Neural Computing and Applications (2023) 35:7963–7974
https://doi.org/10.1007/s00521-022-07516-7

S.I.: TAM-LHR

Transfer learning for histopathology images: an empirical study

Tayyab Aitazaz[1] · Abdullah Tubaishat[2] · Feras Al-Obeidat[2] · Babar Shah[2] · Tehseen Zia[1] · Ali Tariq[3]

Received: 6 March 2022 / Accepted: 7 June 2022 / Published online: 5 July 2022

Abstract
Histopathology imaging is one of the key methods used to determine the presence of cancerous cells. However, determining the results from such medical images is a tedious task because of their size, which may cause a delay in results for days. Even though CNNs are widely used to analyze medical images, they can only learn short-term dependency and ignore long-term dependency, which could be crucial in processing higher dimensional histology images. Transformers, however, make use of a self-attention mechanism, which might be helpful to learn dependencies across an entire set of features. To process histology images, deep learning models require a large number of images, which is usually not available. Transfer learning, which is often used to deal with this issue, involves fine-tuning a trained model for use with medical images by adding features. In context, it is essential to analyze which CNNs or transformers are more conducive to transfer learning. In this study, we performed an empirical study to evaluate the performance of different pre-trained deep learning models for the classification of lung and colon cancer on histology images. Vision transformer and CNN models pre-trained on image-net are analyzed for the classification of histopathology images. We performed an experiment on the LC25000 dataset for the evaluation of models. The dataset consists of five classes, two belong to colon and three belong to lung cancer. The insights and observations obtained from an ablation study performed on different pre-trained models show vision transformers perform better than CNN based models for histopathology image classification using transfer learning. Moreover, the vision transformer with more layers of ViT-L32 performs better than ViTB32 with fewer layers.

Keyword Histopathology image classification · Vision transformer · Lung and colon cancer · Whole slide image

1 Introduction

As, reported by the World Health Organization, Cancer is one of the deadly diseases with a high mortality rate, especially affecting underdeveloped countries [1]. On-time diagnoses with proper examination and treatment can help to fight the disease. Some types of cancers have restrained growth, which affects the other body parts of a human. According to the WHO international agency of cancer research in 2020, 178388 cases are reported in Pakistan [2], with Lung cancer among the top 3 and colon in the top 10.

The diagnosis of cancer with a single test may not be possible, as it requires a comprehensive analysis of a patient with a physical examination by conducting numerous tests and thorough history of a patient. The

✉ Tehseen Zia
tehseen.zia@comsats.edu.pk

Tayyab Aitazaz
tayyab.aitazaz@gmail.com

Abdullah Tubaishat
abdallah.tubaishat@zu.ac.ae

Feras Al-Obeidat
feras.al-obeidat@zu.ac.ae

Babar Shah
babar.shah@zu.ac.ae

Ali Tariq
alitariq.bsc06@gmail.com

[1] COMSATS University Islamabad, Islamabad, Pakistan

[2] College of Technological Innovation, Zayed University, Abu Dhabi, UAE

[3] Medical Imaging and Diagnostic Lab, National Center for Artificial Intelligence, Comsats University, Islamabad, Pakistan

symptoms of cancer are much similar to several other diseases [3]. The medical practitioners perform several tests to diagnose possible diseases with symptoms similar to cancer. For effectual examination, tests are not only performed for the diagnosis of disease but also used to monitor the disease, to evaluate how effective the treatment is and sometimes for confirmation of a disease diagnosed using poor quality of the sample. The early diagnosis of cancer is crucial as it increases the survival rate of a patient [4]. Usually, cancer is diagnosed in its later stages until that time, disease affects the vital organs, which malfunction the body and decrease the survival rate of the patient. After the diagnosis of cancer, other tests are performed to examine the growth of cancerous cells and stage of the disease to design a medication scheme.

There are different procedures to diagnose cancer including lab tests, Magnetic Resonance Imaging (MRI), Ultrasound, Computed Tomography scans (CT), Posit-ron imaging tomography, biopsy, and X-ray [5]. Among all procedures, the medical practitioners often recommend a biopsy for the confirmation and detailed analysis of affected cells after performing a physical examination [6] and all other tests which eliminate other diseases to identify a possible cancer case. The analysis of histopathology images is a tedious and time taking process. Pathologists need to process several slides of a patient [7], this laborious task inclined them to inconsistencies. The pathology report provides a large amount of information, and the increase in the number of expected cancer patients increases the burden on pathologists. Nearly 80% of expected cases are benign, and pathologists have to spend much time on analysis of such biopsies. The CAD systems are effective at automating the tedious task of analyzing histopathology images.

Researchers applied both machine learning and deep learning-based techniques for disease diagnoses [8]. In the last decade, Deep Learning has changed the research paradigm. Traditional machine learning algorith-ms stop at a specific point and their performance may not increase as we increase the amount of data. However, deep learning-based algorithms are data-hungry, requiring more data but increasing their performance as we increase the amount of data. In contrast with ML, deep learning, maximum utilize unstructured data, automatic feature learning, works effectively on unlabeled data and produces high-quality results. Some of the widely used deep learning models are CNN, RNN [9] and LSTM [10]. Since the convolutional neural networks outperform other deep learning models in computer vision tasks, researchers widely used them in different domains. The CNN variants AlexNet [11], ResNet [12], YOLO [13] and U-Net [14] are the state-of-the-art models for several computer vision tasks like detection, segmentation, and classification. The healthcare domain produces a large amount of data about medical, which helps medical practitioners in decision-making procedures. The SOTA deep learning models are used for disease diagnoses, classification, organ detection and segmentation of affected organs to make the decisions. Medical imaging is a rapidly growing research area where these models are widely used not for just accuracy but efficiency as well. These STOA models produce results in less time compared to humans without compromising their accuracy, which increases the survival rate of a patient by early diagnosis of disease.

In this study, we will focus on lung and colon cancer diagnoses using histopathology images. Researchers used several deep learning models for lung and colon cancer diagnosis. Convolutional Neural Networks learns structured data and nifty representations of an image due to which, it is widely used in the domain. [15, 16] applied different variants of pre-trained CNN models, ResNet-50, VGG19 [17], ResNet-18, AlexNet, ResNet-34 and ResNet-101 to classify lung and colon cancer. [18] applied an image processing technique for feature extraction and a classical machine learning approach for classification. Most of the researchers proposed models for one type of cancer. The pretrained CNN models outperformed other models in both lung and colon cancer.

