Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-19667-x

Vision transformer based convolutional neural network for breast cancer histopathological images classification

Mouhamed Laid ABIMOULOUD[1,6] · **Khaled BENSID**[2] · **Mohamed Elleuch**[3,6] · **Mohamed Ben Ammar**[4] · **Monji KHERALLAH**[5,6]

Received: 13 February 2024 / Revised: 1 June 2024 / Accepted: 10 June 2024

Abstract

Breast cancer (BC) is a widespread and lethal cancer affecting women world- wide. Early diagnosis plays a pivotal role in ensuring survival, as late detection can result in a fatal outcome. Convolutional neural networks (CNNs) have made significant contributions to the task of medical imaging modalities and have dis- played promise in addressing this challenge. Recently, the success of the vision transformer (ViT) architecture has encouraged the use of the attention mecha- nism in computer-aided diagnosis (CAD) tasks. However, the ViT is known for its data-intensive nature and a substantial number of parameters and needs power- ful computer resources when training, which often leads to the same performance compared to CNNs. These challenges are particularly evident in tasks involving medical image datasets with complex images and limited data. This problem- atic situation led to the suggestion three of low-weight parameter systems based on convolution and attention techniques: vision transformer base model (ViT), compact convolution transformers (CCT), and lightweight mobile vision transformers (MVIT). These systems are developed by using the BreakHis dataset, which includes images captured at different magnification levels (40x, 100x, 200x, 400x), for both binary and multi classification of breast cancer subtypes. These low-weight hybrid ViT-CNN networks operate directly on input patches and convolution layers, to improve feature extraction and attention layers to train patches in all networks. This approach results in lower training time and fewer parameters while achieving accurate breast tumors classification. The proposed method is based on splitting the input image into patches and then focusing them on the area of cancerous lumps, providing a sequence of linear embedding of these patches as input. Second, we applied a convolution layer directly to the histopathology input patches, with the fewest possible modifications. Finally, we train patches in all transformer encoder layers to evaluate the performance of the classification of breast subtypes. The performance accuracies of our suggested models are 98.64% for VIT, 96.99% for CCT and 97.52% for MVIT. Moreover, the proposed models were compared with state-of-the-art models using the same dataset. Our study demonstrates how convolution and attention mechanisms can minimize computational training resources and decision time, to develop high- performing computer-aided analyses for breast cancer diagnosis. The source codes are accessible at https://github.com/abimouloud/ViT-CNN.

Keywords Breast cancer · Convolutional vision transformer · Computer-aided diagnosis · Light weight vision transformer · Histopathological images · Multi classification

Extended author information available on the last page of the article

Published online: 03 July 2024

1 Introduction

Breast cancer is one of the most common types of cancer in women worldwide, and it is expected to become the most common disease in the coming years. The impor- tance of early diagnosis is underscored by the strong correlation between the stage of diagnosis (tumor size) and survival rate [1]. According to data from The International Agency for Research on Cancer (IARC), it is estimated that breast cancer affected 2.3 million women worldwide in 2020, resulting in 7.8 million women having been diagnosed with breast cancer in the five years prior to the end of 2020. Manual med- ical diagnosis, while important, is time-consuming and places a substantial burden on pathologists. In addition, unskilled pathologists who misdiagnose diseased tissues are possible [2]. Although several imaging techniques, such as mammography, X-rays, ultrasound, and MRI, are used by pathologists for tumor detection, their applications are often costintensive, require skilled practitioners, and exhibit limited specificity. In addition, histopathology images can be very beneficial for early cancer treatment and are more sensitive for identifying and categorizing tissue [3] [4]. In clinical practice, the integration of deep learning techniques and computer-aided diagnostic methods pro- vides specialists and clinicians with more effective speed, efficiency, cost, and precise diagnostic outcomes [2] [5]. Thus, CNNs have a role in medical imaging tasks, including extracting features and classifying tumor lumps [6]. The CNN architecture consists of multiple layers, including convolutions, rectified linear activation functions (Re LUs), pooling layers, fully connected layers, and dropout layers. However, a notable limi- tation of CNNs is their inability to handle rotation and scale in variance inherently, necessitating techniques such as data augmentation, feature extraction, and encod- ing relative spatial information [7]. These challenges emerge due to using the entire breast image rather than focusing on The use of specific regions of interest (ROIs), has certain drawbacks. A vital issue that CNN- based deep learning models face is effec- tively localizing tumor regions [8]. The development of models capable of extracting features from extensive datasets is essential in medical image visual tasks, and both ViT and CNN face challenges in achieving satisfactory performance, especially when confronted with limited data availability and constrained computational resources for building robust neural networks[9].

Vision transformers based on encoder-decoder architecture have recently become the standard model for natural language processing (NLP) [10]. Under high-data regimes, CNNs have been replaced by ViTs in the computer vision domain, which can han- dle high-resolution images [11]. Moreover, the utilization of self-attention layers based on the attention mechanism by splitting the image input into various patches that are subsequently linearly embedded in medical imaging is one of the most critical research areas, and has attracted the attention of many researchers for examina- tion of their advantages, as well as, their functionality in diverse situations with the aim of early detection of breast cancer to limit its spread [9]. This approach offers distinct benefits. First, it excels in capturing long-range relationships among pixels within an image. Second, its adaptive modelling capability, driven by dynamic self- attention weight computations, empowers the network to emphasize pertinent image regions effectively [12]. This contributes to the enhancement of tumor identification and localization. Finally, the capacity of the attention mechanism to generate atten- tion maps provides crucial insight into specific areas of interest within an image [13]. The incorporation of this fundamental transformer approach presents a multitude of advantages compared to convolution methods [6]. Recently, there has been a surge of interest in hybridized Deep Neural Network-based frameworks for data analysis

[14]. Combine the advantages of CNN models with ViT architectures, particularly for high-performance analyses in cancer research. The aim is to address challenges associated with reducing the time and computational resources required for model training. This fusion approach holds promise for more efficient medical image diagnosis. [15].

In this paper, we propose hybrid networks and lightweight models aimed at exploiting the feature extraction capabilities of CNN layers and the attention mechanism of ViTs, to create models that excel in recognizing complex patterns revealing various cancer conditions, with low decision time and computing resources. necessary for the classification of histopathological images. To our knowledge, only a few studies have investigated the application of CNN-based vision transformer techniques in breast cancer classification using histopathological images. Our method is based on attention techniques to partition breast images into patches, focusing on the size of the patch in the tumor region to generate histopathological patches at different magnification levels, such as 40X, 100X, 200X, and 400X, for efficient and accurate classification of breast cancer subtype cells and distinguishing them from healthy tissue cells. Addition- ally, this topic has been approached as binary classification in most previous research. Our suggested methodology for breast cancer diagnosis includes both binary and mul- ticlass classification, going beyond binary classification. This approach allows us to comprehensively address the complexities of sub-cancer type detection and classifica- tion problems and provide more robust solutions. Our study makes four contributions to the literature as follows:

- We examined how well the current self-attention model could classify tumours (benign vs malignant).
- We used histopathological images with different factors for higher resolution to classify subtypes of breast cancer.
- We observed the possibility of using the vision transformer technique with convolutional neural network (CNN) for medical data analysis.
- The combination of Convolutional Neural Networks (CNN) and Vision Transformer (ViT) models aims to reduce computational resources and analysis time for classifying breast cancer in histopathological images, thereby facilitating fast and accurate decision-making.
- The mobile VIT model was explored as a novel approach to develop lightweight ViT models for real-time diagnostic applications.

It is imperative to underscore that our primary objective is not to attain state-of- the-art performance but to delve into the impact of leveraging ViT for classifying histopathology images, particularly in complex datasets. Furthermore, our approach is grounded in integrating attention techniques and convolution layers to classify histopathological images. paving the way for a novel approach with potentially enhanced performance in medical data analysis using computer-aided diagnosis.

We scope extends beyond the theoretical realm, aiming to provide practical advancements. It involves the development of lightweight neural networks based on self- attention convolution layers. These networks are designed to be seamlessly integrated into medical equipment, offering accurate performance and low decision time in real-time diagnosis. The paper is organized as follows: Section 2discusses breast cancer-related works employing the Breakhis dataset. Section 3 describes the proposed methods. Section 4 describes the experimental results. Section 5 discusses the collected results. Section 6 finally closes with the conclusion.

1.1 Related works

The development of Internet of Things (IoT)-based healthcare application services in the medical field is a crucial task. This requires several important factors, includ- ing the collection of data using body sensor networks (BASN)[16], and the secure transmission of this medical data, in order to use this data to develop deep neural net- work algorithms for diagnosis in real time [17]. CNNs are extensively used for medical image diagnosis across various imaging modalities, such as MRI, ultrasound, X-ray, and histopathological images. However, a notable limitation of CNNs is their inabil- ity to handle rotation and scale invariance inherently, as well as imbalanced datasets, necessitating techniques such as data augmentation, feature extraction, and encoding relative spatial information. Saeed Iqbal et al. [18] have addressed these challenges by proposing a framework that can adaptively self-learn using different image modalities, such as X-rays of chest abnormalities and breast cancer classification using histopatho- logical images, and skin lesions. They first apply preprocessing techniques and acquire both manual feature extraction techniques, using a Global–Local Pyramid Pattern (GLPP) based on Local Binary Pattern and radiomics features methodologies, and pertinent features based on pretrained CNNs. Then, they combine manual feature extraction and pretrained CNN models to enhance performance. More recently, deep learning techniques have been utilized for automatic diabetic retinopathy detection. Ghulam Ali et al. [19]. proposed IR-CNN model (InceptionV3 ResNet50 Convolutional Neural Networ) for diabetic retinopathy classification. The authors employed an end- to-end mechanism that utilizes both InceptionV3 and ResNet50 for feature extraction from fundus images of diabetic retinopathy. The features extracted from both models are concatenated and input into the IR-CNN model for classification. To enhance the performance of the proposed model, authours used a preprocessing steps such as data augmentation techniques and histogram equalization intensity normalization to opti- mize image quality.

Several studies have investigated breast cancer classification using the Breast Cancer Histopathological Image (BreakHis) dataset, as summarized in Table 1, that addressing the same topic as ours. Wang Pin et al. [20] proposed an automatic classification method based on deep feature fusion and enhanced routing. They designed a novel network with two parallel channels capable of extracting capsule features and convolution features simultaneously. For deep feature fusion, they employed a novel fusion method that combines semantic features extracted by CNN with spatial features extracted by CapsNet into capsules. The fusion of semantic and spatial information resulted in enhanced features, further improving classification accuracy and stability.

Albashish Dheeb et al. [21] proposed a transfer learning approach based on the Visual Geometry Group's 16-layer deep model architecture (VGG16) to extract high-level features from the BreaKHis dataset.Subsequently, various machine learning models, with a focus on Radial Basis Function Support Vector Machine (RBF-SVM) classifiers, were employed to address different Breast Cancer (BC) classification tasks, including both binary and multiclass classification involving eight classes. The authors removed the last fully connected layers in the VGG16 model. Following this, the extracted fea- tures were classified using a series of heterogeneity classifiers. This study demonstrates the effectiveness of utilizing the features extracted via the VGG16 transfer learning model in combination with polynomial and RBF SVM classifiers.

Al-Jabbar Mohammed et al. [22] presented a hybrid CNN system technique that combines AlexNet and GoogLeNet for feature extraction, followed by machine learning

Table 1 Comparative of the Literature's Review

Authors	Method	Data	calssification
Saeed Iqbal et al. [18]	feature extraction (GLPP-LBP-RF) pre-trained CNNs	BreakHis NIH X-ray ISIC	BC covid19 skin
Ghulam Ali et al. [19]	IR-CNN for feature extraction based on (InceptionV3 and ResNet50)	open source dataset	diabetic retinopathy
Wang Pin et al. [20]	CapsNet model based on fusion of semantic and spatial information	BreakHis	BC
Albashish Dheeb et al. [21]	VGG16 model in combination with polynomial and RBF, SVM classifiers	BreakHis	BC
Al Mohammed et al. [22]	feature extraction (AlexNet and GoogLeNet) SVM classifiers	BreakHis	BC
A Muhammad Sadiq et al. [23]	FabNet model to learn fine-to-coarse structural and textural features	BreakHis NCT-CRC-HE-100	BC
Hao Yan et al. [24]	feature extraction (DenseNet201) SVM classifiers	BreakHis	BC
S Mahati Munikoti et al. [25]	feature extraction (Inception, ResNetV2, ResNet5), LSTM classifiers	BreakHis	BC
Mahmud M et al. [26]	feature extraction (AlexNet, Vgg16) SVM classifiers	BreakHis	BC
Abunasser Basem et al. [27]	fine-tune deep learning models	BreakHis	BC
Sriwastawa Asmi et al. [30]	fine-tune vision transformer models	BreakHis IDC	BC
Our Approach	feature extraction (CNN-ViT) models	BreakHis	BC

using SVM to achieve precise classification. Fusion feature vectors were generated by integrating CNN with custom features. The authors proposed a method that hybridizes CNN (AlexNet and GoogLeNet) models to extract and classify features using the support vector machine (SVM). Consequently, all Breast Cancer (BC) datasets were diagnosed using both AlexNet + SVM and GoogLeNet + SVM. In the second pro- posed method, all BC datasets were diagnosed using an Artificial Neural Network (ANN) based on a combination of CNN features with handcrafted features extracted using the fuzzy colour histogram (FCH), local binary pattern (LBP), and gray level co-occurrence matrix (GLCM), collectively referred to as fusion features. Finally, the fusion features were fed into an ANN for classification.

