Assessment of a deep-learning system for colorectal cancer diagnosis using histopathology images

Purna Kar[1] **and Sareh Rowlands**[1,*]

[1]College of Engineering, Mathematics, and Physical Sciences,University of Exeter, Exeter, EX4 4QG, United Kingdom
[*]S.Rowlands@exeter.ac.uk

ABSTRACT

Colorectal Cancer is one of the most common forms of cancer hence, an early and accurate detection is crucial. Manual diagnosis is a tedious and time-consuming job which is prone to human errors as it involves visual examinations of pathological images. Therefore, it is imperative to use computer-aided detection (CAD) systems to interpret the medical images for a quicker and more accurate diagnosis. The traditional methods for diagnosis comprise extraction of features based on texture, pattern, illumination etc. from pathological images and then use these features in a Machine Learning model for binary classification i.e., cancerous, or non-cancerous. Deep-learning approaches like the Convolutional neural networks (CNNs) have proved to be very effective in classifying and predicting cancer from pathological images. In this study, we have assessed several CNN-based techniques for cancer diagnosis on digitized histopathology images. We have also compared the results of traditional methods for diagnosis with the deep-learning models. Moreover, we have proposed a new model by borrowing the idea from Xception architecture (Xception+), which outperforms the existing architectures. We have studied the effect of transfer learning technique by using models pre-trained on unrelated histopathology images, performed segmentation of glands in the histology images and also performed cancer grade classification.

Introduction

According to the American Institute of Cancer Research[1] Colorectal Cancer (CRC) is the third most common form of cancer after lung and breast cancers, it constitutes almost 10 percent of the total cancer cases worldwide. The mortality rate of colorectal cancer is 9 percent of all the deaths attributed to cancer. However, the five-year survival rate when detected at an early stage (regionalized stage) in the USA is as high as 70 percent, this makes it vital for an early and accurate detection. The diagnosis of Colorectal Cancer demands a thorough visual examination of digital whole-slide images (WSIs) of H and E-stained histology images. A pathologist determines the stage of cancer by counting the number and size of tumors in the section of the images. This makes the task extremely monotonous and prone to errors as the size of tumor cells are minute they can be easily overlooked, sometimes it requires substantial amount of amplifying of the image to determine a tumor cell. According to the American Institute of Cancer Research[1] the cases for colorectal cancer world-wide are expected to rise by 60 percent over the next 15 years, therefore, the need for diagnosis will also increase rapidly which would prove disastrous if pathologists only relied on manual examinations. Thus, it is essential to take the help of computer-aided detection (CAD) systems to improve the precision as well as diminish the time and manual effort. With the advancement of image processing and computer vision disciplines there has been a huge improvement in computer-aided detection (CAD), state-of-the-art deep neural networks have replaced the traditional methods of feature extraction and classification. Traditional methods comprise two steps, first an image descriptor is used to encode the texture and patterns in an image called ’features’ into a feature matrix, then this feature matrix is used in a supervised machine learning based classifier like SVM to classify the images into cancerous and non-cancerous classes. Various researches[2] have incorporated the traditional supervised methods, but the time taken to extract features and classify images is very high and the precision of these traditional methods are very mediocre. The advent of Artificial Neural networks has revolutionized the field of machine learning and computer vision, Convolutional Neural Networks (CNNs) have been since used in most image processing problems[3]. The performance of these CNN based models has also vastly improved from the traditional methods. Nowadays, CNN based deep-learning models are used in most computer vision problems like pattern recognition, object recognition, tracking etc. Deep-learning techniques like the Convolutional neural networks (CNNs) have proved to be effectual in the prediction and classification problems[3]. Of late CNN based deep-learning techniques have proved to be very effective in analyzing various pathological images for various oncology and clinical studies for cancer. Post cancer diagnosis studies have incorporated techniques like grade classification[4–6] , tumor cell detection[7,8] , gland segmentation[5] and even speculation of the patient survivorship[9]. A small number of the researches in the field of cancer diagnosis use CNN based deep-learning methods. Some related works have aimed to do comparative studies of the CNN architectures[10,11] , some proposed new adaptive CNN models from scratch[10,11] which performed better than the state-of-the-art models for the dataset used by them, some studied the effect of transfer learning[10]. Segmentation of glands also have been incorporated by researches[5] to quantify morphology of glands which helps pathologists to perform clinical diagnosis in a better manner. One shortcoming of the deep-learning models is the requirement of massive amount of labelled data to train the model. In context of pathological datasets this is a huge challenge as datasets with proper labels is in short supply due to various reasons. Proper labelling of images is very costly as it requires visual scrutiny by pathologists which is strenuous and time consuming. There is also the concern of privacy, one needs to be very careful not to breach any privacy policies and must ensure that any data or information cannot be traced back to patients whose images are used. This is a huge setback in using deep-learning models for diagnosis in bio-medical field. Despite the challenges deep-learning methods are finding extensive use in several bio-medical problems[12] due to the accuracy of prediction and classification. To diminish the shortage of labelled dataset the technique of patch generation[13] from a single image is very useful, it increases the number of labelled images thus increasing the size of dataset considerably. Most researches use patch-based techniques[13] to expand the dataset. In this study we have used a huge dataset[14] (almost 1,00,000) of digitized HE-stained colorectal cancer histopathology images annotated into 9 tissue classes to detect cancer using 8 CNN based models. Some of the models are simpler like AlexNet, other state-of-the-art models like GoogLeNet, Inception v3, Xception, MobileNet etc. have also been used. We have addressed two sets of problems in this study, firstly, we have used several CNN based models for Cancer Detection, i.e., determine whether an image belongs to cancerous or non-cancerous class by predicting whether images have cancerous tumor present in it or not. Secondly, we have performed classification of tissues present in the dataset which consists of 9 tissue annotations (Supplementary Figure 3); we have designed models to determine which tissue class the image belongs to. We have proposed a new CNN based model by making slight modifications to the Xception architecture (henceforth will be called Xception+). Additionally, we have also performed an experimental analysis of the effect of Transfer Learning on performance of the models by using a model which was pre-trained on lung cancer images. Moreover, we have aimed to perform semantic segmentation of glands on the histology images. In computer vision image segmentation aims to separate an image into numerous image segments. This technique is very useful in multiple problems like object recognition, facial recognition etc. In bio-medical field this technique has immense significance as it can help in separating tissues, glands, nuclei from muscles. The colon consists of glands and other tissues, colorectal adenocarcinoma or colon cancer causes deformation of the glandular structures thus the separation of glands from background tissues from the histology images can depict the changed morphology of glandular parts afflicted by cancer. A quality segmentation can assist in grading Colorectal Cancer from the tissue slides. In this experiment we have aimed to perform a segmentation of the colorectal cancer images, the segmented images will show glands and background tissues in different illumination. Finally, we have performed cancer grade classification i.e., determining the grade of colon cancer from histopathology tissue slides using CNN based models which performed well in cancer detection and cancer classification problems like GoogLeNet, Xception and our proposed model (Xception+).