In recent years, transformers have outperformed the RNN's and CNN's in NLP and computer vision tasks. As compared to other STOA deep learning models, vision transformers tend to perform better on a large scale. CNN captures only local features in lower layers, while ViT captures both local and global features. This work proposes a framework for diagnosing cancer from histopathology images with better performance than is currently available in CAD systems to help medical practitioners. Histopathology images are widely used by medical practitioners for cancer diagnosis. The detection of abnormal cells is a tedious task and involves human error, therefore a computer aided diagnosis system is used to obtain better results. Majority of techniques used transfer learning due to lack of data. CNNs are widely used models for analyzing histopathology images. However, as CNN models correlations among neighboring pixels of an image [19, 20], it may not capture long range relationships between abnormal cells, which could be crucial for analyzing very high dimensional histopathology slides. Further, studies [21] have reported that CNNs are not very effective for transfer learning from ImageNet to medical images. In this study, we propose a computer-aided diagnosis (CAD) framework for detecting lung and colon cancers in histopathology pictures. As CNN and transformers are widely used image processing models, we analyzed the efficacy of these models for devising the aforementioned framework. Researchers presented various techniques to detect

abnormalities in histopathology whole slide image patch. Although, a lot of data are produced nowadays in the medical domain. The data are not annotated and requires pre-processing. Among the main contributions of our study, we conducted an empirical study and analyzed the efficacy of vision transformers and CNN models to classify lung and colon cancer. A vision transformer-based model is proposed for the classification of histopathology images. For the performance analysis of different algorithms for the histo-pathology image analysis following research questions are addressed:

1. Among contemporary CNN and transformer models, which one has better efficacy to classify histopathology images?
2. What is the effect of the increasing number of images in a dataset to fine-tune vision transformer and CNN to classify histopathology images?
3. Do pre-training vision transformers help to improve their efficacy to classify histopathology images?

The remainder of this paper is structured as follows: the deep learning based most relevant approaches for the diagnosis of cancer using the histopathology images are presented in Sect. 2. The methodology for the analysis of performance of different algorithms for histopathology image analysis is presented in Sect. 3. The results obtained from the study are discussed in Sect. 4. Section 5 contains the conclusion and future direction of the study.

2 Related work

Pathologists examine the tissue to find abnormal cells to classify the sample as cancerous or noncancerous. The pathologists examine the histopathology slides, which consist of complex visual patterns. These patterns make it difficult for a human to find abnormalities and require expertise with hard work. The diagnosis of abnormal cells from complicated patterns is difficult. The CAD systems are developed to assist medical practitioners in such circumstances. Researchers applied various machine learning techniques for the diagnosis of cancer.

To classify nevus and melanoma cancer, [22] proposed a deep learning-based model which consists of ResNet50. The proposed model was trained on 595 different histopathology images. The model is evaluated on a sample image slide which achieves better classification as compared to 11 pathologists with the same image slide. [23] proposed an ensemble based boos-ted Convolution neural network for breast cancer classification. The author used a deep convolution neural network for feature extraction along with augmentation techniques to increase the dataset size. The model is evaluated on the H , E dataset and outperforms all state-of-the-art models. For the classification of breast cancer [24] proposed a CNN-based transfer learning models. The author applied the pre-processing technique to enhance image quality using color normalization. ResNet, VGGNet and Google-Net are used as a classifier, which is evaluated on test samples and shows improved classification accuracy as compared to other models. Pre-trained CNN models are widely used for feature extraction as well as for classification. [25] used pre-trained CNN models for feature extraction from breast cancer images. Extracted features from ResNet50, Xception, VGG16 and VGG19 were then fed to SVM and logistic regression for classification. The proposed model is evaluated on BreakHis Dataset and shows improved classification performance. [26] IL-MCAM is a deep learning framework based on attentional mechanisms and interactive learning. The proposed framework classifies colorectal histopathological images. Based on an experiment conducted on the HE-NCT-CRC-100K dataset, the proposed model achie-ved 98.98% and 99.77% accuracy.

Diagnosing cancer from histopathology whole slide images is a time-consuming process, but it helps to diagnose cancer more accurately than other image modalities. [27] proposed pretrained CNN models for faster diagnoses of cancer from whole slide images. Data augmentation technique is applied for better results are dataset only consists of 306 benign and 315 malignant cells. The proposed model is evaluated on the original dataset as well as on the augmented dataset and shows better performance than other models while using the augmented dataset. A CNN-based DBLCNN network [28] is proposed to acheive better feature representation. Furthermore, MobileNet is redesigned to obtain better recognition result on BreakHis histopathology dataset. [29] proposed a CNN model for the classification of Lung and colon cancer. The proposed model first extract 2D Fourier and 2D wavelet features from images and then combined features are then fed to the proposed CNN architecture for cancer classification. The model is evaluated in the LC25000 dataset which consists of 5 classes, three for lung and 2 for colon cancer, with 96.3% of classification accuracy. A hybird convolution neural network [30] is proposed to classify cancerous histopathology images. The histopathology images were first preprocessed, and then a Hybird network based on two CNN models was used to learn better features for classification.

Different pre-trained CNN models, Alex-Net, VGG-19, ResNet-18, ResNet-34, ResNet-50 and ResNet-101 are applied in [15] for the classification of lung cancer. The pre-trained models are evaluated on the LC25000 dataset on only lung cancer classes and achieve better classification accuracy as compared to other models. [16] applied pre-trained CNN models ResNet-18, ResNet-30, and Res-

Net50 for classification of colon cancer. An experiment was conducted on the LC25000 dataset with colon images only. Results show better classification performance as compared to other CNN models. A dual horizontal squash capsule network is proposed [31], for lung and colon cancer classification. Encoder feature fusion and horizontal squash function are applied which extract important features from images to classify histopathology images. The proposed model is evaluated on the LC25000 dataset which outperforms state of the art models. The MV-CRecNet model [32] uses CT scans to identify lung cancer nodules. [33] applied eight pre-trained CNN models, NasNetMobile, InceptionV3, InceptionResNetV2, ResNet50, VGG16, MobileNet and Dense-Net169 for lung and colon cancer histopathology image classification . SmoothGrad and GradCAM are applied for visualization. The models are evaluated on the LC25000 dataset which shows better performance than other models.

