Image based Colorectal Cancer Diagnosis

Purna Kar
Dept. of Computer Science
University of Exeter,
Exeter, United Kingdom.
pk405@exeter.ac.uk

Supervision: Dr. Sareh Rowlands
College of Engineering, Mathematics, and Physical Sciences
University of Exeter,
Exeter,United Kingdom.

Abstract— **Colorectal Cancer is the one of the most common forms of cancer hence it very crucial for an early and accurate detection. Manual diagnosis is a tedious and time-consuming job which is prone to human errors as it involves visual examinations of pathological images. It is imperative to use computer-aided detection (CADe) systems to interpret the medical images for a quicker and more accurate diagnosis. The traditional methods for diagnosis comprise extraction of features from pathological images which are based on different factors like texture, illumination, repetitive patterns etc. and then use these features in a Machine Learning model for classification into two classes i.e., cancerous, and 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 paper, I have used several deep learning based CNNs for cancer diagnosis on digitized histopathology images. I have performed a comparative study of the different architectures and proposed the best model based on statistical evaluation. I have also compared two traditional methods for diagnosis with the deep-learning models. Moreover, I have conducted various experimentations to broaden the scope of solving the problem like fine-tuning one of the CNN architectures, implementing transfer learning technique by using models trained on unrelated histopathology images, performing segmentation of glands in the histology images, and using these segmented images for diagnosis.**

Keywords—Colorectal cancer (CRC), Convolutive neural networks (CNNs), deep-learning, Artificial Intelligence (AI), Transfer Learning, Semantic Segmentation.

I. 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&E-stained histology images. A pathologist determines the stage of cancer by counting the number and size of tumors in the section of images. This makes the task extremely monotonous and prone to errors e.g., the size of tumor cells are minute which requires substantial amount of amplifying of the image which is prone to being overlooked. Thus, this is a very tedious and time-consuming task and prone to human errors. 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 (CADe) 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 (CADe), 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 use this feature matrix in a supervised machine learning based classifier to classify the images into cancerous and non-cancerous classes. The time taken to extract features and classify images is very high and the efficiency is also 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 [2]. 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.

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 [3] due to the accuracy of prediction and classification. To diminish the shortage of labelled dataset the technique of patch generation [4] 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.

In this project I have used a huge dataset [5] (almost 1,00,000) of digitized H&E-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. I have addressed two sets of problems in this project, firstly, I have used the several CNN based models to detect cancer, i.e., predict whether images have cancerous tumor present in it or not. Secondly, I have performed prediction of the different tissue classes present in the dataset.

It was observed that Xception model performs the best cancer detection and GoogLeNet model has best accuracy rate for classifying the 9 tissue classes from the images. The overall performance of the deep-learning models is very competitive and are superior to that of the traditional models. Precision, Recall, accuracy, F1 score, and AUC are the major metrics used for comparing the models for the first problem. Accuracy is the primary metric used for comparing the performance of the models for the second problem. I 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. It was found that the performance of the model was slightly better when transfer learning technique was incorporated. In addition to these I have studied the effect of fine-tuning the Xception architecture to omit few layers in the classification problem. Finally, I have aimed to perform semantic segmentation of glands on the histology images, the segmented images were then used in the deep-learning model for cancer detection.

II. Related work

Deep learning techniques like the Convolutional neural networks (CNNs) have proved to be effectual in the prediction and classification problems [2]. 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 [6,13], tumor cell detection [7,8], gland segmentation [13] 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. One of the crucial challenges faced by researchers is the unavailability of properly labelled dataset. Also, the developed models are tested against a typical dataset and not a variety of datasets. Most researches use patch-based techniques [4] to expand the dataset. Some related works have aimed to do comparative studies of the CNN architectures [11, 12], some proposed new adaptive CNN models from scratch [11,12] which performed better than the state-of-the-art models for the dataset used by them, some studied the effect of transfer learning [11]. Segmentation of glands also have been incorporated by researches [13] to quantify morphology of glands which helps pathologists to perform clinical diagnosis in a better manner.

Various researches [10] have implemented traditional supervised methods that involve feature extraction followed by training a classifier (like SVM) over the extracted features to detect cancerous cells. The precision of these traditional methods are much lower than the state-of-the-art deep learning methods.