Amin Muhammad Sadiq et al. [23]. introduced the FabNet model, which leverages CNNs and integrates datasets from BreakHis and NCT-CRC-HE-100 at various magnification levels. The model adopts an accretive network architecture, amalgamating hierarchical characteristics to achieve a high level of classification accuracy. FabNet model proposes to learn fine-to-coarse structural and textural features of multiscale histopathological images through its accretive network architecture, which consolidates hierarchical feature maps to attain significant classification accuracy. The authors demonstrate that a lightweight network architecture with fewer parameters proposed leads to improved classification accuracy by incorporating deep and close integration, thereby finely combining features across layers and expanding upon the conventional convolutional neural network architecture.

Hao Yan et al. [24]. demonstrated a method based on gray-level co-occurrence matrix (GLCM) characteristics and deep semantic features. They used a pre-trained DenseNet201 as the foundational model and applied support vector machines (SVM) to classify data by extracting deep semantic features from the last dense block's con- volutional layer features, combined with three-channel GLCM features.

Srikantamurthy Mahati Munikoti et al. [25] present a hybrid model for classifying breast cancer subtypes, employing convolutional neural networks (CNN) and long short-term memory recurrent neural networks (LSTM RNN). The authors propose a hybrid CNN-LSTM model comprising two main modules: the CNN input shape and an independent RNN module. The CNN passes through a pre-trained transfer learning model (Inception, ResNetV2, ResNet50) until it reaches the final convolutional layer, which contains the bottleneck features. Meanwhile, the independent RNN module con- sists of 2 LSTM layers. The outputs of both modules are merged using element-wise multiplication, and this output is then fed into the classification layer.

Mahmud M et al. [26]. employed transfer learning and deep feature extraction methods, utilizing AlexNet for additional fine-tuning. They modified pre-trained CNN mod- els AlexNet and Vgg16 for feature extraction and employed support vector machines (SVM) for feature classification.

Abunasser Basem et al. [27] proposed a deep learning model with additional fine- tuned deep learning models, including Xception, InceptionV3, VGG16, MobileNet, ResNet50, and BCCNN.

Simultaneously, ViTs have gained challenge in the realm of computer vision outperforming CNNs in tasks that involve medical image classification [28]. For this task, in 2021 Matsoukas Christos et al. [6] proposed investigating whether Vision Trans- formers (ViTs) could serve as viable replacements for CNNs in medical imaging tasks.

Their research aimed to determine that ViTs can achieve the same performance levels comparable to CNNs when working with small medical datasets.

In 2022, Henry Emerald et al. [9] observed the application of transformer architec- tures across various imaging modalities in medical imaging. Their research focuses on a comparative analysis of the performance of both architectures, exploring their strengths, weaknesses, and performance in various scenarios. However, their obser- vation revealed a deficiency in clear and detailed comparisons between transformer VIT and CNN counterparts. In 2023, He Kelei al. [29] aimed to highlight the applica- tions of transformers in medical image analysis. They underscored that many existing transformer-based approaches can be easily adapted to address a variety of medical imaging challenges with minimal modifications.

In recent studies, that focused on the combination of CNN and ViT models in breast cancer histopathological images classification, Sriwastawa Asmi et al. [30]. proposed a comparison of eight Vision Transformer (ViT) models using BreakHis and IDC datasets for binary classification. Their experiments initially involved training the models from scratch. Subsequently, the models trained on the BreakHis dataset were considered pre-trained models and then fine-tuned on the IDC dataset. In their exper- iments, the authors calculated evaluation metrics such as the number of epochs, the time taken to train the model, accuracy, the number of parameters in the model, speci- ficity, precision, recall, F1-score, and ROC-AUC score.

While previous studies have made significant progress in classifying breast cancer histopathological images, the need to achieve high performance and accuracy has led to the application of various preprocessing methodologies and techniques, such as the global–local pyramid model (GLPP), a fuzzy color histogram (FCH), local binary model (LBP), and gray level co-occurrence matrix (GLCM). Additionally, hybrid approaches integrating CNNs with LSTM and SVM classifiers, as well as transfer learning techniques using models like VGG16 and fine-tuning, were used.

Furthermore, recent works have proposed models combining CNNs with Vision Trans- former (ViT) models, which often have a high number of parameters and high complexity, requiring significant computational resources and long training times. This paper aims to fill these gaps by introducing hybrid lightweight CNN-ViT models. These models capitalize on the feature extraction capabilities of CNN layers and the attention mechanism of ViT while maintaining a lower number of parameters and shorter training times without applying any preprocessing techniques and training all models from scratch. Through meticulous comparisons with similar recent studies, the novelty and impact of the proposed methodology can be effectively established in the field of breast cancer histopathological image classification.

2 Materials and methods

2.1 Histopathological images

Histopathology is the microscopic examination of samples to determine the location and classification of cancer [31]. Histopathologists evaluate the regularities of cell shapes and tissue distributions visually during histological image processing to diag- nose cancer [32]. Nuclear pleomorphism, tubular development, and mitotic activity are the three criteria used to grade breast cancer [33]. Such histopathology studies have been widely used for cancer detection and classification to ascertain the the degree of malignancy and whether tissue areas are malignant [34].

Table 2 The BreakHis dataset by magnification level and class

MAGNIFICATION LEVEL	400×	200×	100×	40×	TOTAL
Malignant	1232	1390	1437	1370	5429
Benign	604	623	644	625	2496
TOTAL	1836	2013	2081	1995	7925

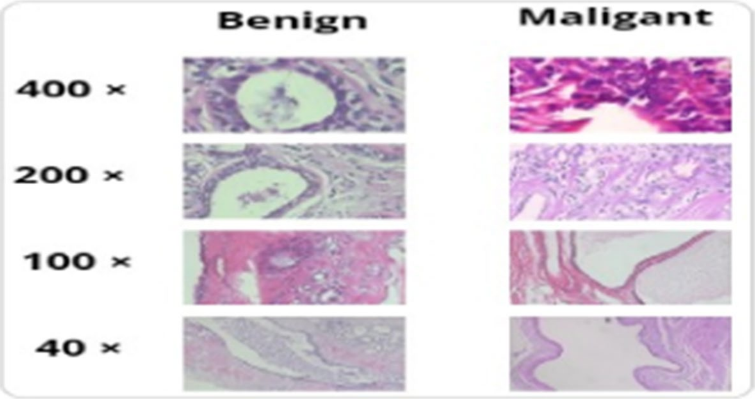


Fig. 1 Images from the BreaKHis dataset

2.2 Dataset

The BreakHis dataset introduced by Spanhol et al. [35]. is a public dataset available at [36]. The dataset contains a total of 7925 microscopic biopsy images, featuring 2,496 benign and 5429 malignant breast tumors. Among benign breast tumors, there are four distinct histopathology subtypes: adenosis (AD), fibroadenoma (FI), phyl- lodes tumor (PH), and tubular adenoma (TU). The malignant breast tumors (breast cancer) are classified into four subtypes: ductal carcinoma (DU), lobular carcinoma (LO), mucinous carcinoma (MU), and papillary carcinoma (PA) [37]. These images were captured at magnifications of 40×, 100×, 200×, and 400×, with a resolution of 700×460 pixels having 3-channel RGB (Red–Green–Blue) True Color representation, providing 24-bit color depth with 8 bits per color channel. Table 2 details the number of benign and malignant image samples at each magnification level from the BreakHis dataset, while Fig. 1 provides sample images from this dataset.

2.3 The proposed methods

This study employed several methods, including a self-attention transformer VIT model, a vision transformer based convolution CCT model, and a lightweight Mobile Vision Transformer model MVIT. These methods were utilized for binary and multiclass classification of breast histopathology subtype tissues.

2.3.1 Vision transformer VIT

The self-attention vision transformers [11] were inspired by the original transformer model employed in natural language processing (NLP) [38]. In the vision transformer model, as presented in Fig. 2 and detailed in the algorithm 1. The input 2D image is represented as $X \in R^{H \times W \times C}$, where H stands for height, W for width, and C for the number of channels. This mechanism based on split image into a sequence of image patches that are denoted as $N_{xp} \in R^{N \times P2C}$, with N demonstrated in Eq. 1.

$$N = \frac{HW}{P^2} \tag{1}$$

and P being the patch size. Each patch is then flattened and projected into a higher-dimensional space D using a trainable linear projection, resulting in embedded patch images known as patch embedding. Additionally, a learnable class embedding Xclass is included in the sequence of embedded patches, serving as inputs for the transformer encoder block. The transformer encoder block employs a forward connection that com- bines the initial input with the output of multi-head attention. The combined output undergoes normalization and is processed through an MLP layer, which comprises a dense layer with dropout. The resulting outputs are then forwarded to the MLP head layer. Ultimately, the MLP head utilizes the output from the transformer encoder lay- ers to generate a probability distribution of labels, for the final prediction of the image class [30].

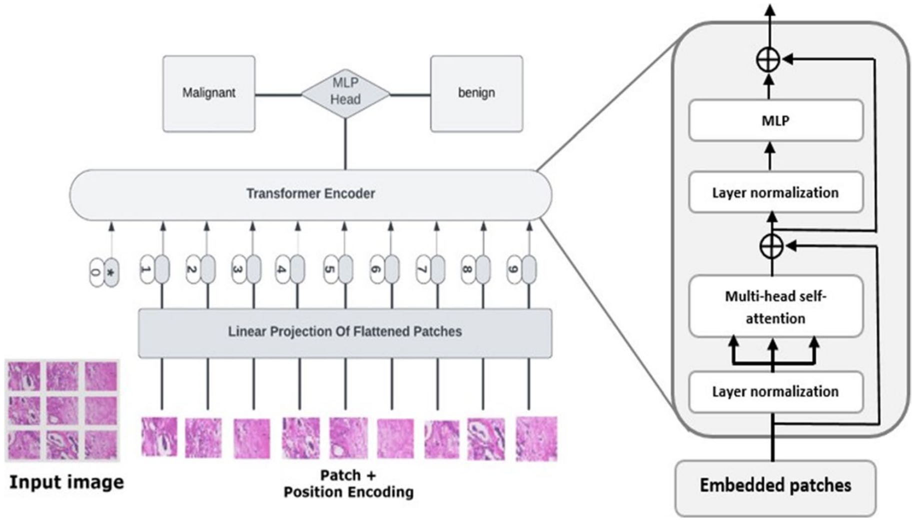


Fig. 2 Vision transformer self-attention classification architecture

Algorithm 1 Vision Transformer (ViT)

```
procedure ViT Classifier
    Input: $X \in R^{H\times W\times C}$    ▷ Input image tensor where H, W, and C represent height, width, and number of channels, respectively
    Compute N, the number of patches, using $N = \frac{HW}{P^2}$ where P is the patch size    ▷ Equation 1
    Encode patches using PatchEncoder
    for i ← 1 to $N_{\text{transformer\_layers}}$ do
        x1 ← Layer normalization on encoded patches
        Compute multi-head attention on x1 with num heads heads and projection_dim key dimension
        x2 ← Add attention output and encoded patches    ▷ Skip connection 1
        x3 ← Layer normalization on x2
        x3 ← Multi-Layer Perceptron (MLP) on x3 with hidden units transformer units and dropout rate
        Update encoded patches with Skip connection 2
    end for
    Perform Layer normalization on encoded patches
    Flatten the representation
    Apply dropout regularization with dropout rate
    Apply MLP to the representation with hidden units mlp head units and dropout rate
    Classify outputs using MLP Head layer
    Output: Y    ▷ Class predictions
end procedure
```

The proposed ViT model is designed with specific parameters, as outlined in Table 3. By using a smaller patch size of 14×14, instead of the 16×16 patch size pro- posed in the original ViT paper, we aimed to enable more focused attention and better capture intricate details within the input patches. The decision to use 8 layers was made to strike a balance between model complexity and performance, resulting in optimal classification accuracy. The total number of parameters in our model is 36376521, which were carefully optimized to ensure efficient training. These parameter choices were selected to find the right balance between model performance and computational resource requirements. Overall, our design decisions for the ViT model focused on achieving attention efficiency, managing model complexity, and optimizing computational resources. These factors collectively contribute to the effectiveness of our proposed approach for breast cancer histopathological image classification.