The capsule network-based model [34] is appplied for classification lung and colon cancer on histopathology images. The proposed model normalizes the images before feeding them to the capsule network, which then classifies images. The proposed model is evaluated on the LC25000 dataset and outperforms CNN models. [18] proposed a two-stage model for cancerous histopathology images, first features are extracted using texture analysis and homology-based image processing and then these features are fed to machine learning algorithms, logistic regression, SVM, decision tree, random forest etc. The experiment is conducted on the LC25000 and private dataset, which shows homology-based image processing gives better performance as compared to texture analysis. [35] proposed a model which consists of pre-trained ResNet-50 and DenseNet-121 for feature extraction and classification SVM, Random Forest, KNN, and XGBoost are applied. The proposed model is evaluated on the LC25000 dataset which shows improved classification performance. In the medical field, the main issue faced by researchers globally is a shortage of clinical data. To overcome this issue, researchers applied data augmentation and transfer learning strategies. Several authors examined transfer learning strategy by finetuning the pre-trained models, [36] used inceptionV3 model. [37] applied several CNN-based pre-trained architectures, namely VGG19, MobileNetV2, and DenseNet201 for breast cancer diagnosis. These pre-trained models are fine-tuned on four datasets, the Break-His, PCam, Bio-imaging, and ICIAR. Recently, the vision transformers escalated in the imaging domain with applications from segmentation, detection, and classification. However, the vision transformers are applied with a focus on segmentation in the medical imaging domain. Most of the work consists of custom architectures with a combination of transformer and convolutions, notably work considers convolution free ViT models. The COVID-VIT proposed by [38] applied vision transformer on CT images for covid diagnosis. The 3D dataset is used for the performance evaluation of a model. The proposed model is compared with CNN-based Dense Net architecture and shows better performance. In recent work, [39] classify shoulder implants using a vision transformer. The vision transformer shows better performance compared to different models of machine learning and CNN-based architectures. [40] proposed transformer-based multiple instances learning TransMIL for the classification of whole slide images. [41] proposed a model that effectively diagnoses tuberculosis from chest X-ray images. The pre-trained vision transformer is applied to boost the performance along with Efficient Net. The experiment was carried out on a large dataset with heterogeneous data resources. The results show better performance as compared to other baseline models.

3 Methodology

In this section, the methodology of the study is described. In Sect. 3.1, the proposed vision transformer-based histopathology image classification model is detailed. In Sect. 3.1, the fine-tunning of models is described.

3.1 Vision transformer

We have employed ViT transformer [42] to classify histopathology images. In this section, the structure of VIT is described. The vision transformer consists of a linear embedding layer, an encoder, and a classifier complete architecture diagram of a model is presented in Fig. 1. In the first step, the dataset set is transformed according to the model, let $M = (X_i, y_i)_{i=1}^{n}$, where n is the total number of images in histopathology images dataset, X_i is the image and y_i is the class label, there are five classes, (lung benign tissue, lung adenocarcinoma, lung squamous cell carcinoma, colon adenocarcinoma, colon benign tissue). The image X_i is 720x720 which is reshaped to X_i' 128x128 pixels. Let's take a reshaped image X, which is passed to the model first, an image is divided into non-overlapping histopathology patches. The image $X \in \mathbb{R}^{(h \times w \times c)}$ where h is the height, w is the width and c the channel which in our case is 1 is divided into patches, each patch is $c \times p \times p$ where p is usually of size 16, while a smaller value of p produces better longer sequence. There are t_p, the total number of patches $(x_1, x_2, \ldots, x_{tp})$ are extracted from a histopathology image X, in transformer every patch x is considered as a token.

subsubsectionLinear embedding layer

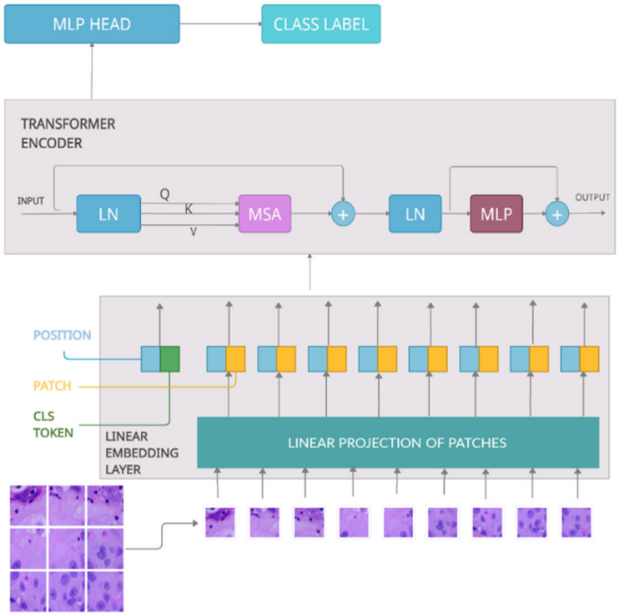

Fig. 1 Vision transformer architecture [42]

The extracted histopathology patches are linearly calculated into a vector of dimension d using embedding matrix E_m then these vectors are fed to an encoder. The learned embedding matrix $E_m \in \mathbb{R}^{(p^2c)\times d}$ project patches to their equivalent linear vectors L_v. The classification token CLS_{token} is concatenated with embedded representation L_v to achieve classification tasks.

$$L_v = [x_1E_m, x_2E_m, \ldots..x_{tp}E_m] \tag{1}$$

$$L_v + cls = [CLS_{token}; x_1E_m, x_2E_m, \ldots..x_{tp}E_m] \tag{2}$$

The position of each patch in a histopathology image is important in an image to keep the position of each patch intact. The 1D positional encoded E_p, where $E_p \in \mathbb{R}^{(m+1)\times d}$, the positional information is concatenated with L_v to form the embedded representation E_b of the patch, $E_b \in \mathbb{R}^{m\times d}$.

$$E_b = [CLS_{token}; x_1E_m, x_2E_m, ...x_{tp}E_m] + E_p \tag{3}$$

3.1.1 ViT encoder

The encoder of the vision transformer consists of u number of blocks. Each block has three components, the normalization layer (LN), multi-head self-attention block (MSA), and Multi-Layer Perceptron (MLP), the fully connected feed-forward dense block, and both MSA and MLP with skip connections. The embedded representation is feed-forward to the encoder, the E_b is normalized and then passed to MSA. The main component of the encoder multi-head self-attention block marks for important information of an embedded patch related to other embedded patches.