III. Aims and Objectives

In my project I divided the colorectal cancer diagnosis into two parts.

A. *Cancer Detection:* Determine whether an image belongs to cancerous or non-cancerous class.

B. *Tissue classification:* The Dataset consists of 9 tissue annotations; design a model to determine which tissue class the image belongs to.

The deep-learning models incorporated have aimed to solve the above two problems.

In this project my goal is to compare the performance of the traditional techniques and the CNN based deep-learning methods to diagnose colorectal cancer from histopathology images. The objectives of this project are as follows:

A. *Identify the best machine learning based algorithm for cancer detection from histopathology images*

To achieve this, I have done a comparative study of the results of few state-of-the-art deep learning techniques with the traditional techniques for feature extraction and classification of the features.

B. *Identify the best deep-learning algorithm for tissue classification*

Determine which CNN based architecture predicts the tissue classes accurately.

C. *Effect of Fine-tuning*

Study the effect of fine-tuning an architecture.

D. *Effectiveness of Transfer learning*

Determine the effect of using transfer learning by training a model on a dataset of dissimilar pathological images and use this pre-trained model to perform cancer diagnosis.

E. *Scope for Image Segmentation*

Perform semantic gland segmentation on the histology images. Use the segmented images for Cancer detection problem by training a CNN based architecture and predicting the results.

IV. Research Context

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 I have used Local Binary Pattern (LBP) [27] and Haralick [28] 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. I 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.

I 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 by me:

A. *AlexNet [14]*

This CNN based architecture has eight layers of learnable parameters. The model comprises five convolutional and max pooling layers and three fully connected layers and a final output layer. Each of the layers in the convolutional blocks and fully connected blocks use ReLu (Rectified Linear Unit) activation functions. This architecture was first used to classify the ImageNet dataset [14].

B. *GoogLeNet [15]*

This architecture is the first to use inception blocks. Inception blocks use convolutions that aim to decrease the number of learnable parameters and in turn makes the architecture go deeper. In an inception block there is a common input through which convolutions of filter sizes 1x1, 3x3, 5x5 and a 3x3 max pooling is performed and the output of each of these modules are stacked together to a single output. Fig 1 shows the architecture of an inception block. GoogLeNet has twenty-two layers in total, three convolutional layers followed by nine inception blocks (these inception blocks are occasionally followed by Max pooling layers) and a final block that consists of average pooling, dropout, and linear layers.

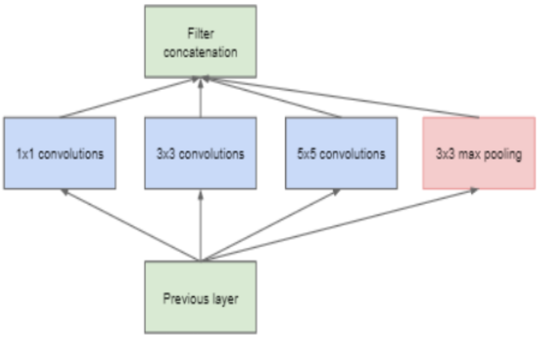


Fig. 1: architecture of an Inception block

C. *Inception V3 [16]*

This is one of the state-of-the-art architectures which is 42 layers deep, its architecture is like GoogLeNet with few improvisations and advancements. The few characteristics of Inception V3 are: (i) Factorizing larger convolutions into smaller ones which reduces the number of parameters and reduces the cost of computation. e.g., a 5x5 convolutional layer is factorized into two 3x3 convolutional layers. (ii) Spatial factorization into asymmetric convolutions: e.g., replacing a 3x3 convolutional layer with a 1x3 and a 3x1 convolutional layers. This reduces the number of parameters and decreases the cost of computation. (iii) Auxiliary classifiers: usage of these classifiers improves the convergence of deep networks, hence improving the accuracy. (iv) Efficient Grid size reduction of feature maps.

D. *ResNet [17]*

This is a 34 layered residual network, the core idea of ResNet is to address the vanishing gradient problem faced by many deep-neural networks by pioneering 'identity shortcut connection' which skips one or more layers. Fig 2 shows are typical residual block.