Results

Performance of traditional methods

The performance of traditional methods LBP[15] and Haralick[16] has been poor. For this problem the binary classes are defined as 0 (non-cancer) and 1 (Cancer). The accuracy of LBP is 76.46 percent and that of Haralick is 75.20 percent. Both methods showed higher accuracy in predicting the cancer class (i.e., class 1). Fig. 1 a and b, plots the confusion matrices for the two methods. A confusion matrix visualizes and summarizes the overall performance of the classification model. It visualizes the True Negative (TN): (actual class is non-cancer and classified as non-cancer) , False Negative (FN): (actual class is cancer but classified as non-cancer), False Positive (FP): (Actual class non-cancer but classified as cancer) and True Positive (TP): (Actual and predicted class as cancer) accuracy rates of the algorithm. The deep-learning method GoogLeNet has performed much better than LBP and Haralick, the accuracy achieved by it is almost 20 percent higher than that of the traditional methods, it has predicted the non-cancer class with slightly better accuracy than the cancerous class. Fig. 1 c, visualizes the accuracy of GoogLeNet and the traditional methods in predicting the binary classes. The overall accuracy achieved by GoogLeNet is 96.98 percent for the same image dataset.

Performance of deep-learning models

Two optimizers were used in the training phase of the deep-learning models: Adam (uses features of AdaGrad and RMSProp algorithms) and SGD (extension of Gradient Descent). Optimizers are algorithms for stochastic gradient descent for training the deep-learning models. The chart in Figure 8 a, shows the performance of each model using these two optimizers for the two problems i.e., Cancer detection and Tissue classification. It is observed the performance of the models using SGD optimizers are slightly better in most cases. The hyperparameters for the deep-learning models have been kept constant for all models. Below are the values for the hyper parameters.

- Learning rate=0.0005
- Weight decay=0.001 (only for Adam Optimizer)
- Momentum=0.9 (only for SGD)
- Epochs=30

For the cancer detection problem, we have used 4 metrics (Accuracy, precision, recall and F1 score) to evaluate the performance of the different models. We have plotted the four metrics for all the deep-learning models for the Cancer Detection problem as shown in Figure 2 a. It is observed that the GoogLeNet architecture has the highest precision of 99.11 percent, whereas the mean accuracy, recall and F1 score of Xception is the highest (99.25 percent, 99.70 percent and 99.14 percent respectively). Overall, Xception has the best performance amongst all the deep-learning models. We have also plotted a Receiver Operating Characteristic Curve (ROC) for the 3 models Xception, AlexNet and Inception v3 respectively shown in Figure 2 b. The ROC curve for a binary classifier shows the performance of the model at various threshold settings. The ROC curve is plotted with TPR (True Positive Rate) vs FPR (False Positive Rate). TPR is nothing but Recall, FPR is defined by the formula: FP/(FP+TN). AUC stands for the Area under the Curve, it is a performance metric widely used for classification problems, higher the AUC the better is the model at predicting the classes. As seen in Figure 2 b. The AUC for Xception is 1.0 which is the ideal value, this implies that this model is highly accurate in classifying the images.

For the tissue classification problem, we have compared the accuracy of the models in classifying the 9 tissue classes. From Figure 3 a, it is observed that GoogLeNet has the highest accuracy for classifying across all tissue classes followed by Xception. The Figure 3 b, summarizes the performance of each deep-learning model for both the problems i.e., Cancer detection and tissue classification. As demonstrated by the results, Xception has the best mean accuracy (99.25 percent) for the Cancer detection problem and GoogLeNet has displayed the best mean accuracy (98.86 percent) in the tissue classification problem. We have plotted the training loss, validation loss and validation accuracy vs the epochs observed in the training phase of GoogLeNet model in Figure 4 b. We can see that the training loss of the model exponentially decreases, the overall loss in the validation also decreases but some occasional spikes in loss which amounts to overfitting which is overcome by several mechanisms used in the algorithm like image augmentations, use of hyper-parameters like weight decay and mechanisms like back propagation. The validation accuracy also gradually increases.

For the various CNN architectures the test accuracy has been recorded for 5 experiments and the mean and the standard deviation of the accuracy has been tabulated in 8 c.

Further results are provided in Supplementary Tables 1–2 in Supplementary Information.

Performance of Proposed Model

The proposed model based on Xception architecture (will be referred henceforth as Xception+) performed exceptionally well, for both the classification problems. It achieved an overall mean accuracy of 99.37 percent and 98 percent for the cancer detection and the tissue classification problems respectively.

Transfer Learning vs Learning from scratch

It has been observed that the accuracy of the model pre-trained on histology images of lung cancer performed slightly better than the traditional approach for training a deep-learning model, i.e., GoogLeNet model trained from scratch on colorectal cancer images only. The accuracy of the model trained on lung cancer images is 99.47 percent, when this model is used for training on colorectal images the predictive accuracy is 98.19 percent whereas the accuracy of the GoogLeNet model trained only on colorectal images have accuracy of 98.15 percent. The accuracies of the models have been plotted in Figure 4 a.

Semantic segmentation of glands

Due to unavailability of the image masks of colorectal cancer we have used histology images of glands and their respective masks[5] to train a UNet based model for the semantic segmentation task. Figure 5 a, shows the glands and their corresponding masks that have been used to train the model. The validation accuracy achieved by the UNet model is 84.11 percent. It has been observed that there are slight differences between the predicted segmented image and the actual mask. After training the UNet model to perform semantic segmentation, we have used this model on the histopathological images of colorectal cancer and predicted the corresponding segmented images. The segmented images show only two morphological segments, the glands, and the background for each patch of image. Figure 5 b, shows the resultant segmented images. Though image masks for colorectal cancer were not used the semantic segmentation results were promising.

Grade Classification

We have used the dataset from[6] which had only 139 histopathology images, consisting of 71 normal, 33 low grade and 35 high grade cancer images. We have used patch generation technique to generate 15,520 patches. We have used 3 CNN models (GoogLeNet, Xception and Xception+) for training and validating on these patches for classification, out of which the best average accuracy of validation of GoogLeNet was 92.86 percent and Xception was 92.86 percent. Xception+ (our proposed model) achieved an accuracy of 94.48 percent.

Discussion

The deep-learning models for image classification have out-performed the traditional methods, the average accuracy of the GoogLeNet based model is 98.15 percent which is almost 20 percent more than that of methods using LBP or Haralick. In comparison with the works of Junaid Malik. et. all[10] the accuracy achieved by them is 79.5 percent for rLBP and 65 percent for Haralick using Linear Kernel for SVM classifier for Cancer detection problem. Pavel Kr´al et. all[2] achieved an accuracy of 84 percent for LBP for detecting cancer on Breast Cancer images while the accuracy achieved in our project is 76.46 percent and 75.20 percent for LBP and Haralick respectively. For deep-learning based models based on the results Xception and GoogLeNet has outperformed rest of the state-of-the-art architectures. For the Cancer detection problem, four metrics accuracy, precision, recall and F1 score were used to compare the efficiency of the model. These metrics are also called "Confusion Metrics" which uses the counts for TN (True Negative), FN (False Negative), FP (False Positive) and TP (True Positive). TP and TN are the observations that are correctly predicted. Accuracy is the ratio of correctly predicted observations to the total observations (useful when there is no class imbalance) whereas precision is the ratio of correctly predicted observations to the total positive observations, in our problem how many actually has cancer. Recall is the ratio of correctly predicted positive observations to all observations in actual class. This metric answers the question of all cancer cases how many we labelled accurately. F1 score is the weighted average of Precision and Recall, this metric is useful if we have unbalanced class. Since the dataset used by us has equal class distribution, the accuracy and F1 score are approximately similar. The best results were shown by Xception (mean Accuracy: 99.25 percent, Precision: 98.60 percent, Recall: 99.70 percent and F1 score: 99.14 percent) followed by GoogLeNet (mean Accuracy: 99.14 percent, Precision: 99.11 percent, Recall: 99.25 percent and F1 score: 99.13 percent). The precision of GoogLeNet is slightly better that that of Xception which would imply GoogLeNet predicts the cancerous class with better perfection. The AUC for the Xception and GoogLeNet models is 1.0, whereas that of DenseNet and ResNet is 0.999 which suggests that all the models are highly efficient in predicting the classes. For tissue classification problem as well GoogLeNet had the best performance with mean accuracy of 98.86 percent, all the tissue classes has been predicted with highest accuracy by GoogLeNet. Xception, Inception v3 and ResNet were also very effective in predicting the 9 tissue classes.