The MSA block in Fig. 2 consists of three components, linear block, self-attention layer (SA), and concatenation layer. The attention mechanism helps the model to focus on important information in the input. The linear block is added on both ends of the MSA block. At the start of MSA, the linear block consists of three linear layers. The linear layer consists of fully connected neurons without the activation function. The input is transformed into Query(Q), Key(K), and Value(V) vectors by calculating the scaled dot product input with the weight matrix W. For each vector Q, K, and V different weight matrices W_q, W_k and W_v are used for computation, $W_q \in \mathbb{R}^{d\times d_q}$, $W_k \in \mathbb{R}^{d\times d_k}$, $W_v \in \mathbb{R}^{d\times d_v}$.

$$Query(Q) = E_bW_q$$
$$Key(K) = E_bW_k$$
$$Value(V) = E_bW_v$$

The similarity between Q and K is obtained by calculating the dot product of both vectors. The resultant matrix is then scaled by dividing each value by the dimension of the K matrix d_k, these scaled values are then squashed between 0 and 1 using SoftMax. The resultant matrix called Attention Matrix A contains very important information, where $A \in \mathbb{R}^{(m\times m)}$. The Attention Matrix is then multiplied with the V matrix, which outputs the filter matrix removing all unnecessary information with a focus on the important features. The h the number of filtered matrices obtained from multi attention heads is then concatenated in the concatenation layer. At the end of MSA, the linear block with one linear layer is used to transform the matrix into the desired dimension by multiplying with weight matrix W_d, where $W_d \in \mathbb{R}^{(h.D_k\times D)}$.

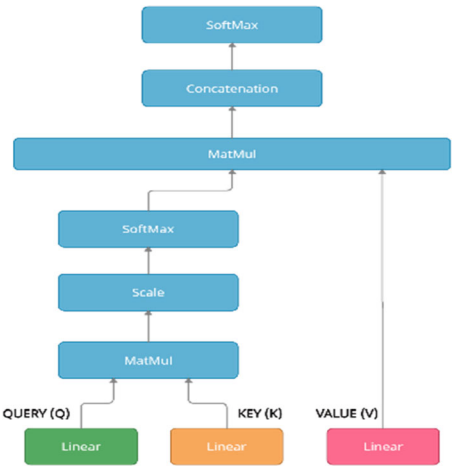

Fig. 2 Vision Transformer MSA Block [42]

$$Attention\ Matrix\ (A) = SoftMax\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{4}$$

$$Filtered\ Matrix\ (F) = A \cdot V \tag{5}$$

$$MSA(E_b) = Concat(F_1(E_b), F_2(E_b), ..., F_h(E_b))W_d \tag{6}$$

$$Y'_u = MSA(LN(E_{bu-1})) + E_{u-1} \tag{7}$$

$$Y_u = MLP(LN(Y'_u)) + Y'_u \tag{8}$$

The resultant matrix generated from the MSA block is normalized (LN) before passing to the MLP block. The MLP block consists of a GeLU activation function applied in between the two linear blocks. As seen in Fig. 3. the relationships among patches are important for the complete analysis of histopathology images. The Fig. 4. shows the CNN models relations among neighbouring pixels, while the vision transformer captures the relationship of one pixel related to all other pixels to classify histopathology images.

The vision transformer has two variants concerning depth, ViT-B base and Vit-L large model. These variants are further subdivided regarding the input patch size of an image, two variants 32x32 and 16x16 patches size is used. The Vit-L and smaller patch sizes are resource exhaustive, as the smaller the patch size and extended depth increase the number of parameters which takes more resources.

3.2 Pre-processing and fine tuning

The proposed scheme for the evaluation of the deep learning models is presented in Fig. 5. The evaluation framework is divided into two steps, preprocessing, which process images before feeding them to the framework. The finetuning step fine-tunes the pre-trained model on histopathology images.

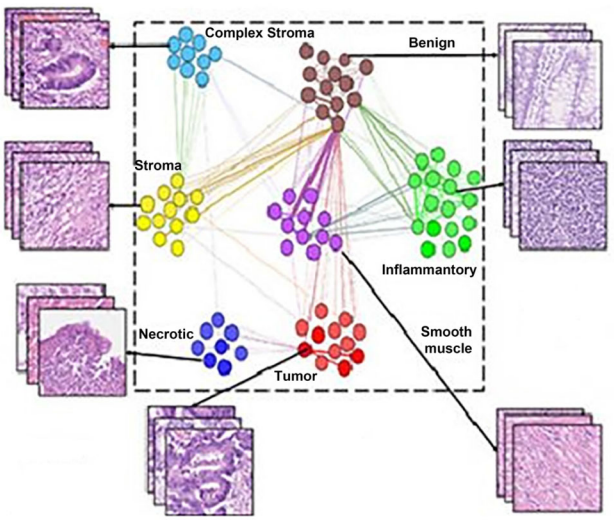

Fig. 3 Relationship among patches in histopathology images [43]

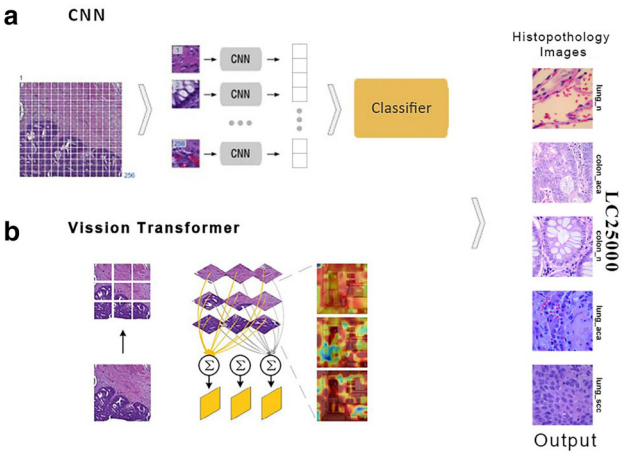

Fig. 4 Vision transformer attention

3.2.1 Preprocessing

Before passing the histopathology images to the pre-trained models, a similar number of images, according to sample size from each class is extracted from the dataset. The extracted sample is divided into three sets, training, validation and test with a ratio of 70, 10, and 20, respectively. The dataset consists of 720x720 size histopathology images. The images are re-shaped to 128x128. The larger size of an input image does not affect the accuracy, but it increases the computational cost by reducing size computational cost is reduced without affecting the accuracy. Furthermore, the training set is divided into a batch size of 40 to reduce the training time.

3.2.2 Fine tuning

The input images are passed to pre-trained models. In this research, vision transformer variants ViT-B32 and ViT-L32, and CNN variants ResNet-50, ResNet-101, InceptionResNetv2, and VGG-16 are applied. The deep learning models are pre-trained on the Imagenet dataset and fine-tuned on the histopathology images. Furthermore, all pre-trained models consist of a flattened layer that fattens the output of a model, followed by three fully-connected layers. The first layer consists of 256 neurons, followed by a layer with 128 neurons both layers consist of the ReLU (rectified linear unit) activation function. The last layers consist of 5 neurons with a softmax activation function to classify histopathology images into five classes.