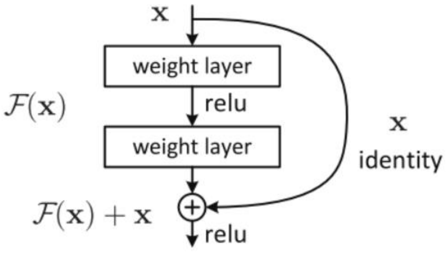

x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Fig. 2: A residual block

E. *MobileNet [18]*

This CNN was designed typically for mobile and embedded vision applications. This architecture uses depthwise separable convolutions that helps in reducing the latency of the application. A depthwise separable convolution is a depthwise convolution succeeded by a pointwise convolution. A depthwise convolution is a channel-wise kxk spatial convolution, whereas a pointwise convolution is simply a 1x1 convolution which is used to change the dimension.

F. *Xception [19]*

Xception stands for “extreme inception.” This architecture uses features of both Inception v3 and MobileNet architectures. It uses a modified depthwise separable convolution i.e., it has a pointwise convolution and then a depthwise convolution (opposite of MobileNet), this feature is inspired by inception v3 where a 1x1 convolution precedes a spatial convolution. Xception does not use any intermediate activation function like ReLu. Also, convolutions are not performed across all the channels, therefore the connections are fewer, and the model is less deep. It uses 'skip connections' like the ResNet.

G. *DenseNet [20]*

These are densely connected convolutional networks with a very narrow set of layers, the salient feature of this architecture is it re-uses its own features thus exploiting the full potential of its own network which reduces the number of learnable parameters by removing the redundancy to re-learn its feature maps. Dense Nets can handle the vanishing gradient problem as the loss function of each layer provides the gradient of the layers.

H. *ResNeXt [21]*

This CNN architecture has features of VGG, ResNet and Inception. It uses repetitive blocks (like VGG), skip connections (as in ResNet) and auxiliary classifiers (like Inception v3). ResNeXt replaces the standard residual blocks with one that leverages a "split-transform-merge" strategy.

I. UNet [25]

It is a CNN architecture for swift and meticulous segmentation of images mainly used for Biomedical Image segmentation. It has a U-shaped architecture, hence the name 'UNet', it comprises a system of encoder and decoder. The encoder increases the number of channels and reduces the spatial dimensions for each layer whereas the decoder reduces the number of channels and increases the spatial dimensions. There are four encoders and four decoders in the UNet architecture.

All the above architectures have been used by me for solving the two classification problems as stated in section III.
I have taken help of [29,30] while building the architectures from scratch. The performance of these architectures have been discussed in section VI.

V. METHOD AND EXPERIMENT DESIGN

A. Data and Resources

The dataset primarily used ('NCT-CRC-HE-100K') [5] contains digitized histopathological images; they are hematoxylin and eosin (H&E)-stained tissue sections. It has 100,000 non-overlapping image patches from H&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 class 'TUM' signifies the cancerous class. Fig. 3 depicts various tissue classes. For Cancer detection problem (refer section III): 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 (refer section III): 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.

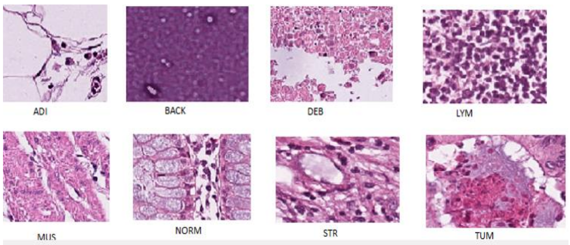


Fig. 3: Sample patches showing the different tissue classes.

For Transfer Learning (Section III, Objectives D) I have used a dataset [24] which consists of 10,000 patches of histopathology images of lung cancer. 5000 images of cancer and non-cancerous classes respectively. For semantic segmentation (Section III, Objectives E), I have used 165 images and their segmentation masks [13]. Pytorch [22] and Keras [23] Libraries have been used to build the deep-learning architectures defined in section IV. For extracting features from images using Haralick method I have used the Mahotas package [26].

B. High Level Project Design

Traditional machine-learning based approaches for Cancer Detection (refer section III) 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. I have used SVM for classification, the hyper parameter for kernel used is linear.
- Once the classifier has been trained I 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 Fig 4.