There might be few reasons why Xception and GoogLeNet outperformed rest of the state-of-the-art neural networks. As the network goes deeper it becomes more difficult to train and beyond a certain point the test loss increases which overfits the models and does a poor generalization which in turn has a detrimental effect on the performance and accuracy of the model. Moreover, deeper networks are more prone to the vanishing gradient problem. By increasing more layers with activation functions like sigmoid, the gradients of the loss function tends to zero making the network difficult to train. The gradients of the loss functions for each layer is computed using backpropagation method which uses chain rule, if the gradients are very low, with each layer the gradient reduces exponentially and by the time it propagates to the initial layer it approaches zero. The main purpose of the backpropagation is to find the optimum amount for changing the weights and biases of the learnable parameters. If the gradient of the initial layers are very low then the learnable parameters of these layers will not be updated properly thus the performance will get saturated. The architecture of GoogLeNet is less deep than all the state-of-the-art models hence GoogLeNet is less prone to vanishing gradient problem and in Figure 4 c we can see with each training cycle (epoch) the accuracy increases faster than most models. Figure 4 c, demonstrates the change of validation accuracy after each epoch for GoogLeNet and Inception v3 models. Xception on the other hand uses skip connections, skip connections allow the gradients to propagate to the initial layers with greater magnitude by skipping few in-between layers thus tackling the vanishing gradient problem. In addition to it in Xception model convolutions are not performed across all the channels which makes the connections lesser and the architecture less deep and massive making the model more trainable and less prone to overfitting.

It is observed that our proposed model Xception+, which is obtained by modifying the Xception network to a smaller size architecture, provides better mean accuracy in both the problems (99.37 percent and 98 percent). This may imply that re-training a huge number of layers might lead to the vanishing gradient problem and overfitting, thus this new model (Xception+) which has fewer layers performs better than the original architecture. The table in Figure 8 b, demonstrates the accuracy obtained by various researchers who have aimed to solve similar problems. It is observed that the Xception and GoogLeNet models we used in our experiments performed better compared to other existing results. The primary reason for better performance of our models could be due to the fact that huge amount of data (almost 30,000) was used by us to train the models, thus the models had very efficient learnable parameters. Moreover, to improve the performance of all the methods, we designed a balanced dataset where we have used equal number of images from each class.

Our experiment demonstrated that transfer learning performs slightly better than the model trained from scratch. Transfer Learning technique is said to improve both performance as well as speeding up the progress. This technique is effective when the size of dataset is small and there is availability of similar or un-related data. Transfer learning ensures the initial model for training on the core dataset is superior as the model is already trained on some different dataset thus making the learning process faster and more accurate. Another benefit of transfer learning is the performance of such models are more accurate even for limited training data of the core problem. In our experiment the model was pre-trained on Lung Cancer histopathology images, the knowledge gained by this model was then used to train the core problem i.e., cancer detection on colorectal cancer images.

For semantic segmentation experiment, due to unavailability of image masks of colorectal cancer there were minor differences in the predicted and actual masks in the validation phase. Thus, when the actual images of colorectal cancer were segmented, the segmentation achieved was not of highest accuracy. Without any access to the required dataset with proper masks, the off-the-shelf pre-trained model still achieves promising results.

For grading of colorectal cancer the average accuracy achieved by our model (Xception+) was 94.48 percent, whereas the accuracy achieved by R. Awan et. all[6] was 91 percent, even here we see our proposed model performed better than existing works.

We have tested all our models primarily on the dataset: 'NCT-CRC-HE-100K'[14]. This work can be expanded to include other similar datasets. It would be interesting to observe the performance of the Xception and GoogLeNet models on a dataset which has histopathology images from multiple sources and patients. The tests can include histology images for breast cancer, skin cancer, brain tumors etc. Different types of pathological images like biopsy, CT scan, X-rays can also be included to train and test the models. Further testing the effect of a variety of image augmentations like horizontal flipping, cropping, rotation etc. can also be performed and the results can be compared with our work. Also, the performance of models on images where blur and noises have been introduced can be tested. There is a huge scope of further studying the effects of Transfer Learning by using pre-trained models trained on similar histology images like breast cancer, skin cancer etc. Also, the performance of using pre-trained models on unrelated datasets like CIFR can be observed. Moreover, we can fine tune various architectures and study the effect on efficiency. Semantic segmentation of glands in histopathology images can be improved with the experimentation with different images and their masks. Other segmentation methods like nuclei segmentation of the images using proper images of cells and tissues with the masks of nuclei can be tried.

Methods

Data and Resources

The dataset primarily used ('NCT-CRC-HE-100K')[14] contains digitized histopathological images; they are hematoxylin and eosin (H and E)-stained tissue sections. It has 100,000 non-overlapping image patches from H and E-stained colorectal cancer and normal tissue. Tissue classes are: Adipose (ADI), background (BACK), debris (DEB), lymphocytes (LYM), mucus (MUC), smooth muscle (MUS), normal colon mucosa (NORM), cancer-associated stroma (STR), colorectal adenocarcinoma epithelium (TUM). The tissue class 'TUM' signifies the cancerous class. Supplementary Figure 3 depicts various tissue classes. For Cancer detection problem the Training and Validation dataset contains 30,000 images (split in 7:3 ratio). 15,000 images of cancer and normal tissues respectively. The Test dataset contains 7200 image, 3600 images of cancer and normal tissues respectively. For Tissue Classification Problem the Training and Validation dataset contains 18,000 images (split in 7:3 ratio). 2000 images of each tissue class. The Test dataset contains 4050 images. 450 images of each tissue class. All the datasets have no class imbalance, there is equal representation of all classes. For Transfer Learning we have used a dataset[17] which consists of 10,000 patches of histopathology images of lung cancer. 5000 images of cancer and non-cancerous classes respectively. For semantic segmentation we have used 165 histology images of glands and their segmentation masks[5]. For cancer grading we have used the dataset[6], which has 139 images, consisting of 71 normal, 33 low grade and 35 high grade cancer images. Pytorch[18] and Keras[19] libraries have been used to build the deep-learning architectures. For extracting features from images using Haralick method we have used the Mahotas package[20].