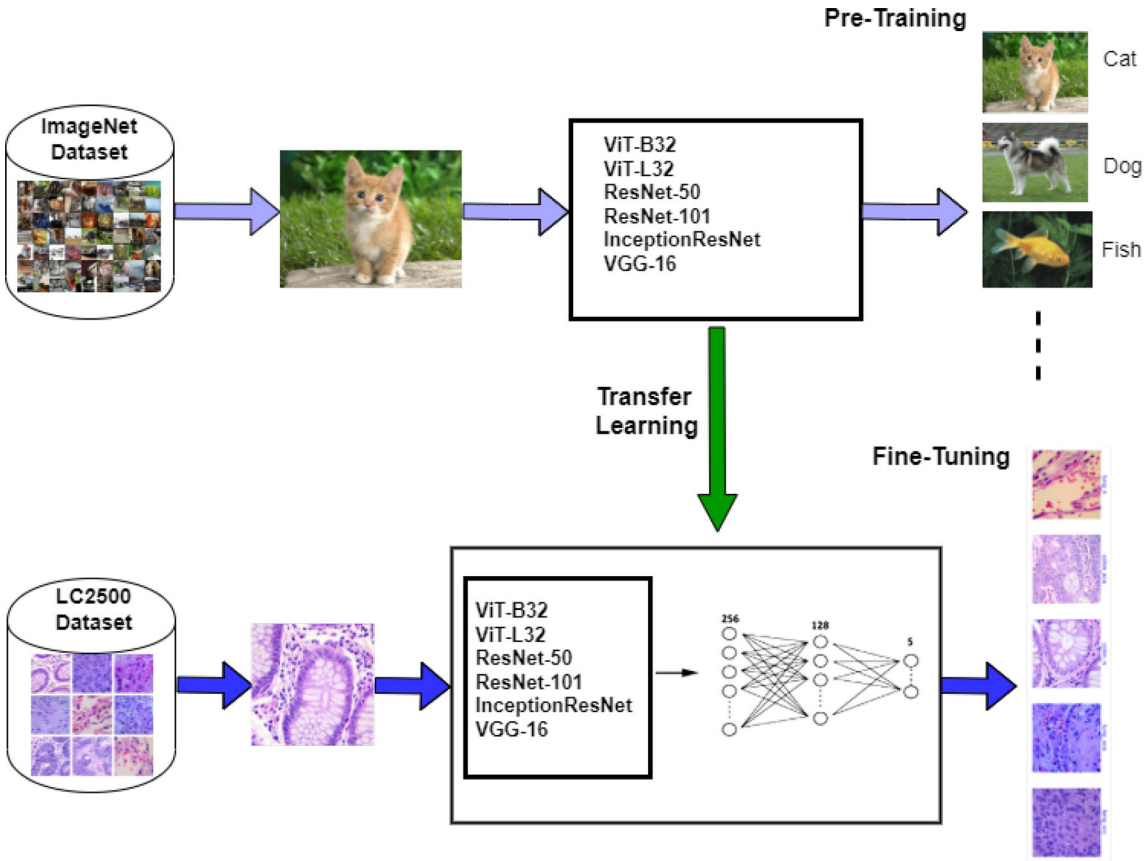

Fig. 5 Scheme for transfer learning of histopathology image classification

4 Experimental results and discussion

The detailed information related to the experimental setup, which includes, the models and platform used in this work and results obtained from the experiment were discussed in this section. All experiments are conducted on Kaggle, with all models implemented in Python using Tensorflow and Keras library. The Kaggle environment consists of 16 GB Nvidia P100 GPU with 12 GB of ram.

In this study, the transfer learning technique is applied on VGGNet, ResNet, Inceptionresnetv2 and vision transformer for the classification of lung and colon cancer on histopathology image modality. In this experiment, the balanced dataset is used to fine-tune all models. The dataset is randomly split into train, validation, and test sets with a ratio of 70:10:20. The training set consists of 70% of the dataset, while 10% for validation and 20% for the test set. Similar, hyperparameters mentioned in Table 1 are applied to all models for evaluation. All models are pre-trained on Imagenet and then fine-tuned on histopathology images. To fine-tune, we train the models for 10 iterations with a batch size of 40. The learning rate of 0.0001 is used while the categorical cross-entropy is a loss function and optimized with the Adamax optimizer method.

4.1 Performance analysis

There are several performance measures to evaluate the models. In this research, a balanced dataset is used, with an equal number of samples in the test set from each class for that classification accuracy Eq. 9 is an effective and extensively used evaluation measure to evaluate model performance. We will also use a confusion matrix for better evaluation of models. The confusion matrix summarizes the correct and incorrect classification along with that, from the confusion matrix, other metrics precision Eq. 10, recall Eq. 11 and F1-score in Eq. 12 can be calculated to evaluate the performance of models on each class.

Table 1 Hyper Parameters

Parameters	Value
Epochs	10
Learning Rate	0.0001
Loss Function	Categorical Crossentropy
Optimizer	Adamax

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{12}$$

4.2 Dataset

In this experiment, the LC25000 [44] dataset is used for the evaluation of models. The dataset consists of 25000 histology images of 768x768 pixels divided into two categories Lung and Colon images. The 15000 lung images with three classes of lung adenocarcinoma, lung squamous cell carcinoma, and benign lung tissue with 5000 images each as seen in Fig. 6. Similarly, 10000 colon images with two classes of colon adenocarcinoma and benign colon tissue with 5000 images each.

4.3 Results

The CNN-based architectures are the state-of-the-art models in the field of computer vision. Recently, vision transformers have also shown encouraging results to solve numerous computer vision problems. In medical imaging, there is a lack of annotated data therefore, transfer learning is applied to solve problems. The CNN architectures are widely used as they are state-of-the-art models but, studies confirm that CNN does not perform well in the medical domain if trained on the natural images dataset ImageNet. The results of vision transformers are compared with VGG16, ResNet50, Res-Net101, and InceptionResNetV2, which are considered the best image classifiers. All models used similar configurations with the same learning rate, optimizer, loss function, number of iterations, dataset sample, and MLP head. For better evaluation, several experiments are performed with different numbers of samples in a dataset. The study aims to evaluate the performance of vision transformers on the different number of samples in a dataset and compare them with the state-of-the-art CNN-based architectures.

Fig. 6 Sample images LC25000 dataset [44]

For the evaluation of deep learning models, five data-set samples were applied to evaluate the performance of learning models with different numbers of samples in Table 2. The LC25000 is a balanced dataset, and all samples contain an equal number of images from each class. For each sample, 20% sampled images belong to the test set, while 10% to validation and the remaining 70% to the training set.