Below steps have been followed in solving the two classification problems (as declared in section III) using the deep-learning algorithms. The entire process can be divided into two phases, the training phase and testing phase.

- In the training phase the training dataset (mentioned in section VA) 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.
- A deep-learning algorithm is trained using the batches of 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 (mentioned in section V.A) is used. Again, Image augmentations are performed on each image and the images are converted to tensor images.
- Images are loaded in batches 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 flowchart in Fig 5 visualizes the algorithm used for training and testing the deep-learning model for classification.

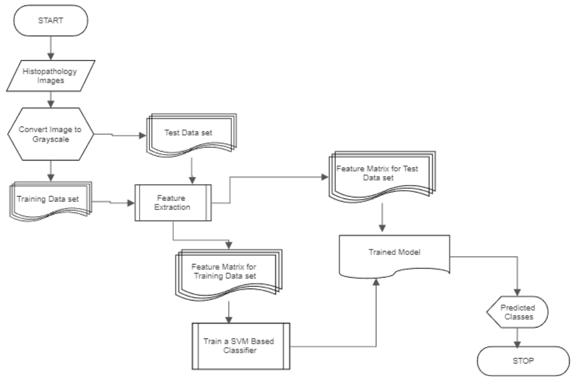

START
Histopathology Images
Convert Image to Grayscale
Test Data set
Training Data set
Feature Extraction
Feature Matrix for Test Data set
Trained Model
Feature Matrix for Training Data set
Predicted Classes
Train a SVM Based Classifier
STOP

Fig. 4: Project Design for Traditional Methods

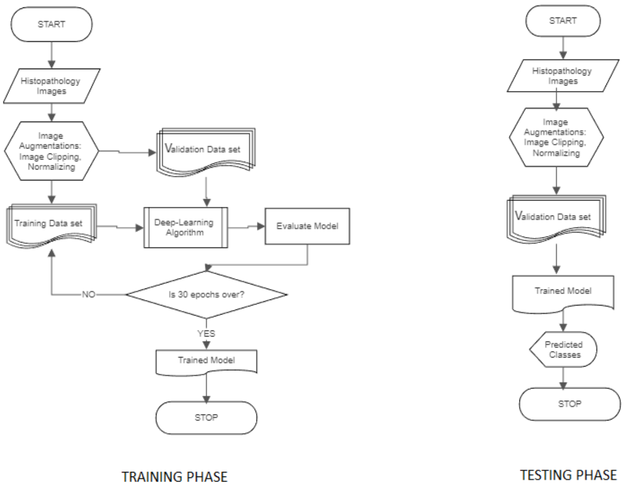


Fig. 5: Project design for deep-learning models

C. *Compare traditional vs deep-learning method*

In this task I 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 (refer section III).

D. *Find the best deep-learning model*

In this task I have aimed to find the best deep-learning model for both the problems Cancer Detection and Tissue classification as defined in section III, objectives A & B. 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.

E. *Experiment 1: Modify Xception Architecture*

In this experiment I have aimed to modify the Xception architecture and observe the efficiency of the resultant model in predicting the classes. Fig 6. portrays the original architecture of the Xception. In this experiment, the model 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 Fig 6. the layers outlined with red rectangles have been excluded from the model.

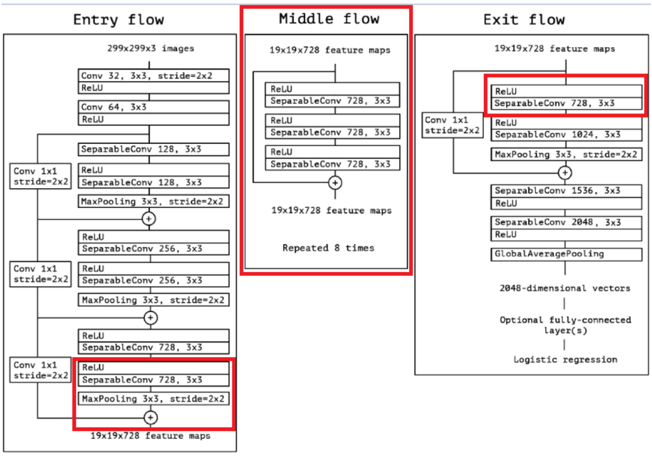


Fig. 6: Modified Xception architecture

F. *Experiment 2: Transfer Learning*

In this experiment I have aimed to understand the effect of transfer learning on deep-learning models (as discussed in section III, objectives: C). To achieve this, I have used a dataset [24] which comprises histopathology images of lung cancer. Below are the steps that were followed:

- A deep-learning model (GoogLeNet) is trained on the 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.