High Level Project Design for Traditional Methods

Traditional methods for cancer detection consists of two steps: first a feature extraction algorithm is used to extract the feature matrix for each image then the feature matrix is used to train a supervised machine learning model which classifies the features into two classes i.e., cancerous, or non-cancerous. For extraction of features from images we have used Local Binary Pattern (LBP)[15] and Haralick[16] techniques. Support Vector Machine (SVM) is used as the classifier to train and predict the classes from the feature matrix. In the LBP method, for an image each pixel has a pre-defined neighborhood of 3x3 or 8 cells. The selected center pixel is thresholded against the neighboring pixels and converted to an 8-bit binary value, i.e., if the value of the center pixel is greater than its neighbor the threshold value of the neighbor is '0' else it is '1'. This 8-bit binary value is converted to a decimal value, once this method is completed for every pixel an LBP array is generated. Finally, a histogram is computed for the frequency of each number from 0 to 255 occurring in the LBP array, thus generating a feature vector of 256 dimensions. Haralick method on the other hand computes features using the Gray-Level-Co-occurrence Matrix (GCLM). This method records how often same adjacent pixels occur in an image. Four directions of adjacency are defined hence four GCLM matrices are computed, the final feature matrix is computed by taking the mean of the four GCLM matrices. The feature matrices extracted by LBP or Haralick method is then used to train a classifier for predicting the labels. We have used Support Vector Machine (SVM) as the classifier, it is a supervised machine learning algorithm that is very popular for solving classification challenges. In this algorithm each feature in a feature vector in plotted in an n-dimensional space and then classification is performed by determining the hyper-plane that best separates the classes. SVM classifier performs best on data where the distribution is unknown (not gaussian) such as images, text etc. also the computational cost of SVM is lower than that of other classifiers like Naive-Bayes. Traditional machine-learning based approaches for Cancer Detection is a step-by-step process, the following steps were followed while incorporating the techniques.

- The dataset consists of 3200 images, with 1600 images each of cancerous and non-cancerous classes.
- All images are converted to grayscale.
- The dataset is then split into a training and test datasets in a proportion of 7:3.
- Feature vectors were extracted from each image from the training and test datasets using techniques like LBP and Haralick. For LBP the feature vector has 256 dimensions, thus the final feature matrix for training dataset has dimensions: 2240x256 and that of the test dataset is of dimensions: 960x256. For Haralick the feature vector has 13 dimensions, hence the training dataset's feature matrix has dimensions: 2240x13 and the test dataset's feature matrix has dimensions 960x13.
- The feature matrix for the training dataset that has been extracted in the previous step is used to train a supervised Machine Learning classifier. We have used SVM for classification, the hyper parameter for kernel used is linear.
- Once the classifier has been trained we have used the feature matrix for the test dataset to predict the classes of the images in the test dataset.

The flowchart for the Project Design has been depicted in Figure 7 a.

High Level Project Design for Deep-Learning Methods

We have used several popular deep-learning architectures that many researchers use not only for diagnosis in bio-medical fields but also in several other fields like object detection, facial recognition, tracking etc. Following are the architectures incorporated in our paper: AlexNet[21], GoogLeNet[22], Inception V3[23], ResNet[24], MobileNet[25], Xception[26], DenseNet[27], ResNeXt[28], UNet[29]. Further description of each of the models is provided in Supplementary Notes 1-9. All the above architectures have been used by us for solving the two classification problems i.e., Cancer Detection and Tissue Classification. Below steps have been followed in incorporating the deep-learning algorithms. The entire process can be divided into two phases, the training phase and testing phase and after completion of test phase the entire experiment (training + testing) is repeated 5 times and the test accuracy is recorded with the mean and standard deviation.

- In the training phase the training dataset is used. In the first step image augmentations like resizing (224x224) and normalization is performed on each image, then the images are converted to tensor images. Image augmentation strategy increases the diversity of the available data thus leading to better prediction accuracy. The dataset is further split into training and validation datasets in the ratio 7:3.
- The training and validation datasets are loaded to the training data loader and validation data loader in pre-defined batches of 10.
- A deep-learning algorithm is trained using the batches of 10 training images. The cross-entropy loss is computed for each batch of images and then the loss gradients of all tensors are computed following which the weights of all learnable parameters are updated using the back propagation method.
- Once all batches of images are trained, the model is used to predict the classes of images using the validation dataset and the accuracy is noted.
- The previous two steps are repeated for 30 epochs or iterations. In the end of 30 epochs, we have a trained deep-learning model that can be used to predict the classes of images.

- The next phase is test phase, here the test dataset is used. Again, Image augmentations are performed on each image and the images are converted to tensor images.
- Images are loaded in batches of 50 images to a test loader. The trained deep-learning model is used to predict the classes of test images.
- The accuracy of the prediction is recorded, this metric is later used for evaluating the performance of the model.
- The above experiment of training and testing is repeated 5 times and the mean and standard deviation of the accuracy obtained is recorded.

The flowchart in Figure 7 b, visualizes the algorithm used for training and testing the deep-learning model for classification.

Compare traditional vs deep-learning method

In this task we have aimed to compare the performance of traditional methods of feature extraction and classification with the deep-learning model (GoogLeNet) for image classification for the Cancer Detection problem.

Find the best deep-learning model

In this task we have aimed to find the best deep-learning model for both the problems Cancer Detection and Tissue classification. The performance of the models have been recorded using both kinds of Optimizers i.e., Adam and SGD. The hyperparameters like learning rate, weight decay and momentum have been experimentally determined. For the cancer detection problem we have used the below 4 metrics to evaluate the performance of the different models.

- Accuracy: represents the number of correctly classified images over the total number of images. Formula is given by (TN+TP)/(TN+FP+TP+FN)
- Precision: is the positive predictive value and is given by the formula TP/(TP+FP)
- Recall: also known as sensitivity or true positive rate. This should be high for a good classifier. The formula is given by: TP/(TP+FN)
- F1 score: this metric considers both precision and recall and is defined as: 2*(Precision*Recall)/(Precision+Recall), this is a better metric than accuracy.

Proposed Model

In this experiment we have aimed to modify the Xception architecture and observe the accuracy of the resultant model in predicting the classes. Figure 6, portrays the original architecture of the Xception. The proposed model (Xception+) has been designed to omit the last layer of the Entry Flow which consists of a depth-wise separable convolution layer , Entire Middle Flow and the first layer of the Exit Flow which consists of another depth-wise separable convolution layer. In Figure 6 the layers outlined with red rectangles have been excluded, the remainder network is our proposed model (Xception+). Large networks are prone to overfitting and incur high computational cost for training the model. Omitting layers is effective in reducing computational cost, overfitting and generalization error thus making the network more efficient. Not every neuron in the neural network contributes to the output some of them are redundant, removing these neurons contribute to a smaller and faster network. A resultant smaller network is better at handling overfitting problem, vanishing gradient problem thus improving the accuracy. Our proposed model (Xception+) is more compact and smaller than the actual Xception model and achieves better prediction accuracy.

Transfer Learning

In this experiment we have aimed to understand the effect of transfer learning on deep-learning models. To achieve this, we have used a dataset[17] which comprises histopathology images of lung cancer. Below are the steps that were followed:

- A deep-learning model (GoogLeNet) is trained on lung cancer images
- The trained model is then used to train on a dataset of 5040 histopathology images of colorectal cancer. There is no class imbalance, there are 2520 images of both classes, i.e., cancerous, and non-cancerous
- The final model is used to test 2160 images of colorectal cancer.
- The performance of this model is compared with traditional method of training the same dataset of colorectal cancer only using GoogLeNet.

Semantic segmentation

In this experiment, we have aimed to perform semantic segmentation of the histopathology images of colorectal cancer. We have trained a dataset containing histology images of gland and their segmented masks[5] on a UNet based model. Using this trained UNet model we have tried to segment the gland on 7200 histopathology images of colorectal cancer containing both classes (cancerous and non-cancerous). Dice loss has been used as the loss function in training phase of the UNet model.