The classification performance of ResNet-50, ResNet-101, InceptionResNetV2, VGG16, and Vision Transformer variants ViT-B32 and ViT-L32 models are evaluated on LC25000 histopathology image dataset. The summary of the result obtained from pre-trained models is presented in graphical figures and tabular form using performance metrics. All pre-trained models are fine-tuned with 10 epochs to classify histopathology images. The training accuracy in Fig. 7 and validation accuracy in Fig. 8 of all fine-tuned models are presented. The training and validation accuracy of VGG-16 is slightly better than other models. The accuracy graph shows that the accuracy of all models remains stable after 4 epochs or shows minor improvement till 10 epochs.

Among contemporary CNN and transformer models, which one has better efficacy to classify histopathology images?

The accuracy and loss of all fine-tuned models are shown in Table 3. According to the results obtained, the variant of vision transformer ViT-L32 produced better classification accuracy on the test set on all samples of a dataset. The ViT-L32 architecture has deeper architecture than ViT-B32 and learns long-range relationships as compared to CNN architectures which only learn nearby correlation among pixels. Among all models, the ViT-L32 model misclassified only 7 images out of 3000 images in a test set of sample 3 of a dataset. However, other models have a higher misclassification rate, the VGG-16 model has a test accuracy of 99.60% highest among models after ViT-L32 misclassified 12 images out of 3000 images of the test set. The ViT-L32 and VGG-16 achieve 100% average precision, recall, and F1-score but, 0.17% higher accuracy is achieved by ViT-L32 than VGG-16 with 99.77% test accuracy.

Table 2 Dataset samples used for evaluation

Dataset	Total images	Training set	Validation set	Test set
Sample 1	5000	3500	500	1000
Sample 2	10,000	7000	1000	2000
Sample 3	15,000	10,500	1500	300
Sample 4	20,000	14,000	2000	4000
Sample 5	25,000	17,500	2500	5000

What is the effect of the increasing number of images in a dataset to fine-tune the vision transformer and CNN to classify histopathology images?

The insights and observations obtained from an ablation study performed on different pre-trained models show that vision transformers perform better than CNN-based models for histopathology image classification using transfer learning. As seen in Fig. 9, the performance of all employed models improved as the number of images increased in fine-tuning. The highest accuracy 99.94% was achieved by vision transformer after fine-tuning on a larger dataset sample of histopathology images. The obtained results show that the larger dataset for fine-tuning achieves better performance as compared to the smaller dataset.

Furthermore, an entire slide image based on the TCGA dataset is also examined with the proposed framework. First, the WSI is divided into 128x128 patches, after which each patch is passed through a proposed vision transformer for classification into corresponding classes. In Fig. 10, a WSI is represented as a combination of all classified patches obtained from proposed framework.

Do pre-training vision transformers help to improve their efficacy to classify histopathology images?

An experiment was conducted on both variants of the vision transformer ViT-B32 and ViT-L32 with and without a transfer learning scheme. The results obtained from the experiment show that the vision transformer performs better in a transfer learning scheme. The vision transfer pre-trained on ImageNet was finetuned on histopathology images, the results in Fig. 11, show there is 9% more performance gain in a transfer learning scheme as compared to a model without a transfer learning scheme.

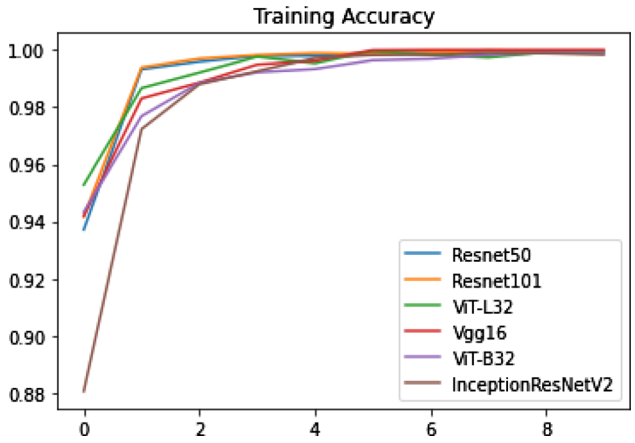

Fig. 7 Training accuracy on sample 3

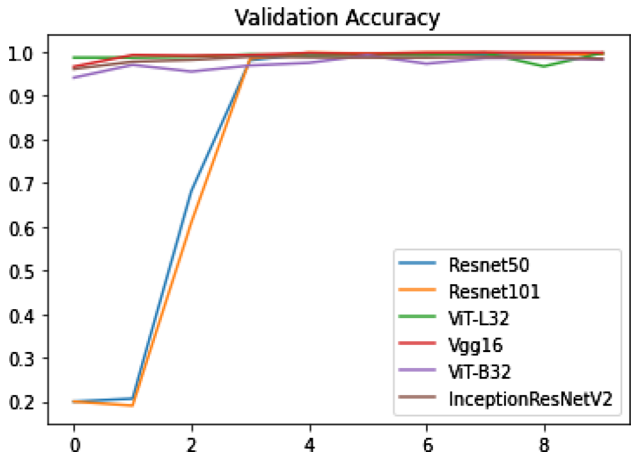

Fig. 8 Validation accuracy on sample 3

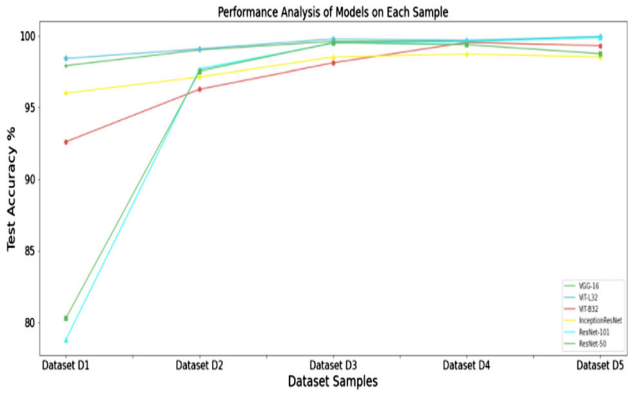

Fig. 9 Performance of models on different dataset samples

Table 3 Comparison of Models Pre-Trained on ImageNet on Sample 3

Classifier	Loss %			Accuracy %		
	Training	Validation	Test	Training	Validation	Test
VGG-16	0.00002	0.87	0.78	100	99.80	99.60
ViT-L32	0.20	0.98	0.64	99.92	99.73	99.77
ViT-B32	0.41	7.79	7.78	99.85	98.20	98.10
InceptionResNetV2	0.54	124.9	237.73	99.85	98.40	98.50
ResNet-101	0.34	1.03	1.42	99.90	99.47	99.47
ResNet-50	0.20	1.40	1.45	99.94	99.53	99.50