Table 3 Details of vision transformer model variants

Model	Image size size	patch size	Layers	Heads	parameters
Our proposed adapted VIT	224×224	14	8	4	36,376,521

2.3.2 Compact convolutional transformers CCT

The self-attention mechanism in the Vision Transformer (ViT) model enables captur- ing long range dependencies within the histopathological images, which is crucial for identifying subtle patterns that indicate different breast cancer subtypes. However, ViT models are known for their data-intensive characteristics and considerable num- ber of parameters. Consequently, the performance improvements come at the cost of a large model size (due to the vision transformer layers), requiring substantial com- putational resources for model training. To minimize the complexity of ViT, we have chosen to use the CCT model. CCT leverages convolutional layers with an atten- tion mechanism, offering a complementary approach to feature extraction and spatial information processing. The application of convolutional layers can help multi-head attention layer in ViT to extract features with fewer layers, potentially leading to a more efficient and lightweight architecture. the proposed CCT model seeks to address the limitations of the data-intensive and computationally expensive ViT architecture, while maintaining the benefits of the attention mechanism for capturing long-range dependencies in histopathological images.

An overview of the compact convolutional transformer [15] architecture is detailed in Fig. 3, and detailed in the algorithm 2. The most significant modification in the model is the replacement of the patch and embedding block in VIT with a basic con- volution block that includes a standard structural convolution, ReLU activation, and a max pool layer. Given a picture with the following dimensions $x \in R^{H\times W\times C}$

$$x_0 = \mathrm{MaxPool}(\mathrm{ReLU}(\mathrm{Conv2d}(x))) \tag{3}$$

Furthermore, by using this convolution block, the model gains flexibility over a model like ViT by not being restricted to input resolutions strictly divisible by the predefined patch size. Additionally, the convolution and max pool procedures can overlap, which could improve performance by infusing inductive biases. This permits the model to retain local spatial information. In this study, we used CCT architecture model with 2 transformer encoder layers, 2 MLP heads, and 2 convolution layers with a 3×3 kernel size.

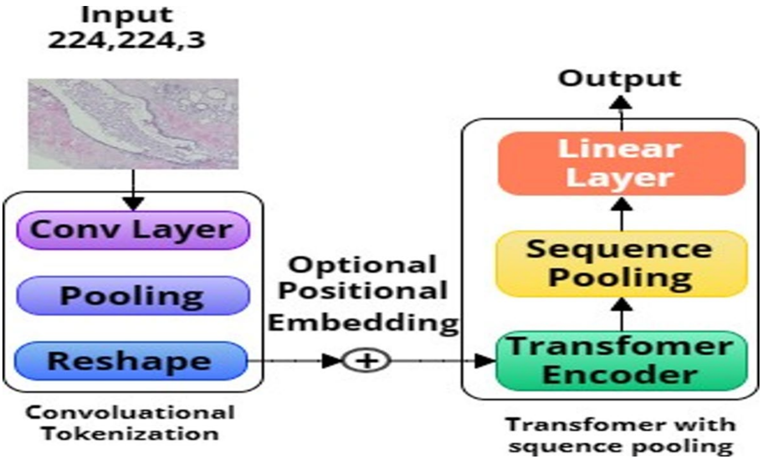


Fig. 3 Architecture of compact convolutional transformers

Algorithm 2 Compact Convolutional Transformer (CCT)

```
procedure CCT_MODEL
    Input: Image x $\in$ $R^{H\times W\times C}$    ▷ where H is height, W is width, and C is number of channels
    encoded_patches ← Apply Convolutional layers to Input
    if positional_emb then
        pos-embed, seq_ length ← Positional embedding based on Convolutional layers
        positions ← Generate positions from 0 to seq_length
        position_embeddings ← Calculate positional embeddings
        encoded_patches ← encoded_patches + position_embeddings
    end if
    dpr ← Calculate Stochastic Depth probabilities
    for i in range(transformer_layers) do
        x1 ← Apply Layer Normalization to encoded_patches
        attention_output ← Apply Multi-Head Attention to x1
        Apply Stochastic Depth with dpr[i] to attention_output
        x2 ← Add attention_output and encoded_patches (Skip Connection 1)
        x3 ← Apply Layer Normalization to x2
        x3 ← Apply Multi-Layer Perceptron (MLP) to x3
        Apply Stochastic Depth with dpr[i] to x3
        encoded_patches ← Add x3 and x2 (Skip Connection 2)
    end for
    Apply sequence pooling to encoded_patches
    logits ← Apply Dense layer for classification
    Output: Y    ▷ Class predictions
end procedure
```

The CCT model parameters are shown in Table 4. The objective was to reduce the number of parameters compared to the standard ViT model Table 3. In the CCT model, we utilized 2 convolutional layers, which can effectively replace the 6 trans- former layers required in the original ViT architecture. This design choice allowed us to use just 2 attention layers, resulting in a significant reduction in the overall model complexity. The total number of parameters in the CCT model is 407 M, which was carefully optimized to balance performance and computational efficiency. This param- eter count is substantially lower than the parameter-intensive ViT model, while still maintaining competitive classification accuracy.

2.3.3 The mobile vision transformer MVIT

Despite the potential advantages of the CCT model in combining CNN and ViT lay- ers, it also has some complications, such asrequire extensive data augmentation and regularization to avoid overfitting, difficulty handling complex data, and long training times. To address these limitations, the lightweight mobileViT (MVIT) model is chosen for its lightweight design, enabling efficient processing of large-scale histopathologi- cal images while maintaining competitive performance

Table 4 Details of the compact convolutional transformer model variants

Model	Conv-Layers	Attention-Layers	Heads	kernel size	Params
CCT	2	2	2	3×3	407 M

to optimize and integrate into task-specific networks. By leveraging the advantages of lightweight CCT and attention mechanism of ViT, the MVIT model aims to strike a balance between model complex- ity and classification performance, making it a promising approach for effective and efficient classification of histopathological images of breast cancer.

For this task, MobileViT [39](or MVIT) displays stronger task level generaliza- tion features than light-weight CNN models (MobileNetv1[40], MobileNetv2[41], MobileNetv3[42]). As shown in Table 5, the main components of the MViT architec- ture are a stride 3×3 Depth-wise convolution 2D layer and Swish as an activation function, followed by the MobileNetv2 (or MV2) block and the MVIT block and 1×1 convolution for modifying the number of channels.

The MV2 blocks in the MViT network play a crucial role in down-sampling by extracting local features from the input image feature $x \in R^{H \times W \times C}$, where R repre- sents the set of real numbers. Here, x is a three-dimensional tensor with dimensions H, W, and C, representing the height, width, and number of channels of the ten- sor, respectively. The process involves expanding the low-dimensional compressed data into higher dimensions and filtering the data with depth-wise separation. This structural design employs compact tensor data in the reasoning process, which subse- quently reduces the demand for embedded hardware for main memory access, leading to improved reaction performance [43].

The MVIT block, on the other hand, is divided into three components: the local information coding module, the global information coding module, and the feature fusion module. The feature tensor is then projected into a high dimensional space $\mathrm{X_L} \in \mathrm{R}^{\mathrm{H \times W \times d}}$, where d represents the number of dimensions, and d > C. This tensor X_L is passed through the Global Representation module. Figure 4 depicts its structure and detailed in algorithm 3.

Following that, $\mathrm{X_{Unfold}}$ is fed into L-stacked Transformers for global information encoding, where the attention mechanism is used to compute inter-column pixel atten- tion, yielding $\mathrm{X_G} \in \mathrm{R}^{\mathrm{P \times N \times d}}$ (P = wh is the number of pixels in the patch with height h and width w, and N is the number of patches).

Finally, the Fold operation is used to generate $\mathrm{X_F} \in \mathrm{R}^{\mathrm{H \times W \times d}}$, which is the same size as X_L. Because each pixel in the Transformer's output contains information from all pixels in the input feature map, the receptive field can be expanded to H W, MVIT block may fully extract image feature information using this approach with less parameters [44].

Table 5 The general structure of MobileViT

Layer	Output size	Output stride	Repeat	XXS	XS	XS
Image	256×256	1				
Conv-3×3, ↓ 2	128×128	2	1	16	16	16
MV2			1	16	32	32
MV2, ↓ 2	64×64	4	1	24	48	64
MV2			2	24	48	64
MV2, ↓ 2	32×32	8	1	48	64	96
MobileViT block (L = 2)			1	(d = 64)	64 (d = 96)	96 (d = 144)
MV2, ↓ 2	16×16	16	1	64	80	128
MobileViT block (L = 2)			1	64 (d = 64)	80 (d = 120)	128 (d = 192)
MV2, ↓ 2	8×8	32	1	80	96	160
MobileViT block (L = 3)			1	80 (d = 96)	96 (d = 144)	160 (d = 240)
Conv-1×1				320	384	640
Global pool	1×1	256	1	1000	1000	1000
Linear						
Nerwork Parameters				1.3 M	2.3 M	5.60 m

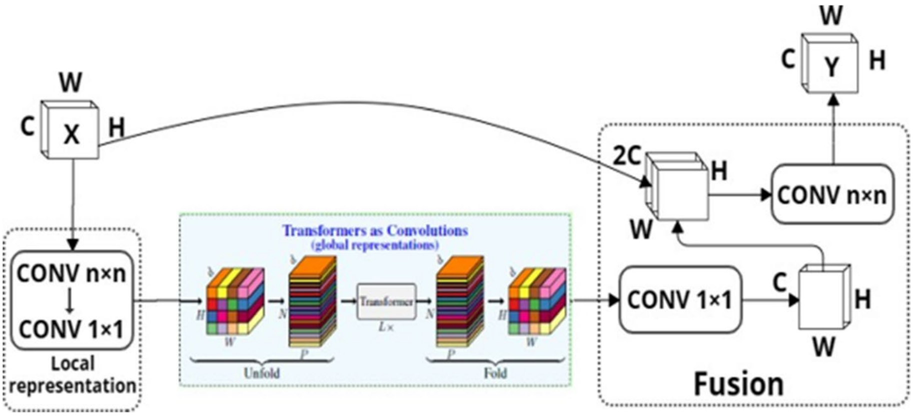


Fig. 4 MobileVit block architecture

3 Experiment and results

3.1 Dataset splitting

This paper aimed to use histopathological images at different magnifications. The 7925 microscopic biopsy images that made up the study's dataset were divided into four groups based on their magnification 40×, 100×, 200×, and 400×. In binary classifi- cation, which was divided into two classes for each magnification level of malignant.