Cancer Grading

In this experiment, we have aimed to perform grading cancer from histopathology tissue slides of colorectal cancer. The dataset has a total of 139 images, consisting of 71 normal, 33 low grade and 35 high grade cancer images. For our study in the first step, we have generated patches from the individual images resulting in a total of 15,520 images. To diminish the shortage of labelled dataset the technique of patch generation from a single image is very useful, it increases the number of labelled images. In this technique a single image is cropped into multiple smaller non-overlapping images or patches, each of the patches inherits the label or the class of the original image, each patch is considered a unique image thus increasing the size of dataset considerably.The patches generated are of 3 grades (classes) namely High, Low and normal. In the second step image augmentations are performed on each image and the images are converted to tensor. The dataset is split into training and validation datasets in the ratio 7:3. A deep-learning algorithm (GoogLeNet, Xception and Xception+) is trained for 100 epochs on these images. The cross-entropy loss is computed for each batch of images and then the loss gradients of all tensors are computed following which the weights of all learnable parameters are updated using the back propagation method. Once all batches of images are trained, the model is used to predict the classes of images using the validation dataset and the best validation accuracy is noted.

Availability of materials and data

All data for performing the experiments have been uploaded in the location:

$$https://drive.google.com/drive/folders/1jWJxDwN1_{8}z95t2aPb1PwZGImo-4hdgL?usp=sharing$$

. The codes are available on the github repository:

$$https://github.com/purnakar18/Image-Based-Colorectal-Cancer-Diagnosis$$

. There is a readme file ‘Read Me instructions.odt’ which has the detailed instructions on how to execute the codes, which datasets to use to reproduce the results. If the above link is inaccessible, all datasets and codes can be produced upon request.

References

1. Diet, nutrition, physical activity and cancer: a global perspective. *American Institute for Cancer Research, Continuous Update Project Expert Report* https://www.wcrf.org/diet-activity-and-cancer/ (2018).
2. Kral, P. & Lenc, L. Lbp features for breast cancer detection. *IEEE Int. Conf. on Image Process. (ICIP)* **2016**, 2643–2647, DOI: 10.1109/ICIP.2016.7532838 (2016).
3. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–900, DOI: https://doi.org/10.1145/3065386 (2017).
4. Sari, C. T. & Gunduz-Demir, C. Unsupervised feature extraction via deep learning for histopathological classification of colon tissue images. ieee trans med imaging. *IEEE Trans Med Imaging* **38(5)**, 1139–1149, DOI: 10.1109/TMI.2018.2879369 (2018).
5. Sirinukunwattana, K. & et al. Gland segmentation in colon histology images: The glas challenge contest. *Med Image Anal* **35**, 489–502, DOI: 10.1016/j.media.2016.08.008 (2017).
6. Awan, R. & et al. Glandular morphometrics for objective grading of colorectal adenocarcinoma histology images. *Sci. Reports* **7**, DOI: https://doi.org/10.1038/s41598-017-16516-w (2017).
7. Sirinukunwattana, K. & et al. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Trans Med Imaging* **35(5)**, 1196–1206, DOI: 10.1109/TMI.2016.2525803 (2016).
8. Chaddad, A. & Tanougast, C. Texture analysis of abnormal cell images for predicting the continuum of colorectal cancer. *Anal Cell Pathol (Amst)* **2017:8428102**, DOI: 10.1155/2017/8428102 (2017).
9. Kather, J. N. & et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS Med* **16(1)**, DOI: 10.1371/journal.pmed.1002730 (2019).

10. Malik, J. & et al. Colorectal cancer diagnosis from histology images: A comparative study. *arXiv* https://doi.org/10.48550/arXiv.1903.11210 (2019).

11. Wang, K. S. & et al. Accurate diagnosis of colorectal cancer based on histopathology images using artificial intelligence. *BMC Med* **19**, 76, DOI: https://doi.org/10.1186/s12916-021-01942-5 (2021).

12. Yamashita, R., Nishio, M., Do, R. K. G. & Togashi, K. Convolutional neural networks: an overview and application in radiology. *Insights Imaging* **9**, 611–629, DOI: https://doi.org/10.1007/s13244-018-0639-9 (2018).

13. Araújo, T. & et al. Classification of breast cancer histology images using convolutional neural networks. *PLoS One* **12**, 6, DOI: 10.1371/journal.pone.0177544 (2017).

14. Kather, J. N., Halama, N. & Marx, A. 100,000 histological images of human colorectal cancer and healthy tissue. *Zenodo* https://doi.org/10.5281/zenodo.1214456 (2018).

15. Ojala, T., Pietikainen, M. & Maenpaa, T. Multiresolution grayscale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis Mach. Intell.* **24**, 971–987, DOI: 10.1109/TPAMI.2002.1017623 (2002).

16. Porebski, A., Vandenbroucke, N. & Macaire, L. Haralick feature extraction from lbp images for color texture classification. *First Work. on Image Process. Theory, Tools Appl.* **2008**, 1–8, DOI: 10.1109/IPTA.2008.4743780 (2008).

17. Borkowski, A. A. Lung and colon cancer histopathological image dataset (lc25000). *arXiv* https://doi.org/10.48550/arXiv.1912.12142 (2019).

18. Paszke, A. & et al. Pytorch: An imperative style, high-performance deep learning library. *Conf. on Neural Inf. Process. Syst. (NeurIPS 2019)* **33**, 8024–8035, DOI: https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf (2019).

19. Abadi, M. & et al. Tensorflow: Large-scale machine learning on heterogeneous systems. *arXiv* https://doi.org/10.48550/arXiv.1603.04467 (2016).

20. Coelho, L. Mahotas: Open source software for scriptable computer vision. *Journal of Open Research Software* http://doi.org/10.5334/jors.ac (2013).

21. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90, DOI: https://doi.org/10.1145/3065386 (2017).

22. Szegedy, C. & et al. Going deeper with convolutions. *IEEE Conf. on Comput. Vis. Pattern Recognit. (CVPR)* **2015**, 1–9, DOI: 10.1109/CVPR.2015.7298594 (2015).

23. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. *IEEE Conf. on Comput. Vis. Pattern Recognit. (CVPR)* **2016**, 2818–2826, DOI: 10.1109/CVPR.2016.308 (2016).

24. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *IEEE Conf. on Comput. Vis. Pattern Recognit. (CVPR)* **2016**, 770–778, DOI: 10.1109/CVPR.2016.90 (2016).

25. Howard, A. G. & et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* https://doi.org/10.48550/arXiv.1704.04861 (2017).

26. Chollet, F. Xception: Deep learning with depthwise separable convolutions. *IEEE Conf. on Comput. Vis. Pattern Recognit. (CVPR)* **2017**, 1800–1807, DOI: 10.1109/CVPR.2017.195 (2017).

27. Huang, G., Liu, Z., Van-Der-Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. *IEEE Conf. on Comput. Vis. Pattern Recognit. (CVPR)* **2017**, 2261–2269, DOI: 10.1109/CVPR.2017.243 (2017).

28. Xie, S., Girshick, R., Dollár, P., Tu, Z. & He, K. Aggregated residual transformations for deep neural networks. *arXiv* https://doi.org/10.48550/arXiv.1611.05431 (2017).

29. Ronneberger, P., O.and Fischer & Brox, T. U-net: Convolutional networks for biomedical image segmentation. *Springer* **9351**, 234–241, DOI: https://doi.org/10.1007/978-3-319-24574-4_28 (2015).

Author contributions statement

P.K. and S.R. planned the experiments, P.K. designed the software codes and recorded all results of the experiments, performed all data analysis. P.K. prepared the draft of manuscript. All authors edited the manuscript.

Additional information

Financial Competing interests: The authors declare no competing financial interests.

