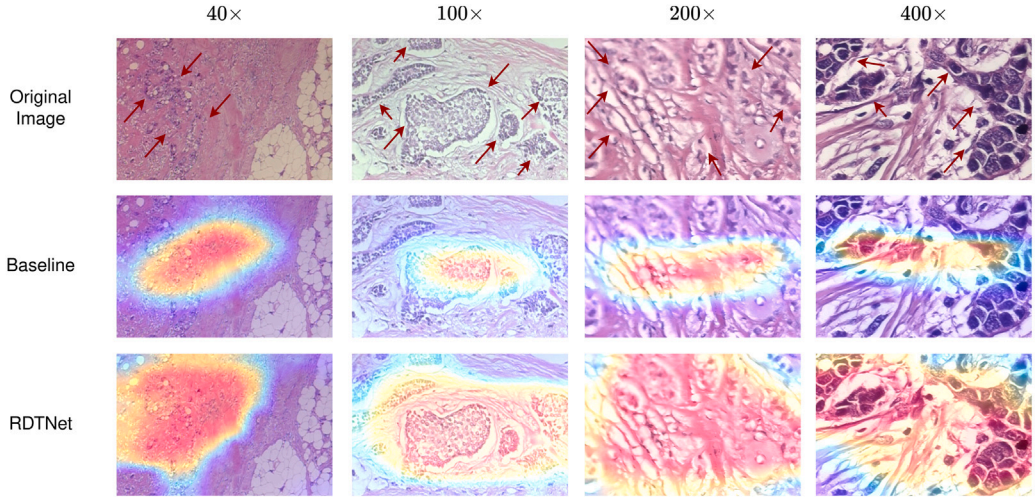


Fig. 5. GradCAM visualization maps of the backbone network and RDTNet for malignant samples at different magnifications.

Table 9
Results of the ablation study on the RDTL component.

Accuracy	Deformable convolution			RC	Magnification factor			
	Q	K	V		40×	100×	200×	400×
A_{IL}	×	×	×	×	92.82	90.63	88.07	91.94
	×	×	×	✓	98.10	98.04	97.74	96.06
	✓	×	×	✓	98.49	97.92	98.50	96.88
	×	✓	×	✓	98.23	97.51	97.17	96.71
	×	×	✓	✓	98.66	98.24	97.63	97.18
	✓	✓	✓	✓	99.00	98.87	98.84	97.80
A_{PL}	×	×	×	×	88.35	86.11	85.73	82.56
	×	×	×	✓	93.68	90.85	91.72	89.63
	✓	×	×	✓	94.19	94.30	92.29	89.10
	×	✓	×	✓	91.74	92.05	92.06	88.05
	×	×	✓	✓	94.09	93.26	91.89	89.68
	✓	✓	✓	✓	96.41	94.82	93.91	91.25

superiority of the proposed RDTNet model compared to state-of-the-art methods. The comparisons with other attention mechanisms further verified the effectiveness of our RDTL. Therefore, the proposed model can be utilized to aid pathologists to cross-check their diagnosis.

In the future, the role of heuristic search optimization algorithms (Mozaffari, Abdy, & Zahiri, 2016; Shahraki & Zahiri, 2021) could be explored for optimized feature representations and hyperparameter selection. Also, this work could be extended to multiclass breast cancer histopathological image classification. It has been observed that the dataset used for validation of our RDTNet is not balanced. Hence, a few potential class imbalance techniques could be explored to further improve the performance at the patient-level.

CRediT authorship contribution statement

Babita: Formal analysis, Data curation, Investigation, Methodology, Software, Writing – original draft. **Deepak Ranjan Nayak:** Conceptualization, Methodology, Software, Validation, Writing – review & editing, Visualization, Supervision, Validation.

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.

Table B.10
List of abbreviations.