Algorithm 3 Mobile Vision Transformer (MVIT)

```
Input: Input feature tensor $x \in R^{H\times W\times C}$
Parameters: Number of dimensions d, Patch size P, Number of patches N,
  Number of pixels P, Height H, Width W
procedure MVIT(x)
    Extract local features using MV2 blocks
    Expand low-dimensional data into higher dimensions
    Apply depth-wise separation to filter the data
    Compact tensor data for reasoning, reducing demand for memory access
    Local Information Coding Module:
    Project feature tensor into high dimensional space $X_L \in R^{H\times W\times d}$ (d > C)
    Global Information Coding Module:
    Project $X_L$ through Global Representation module
    Unfold $X_L$ into patches $X_{\text{Unfold}} \in R^{P\times N\times d}$ (P = wh, N patches)
    Apply L-stacked Transformers for global information encoding
    Compute inter-column pixel attention using attention mechanism
    Obtain global feature tensor $X_G \in R^{P\times N\times d}$
    Feature Fusion Module:
    Fold operation to generate $X_F \in R^{H\times W\times d}$
    Receptive field expanded to H × W for full image feature extraction
    Output: Feature tensor $X_F \in R^{H\times W\times d}$
end procedure
```

and benign tumours. The multiclassification task involves classifying images into eight subtypes of tumors (AD, DU, FI, LO, MU, PA, PH, and TU) for each magnification level. The dataset was split into 80% for the training phase, while the remaining 20% were used for testing. Furthermore, within the training set, a further division was made, with 80% allocated for the training process, and the remaining 20% allocated for validation to ensure that the model was not overfiting. The models were developed in a computer workstation HP Z8 G4:

- Memory (RAM): 96.00 GB
- Processor: Intel(R) Xeon(R) Silver 4108 CPU @ 1. 80 GHz 1. 80 GHz.
- Graphic Processing Unit: (GeForce RTX 2080 Ti, GeForce RTX 3090)
- System type: 64-bit operating system, × 64 processor.
- Python 3.11 programming language.

Subsequently, all codes executions were carried out utilizing the execution environments provided by Google Colab Pro.

3.2 Dataset preprocessing

In this work, we carefully considered the hyperparameters to strike an optimal balance between models performance and computational complexity. However, practically, the selection of input image sizes and patch sizes plays a crucial role in the performance and efficiency of vision transformer architectures. For the CCT and MobileViT MVIT models, we followed the original Transformer (ViT) method by Vaswani et al. [11], resizing input histopathological images to 256×256 input image into 16×16 patches yielded 256 input patches. This patch size, combined with convolutional feature extraction techniques employed in these models, demonstrated high performance results. However, for our proposed self-attention ViT model, we opted for a smaller patch size of 14×14 while splitting input images of 224×224 pixels, which also yielded 256 input patches, the same as the CCT and MVIT models. The rationale behind the 14×14 patch size is that it allows the transformer encoder's attention mechanism to oper-ate more efficiently and capture intricate details within each patch's pixels. Although smaller patch sizes increase the number of patches and computational complexity, we found that the 14×14 patch size provided an optimal trade-off between performance and computational efficiency for the proposed self-attention ViT model. Additionaly, it is essential to note that the BreakHis dataset is imbalanced. Addressing this imbal- ance is crucial to developing a robust and unbiased classification model, especially in the medical field. Failure to balance the dataset can result in poor generalization and a higher likelihood of misclassification, particularly for minority classes. Misclassifying minority classes can have more severe consequences in a medical context, as these often represent patients at higher risk. To solve this problem, first, one of the common approaches to mitigate data scarcity is by applying the data augmentation technique. This approach aligns with the idea that pathologists can interpret breast histopathol- ogy images from different angles, sizes, and orientations with sizes detailed in Table 6. Second, for multiclassification tasks, we use a three-K-fold cross-validation training strategy. The hyperparameter values used for training are detailed in Table 7, while Fig. 5 illustrates the block diagram of our experimental approach.

Table 6 Data augmentation technique

Data Augmentation Technique	Value
Rotation range	5
Width shift range	0.1
Height shift range	0.1
Zoom range	0.001
Fill mode	0.001

Table 7 Hyperparameter

Hyperparameter	Value
Batch size	16 32 64
Number of epochs	70 80 90
Optimizer	Adam
Loss function	Binary Cross-entropy Categorical Cross-entropy

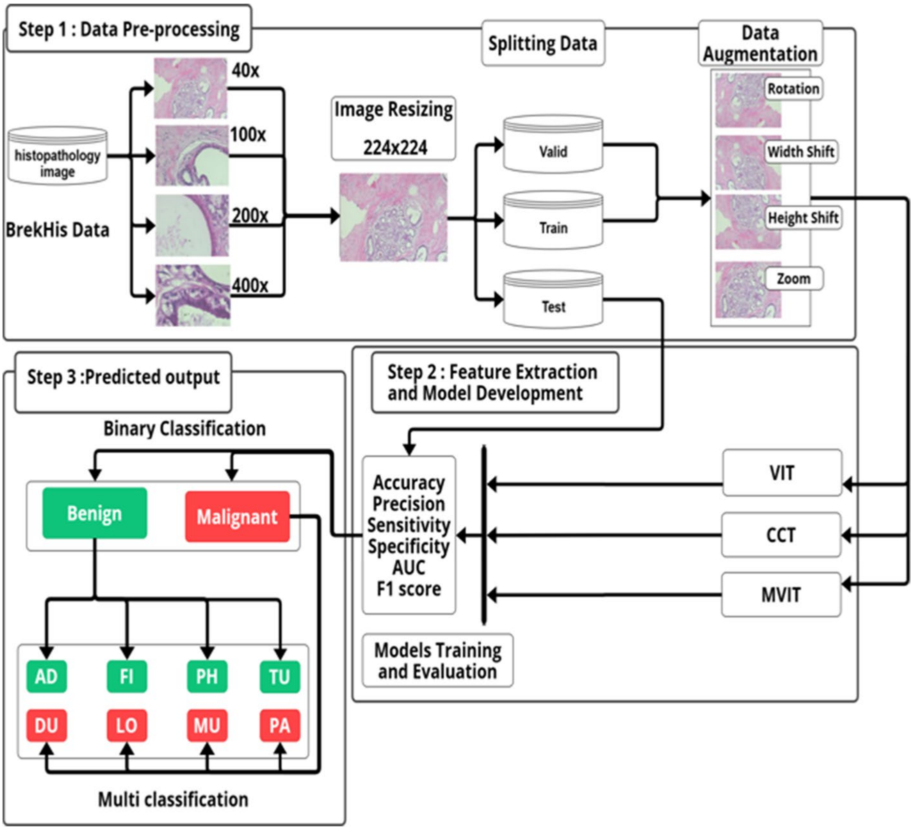


Fig. 5 Breast cancer diagnosis block diagram

3.3 Evaluation metrics

The systems' performance was assessed using accuracy, precision, sensitivity, speci- ficity, AUC, F1 score, Total Training Time in second and Average Training Time per epoch. The variables in these equations were obtained from the confusion matrix generated by the systems. The confusion matrix provides information on correctly classified images (True Positives [TP]are classified correctly as malignant breast can- cers tumors and True Negatives [TN] are classified correctly as benign breast cancers tumors and False Positives [FP]are benign tumors classified as malignant breast can- cers tumors and False Negatives [FN]are malignant tumors classified as benign breast cancers tumors) [45], The Equations are detailed as follows:

$$\text{Accuracy}(\%) = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3}$$

$$\text{Sensitivity}(\%) = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

$$\text{Precision}(\%) = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$Specificity(\%) = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{6}$$

$$F1score(\%) = \frac{2 \times [\text{Precision} \times \text{Recall}]}{\text{Precision} \times \text{Recall}} \tag{7}$$

$$\text{AUC}(\%) = \frac{Sensitivity}{Specificity} \tag{8}$$

$$Total - \text{Training} - \text{Time(S)} = \text{endtime} - start - \text{time} \tag{9}$$

$$\text{Avg} - \text{Training} - \text{Time} - \text{per} - \text{epoch} = \frac{Total\_Training\_Time}{Num\_epochs} \tag{10}$$

3.4 Results

This section presents the results of various experiments conducted using the pro- posed vision transformer models. The experiments are organized into binary and multi-classification tasks in order to address the research questions.

3.5 Experiment 1: RQ1. How can we employ a highly efficient and accurate technique to utilise vision transformer models for early breast *cancer* diagnosis? as follows.

3.5.1 Binary classification

The proposed vision-transformer models performed well for binary classification (benign or malignant) of histopathological breast cancer images, as shown in Table 8. All three models used were based on vision transformers, and convolution architectures performed consistently across all the performance metrics. As a result, the proposed models achieved high accuracy, precision, sensitivity, specificity, AUC, and F1 score at different magnifications on the Brekhis dataset. This provides strong evidence that vision transformer can be used to effectively improve DL approaches for obtaining breast histopathology images, thereby improving early diagnosis techniques for breast cancer.

3.5.2 Multi classification

One of the primary objectives of this paper was to enhance the performance specif- ically in the classification of breast cancer subtypes after binary classification. to investigate the performance of the proposed transformer models in multi classification tasks across the entire dataset, encompassing different magnification levels, including 40×, 100×, 200×, and 400×. A three K-fold cross-validation technique is used to obtain the multi classification results. This evaluation approach ensures robustness and reliability in the assessment of models performance across different magnification levels. The best performances of the proposed systems are presented in Table 9.

The results obtained from the first experiment indicate that the vision transformer significantly enhances the strong performance of deep learning approaches in medical image analysis and breast cancer diagnosis.

Table 8 Binary classification Performances metrics for the different models at various magnifications

Magnification		Accuracy	Precision	Sensitivity	Specificity	AUC	F1 score
VIT	400X	98.56%	98.35%	98.35%	99.47%	98.45%	98.56%
	200X	97.87%	98.70%	98.70%	99.85%	98.27%	97.87%
	100X	94.66%	95.18%	95.18%	98.77%	94.91%	94.66%
	40X	96.50%	95.88%	95.88%	99.34%	96.18%	96.50%
CCT	400X	93.12%	93.12%	85.95%	97.38%	93.12%	93.12%
	200X	94.33%	95.68%	87.2%	96.63%	95.0%	94.33%
	100X	92.26%	93.52%	83.09%	95.52%	92.88%	92.26%
	40X	97.81%	97.81%	95.2%	98.93%	97.81%	97.81%
MOBILE VIT	400X	92.03%	89.91%	89.91%	97.17%	90.85%	92.03%
	200X	96.91%	97.32%	97.32%	99.69%	97.11%	96.91%
	100X	92.00%	92.49%	92.49%	97.03%	92.24%	92.00%
	40X	94.33%	93.34%	93.34%	98.30%	93.81%	94.33%

Table 9 Multi classification performance metrics for the different models at various magnifications

Magnification		Accuracy	Precision	Sensitivity	Specificity	AUC	F1 score
VIT	400x	89.29%	89.53%	89.29%	98.2%	97.02%	89.03%
	200x	90.86%	90.93%	90.86%	98.44%	96.8%	90.73%
	100x	91.88%	92.1%	91.88%	98.64%	94.13%	91.81%
	40x	94.8%	94.97%	94.8%	99.14%	97.73%	94.79%
CCT	400x	75.79%	74.58%	75.79%	95.93%	93.74%	73.77%
	200x	76.80%	77.63%	76.80%	95.90%	94.21%	74.34%
	100x	83.51%	83.39%	83.51%	97.22%	97.08%	83.02%
	40x	84.63%	84.8%	84.63%	97.4%	97.1%	84.4%
MOBILE-VIT	400x	86.77%	86.21%	86.77%	97.71%	96.87%	86.18%
	200x	83.96%	84.46%	83.96%	97.17%	97.37%	83.49%
	100x	86.34%	86.54%	86.34%	97.73%	96.44%	85.98%
	40x	87.84%	87.9%	87.84%	97.97%	98.01%	87.55%

3.6 Experiment 2: RQ2. Is it feasible to combine convolution techniques with vision transformers to develop more streamlined approaches for medical data analyses? as follows.

3.6.1 Binary classification

Table 10 shows the best performances of three different models (VIT, CCT and MVIT) for a binary classification task. The performance measures calculated are accuracy, precision, sensitivity (recall), specificity and F1 score. Plus, total training time for each model.

- Accuracy: The accuracy metric represents the overall ability of the models to cor- rectly classify the binary samples. The results show that the VIT model achieves the highest accuracy at 98.64%, indicating it has the best overall classification per- formance. The MVIT model follows closely with an accuracy of 97.52%, while the CCT model has the lowest accuracy at 96.99%.