G. *Experiment 3: Semantic segmentation*

In this experiment, I have aimed to perform semantic segmentation of the histopathology images of colorectal cancer (Section III, Objectives D). 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 section I have aimed to perform a segmentation of the colorectal cancer images, the segmented images will show glands and background tissues in different illumination, then use these segmented images to train a classifier to perform Cancer Detection task.

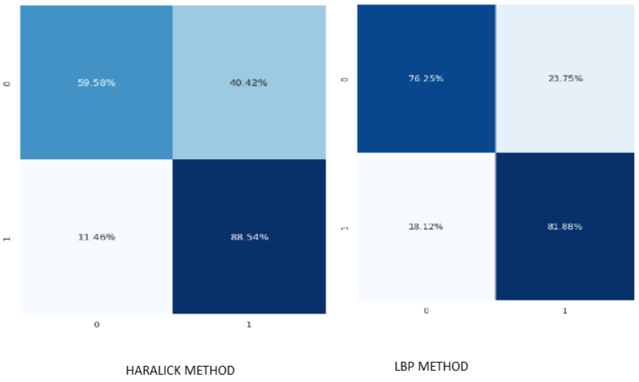
Below are the objectives of this experiment:

- I have trained a dataset containing histology images of gland and their segmented masks [13] on a UNet model.
- Using this trained UNet model I have tried to segment the gland on 7200 histopathology images of colorectal cancer containing both classes (cancerous and non-cancerous)
- In the next step I have trained the segmented images on a deep-learning model (GoogLeNet) and aimed to predict the classes.
- Binary-cross entropy loss has been used as the loss function in training phase of the UNet model.

VI. Results and Analysis

A. *The traditional methods vs deep-learning method*

The performance of traditional methods LBP and Haralick has been mediocre. For this problem the binary classes are defined as 0 (non-cancer) and 1 (Cancer). The accuracy of LBP is 76.46% and that of Haralick is 75.20%. Both methods showed higher accuracy in predicting the cancer class (i.e., class 1). Fig 7. 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. GoogLeNet has performed much better than LBP and Haralick, the accuracy achieved by this deep-learning model is almost 20% higher than that of the traditional methods, it has predicted the non-cancer class with slightly better accuracy than the cancerous class. Fig. 8 visualizes the accuracy of GoogLeNet and the traditional methods in predicting the binary classes.



HARALICK METHOD LBP METHOD

Fig.7:Confusion matrix showing the prediction results of traditional methods

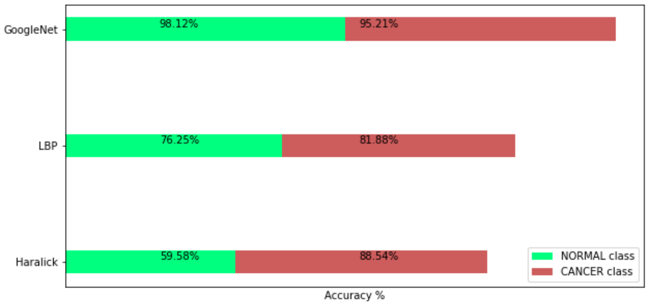
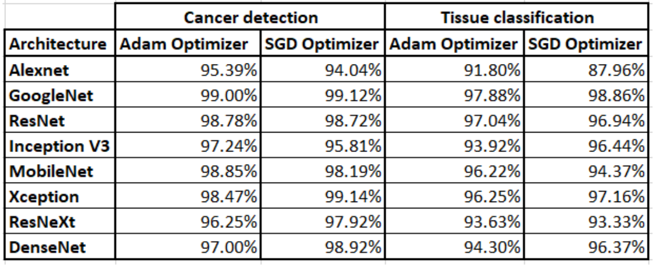


Fig. 8: Traditional methods vs GoogLeNet.