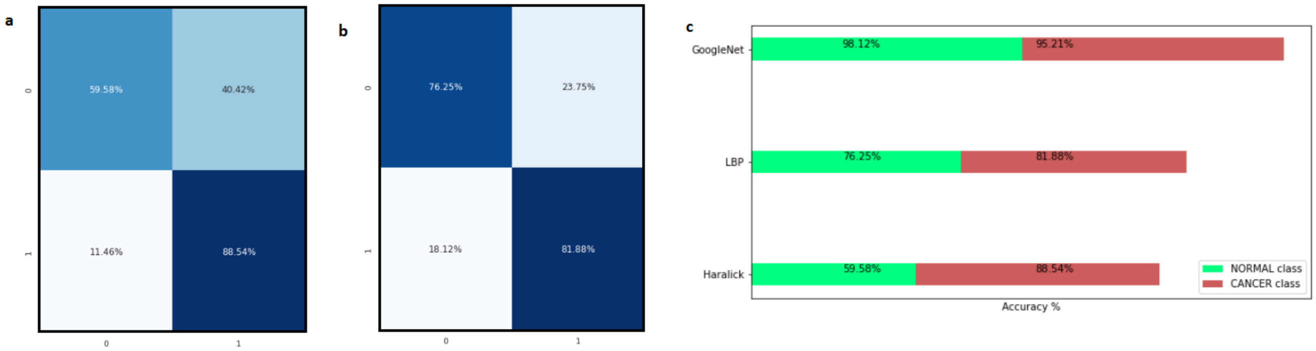


Figure 1. Tradition Methods vs Deep-Learning Method (a) Confusion Matrix showing the prediction results of Haralick Method, (b) Confusion Matrix showing the prediction results of LBP Method, (c) Accuracy of Traditional methods and GoogLeNet

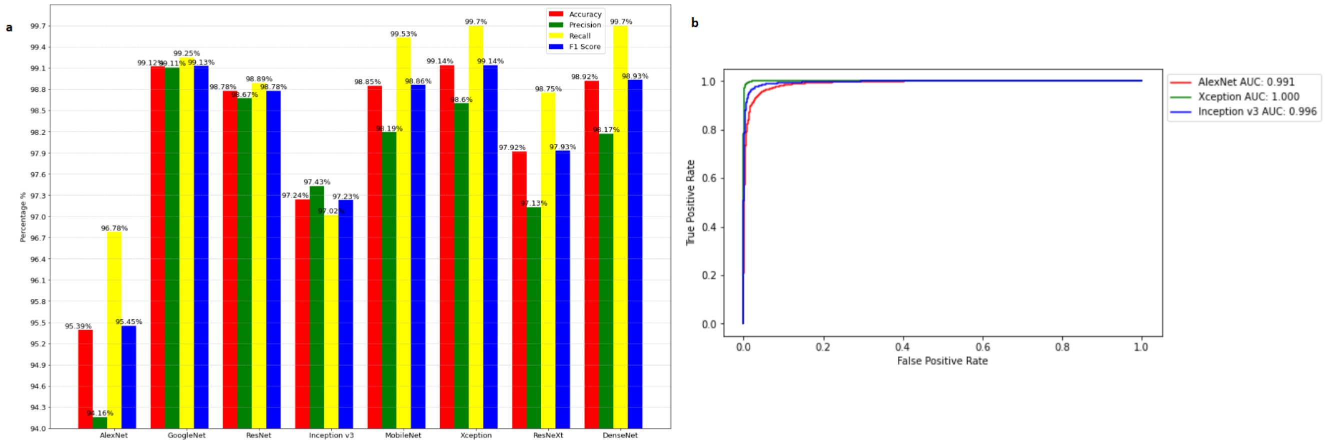


Figure 2. Cancer Detection Problem (a) Performance comparison of the deep-learning algorithms, (b) ROC curve for few deep-learning models

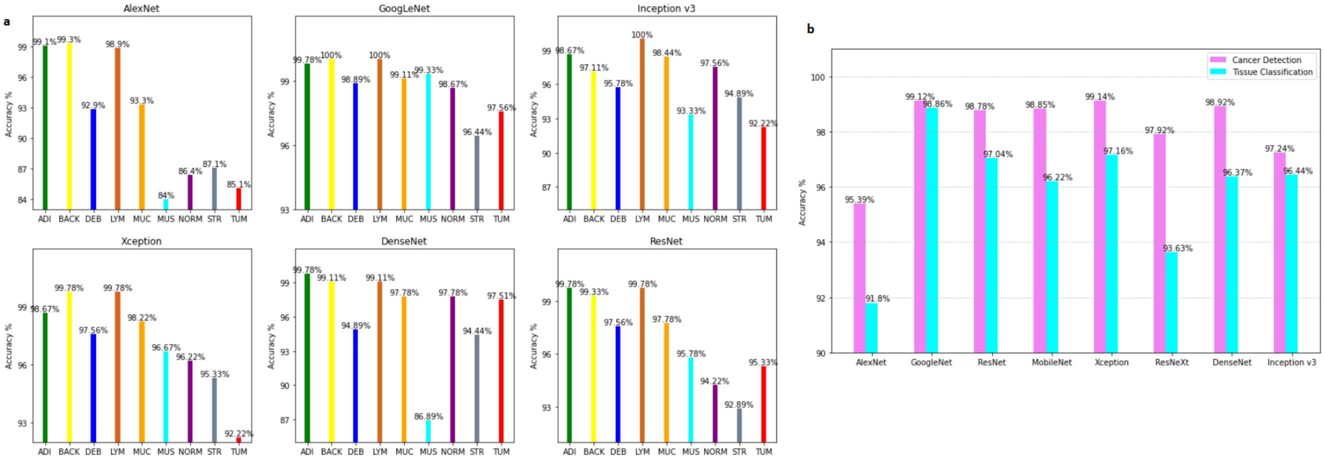


Figure 3. (a) Accuracy of Models for each class: Tissue Classification, (b) Comparison of accuracy of each model for the two problems.