Abbreviation	Expansion
BAM	Bottleneck Attention Module
BS	Batch Size
CAB	Category Attention Block
CAD	Computer-Aided Diagnosis
CAGA	Chain-like Agent Genetic Algorithm
CBAM	Convolutional Block Attention Module
CLBP	Completed Local Binary Pattern
CNN	Convolutional Neural Network
CSE	Component Selective Encoding
DSA	Deformable Self-Attention
FV	Fisher Vector
GAP	Global Average Pooling
GLCM	Gray Level Co-occurrence matrix
KPCA	Kernel-based Principal Component Analysis
LBP	Local Binary Pattern
LPQ	Local Phase Quantization
LR	Logistic Regression
MDSA	Multi-head Deformable Self-Attention
MPCS	Magnification Prior Contrastive Similarity
MRI	Magnetic Resonance Imaging
PFTAS	Parameter-Free Threshold Adjacency Statistics
QDA	Quadratic Linear Analysis
RC	Residual Connection
RDTL	Residual Deformable Transformer Layer
RDTNet	Residual Deformable attention-based Transformer Network
RF	Random Forest
SA	Self-Attention
SE	Squeeze-and-Excitation
SVM	Support Vector Mahine
TL	Transfer Learning
ViT	Vision Transformer

Data availability

The source of the code and data has been provided in the article.

Appendix A. Dataset availability

The BreaKHis dataset is publicly available at https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis and the UCSB dataset is openly available at http://bioimage.ucsb.edu/biosegmentation.

Appendix B. Abbreviations

See Table B.10.

References

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. arXiv preprint arXiv:1607.06450.

Bayramoglu, N., Kannala, J., & Heikkilä, J. (2016). Deep learning for magnification independent breast cancer histopathology image classification. In *International conference on pattern recognition* (pp. 2440–2445).

Benhammou, Y., Tabik, S., Achchab, B., & Herrera, F. (2018). A first study exploring the performance of the state-of-the art CNN model in the problem of breast cancer. In *International conference on learning and optimization algorithms: theory and applications* (pp. 1–6).

Chattopadhyay, S., Dey, A., Singh, P. K., & Sarkar, R. (2022). DRDA-Net: Dense residual dual-shuffle attention network for breast cancer classification using histopathological images. *Computers in Biology and Medicine*, *145*, Article 105437.

Chhipa, P. C., Upadhyay, R., Pihlgren, G. G., Saini, R., Uchida, S., & Liwicki, M. (2023). Magnification prior: a self-supervised method for learning representations on breast cancer histopathological images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 2717–2727).

Chiang, T.-C., Huang, Y.-S., Chen, R.-T., Huang, C.-S., & Chang, R.-F. (2018). Tumor detection in automated breast ultrasound using 3-D CNN and prioritized candidate aggregation. *IEEE Transactions on Medical Imaging*, *38*(1), 240–249.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *IEEE conference on computer vision and pattern recognition* (pp. 1251–1258).

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., et al. (2017). Deformable convolutional networks. In *IEEE international conference on computer vision* (pp. 764–773).

Das, D., & Nayak, D. R. (2023). GS-Net: Global self-attention guided CNN for multi-stage glaucoma classification. In *2023 IEEE international conference on image processing* (pp. 3454–3458).

Das, D., Nayak, D. R., & Pachori, R. B. (2023). CA-Net: A novel cascaded attention-based network for multi-stage glaucoma classification using fundus images. *IEEE Transactions on Instrumentation and Measurement*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., & Weissenborn, D. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International conference on learning representations*.

Drelie Gelasca, E., Obara, B., Fedorov, D., Kvilekval, K., & Manjunath, B. (2009). A biosegmentation benchmark for evaluation of bioimage analysis methods. *BMC Bioinformatics*, *10*, 1–12.

Gheflati, B., & Rivaz, H. (2022). Vision transformers for classification of breast ultrasound images. In *44th annual international conference of the IEEE engineering in medicine & biology society* (pp. 480–483).

Guida, F., Kidman, R., Ferlay, J., Schüz, J., Soerjomataram, I., Kithaka, B., et al. (2022). Global and regional estimates of orphans attributed to maternal cancer mortality in 2020. *Nature Medicine*, *28*(12), 2563–2572.