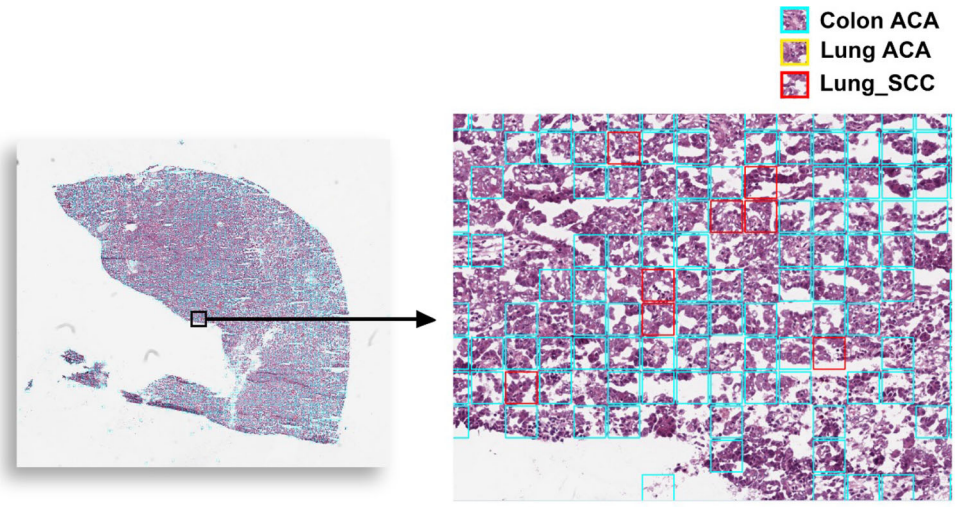

Fig. 10 Whole Slide Image Patch Wise Classification

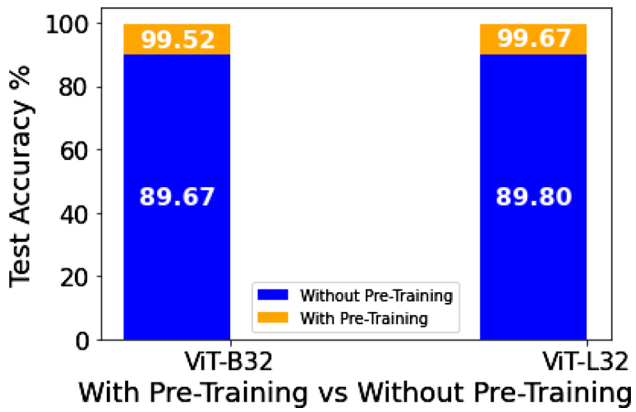

Fig. 11 Vision Transformer with pre-training vs without pre-training

5 Conclusion

The CAD systems are sensitive systems that require high performance in the diagnosis of disease. The CAD systems help medical practitioners to identify and provide medication for the diagnosed disease. Therefore, a CAD system should have low false negatives and positives to be trusted by patients and medical practitioners. In this study, an extensive evaluation of different pre-trained deep learning models is performed to classify lung and colon cancer using histopathology images. The publicly available dataset LC25000 is used which consists of 25000 images, 15000 images of Lung cancer belonging to 3 classes and 10000 colon cancer images belonging to 2 classes. The SOTA CNN models, ResNet50, ResNet-101, InceptionResNetV2, VGG16, and two variants of vision transformer ViT-L32 and ViT-B32 are applied. Numerous performance evaluation metrics are applied, such as accuracy, precision, recall and F1-Score. The annotated datasets in medical imaging are limited therefore, the transfer learning scheme is applied. All models are pre-trained on Imagenet and finetuned on the LC25000 dataset. The vision transformer variant ViT-L32 achieved the best accuracy in all five experiments conducted with a different number of samples. This study proves that the vision transformer works better than CNN models on histopathology images, pre-trained on ImageNet. The vision transformer achieves higher accuracy than CNN when the transfer learning scheme is applied.

6 Future work

In future, we will apply the vision transformer to other imaging modalities. The vision transformer can be applied to other computer vision tasks. Moreover, the histopathology images datasets are limited, and only one dataset, the LC25000 is available publically that contains patches of whole slide images a lot of work require in data acquisition of histopathology cancer images.

Declarations

Conflict of interest Tayyab Aitazaz, Abdullah Tubaishat, Feras Al-Obeidat, Babar Shah, Tehseen Zia and Ali Tariq declare that they have no conflict of interest.

References

1. Farmer P, Frenk J, Knaul FM, Shulman LN, Alleyne G, Armstrong L, Atun R, Blayney D, Chen L, Feachem R et al (2010) Expansion of cancer care and control in countries of low and