Table 10 Best performance in binary classification for each model

Model	VIT	CCT	MVIT
Accuracy	98.64%	96.99%	97.52%
Precision	98.56%	97.81%	96.91%
Sensitivity	98.35%	97.81%	97.32%
Specificity	98.35%	95.2%	97.32%
AUC	99.47%	98.93%	99.69%
F1 score	98.45%	97.81%	97.11%
Best Magnification	400×	40×	200×
Total Training Time (s)	521.03	5796.42	3218.72
Avg. Training Time per Epoch (s)	6.51	64.40	35.76
Total parameters	36,376,521	407,107	1,488,722

- Precision: Precision measures the proportion of true positive predictions out of all positive predictions made by the model. The ViT model exhibits the highest preci- sion at 98.56%, suggesting it has the lowest false positive rate and can make the most reliable positive predictions. The CCT model follows with a precision of 97.81%, and the MVIT model has the lowest precision at 96.91%.
- Sensitivity: This metric represents the proportion of true positive samples that are correctly identified by the model. The VIT model achieves the highest sensitivity at 98.35%, indicating it can successfully detect the most positive samples. The CCT model has the second-highest sensitivity at 97.81%, outperforming the MVIT model at 97.32%.
- Specificity: Specificity measures the proportion of true negative samples that are correctly identified by the model. The VIT model demonstrates the highest speci- ficity at 98.35%, meaning it can accurately detect negative samples. The MVIT model has the second-highest specificity at 97.32%, while the CCT model has the lowest specificity at 95.2%.
- AUC: The Area Under the Receiver Operating Characteristic (ROC) Curve is a comprehensive performance metric that measures the overall discriminative ability of the model. The MVIT model has the highest AUC at 99.69%, suggesting it has the best trade-off between true positive rate and false positive rate among the three models. The VIT model follows with an AUC of 99.47%, and the CCT model has the lowest AUC at 98.93%.
- F1-score: The F1-score is the harmonic mean of precision and sensitivity, providing a balanced evaluation of the model's performance. The ViT model demonstrates the highest F1-score at 98.45%, indicating a good balance between precision and recall. The CCT model also exhibits a high F1-score of 97.81%, outperforming the MVIT model at 97.11%.
- Total Training Time: The total training time metric provides insights into the computational efficiency of the models. The VIT model has the shortest total training time at 521.03 s, suggesting it is the most computationally efficient among the three models. The MVIT model has the second-shortest total training time at 3218.72 s, while the CCT model has the longest total training time at 5796.42 s.
- Average Training Time per Epoch: The average training time per epoch metric reflects the convergence speed of the models during the training process. The VIT model has the shortest average training time per epoch at 6.51 s, indicating it converges faster compared to the CCT model at 64.40 s and the MVIT model at 35.76 s.
- Total Parameters: The total parameters metric represents the complexity of the models. The CCT model has the fewest total parameters at 407,107 suggesting a more compact and efficient architecture compared to the ViT model with 36,376,521 parameters and the MVIT model with 1,488,722 parameters.

Figure 6 illustrates the confusion matrices for binary classification at 400×mag- nification. The high number of true positives (TP: 245) and true negatives (TN: 118) in the ViT model indicates its superior performance in distinguishing between malig- nant and benign samples. MobileVIT also performs well, with (TP: 238) true positives and (TN: 101) true negatives. Conversely, the CCT model shows a higher number of false positives (FP: 17), suggesting that it may be more prone to misclassifying benign samples as malignant.

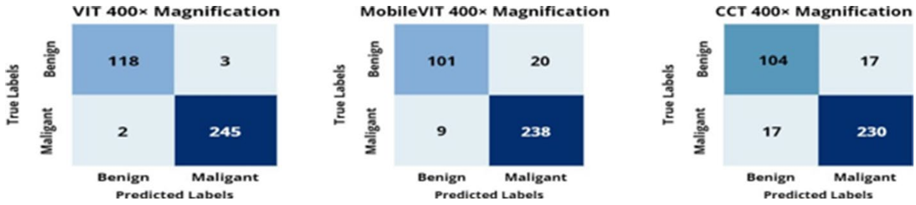


Fig. 6 The confusion matrices for classifying benign and malignant cancer types using VIT, and MobileVIT, and the CCT model, applied to the Brekhis images dataset at a 400×magnification

Figure 7 presents the confusion matrices for binary classification at 200×magni- fication using the VIT, MobileVIT, and CCT models. The VIT model continues to exhibit excellent performance, with a high number of true positives (TP: 273) and true negatives (TN: 124), indicating its ability to accurately classify both malignant and benign samples at this magnification level. The MobileVIT model also demonstrates strong results, with (TP: 272) true positives and (TN: 121) true negatives, although it has slightly more false positives (FP: 4) and false negatives (FN: 6) compared to the VIT model. In contrast, the CCT model shows a higher tendency toward misclassifi- cation, with (FP: 16) false positives, suggesting it may be more prone to incorrectly labeling benign samples as malignant, and it has (FN: 12) false negatives, indicating a higher rate of misclassifying malignant samples as benign compared to VIT and MVIT models at this magnification level.

Figure 8 illustrates the confusion matrices for binary classification at 100×mag- nification using three different models: VIT, MobileVIT, and CCT. The VIT model demonstrates exceptional performance, with a high number of true positives (TP: 282) and true negatives (TN: 128), indicating its effectiveness in accurately classifying both malignant and benign samples at this magnification level. The MobileVIT model also exhibits promising results, with (TP: 277) true positives and (TN: 123) true negatives, albeit with slightly more false

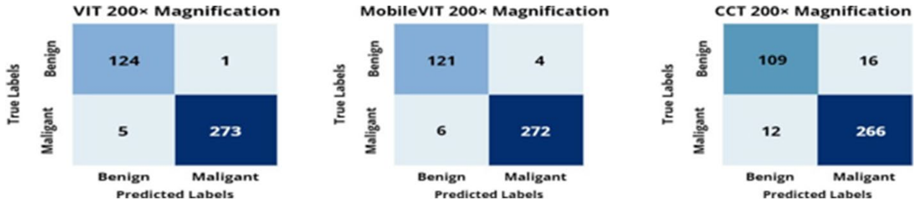


Fig. 7 The confusion matrices for classifying benign and malignant cancer types using VIT, and MobileVIT, and the CCT model, applied to the Brekhis images dataset at a 200×magnification

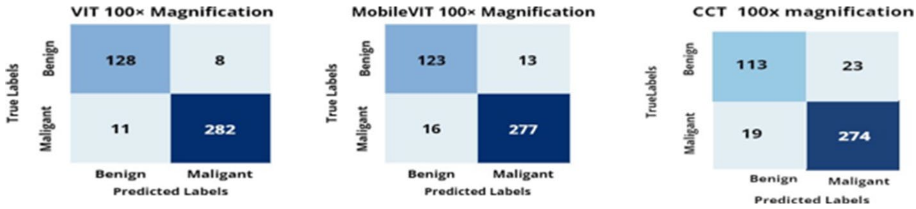


Fig. 8 The confusion matrices for classifying benign and malignant cancer types using VIT, and MobileVIT, and the CCT model, applied to the Brekhis images dataset at a 100×magnification

positives (FP: 13) and false negatives (FN: 16) com- pared to the VIT model. However, the CCT model shows a higher tendency toward misclassification, with (FP: 23) false positives, suggesting it may be more prone to incorrectly labeling benign samples as malignant, and it has (FN: 19) false negatives, indicating a higher rate of misclassifying malignant samples as benign compared to the other two models.

Figure 9 presents the confusion matrices for binary classification at 40×magni- fication. The VIT model continues to demonstrate strong performance, with a high number of true positives (TP: 269) and true negatives (TN: 117), indicating its ability to accurately classify both malignant and benign samples at this magnification level. The MobileVIT model also exhibits promising results, with (TP: 266) true positives and (TN: 112) true negatives, although it has slightly more false positives (FP: 13) and false negatives (FN: 8) compared to the VIT model. In contrast, the CCT model shows a higher tendency toward misclassification, with (FP: 6) false positives and (FN: 6) false negatives, suggesting it may struggle more with accurately classifying samples at this magnification level compared to the other two models.

Figure 10 shows the ROC curves representing the best performance of each model for classifying benign and malignant subtypes. A comparison of these ROC curves shows that our Vit Proposed model yields good classification performance (Benign = 99% Malignant 99%). Additionally, using the transformer-based convolution CCT model also delivers a significant classification performance (Benign = 98% Malignant 99%). However, the classification performance is higher substantially improved in the MVIT model with (Benign = 100% Malignant 100%).

3.6.2 Multi classification

Table 11 shows the best performances of three different models (VIT, CCT and MVIT) for a multi classification task.

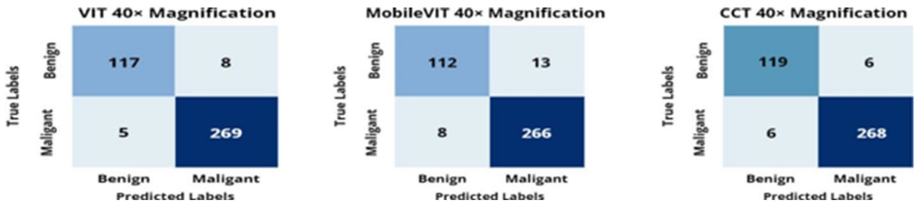


Fig. 9 The confusion matrices for classifying benign and malignant cancer types using VIT, and MobileVIT, and the CCT model, applied to the Brekhis images dataset at a 40×magnification

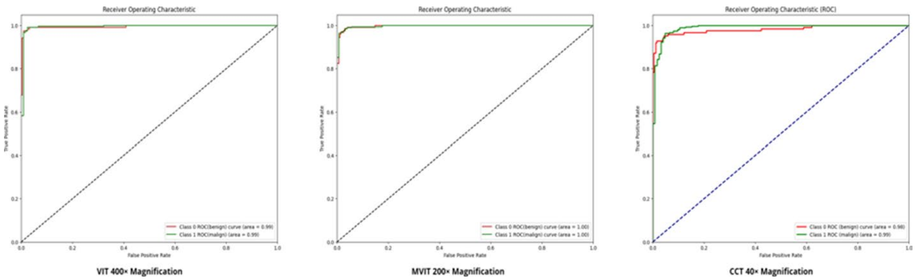


Fig. 10 Illustration of ROC curves of Vision transformers, Mobile vision transformers, and Compact vision transformer with best magnification level

Table 11 Best performance in Multi classification for each model

Model	VIT	CCT	MVIT
Accuracy	94.80%	84.60%	87.84%
Precision	94.97%	84.80%	87.90%
Sensitivity	94.80%	84.63%	87.84%
Specificity	99.14%	97.40%	97.97%
AUC	97.73%	97.1%	98.01%
F1 score	94.79%	84.4%	87.55%
Best Magnification	40x	40x	40x

- Accuracy: The VIT model achieves the highest accuracy at 94.80%, indicating it has the best overall classification performance among the three models. The MVIT model follows closely with an accuracy of 87.84%, while the CCT model has the lowest accuracy at 84.60%.
- Precision: The VIT model exhibits the highest precision at 94.97%, suggesting it has the lowest false positive rate and can make the most reliable positive predictions. The MVIT model follows with a precision of 87.90%, and the CCT model has the lowest precision at 84.80%.
- Sensitivity: The VIT model achieves the highest sensitivity at 94.80%, indicating it can successfully detect the most positive samples across the multiple classes. The MVIT and CCT models both have a sensitivity of 87.84%, 84.63%, which is lower than the VIT model.
- Specificity: The VIT model demonstrates the highest specificity at 99.14%, meaning it can accurately detect negative samples. The MVIT model has the second-highest specificity at 97.97%, while the CCT model has the lowest specificity at 97.40%.
- F1-score: The VIT model achieves the highest F1-score at 94.79%, indicating a strong trade-off between precision and sensitivity. The MVIT model follows with an F1-score of 87.55%, and the CCT model has the lowest F1-score at 84.40%.
- Best Magnification: The Best Magnification column shows the optimal magnification level for each model. All three models have a Best Magnification of 40x.