Gupta, V., & Bhavsar, A. (2017). Breast cancer histopathological image classification: Is magnification important? In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 17–24).

Gupta, V., & Bhavsar, A. (2018). Sequential modeling of deep features for breast cancer histopathological image classification. In *IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 2335–23357).

Gupta, V., & Bhavsar, A. (2019). Partially-independent framework for breast cancer histopathological image classification. In *IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 1123–1130).

Gurcan, M. N., Boucheron, L. E., Can, A., Madabhushi, A., Rajpoot, N. M., & Yener, B. (2009). Histopathological image analysis: A review. *IEEE Reviews in Biomedical Engineering*, *2*, 147–171.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., et al. (2022). A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(1), 87–110.

He, A., Li, T., Li, N., Wang, K., & Fu, H. (2020). CABNet: Category attention block for imbalanced diabetic retinopathy grading. *IEEE Transactions on Medical Imaging*, *40*(1), 143–153.

He, Q., Yang, Q., & Xie, M. (2023). HCTNet: A hybrid CNN-transformer network for breast ultrasound image segmentation. *Computers in Biology and Medicine*, *155*, Article 106629.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., et al. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *IEEE conference on computer vision and pattern recognition* (pp. 7132–7141).

Huang, S., Li, J., Xiao, Y., Shen, N., & Xu, T. (2022). RTNet: Relation transformer network for diabetic retinopathy multi-lesion segmentation. *IEEE Transactions on Medical Imaging*, *41*(6), 1596–1607.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *IEEE conference on computer vision and pattern recognition* (pp. 4700–4708).

Ijaz, A., Raza, B., Kiran, I., Waheed, A., Raza, A., Shah, H., et al. (2023). Modality specific CBAM-VGGNet model for the classification of breast histopathology images via transfer learning. *IEEE Access*, *11*, 15750–15762.

Krishna, S., Suganthi, S., Bhavsar, A., Yesodharan, J., & Krishnamoorthy, S. (2023). An interpretable decision-support model for breast cancer diagnosis using histopathology images. *Journal of Pathology Informatics*, *14*, Article 100319.

Liu, M., Hu, L., Tang, Y., Wang, C., He, Y., Zeng, C., et al. (2022). A deep learning method for breast cancer classification in the pathology images. *IEEE Journal of Biomedical and Health Informatics*, *26*(10), 5025–5032.

Mo, Y., Han, C., Liu, Y., Liu, M., Shi, Z., Lin, J., et al. (2023). Hover-trans: Anatomy-aware hover-transformer for roi-free breast cancer diagnosis in ultrasound images. *IEEE Transactions on Medical Imaging*.

Mozaffari, M. H., Abdy, H., & Zahiri, S. H. (2016). IPO: an inclined planes system optimization algorithm. *Computing and Informatics*, *35*(1), 222–240.

Park, J., Woo, S., Lee, J.-Y., & Kweon, I. S. (2018). BAM: Bottleneck attention module. arXiv preprint arXiv:1807.06514.

Połap, D., Jaszcz, A., Wawrzyniak, N., & Zaniewicz, G. (2023). Bilinear pooling with poisoning detection module for automatic side scan sonar data analysis. *IEEE Access*.

Połap, D., Srivastava, G., & Jaszcz, A. (2023). Energy consumption prediction model for smart homes via decentralized federated learning with LSTM. *IEEE Transactions on Consumer Electronics*.

Pramanik, S., Banik, D., Bhattacharjee, D., Nasipuri, M., Bhowmik, M. K., & Majumdar, G. (2018). Suspicious-region segmentation from breast thermogram using DLPE-based level set method. *IEEE Transactions on Medical Imaging*, *38*(2), 572–584.

Qi, Q., Li, Y., Wang, J., Zheng, H., Huang, Y., Ding, X., et al. (2019). Label-efficient breast cancer histopathological image classification. *IEEE Journal of Biomedical and Health Informatics*, *23*(5), 2108–2116.