B. *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 below chart in Fig. 9 shows the performance of each model using these two optimizers for the two problems i.e., Cancer detection & Tissue classification. (Refer section III, problems A&B).



	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%

Fig. 9: Performance of models for optimizer type

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=0.0002
- Weight decay=0.001 (only for Adam Optimizer)
- Momentum=0.9 (only for SGD)
- Epochs=30

For the cancer detection problem (refer section III) I 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 $\frac{TN+TP}{TN+FP+TP+FN}$
- Precision: is the positive predictive value and is given by the formula $\frac{TP}{TP + FP}$
- Recall: also known as sensitivity or true positive rate. This should be high for a good classifier. The formula is given by: $\frac{TP}{TP + FN}$
- F1 score: this metric considers both precision and recall and is defined as: $2 * \frac{Precision * Recall}{Precision + Recall}$ this is a better metric than accuracy.

I have plotted the four metrics for all the deep-learning models for the Cancer Detection problem shown in Fig 10. It is observed that the GoogLeNet architecture has the highest precision 99.11%, whereas the accuracy, recall and F1 score of Xception is the highest (99.14%, 99.70% and 99.14 % respectively). Overall, Xception has the best performance amongst all the deep-learning models

I have also plotted a Receiver Operating Characteristic Curve (ROC) for the 3 models Xception, AlexNet and Inception v3 respectively shown in Fig 11. 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: $\frac{FP}{FP + TN}$. 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.

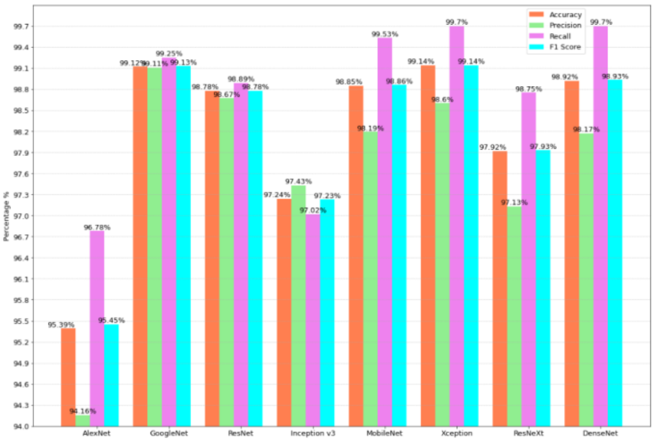


Fig. 10: Performance comparison of the deep-learning algorithms

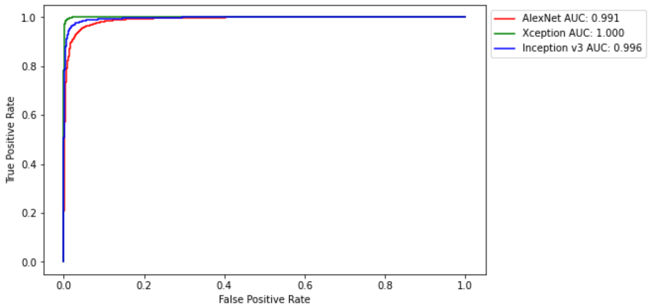


Fig. 11: ROC curve for few models.

As seen in Fig 11. The AUC for Xception is 1.0 which is the ideal value, this implies that this model is highly accurate in classifying the images. AUC for GoogLeNet is also 1.0

For the tissue classification problem (refer section III, problem B), I have compared the accuracy of the models in classifying the 9 tissue classes. From Fig. 12 it is observed that GoogLeNet has the highest accuracy for classifying across all tissue classes followed by Xception.