Figure 11 illustrates the confusion matrices generated during the multi- classification task for distinguishing between adenosis (AD), ductal carcinoma (DU) fibroadenoma (FI), lobular carcinoma (LO), mucinous carcinoma (MU),), papillary carcinoma (PA), phyllodes tumour (PH), tubular adenoma (TU) types respectively, at 400×magnification. The VIT model demonstrates reasonably high accuracy across most classes, with notable true

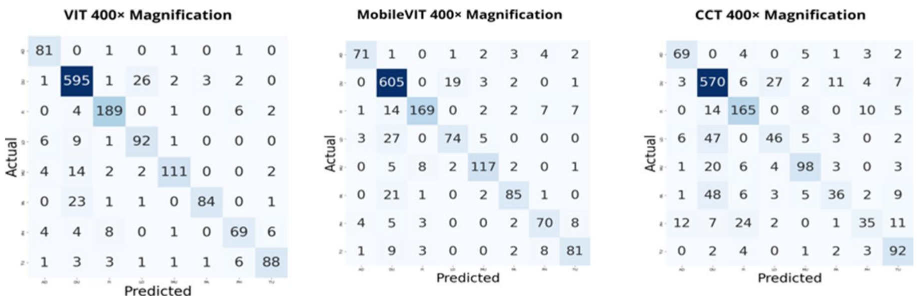


Fig. 11 Confusion matrices at the 400×magnification level of multi classification

positives for adenosis, fibroadenoma, lobular carci- noma, and tubular adenoma. However, it struggles to distinguish ductal carcinoma from mucinous carcinoma and tubular adenoma. The MobileVIT model exhibits sim-ilar performance, with high true positives for fibroadenoma and tubular adenoma, but confuses ductal carcinoma with mucinous carcinoma and phyllodes tumor. The CCT model tends to misclassify more samples across various classes. While it cor- rectly identifies some fibroadenoma and tubular adenoma cases, it shows higher false positives for ductal carcinoma and phyllodes tumor, suggesting lower overall accuracy compared to the other two models for this multi classification task.

Figure 12 presents the confusion matrices for multi-class classification at 200 × magnification across the three models: VIT, MobileVIT, and CCT. The VIT model shows high true positives for adenosis and reasonably accurate classification of other classes like fibroadenoma and lobular carcinoma. However, it struggles with distin- guishing ductal carcinoma from mucinous carcinoma. The MobileVIT model exhibits similar performance, accurately classifying adenosis and fibroadenoma but confusing ductal carcinoma with mucinous carcinoma. The CCT model demonstrates lower over- all accuracy, with higher mis-classifications across multiple classes, suggesting it may be less effective at this magnification level for multi-class classification compared to VIT and MVIT models.

Figure 13 shows the confusion matrices for multi-class classification at 100 × magnification across the VIT, MobileVIT, and CCT models. The VIT model accurately classifies adenosis and fibroadenoma but struggles with ductal carcinoma, confusing it with mucinous carcinoma. MobileVIT performs similarly, correctly identifying adeno- sis and fibroadenoma but misclassifying ductal carcinoma as mucinous carcinoma. The CCT model Wexhibits lower overall accuracy, with higher mis-classifications across multiple classes, indicating potential challenges in distinguishing lesion types at this magnification level compared to the other models.

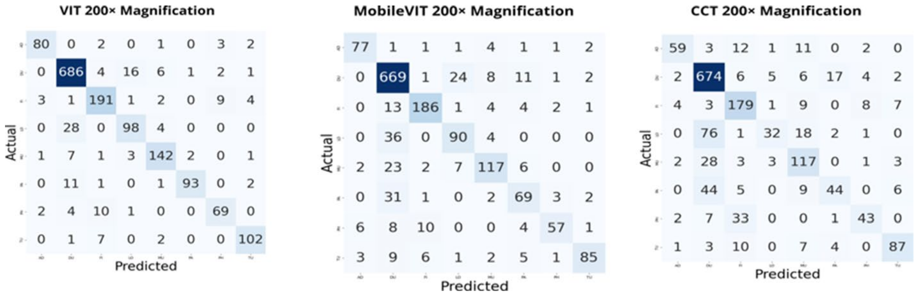


Fig. 12 Confusion matrices at the 200 × magnification level of multi classification

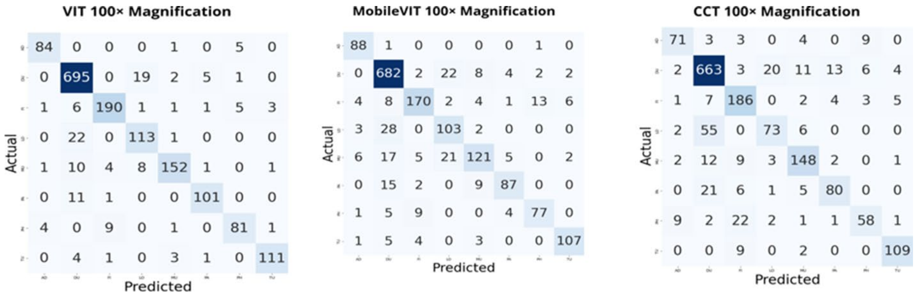


Fig. 13 Confusion matrices at the 100 × magnification level of multi classification

Figure 14 displays the confusion matrices for multi-class classification at 40x magnification across the three models: VIT, MobileVIT, and CCT. The VIT model shows high accuracy in classifying adenosis and fibroadenoma, but confuses ductal carci- noma with mucinous carcinoma. MobileVIT performs well in identifying adenosis and fibroadenoma, but also misclassifies some ductal carcinoma cases as mucinous carcinoma. The CCT model exhibits lower overall performance, with higher mis- classifications across multiple classes, suggesting potential challenges in accurately distinguishing lesion types at this magnification level compared to the other two models.

Figure 15 shows the ROC curves representing the optimal performance of each model for the multi classification of breast cancer sub types.

Within this section, our focus centred on addressing the second research question, leading to the derivation of several insightful observations. Our analysis delved into the examination of confusion matrices generated during both binary and multi clas- sification tasks. Concurrently, we scrutinized the ROC curves presented in Fig. 10, for binary classification, and Fig. 15, for multi classification. Intriguingly, Our findings underscore the significance of integrating convolutional and transformer techniques within the MVIT and CCT models. Notably, while previous studies have predomi- nantly focused on binary classification and used different preprocessing techniques and training on pre-training models, our work delves into multi classification using the BreakHis dataset without preprocessing techniques and training from scratch. This approach allows our model to concentrate on specific regions, thereby enhancing its ability to capture crucial features from patches and ensure stable training. As a result, we achieve better performance with fewer parameters and basic augmentation, leading to lower computational complexity problem in the vision transformer model.

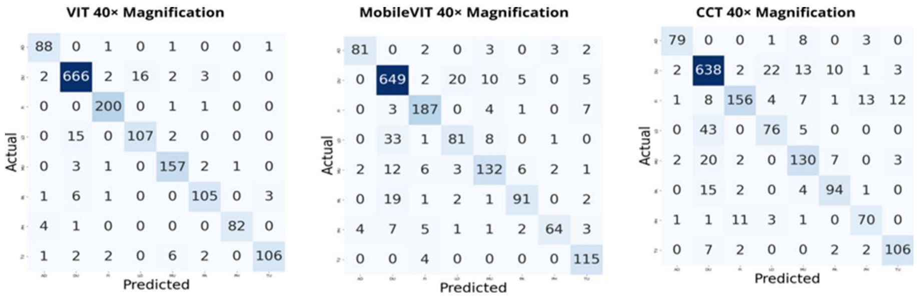


Fig. 14 Confusion matrices on the 40 × magnification level of Multi-classification

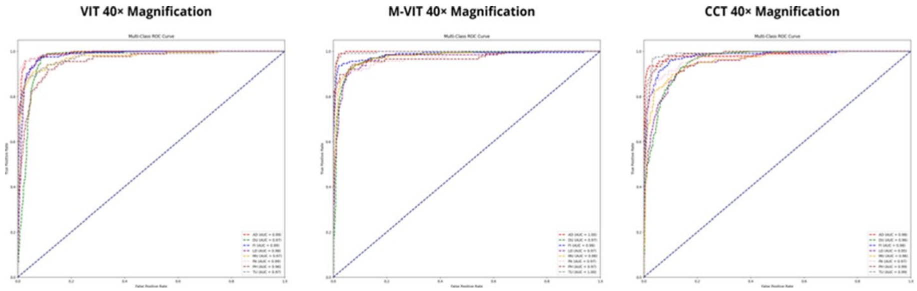


Fig. 15 Illustration of ROC curves of Vision transformers, Mobile vision transformers, and Compact vision transformer with best magnification level in multi classification

3.7 Experiment 3: RQ3. Should we consider transitioning to vision transformer models, or is it more prudent to continue with CNNs? as follows

In this section, our attention is directed towards addressing the third research question, encompassing two distinct parts. The first part involves a comprehensive analysis of results, drawing insights from the two previous research questions. Our objective is to integrate the VIT model against the CCT and MVIT models, providing a holis- tic understanding of how effective the amalgamation of transformer and convolution architectures is. The second part of our exploration seeks to evaluate whether the combined use of VIT and CNN architectures is a proposed approach. This assessment serves as a pivotal consideration in determining the feasibility of transforming vision transformer models.

The evaluation metrics presented in Table 10 offer a comprehensive overview of the performance assessment for the proposed models in binary classification tasks. The superior performance of the ViT model at the 400 × magnification level is remarkable. This suggests that the ViT architecture is particularly well suited to capturing the complex visual features and patterns present in high-resolution histopathological images, leveraging its powerful attention mechanism to discern subtle tumor-related signals. In contrast, the CCT model, while achieving competitive results at 40 × mag- nification, exhibited slower training time compared to the MobileViT and ViT models. That shows the convolution layers used in this model don't show the same performance to capture the complex visual features and patterns present in high-resolution images like ViT. This trade-off between performance and training efficiency is an important consideration, as the ability to train models with limited computing resources is a cru- cial requirement for practical clinical deployment. On the other hand, the MobileViT model demonstrated the best performance at 200 × magnification, outperforming the CCT and close to the Vit model. This finding indicates that MobileViT combination of Depth-wise convolution 2D layers and Swish as an activation function and atten- tion mechanisms is effective in extracting features from medium-resolution images while maintaining a more efficient and lightweight architecture than that of ViT. The differences in performance between different magnification levels highlight the impor- tance of evaluating model performance under various imaging conditions. The optimal model architecture may vary depending on the structure of the neural network and the available image resolution of the histopathological images.

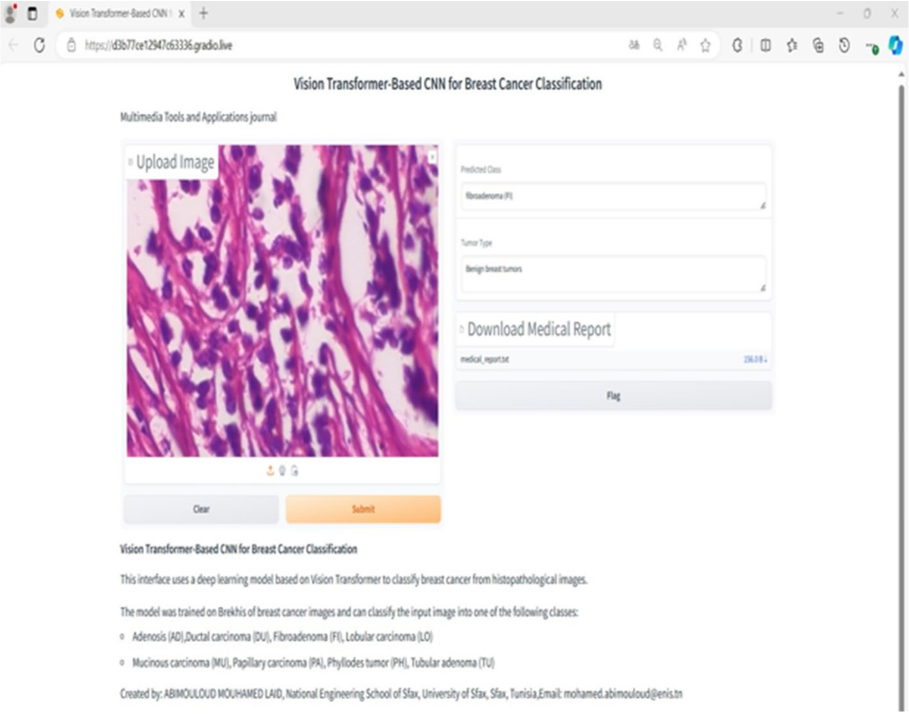
Notably, the evaluation of the models in the multi-classification task, as shown in Table 11, reveals that all models achieved their best accuracy at the 40 × magnification level. The rationale for this observation can be attributed to the limited number of samples and lower resolution in other magnifications, resulting from the problem of limited data available for multi classification. This sparsity of data may not have provided enough information for models to effectively extract features from high-resolution magnifications. In this context, the results suggest that both CNNs and ViTs have limitations in analyzing complex medical data, particularly in multi-classification tasks when the number of samples is limited. The performance of these models appears to be constrained by insufficient training data, which hinders their ability to capture the intricate visual patterns and subtle cues present in the high-resolution histopathological images.