Rubin, R., Strayer, D. S., Rubin, E., et al. (2008). *Rubin's pathology: clinicopathologic foundations of medicine*.

Saini, M., & Susan, S. (2022). VGGIN-Net: Deep transfer network for imbalanced breast cancer dataset. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *20*(1), 752–762.

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).

Shahraki, N. S., & Zahiri, S. H. (2021). DRLA: Dimensionality ranking in learning automata and its application on designing analog active filters. *Knowledge-Based Systems*, *219*, Article 106886.

Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., et al. (2023). Transformers in medical imaging: A survey. *Medical Image Analysis*, Article 102802.

Sharma, P., Nayak, D. R., Balabantaray, B. K., Tanveer, M., & Nayak, R. (2023). A survey on cancer detection via convolutional neural networks: Current challenges and future directions. *Neural Networks*.

Song, Y., Chang, H., Gao, Y., Liu, S., Zhang, D., Yao, J., et al. (2018). Feature learning with component selective encoding for histopathology image classification. In *International symposium on biomedical imaging* (pp. 257–260).

Spanhol, F. A., Oliveira, L. S., & Cavalin, P. R. (2017). Deep features for breast cancer histopathological image classification. In *IEEE international conference on systems, man, and cybernetics* (pp. 1868–1873).

Spanhol, F. A., Oliveira, L. S., & Petitjean, C. (2015). A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, *63*, 1455–1462.

Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2016). Breast cancer histopathological image classification using convolutional neural networks. In *International joint conference on neural networks* (pp. 2560–2567).

Sung, H., Ferlay, J., & Siegel, R. L. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *A Cancer Journal for Clinicians*, *71*, 209–249.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *IEEE conference on computer vision and pattern recognition* (pp. 2818–2826).

Toğaçar, M., Özkurt, K. B., Ergen, B., & Cömert, Z. (2020). BreastNet: A novel convolutional neural network model through histopathological images for the diagnosis of breast cancer. *Physica A. Statistical Mechanics and its Applications*, *545*, Article 123592.

Wang, P., Hu, X., Li, Y., Liu, Q., & Zhu, X. (2016). Automatic cell nuclei segmentation and classification of breast cancer histopathology images. *Signal Processing*, *122*, 1–13.

Wang, S., Sun, K., Wang, L., Qu, L., Yan, F., Wang, Q., et al. (2021). Breast tumor segmentation in DCE-MRI with tumor sensitive synthesis. *IEEE Transactions on Neural Networks and Learning Systems*.

Wang, Y., Wang, Z., Feng, Y., & Zhang, L. (2021). WDCCNet: Weighted double-classifier constraint neural network for mammographic image classification. *IEEE Transactions on Medical Imaging*, *41*(3), 559–570.

Wang, P., Wang, J., Li, Y., Li, P., Li, L., & Jiang, M. (2021). Automatic classification of breast cancer histopathological images based on deep feature fusion and enhanced routing. *Biomedical Signal Processing and Control*, *65*, Article 102341.

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In *Proceedings of the European conference on computer vision* (pp. 3–19).

Wu, P., Qu, H., Yi, J., Huang, Q., Chen, C., & Metaxas, D. (2019). Deep attentive feature learning for histopathology image classification. In *IEEE 16th international symposium on biomedical imaging* (pp. 1865–1868).

Xu, B., Liu, J., Hou, X., Liu, B., & Garibaldi, J. (2019). Look, investigate, and classify: A deep hybrid attention method for breast cancer classification. In *IEEE 16th international symposium on biomedical imaging* (pp. 914–918).

Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2019). Self-attention generative adversarial networks. In *International conference on machine learning* (pp. 7354–7363).

Zhang, Y., Zhang, B., Coenen, F., Xiao, J., & Lu, W. (2014). One-class kernel subspace ensemble for medical image classification. *EURASIP Journal on Advances in Signal Processing*, *2014*, 1–13.