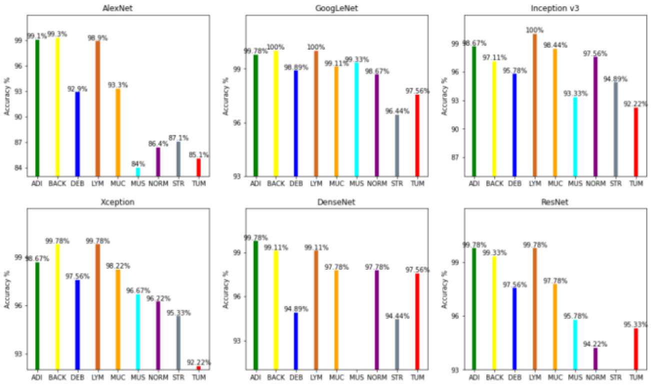


Fig. 12: Accuracy of Models for each class: Tissue Classification

The Fig 13 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 performance (99.14%) for the Cancer detection problem and GoogLeNet has displayed the best accuracy (98.86%) in the tissue classification problem.

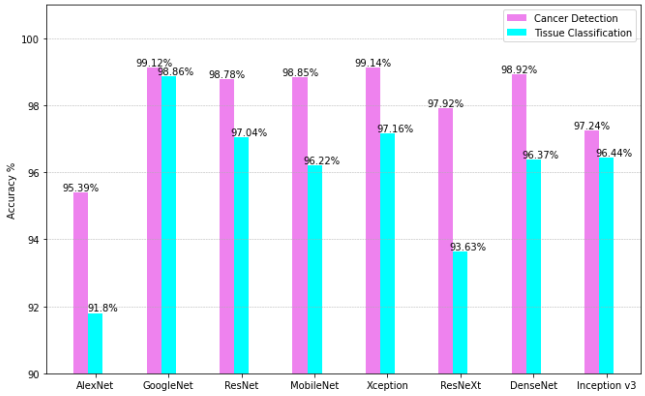


Fig. 13: Comparison of accuracies of each model for the two problems

I have plotted the training loss, validation loss and validation accuracy vs the epochs observed in the training phase of GoogLeNet model in Fig 14. 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. We can see that the issue of overfitting 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.

C. *Performance of Modified Xception.*

The modified Xception architecture performed exceptionally well, for both the classification problems. It achieved an overall accuracy of 99.46% and 98.40% for the cancer detection and the tissue classification problems respectively.

D. *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%, when this model is used for training on colorectal images the predictive accuracy is 98.19% whereas the accuracy of the GoogLeNet model trained only on colorectal images have accuracy of 98.15%. The accuracies of the models have been plotted in Fig. 15

E. *Semantic segmentation of glands.*

In this experiment I have used histology images of glands and their respective masks [13] to train a UNet model for the semantic segmentation task. Fig. 16 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%, Fig. 17 shows the validation image, with its corresponding actual mask and the predicted mask (segmented image). 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, I 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. Fig 18 shows the resultant segmented images.

After segmenting the 7200 histopathology images of colorectal cancer, I have trained a deep-learning model (GoogLeNet architecture) to solve the cancer detection problem. The classes of the original images had been inherited by the segmented images. It has been observed the overall accuracy of this image classifier is 78.61% whereas the accuracy of the GoogLeNet model tested on original images is 98.15%. The dataset used to train UNet model had only 165 images and its corresponding masks which is not sufficient to train and validate a model. The lack of adequate amount of images and masks have limited the performance of this experiment.

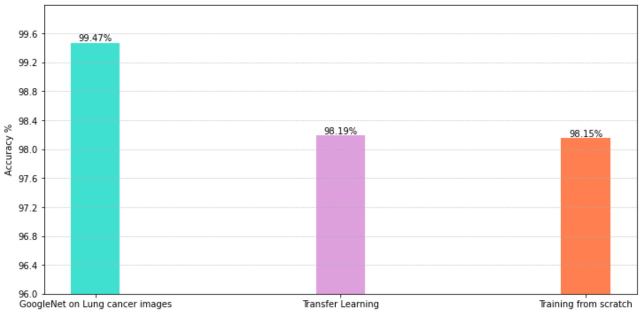


Fig. 15: Transfer Learning vs Training from scratch

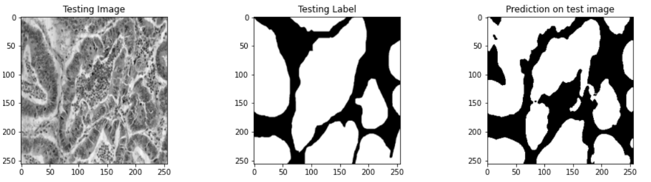


Fig. 17: .Results of the segmentation on the histology images of glands

VII. DISCUSSIONS

In this section I have revisited the aims and objectives as set in Section III and analyzed how far this project has delivered on the objectives. Below are the summarized results of all the experiments performed.