3.8 Experiment 4: Models development and implementation for real-world clinical settings.

Explainable artificial intelligence (EXAI) is a technique for AI-enabled diagnosis and analysis, enabling result tracing and model improvement in healthcare through feature extraction for model explainability and interpretability. EXAI provides a framework to understand and predict the behavior of ML/DL models. It finds applications in var- ious domains, including clinical support systems, disease detection and classification, medical image segmentation, and robotic-assisted surgery in healthcare [46]. Assess- ing the generalizability and applicability of EXAI and these deep learning models in real-world clinical settings is crucial for their successful adoption and implementation in healthcare tasks. Validating their ability to provide accurate decisions in practical clinical environments is essential for realizing the potential of these models to improve patient care and decision-making [47]. One of the main interests in computer health diagnosis is the"Internet of Medical Things" (IoMT). The IoMT encompasses all elements of healthcare-related objects, sensors, and devices. It allows for remote mon- itoring and treatment of patients through the interconnection of intelligent devices and applications that gather, transmit, and analyze medical data and signals without human involvement. The Internet of Medical Things (IoMT) is a subset of the Inter- net of Things (IoT) that specifically deals with the integration and interoperability of medical equipment [48]. Moreover, it offers exceptional prospects for gathering, exam- ining, and sharing medical data, revolutionizing the provision of healthcare services. Medical care has become more individualized and precise in the Internet of Medical Things (IoMT) age. The integration of hardware accelerators like field-programmable gate arrays (FPGAs) or application-specific integrated circuits (ASICs) tailored for efficient inference could mitigate the computational resource constraints in clinical settings based on IoMT. To this end, we developed a web interface using the Gra- dio platform, which enables seamless integration and evaluation of our trained Vision Transformer model in a real-world clinical environment, as shown in Fig. 16. The main concept behind this interface is to provide an easy-to-use platform for pathol- ogists to analyze breast cancer histopathology images and obtain relevant diagnostic information. The interface workflow is as follows:

- The pathologist-user uploads a test image of a breast tissue sample.
- The trained Vision Transformer model analyzes the uploaded image.
- The model predicts the class of breast tumor present in the image.
- The prediction results are presented in two forms: Predicted Class: This indicates the specific subtype of breast cancer from the eight classes the model was trained on. Tumor Type: This classifies the tumor as either benign or malignant class.
- Finaly,the user can download a medical report in a text file format, which includes the Image ID, Predicted Class, and Tumor Type, providing a comprehensive summary of the analysis.

Currently, the growing prevalence of the Internet of Medical Things (IoMT) environment can be primarily attributed to its effective data management, remote patient monitoring, and the integration of networked medical devices (NMDs) for facili- tating informed decision-making and supporting pathologists in the diagnosis and management of breast cancer.


Vision Transformer-Based CNN f
https://d3b77ce12947c63336.gradio.live
Upload Image
Download Medical Report
medical_report.txt
Flag
Clear
Submit

Vision Transformer-Based CNN for Breast Cancer Classification

Multimedia Tools and Applications journal

Predicted Class: fibroadenoma (FI)

Tumor Type: Benign breast tumors

Vision Transformer-Based CNN for Breast Cancer Classification

This interface uses a deep learning model based on Vision Transformer to classify breast cancer from histopathological images.

The model was trained on Brekhis of breast cancer images and can classify the input image into one of the following classes:

- Adenosis (AD),Ductal carcinoma (DU), Fibroadenoma (FI), Lobular carcinoma (LO)
- Mucinous carcinoma (MU), Papillary carcinoma (PA), Phyllodes tumor (PH), Tubular adenoma (TU)

Created by: ABIMOULOUD MOUHAMED LAID, National Engineering School of Sfax, University of Sfax, Sfax, Tunisia,Email: mohamed.abimouloud@enis.tn

Fig. 16 Web Interface of breast cancer Image Classification

4 Discussion

In this study, we explored how ensemble vision transformer models, VIT, CCT, and MVIT, can be used to effectively classify breast cancer from histopathological images at different magnifications (40×, 100×, 200×, and 400×), for binary and multi classification of breast cancer subtypes. Our experiments show that the VIT model achieves a binary classification accuracy of 98.64% with a training time of 521.03 s. To enhance this approach, we introduce ViT-based convolution models lightweight MobileVIT and CCT, which achieve test accuracies of 97.52% and 96.99%, respectively, with training times of 3218.72 and 5796.42 s, respectively. In multi classification tasks, the VIT model demonstrates notable proficiency with an accuracy of 94.80%, highlighting its effectiveness in accurately classifying instances across mul- tiple classes. However, the lightweight MobileVIT and CCT systems achieve slightly lower accuracies of 87.84% and 84.63%, respectively. Comparative analysis reveals insights into the strengths and weaknesses of each model in multi classification tasks. Our research suggests that vision transformer models benefit from the integration of convolution blocks. The combination allows for the acquisition of global informa- tion from convolution layers and thorough feature analysis in each patch through the deep self-attention mechanism. The proposed self-attention ViT outperforms the CCT and MobileViT models due to its smaller 14×14 patch size. This smaller patch size enables the transformer encoder's attention mechanism to extract features from patch pixels to operate more efficiently. Empirical experiments demonstrated that the selected patch sizes provided an optimal trade-off between performance and compu- tational efficiency for each model architecture. The 16×16 patch size used in CCT and MobileViT allowed efficient feature extraction while maintaining high accuracy. However, the 14×14 patch size in the adapted ViT model, combined with its 8 lay- ers and higher number of parameters (36,376,521), leveraged the attention mechanism more effectively to capture intricate details, leading to improved overall performance. The selection of patch size aimed to strike a balance between computational efficiency and model performance. Smaller patch sizes, such as 14×14, increase the number of patches, potentially enhancing the capture of features details but also increasing com- putational complexity. Conversely, larger patch sizes like 16×16 reduce the number of patches, which may sacrifice some detail but improve computational efficiency.

Despite its superior performance, ViT has several drawbacks, such as being data intensive and computationally expensive due to its large architecture and large num- ber of parameters **(36,376,521)**. MVIT and CCT, on the other hand, training with 1,488,722, and 407,107 parameters is less computationally complex than training with VIT. It also shows promising results in merging convolution layers with the attention mechanism. However, further performance improvement is needed to reduce training time and improve the accuracy of multi classification. One drawback of combining attention and convolution layers is the extended training time, as seen in the CCT model. To address this, we introduced the lightweight MobileVIT method, which substantially reduces the training time to 3218.72 s while achieving higher accuracy than CCT. The proposed lightweight MobileVIT method exhibited superior performance in terms of results and lower computational cost and achieved the same performance as the VIT and CCT models for breast tumor histopathology image clas- sification. The main reason for this difference in performance is its architecture, which was inspired by a lightweight CNN, 3×3 Depthwise convolution 2D layer and Swish as an activation function, followed by MobileNetv2 blocks and Mobile-ViT blocks. Using this method, MobileViT achieves better performance with fewer parameters and basic augmentation, thereby providing less computational complexity, making it preferable when combined transformer and convolution layers.

Compared with previous studies, our approach yields significantly improved results, indicating the successful fusion of Vision Transformers (ViTs) and Convolutional Neu- ral Networks (CNNs) in identifying breast tumors in histopathological images using BreakHis dataset, especially in lightweight vision transformer models. First, when we compare our outcomes with studies [20–26], that were based on fea- ture extraction by CNN filters, preprocessing steps, and deep learning approaches that prevent overfitting (such as transfer learning, data augmentation, and SVM classifiers), it is evident that we had superior outcomes when training our models from scratch without preprocessing steps. Second, even when compared to studies [30], based on ViT models, our results surpass previous findings. The MaxVIT model, which was considered the best model in that study, achieved an accuracy of 91.57% and increased to 92.12% after fine-tuning with 8.83 h of training time and 79,213,918 number of parameters. In comparison, our outcomes exceed these results. We conducted a results comparative of our approach against models presented in previously published studies and observed that our method yielded the best performance results, as evidenced in Table 12. Finally, it is important to discuss the limitations of our proposed approach, as well as how these limitations might affect the ability to apply it successfully in practical clinical applications.

Table 12 Assessment of the performance of previously published works on histopathological breast cancer detection using the Brekhis dataset

Reference	Method	Classification	Performance	Magnification level
Sriwastawa Asmi et al. [30]	MaxViT	Binary	ACC 92.12%, Time(H) = 8.83	
	ViT		ACC 86.86%, Time(H) = 1.46	
	PiT		ACC 89.01%, Time(H) = 8.49	
	CvT		ACC 91.41%, Time(H) = 5.46	
	CrossFormer		ACC 90.39%, Time(H) = 8.83	
	CrossViT		ACC 88.11%, Time(H) = 1.46	
	NesT		ACC 91.66%, Time(H) = 6.75	
	SepViT		ACC 87.45%, Time(H) = 4.44	
Wang Pin et al. [20]	Transfer learning	Binary	ACC 94.52%, Sens 95.16%	100x
Albashish Dheeb et al. [21]	VGG16 / RBF-SVM	Binary	ACC 96.0%	400x
Al-Jabbar Mohammed et al. [22]	AlexNet, GoogleNet	Binary	ACC 98.8%, Prec 98.5%	100x
Amin Muhammad Sadiq et al. [23]	FabNet	Multi	ACC 97.20%, Prec 89.947%	400x
Hao Yan et al. [24]	DenseNet201	Binary	ACC 99.00%, Prec 98.99%	40x
		Binary	ACC 96.75%	40×
Srikantamurthy Mahati Munikoti et al. [25]	CNN-LSTM	Multi	ACC 96.3%, Prec 97.0%	40x
Mahmud et al. [26]	AlexNet Vgg16	Binary	ACC 91.37%	200x
our systems				
VIT		Binary	ACC 98.64% Time(H) = 0.14	400x
CCT		Binary	ACC 96.99% Time(H) = 1.61	40x

Table 12 (continued)

Reference	Method	Classification	Performance	Magni-fication level
MVIT		Binary	ACC 97.52%	200x
			Time(H) = 0.89	
VIT		Multi	ACC 94.80%	400x
CCT		Multi	ACC 84.60%	40x
MVIT		Multi	ACC 87.84%	200x

- While though we used a powerful GPU for training the models, implementing these systems for practical applications in standard medical computer-aided diagnosis (CAD) tools may be challenging because such powerful computing resources are often unavailable in clinical environments. This limitation could potentially cause a decrease in real-time diagnostic accuracy when running the models on less capable hardware.
- Validate that the models are robust to various factors like image artifacts, noise, and variations in medical imaging protocols across different magnification levels is crucial for reliable real-time diagnosis using our proposed approach.
- While we used a sizeable BreakHis dataset, the imbalance in the number of benign and malignant samples at each magnification level, as shown in Table 2, created certain limitations, particularly for the task of classifying eight different subtypes of breast cancer. This imbalance could bias the model towards classifying malignant cases more accurately.
- The general evaluation of the confusion matrices for the multi-classification task, as shown in Figs. 11, 12, 13, and 14, revealed that the models exhibited a bias towards the malignant ductal carcinoma (DU) class. This can be explained by the high number of samples in this class compared to the limited number of samples in other classes, resulting from the lack of available data for the multi-classification task. When working with limited or imbalanced datasets, the models may struggle to generalize effectively, leading to overfitting and mis-predection.

The results of our proposed systems suggest that both convolutional neural networks (CNNs) and vision transformers (ViTs) have limitations in analyzing complex medical data, particularly for multi-classification tasks when the number of samples is limited.

5 Conclusions

We have presented an ensemble of three vision transformer models based on self- attention and the attention-convolution approach for breast cancer classification using histopathological images from the BreaKHis dataset. The VIT model was superior at 400× magnification. The MVIT and CCT models were superior at 200× 40× magnification, respectively without using any pre-processing image processes. Con- sequently, we found that the lightweight Mobile vision transformer model provides superior performance with less computational complexity. Therefore, the promising results demonstrate that Vision Transformers are strong and better with CNN for building high-performing deep attention-convolution models in the medical image field. We summarize our experiments as follows:

- The transformer self attention model does not provide many advantages compared to CNNs because it is more complex.
- The hybrid attention convolution approach can offer a new approach that harnesses the benefits of both techniques in medical vision diagnosis.
- Lightweight attention convolution models present a promising solution for medical applications with limited training data and computational GPU resources.