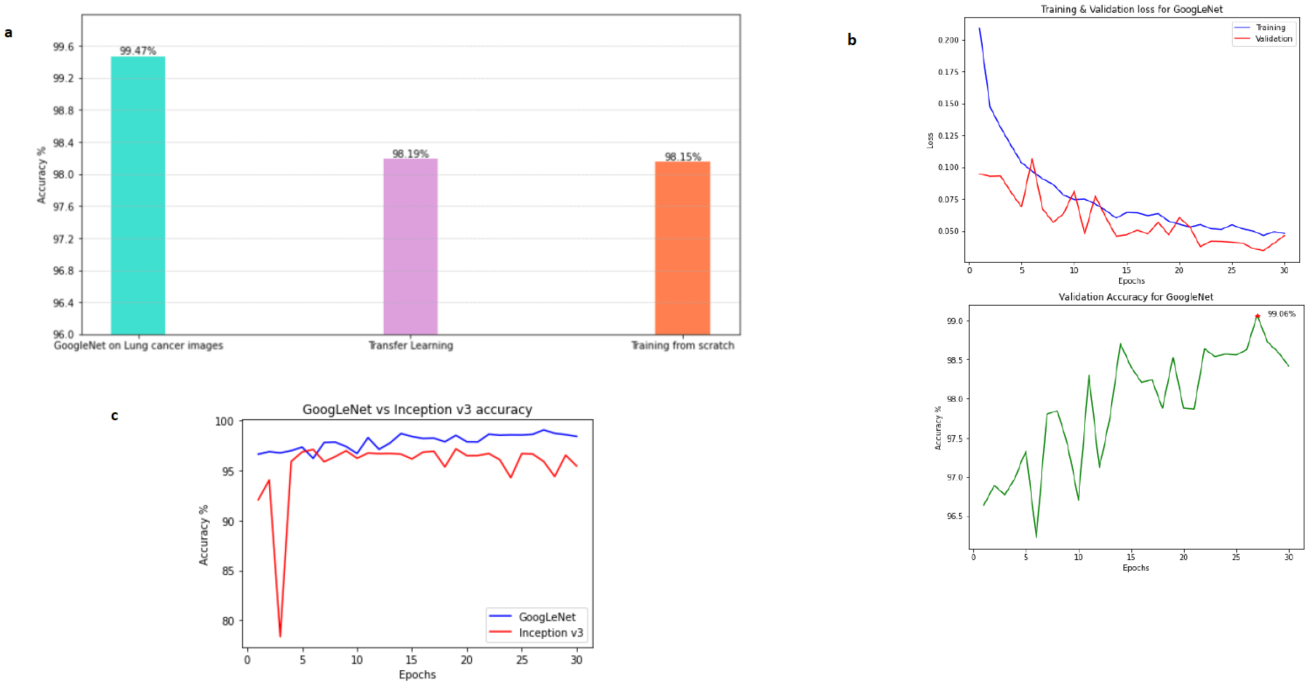


Figure 4. (a) Transfer Learning vs Training from scratch, (b) Performance of GoogLeNet during training phase, (c) GoogLeNet vs Inception v3 validation accuracy

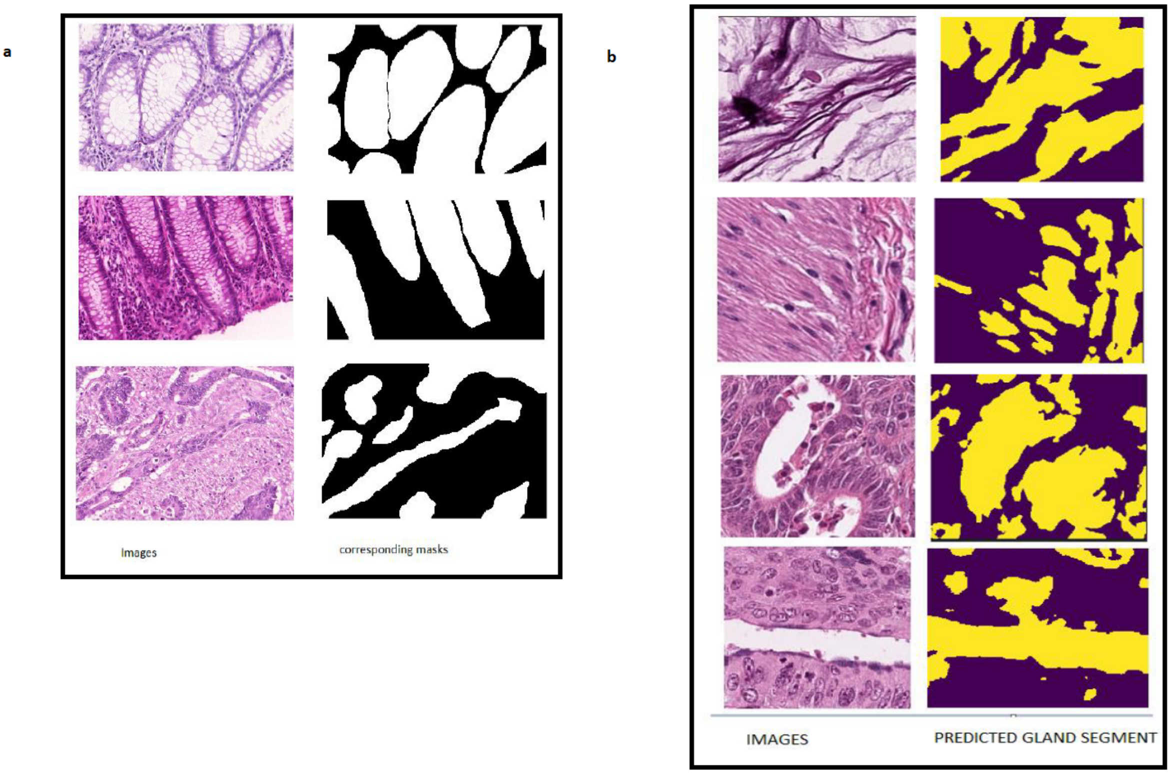


Figure 5. (a) Histology images of gland and their corresponding masks, (b) Gland Segmentation results on the Colorectal cancer images.

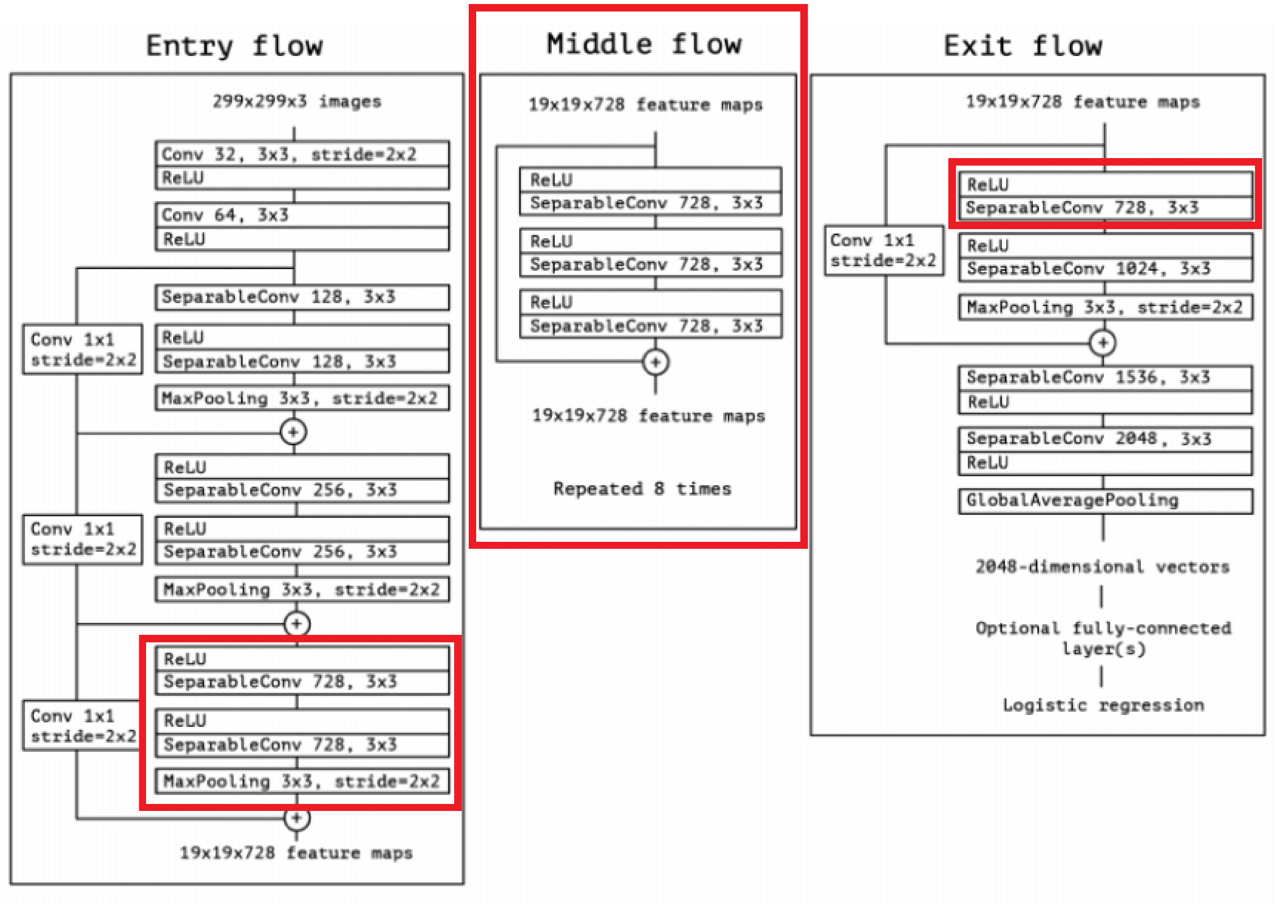


Figure 6. Architecture of the proposed model.[26]