middle income: a call to action. The Lancet 376(9747):1186–1193
2. Pakistan G (2021) "International Agency on Research for Cancer." https://gco.iarc.fr/today/data/factsheets/populations/586-pakistan-fact-sheets.pdf/, [Online; accessed 05-October-2021]
3. Scheel BI, Holtedahl K (2015) Symptoms, signs, and tests: The general practitioner's comprehensive approach towards a cancer diagnosis. Scandinavian journal of primary health care 33(3):170–177
4. Whitaker K (2020) Earlier diagnosis: the importance of cancer symptoms. The Lancet Oncology 21(1):6–8
5. Thakor AS, Gambhir SS (2013)"Nanooncology: the future of cancer diagnosis and therapy," CA: a cancer journal for clinicians, vol.63, no.6, pp.395–418
6. Frysh P (2021)"Can You Find Cancer Without a Biopsy?." https://www.webmd.com/cancer/detect-cancer-without-biopsy, [Online; accessed 05-October-2021]
7. Santoso JT, Coleman RL, Voet RL, Bernstein SG, Lifshitz S, Miller D (1998) Pathology slide review in gynecologic oncology. Obstetrics & Gynecology 91(5):730–734
8. Bhatt C, Kumar I, Vijayakumar V, Singh KU, Kumar A (2021) The state of the art of deep learning models in medical science and their challenges. Multimedia Systems 27(4):599–613
9. Lipton ZC, Berkowitz J, Elkan C (2015)"A critical review of recurrent neural networks for sequence learning," arXiv preprint arXiv:1506.00019
10. Lipton ZC, Kale DC, Elkan C, Wetzel R (2015)"Learning to diagnose with lstm recurrent neural networks," arXiv preprint arXiv:1511.03677
11. Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. Communications of the ACM 60(6):84–90
12. He K, Zhang X, Ren S, Sun J (2016) "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, (Las Vegas, Nevada, USA.), pp.770–778
13. Redmon J, Divvala S, Girshick R, Farhadi A (2016)"You only look once: Unified, real-time object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, (Las Vegas, Nevada, USA.), pp. 779–788
14. Ronneberger O, Fischer P, Brox T (2015) "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention, (Munich, Germany), pp. 234–241, Springer
15. Abbas MA, Bukhari SUK. Syed A, Shah SSH (2020)"The histopathological diagnosis of adenocarcinoma & squamous cells carcinoma of lungs by artificial intelligence: A comparative study of convolutional neural networks," medRxiv, https://doi.org/10.1101/2020.05.02.20044602
16. Bukhari SUK, Asmara S, Bokhari SKA, Hussain SS, Armaghan SU, Shah SSH (2020)"The histological diagnosis of colonic adenocarcinoma by applying partial self supervised learning," medRxiv, https://doi.org/10.1101/2020.08.15.20175760
17. Simonyan K, Zisserman A (2014)"Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556
18. Nishio M, Nishio M, Jimbo N, Nakane K (2021) Homology-based image processing for automatic classification of histopathological images of lung tissue. Cancers 13(6):1192
19. Fan J, Lee J, Lee Y (2021)"A transfer learning architecture based on a support vector machine for histopathology image classification, " Applied Sciences, vol. 11, no.14, p.6380
20. Toğaçar M (2021) Disease type detection in lung and colon cancer images using the complement approach of inefficient sets. Computers inBiology and Medicine 137:104827
21. Raghu M, Zhang C, Kleinberg J, Bengio S (2019) "Transfusion: Understanding transfer learning for medical imaging," arXiv preprint arXiv:1902.07208
22. Hekler A, Utikal JS, Enk AH, Solass W, Schmitt M, Klode J, Schadendorf D, Sondermann W, Franklin C, Bestvater F et al (2019) Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. European Journal of Cancer 118:91–96
23. Vo DM, Nguyen N-Q, Lee S-W (2019) Classification of breast cancer histology images using incremental boosting convolution networks. Information Sciences 482:123–138
24. Khan S, Islam N, Jan Z, Din IU, Rodrigues JJC (2019) A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. Pattern Recognition Letters 125:1–6
25. Gupta K, Chawla N (2020) Analysis of histopathological images for prediction of breast cancer using traditional classifiers with pre-trained cnn. Procedia Computer Science 167:878–889
26. Chen H, Li C, Li X, Rahaman MM, Hu W, Li Y, Liu W, Sun C, Sun H, Huang X et al (2022) Il-mcam: An interactive learning and multi-channel attention mechanism-based weakly supervised colorectal histopathology image classification approach. Computers in Biology and Medicine 143:105265
27. Teramoto A, Yamada A, Kiriyama Y, Tsukamoto T, Yan K, Zhang L, Imaizumi K, Saito K, Fujita H (2019) Automated classification of benign and malignant cells from lung cytological images using deep convolutional neural network. Informatics in Medicine Unlocked 16:100205
28. Wang C, Gong W, Cheng J, Qian Y (2022) Dblcnn: Dependency-based lightweight convolutional neural network for multi-classification of breast histopathology images. Biomedical Signal Processing and Control 73:103451
29. Masud M, Sikder N, Nahid A-A, Bairagi AK, AlZain MA (2021) A machine learning approach to diagnosing lung and colon cancer using a deep learning-based classification framework. Sensors 21(3):748
30. Angayarkanni SP (2022)"Hybrid convolution neural network in classification of cancer in histopathology images," Journal of Digital Imaging, pp.1–10
31. Adu K, Yu Y, Cai J,Owusu-Agyemang K, Twumasi BA, Wang X (2021)"Dhs-capsnet: Dual horizontal squash capsule networks for lung and colon cancer classification from whole slide histopathological images," International Journal of Imaging Systems and Technology
32. Naeem Abid MM, Zia T, Ghafoor M, Windridge D (2021) Multi-view convolutional recurrent neural networks for lung cancer nodule identification. Neurocomputing 453:299–311
33. Garg S, Garg S (2020)"Prediction of lung and colon cancer through analysis of histopathological images by utilizing pre-trained cnn models with visualization of class activation and saliency maps," in 2020 3rd Artificial Intelligence and Cloud Computing Conference, (Kyoto, Japan.), pp.38–45
34. Roy Medhi BB (2020) Lung Cancer Classification from Histologic Images using Capsule Networks. PhD thesis, Dublin, National College of Ireland
35. Sarwinda D, Bustamam A, Paradisa RH, Argyadiva T, Mangunwardoyo W (2020)"Analysis of deep feature extraction for colorectal cancer detection," in 2020 4th International Conference on Informatics and Computational Sciences (ICICoS), (Semarang, Indonesia.), pp.1–5, IEEE
36. Chougrad H, Zouaki H, Alheyane O (2017) "Convolutional neural networks for breast cancer screening: transfer learning with exponential decay," arXiv preprint arXiv:1711.10752
37. Kassani SH, Kassani PH, Wesolowski MJ, Schneider KA, Deters R (2019)"Classification of histopathological biopsy images using

ensemble of deep learning networks," arXiv preprint arXiv:1909.11870
38. Gao X, Qian Y, Gao A (2021)"Covid-vit: Classification of covid-19 from ct chest images based on vision transformer models," arXiv preprint arXiv:2107.01682
39. Zhou M, Mo S (2021)"Shoulder implant x-ray manufacturer classification: Exploring with vision transformer," arXiv preprint arXiv:2104.07667
40. Shao Z, Bian HY, Chen Wang Y, Zhang J, Ji X, Zhang Y (2021)"Transmil: Transformer based correlated multiple instance learning for whole slide image classication," arXiv preprint arXiv:2106.00908
41. Duong LT, Le NH, Tran TB, Ngo VM, Nguyen PT (2021) Detection of tuberculosis from chest x-ray images: Boosting the performance with vision transformer and transfer learning. Expert Systems with Applications 184:115519
42. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D,Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold, G, Gelly S et al (2020) "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929
43. Javed S, Mahmood A, Fraz MM, Koohbanani NA, Benes K, Tsang Y-W, Hewitt K, Epstein D, Snead D, Rajpoot N (2020) Cellular community detection for tissue phenotyping in colorectal cancer histology images. Medical Image Analysis 63:101696
44. Borkowski AA, Bui MM, Thomas LB, Wilson CP, DeLand LA, Mastorides SM (2019) Lung and colon cancer histopathological image dataset (lc25000). arXiv preprint arXiv:1912.12142