In conclusion, the findings of our research will serve as a foundation for future investigations aimed at enhancing breast tumor type classification outcomes. This will contribute to the development of more precise Computer-Aided Diagnosis (CAD) systems for breast cancer. Future work will focus on several key areas:

- Real-time diagnostic interfaces: Explore the integration of user-friendly interfaces and software tools for real-time application of the models in clinical settings. We plan to deploy the models on embedded hardware like Raspberry Pi, NVIDIA Jetson Nano, and FPGAs for on-device inference and low-latency real-time analysis. These compact, low-power solutions can enable point-of-care diagnosis in resource-limited settings.
- Adaptability to other cancers: Investigate the adaptability of the proposed models to other cancers, such as lung and prostate cancers. This will involve retraining the models on new datasets and evaluating their performance across different cancer types, and extend the models to detect and classify cardiovascular conditions using imaging modalities like MRI, CT scans,mammography and echocardiograms.
- Reducing computational complexity: Investigate approaches such as model quantization, pruning, and knowledge distillation to reduce the computational complexity and memory footprint of the proposed models, enabling their adaptation for use in mobile applications and enabling point-of-care diagnostics, especially in resource-limited settings.
- Model refinement and clinical deployment: Further refine the model's performance, integrate with electronic health record systems, and ultimately deploy in actual clinical settings under the supervision of healthcare professionals.
- Multi-view medical scan analysis: Consider different views of a medical scan image for a comprehensive final decision, taking into account other organs besides the breast.
- Dataset balancing and noise removal: Enhance the dataset by adding more databases, employing noise removal techniques, and leveraging diffusion models and GANs to synthesize realistic medical images for data augmentation. This can improve model performance, robustness, and generalization, while mitigating class imbalance and overfitting.
- Lightweight CNN-ViT models: Explore the integration of lightweight CNN-ViT (Convolutional Neural Network-Vision Transformer) models based on tokenization techniques and lightweight vision language models (LVLMs) for medical applications with limited training data and computational GPU resources.

By addressing these areas, we aim to advance the development and practical appli- cation of deep learning models in medical diagnostics, ultimately improving patient outcomes and supporting healthcare professionals in their decision-making processes.

Acknowledgements The authors extend their appreciation to the Deanship of Sci- entific Research at Northern Border University, Arar, KSA for funding this research work through the project number " NBU-FFR-2024-2439-05.

Funding This study was not funded. The authors have no relevant financial or non-financial interests to disclose.

Data availability The dataset analysed during the current study are available in: https://web.inf.ufpr.br/vri/databases/ breast-cancer-histopathological-database-breakhis/

Declarations

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Competing interests Authors declare that they have no conflict of interest.

References

1. Youlden DR et al (2012) The descriptive epidemiology of female breast cancer: an international comparison of screening, incidence, survival and mortality. Cancer Epidemiol 36:237–248
2. Sohns C, Angic BC, Sossalla S, Konietschke F, Obenauer S (2010) Cad in full-field digital mammography—influence of reader experience and application of cad on interpretation of time. Clin Imaging 34:418–424
3. Saba T (2020) Recent advancement in cancer detection using machine learning: sys- tematic survey of decades, comparisons and challenges. J Infect Public Health 13:1274–1289
4. Nassif AB, Talib MA, Nasir Q, Afadar Y, Elgendy O (2022) Breast cancer detection using artificial intelligence techniques: a systematic literature review. Artif Intell Med 127:102276
5. Aggarwal R et al (2021) Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ digital medicine 4:65
6. Matsoukas C, Haslum JF, S¨oderberg M, Smith K (2021) Is it time to replace cnns with transformers for medical images? arXiv:2108.09038. Accessed 19 Jun 2023
7. Mohamed EA, Rashed EA, Gaber T, Karam O (2022) Deep learning model for fully automated breast cancer detection system from thermograms. PLoS ONE 17:e0262349
8. Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R (2019) A deep learning mammography-based model for improved breast cancer risk prediction. Radiology 292:60–66
9. Henry EU, Emebob O, Omonhinmin CA (2022) Vision transformers in medical imaging: a review. arXiv:2211.10043. Accessed 19 juin 2023
10. Dey RK, Das AK (2023) Modified term frequency-inverse document frequency based deep hybrid framework for sentiment analysis. Multimed Tools Appl 82:32967–32990
11. Dosovitskiy A. et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv:2010.11929. Accessed 15 Jun 2023
12. Zhu X, Cheng D, Zhang Z, Lin S, Dai J (2019) An empirical study of spatial attention mechanisms in deep networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 6688–6697
13. Masood A, Naseem U, Kim J (2023) Multi-Level swin transformer enabled automatic segmentation and classification of breast metastases. In 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Sydney, pp 1–4. https://doi.org/10.1109/EMBC40787.2023.10340831
14. Dey RK, Das AK (2024) Neighbour adjusted dispersive flies optimization based deep hybrid sentiment analysis framework. Multimed Tools Appl. https://doi.org/10.1007/s11042-023-17953-8
15. Hassani A. et al (2021) Escaping the big data paradigm with compact transformers. arXiv:2104.05704
16. Faheem M et al (2019) A multiobjective, lion mating optimization inspired routing protocol for wireless body area sensor network based healthcare applications. Sensors 19:5072
17. Alarood AA, Faheem M, Al-Khasawneh MA, Alzahrani AI, Alshdadi AA (2023) Secure medical image transmission using deep neural network in e-health applications. Healthcare Technol Lett 10:87–98
18. Iqbal S, Qureshi AN, Aurangzeb K, et al (2023) AMIAC: adaptive medical image analyzes and classification, a robust self-learning framework. Neural Comput & Applic. https://doi.org/10.1007/s00521-023-09209-1
19. Ali G, Dastgir A, Iqbal MW, Anwar M, Faheem M (2023) A hybrid convolutional neural network model for automatic diabetic retinopathy classification from fundus images. In IEEE Journal of Translational Engineering in Health and Medicine 11:341–350. https://doi.org/10.1109/JTEHM.2023.3282104
20. Wang P et al (2021) Automatic classification of breast cancer histopathological images based on deep feature fusion and enhanced routing. Biomed Signal Process Control 65:102341

21. Albashish D, Al-Sayyed R, Abdullah A, Ryalat MH, Ahmad Almansour N (2021) Deep CNN Model based on VGG16 for breast cancer classification. In 2021 International Conference on Information Technology (ICIT), Amman, pp 805–810. https://doi.org/10.1109/ICIT52682.2021.9491631
22. Al-Jabbar M, Alshahrani M, Senan EM, Ahmed IA (2023) Multi-method diagnosis of histopathological images for early detection of breast cancer based on hybrid and deep learning. Mathematics 11:1429
23. Amin MS, Ahn H (2023) Fabnet: A features agglomeration-based convolutional neural network for multiscale breast cancer histopathology images classification. Cancers 15:1013
24. Hao Y et al (2022) Breast cancer histopathological images classification based on deep semantic features and gray level co-occurrence matrix. PLoS ONE 17:e0267955
25. Srikantamurthy MM, Rallabandi V, Dudekula DB, Natarajan S, Park J (2023) Classification of benign and malignant subtypes of breast cancer histopathology imaging using hybrid cnn-lstm based transfer learning. BMC Med Imaging 23:1–15
26. Mahmud MI, Mamun M, Abdelgawad A (2023) A deep analysis of transfer learning based breast cancer detection using histopathology images. In 2023 10th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, IEEE, pp 198–204, https://doi.org/10.1109/SPIN57001.2023.10117110
27. Abunasser BS, Al-Hiealy MRJ, Zaqout IS, Abu-Naser SS (2023) Con- volution neural network for breast cancer detection and classification using deep learning. Asian Pac J Cancer Preven: APJCP 24:531
28. Ayana G et al (2023) Vision-transformer-based transfer learning for mammogram classification. Diagnostics 13:178
29. He K et al (2023) Transformers in medical image analysis. Intell Med 3:59–78
30. Sriwastawa A, Arul Jothi JA (2024) Vision transformer and its variants for image classification in digital breast cancer histopathology: a comparative study. Multimed Tools Appl 83:39731–39753. https://doi.org/10.1007/s11042-023-16954-x
31. He L, Long LR, Antani S, Thoma GR (2012) Histology image analysis for carcinoma detection and grading. Comput Methods Programs Biomed 107:538–556
32. ahmed IMb, Maalej R, Kherallah M (2023) MobileNet-Based model for histopathological breast cancer image classification. In: Abraham A, Hong TP, Kotecha K, Ma K, Manghirmalani Mishra P, Gandhi N (eds) Hybrid intelligent systems. HIS 2022. Lecture Notes in Networks and Systems, vol. 647. Springer, Cham. https://doi.org/10.1007/978-3-031-27409-1_58
33. Rulaningtyas R, Hyperastuty AS, Rahaju AS (2018) Histopathology grading identification of breast cancer based on texture classification using GLCM and neural network method. In Journal of Physics: Conference Series, vol. 1120, IOP Publishing, p 012050. https://doi.org/10.1088/1742-6596/1120/1/012050
34. He L, Long LR, Antani S, Thoma G (2010) Computer assisted diagnosis in histopathology. Sequenc Genome Anal: Methods Appl 15:271–287
35. Spanhol FA, Oliveira LS, Petitjean C, Heutte L (2015) A dataset for breast cancer histopathological image classification. IEEE Trans Biomed Eng 63:1455–1462
36. Breakhis - breast histopathology images dataset. https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/. Accessed 5 Jun 2023
37. Tummala S, Kim J, Kadry S (2022) Breast-net: multi-class classification of breast cancer from histopathological images using ensemble of swin transformers. Mathematics 10:4109
38. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30. https://doi.org/10.48550/arXiv.1706.03762
39. Mehta S, Rastegari M (2021) Mobilevit: light-weight, general-purpose, and mobile- friendly vision transformer. arXiv:2110.02178. Accessed 22 May 2023
40. Howard AG, et al (2017) Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861. Accessed 21 May 2023
41. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC (2018) Mobilenetv2: inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4510–4520. https://doi.org/10.48550/arXiv.1801.04381
42. Howard A, Sandler M, Chu G, Chen L-C, Chen B, Tan M, Wang W, Zhu Y, Pang R, Vasudevan V, Le QV, Adam H (2019) Searching for MobileNetV3. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)
43. Cheng Q, Li X, Zhu B, Shi Y, Xie B (2023) Drone detection method based on mobilevit and ca-panet. Electronics 12:223

44 Zou W, Xie K, Lin J (2023) Light-weight deep learning method for active jamming recognition based on improved mobilevit. Sonar & Navigation, IET Radar

45. Ahmed IA et al (2022) Eye tracking-based diagnosis and early detection of autism spectrum disorder using machine learning and deep learning techniques. Electronics 11:530

46 Saraswat D et al (2022) Explainable ai for healthcare 5.0: opportunities and challenges. IEEE Access 10:84486–84517

47. Chaddad A, Peng J, Xu J, Bouridane A (2023) Survey of explainable ai techniques in healthcare. Sensors 23:634

48. Wani NA, Kumar R, Bedi J, Rida I, et al (2024) Explainable AI-driven IoMT fusion: unravelling techniques, opportunities, and challenges with explainable AI in healthcare. Inf Fusion 102472. https://doi.org/10.1016/j.inffus.2024.102472

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

Authors and Affiliations

Mouhamed Laid ABIMOULOUD[1,6] · Khaled BENSID[2] · Mohamed Elleuch[3,6] · Mohamed Ben Ammar[4] · Monji KHERALLAH[5,6]

✉ Mouhamed Laid ABIMOULOUD
mohamed.abimouloud@enis.tn

Khaled BENSID
bensid.khaled@univ-ouargla.dz

Mohamed Elleuch
mohamed.elleuch@fss.usf.tn

Mohamed Ben Ammar
Mohammed.Ammar@nbu.edu.sa

Monji KHERALLAH
monji.kherallah@fss.usf.tn

[1] National Engineering School of Sfax, University of Sfax, Sfax, Tunisia

[2] Laboratory of Electrical Engineering (LAGE), University of KASDI Merbah Ouargla, 30000 Ouargla, Algeria

[3] National School of Computer Science (ENSI), University of Manouba, Manouba, Tunisia

[4] Department of Information Systems, FacultyofComputingandInformationTechnology, Northern Border University, Rafha, Saudi Arabia

[5] Faculty of Sciences, Sfax, Tunisia

[6] Advanced Technologies for Environment and Smart Cities (ATES Unit), Sfax University, Sfax, Tunisia