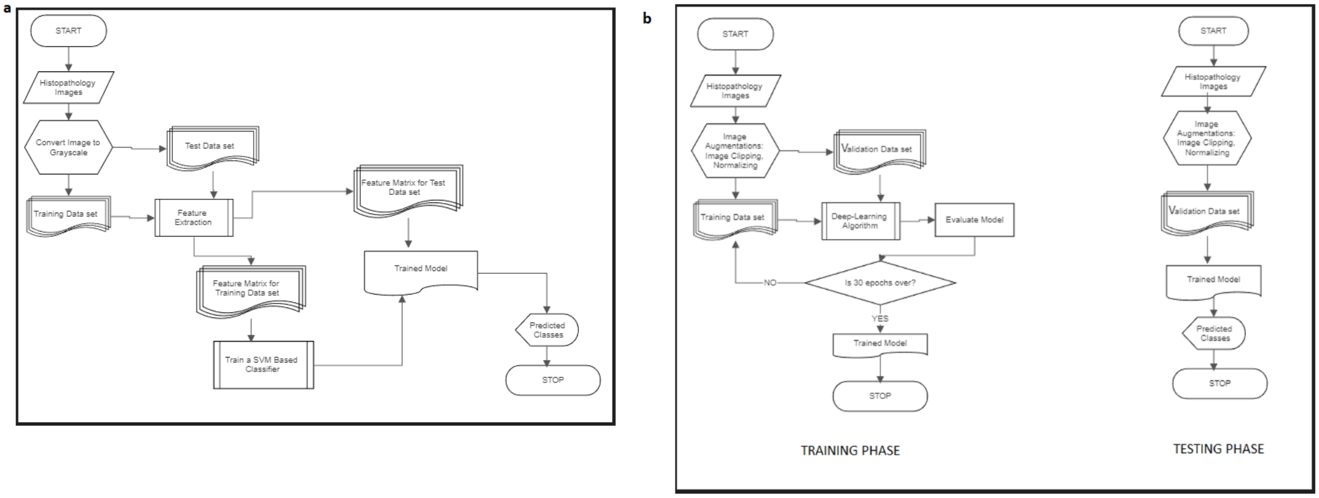
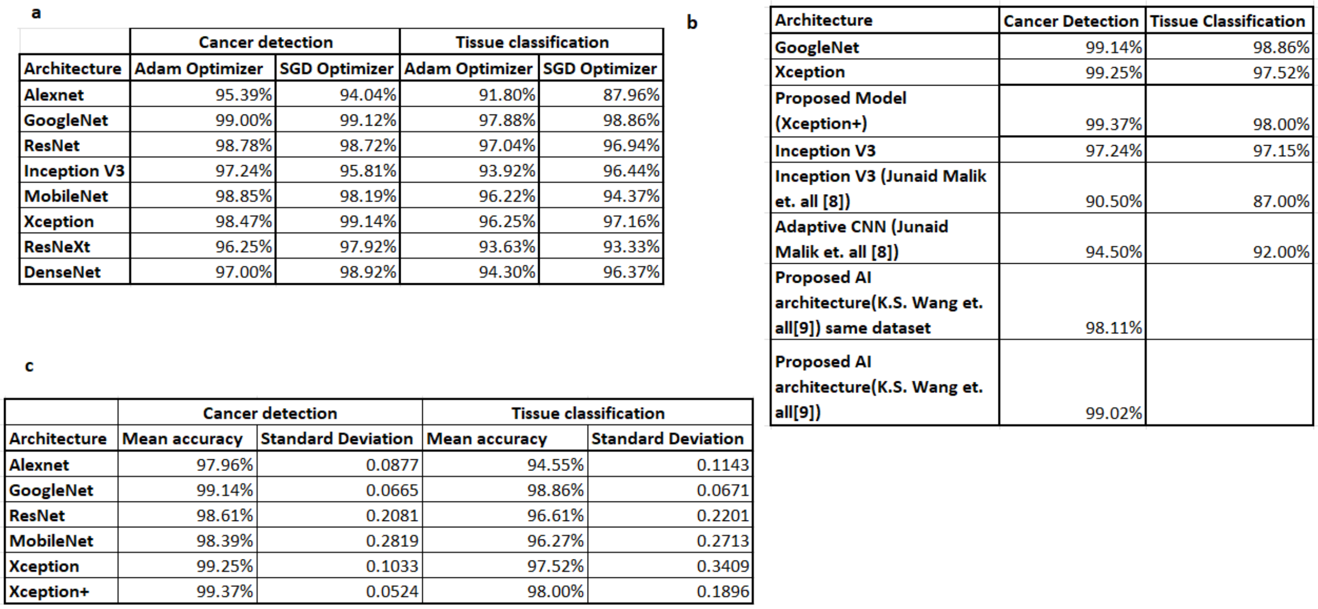


Figure 7. (a) Project Design for Traditional Methods, (b) Project design for deep-learning models



a

	Cancer detection		Tissue classification	
Architecture	Adam Optimizer	SGD Optimizer	Adam Optimizer	SGD Optimizer
Alexnet	95.39%	94.04%	91.80%	87.96%
GoogleNet	99.00%	99.12%	97.88%	98.86%
ResNet	98.78%	98.72%	97.04%	96.94%
Inception V3	97.24%	95.81%	93.92%	96.44%
MobileNet	98.85%	98.19%	96.22%	94.37%
Xception	98.47%	99.14%	96.25%	97.16%
ResNeXt	96.25%	97.92%	93.63%	93.33%
DenseNet	97.00%	98.92%	94.30%	96.37%

b

Architecture	Cancer Detection	Tissue Classification
GoogleNet	99.14%	98.86%
Xception	99.25%	97.52%
Proposed Model (Xception+)	99.37%	98.00%
Inception V3	97.24%	97.15%
Inception V3 (Junaid Malik et. all [8])	90.50%	87.00%
Adaptive CNN (Junaid Malik et. all [8])	94.50%	92.00%
Proposed AI architecture(K.S. Wang et. all[9]) same dataset	98.11%	
Proposed AI architecture(K.S. Wang et. all[9])	99.02%	

c

	Cancer detection		Tissue classification	
Architecture	Mean accuracy	Standard Deviation	Mean accuracy	Standard Deviation
Alexnet	97.96%	0.0877	94.55%	0.1143
GoogleNet	99.14%	0.0665	98.86%	0.0671
ResNet	98.61%	0.2081	96.61%	0.2201
MobileNet	98.39%	0.2819	96.27%	0.2713
Xception	99.25%	0.1033	97.52%	0.3409
Xception+	99.37%	0.0524	98.00%	0.1896

Figure 8. (a) Performance of models for optimizer type, (b) Comparison of similar works Comparison of similar works, (c) Mean and standard deviation of average test accuracy from 5 experiments.