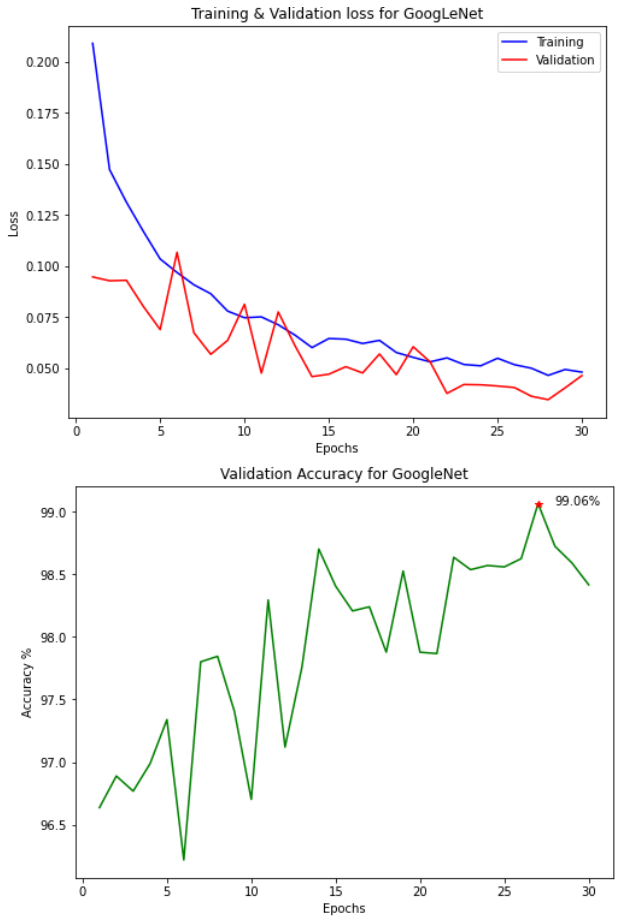


Fig. 14: Performance of GoogLeNet during training phase

A. *Deep-Learning methods perform better than tradional methods.*

The deep-learning models for image classification have out-performed the traditional methods, the average accuracy of the GoogLeNet based model is almost 20% more than that of methods using LBP or Haralick. In comparison with the works of Junaid Malik. et. all [11] the accuracy achieved by them is 79.5% for rLBP and 65% for Haralick using Linear Kernel for SVM classifier for Cancer detection problem. Pavel Kr´al et. all [10] achieved an accuracy of 84% for LBP for detecting cancer on Breast Cancer images while the accuracy achieved in my project is 76.46% and 75.20% for LBP and Haralick respectively.

B. *Identify the best deep-learning model for Cancer detection and Tissue classification*

Based on the results Xception and GoogLeNet has outperformed rest of the state-of-the-art architectures. For the Cancer detection problem, the best results were shown by Xception (Accuracy: 99.14%, F1 score: 99.14%) followed by GoogLeNet, DenseNet and ResNet whose accuracies are 99.11%, 98.92% and 98.78% respectively. The AUC for the Xception and GoogLeNet models is 1.0, whereas that of DenseNet and ResNet is 0.999 which would imply that all the models are highly efficient in predicting the classes. For tissue classification as well GoogLeNet had the best performance with overall accuracy of 98.86%. 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 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 we can see with each training cycle (epoch) the accuracy increases faster than most models in Fig 19. which 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.

C. *Effect of skipping layers in Xception*

It is observed that omitting layers from the Xception architectures provides better accuracy in both the problems (99.46% and 98.40%). This may imply that re-training a huge number of layers might lead to the vanishing gradient problem and overfitting, thus this fine-tuning performs better than the original architecture.

D. *Effect of Transfer learning on performance*

My experiment (refer section V, F) 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 speed 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 my experiment the model was pre-trained on Lung Cancer images, the knowledge gained by this model was used to train the core problem i.e., cancer detection on colorectal cancer images.

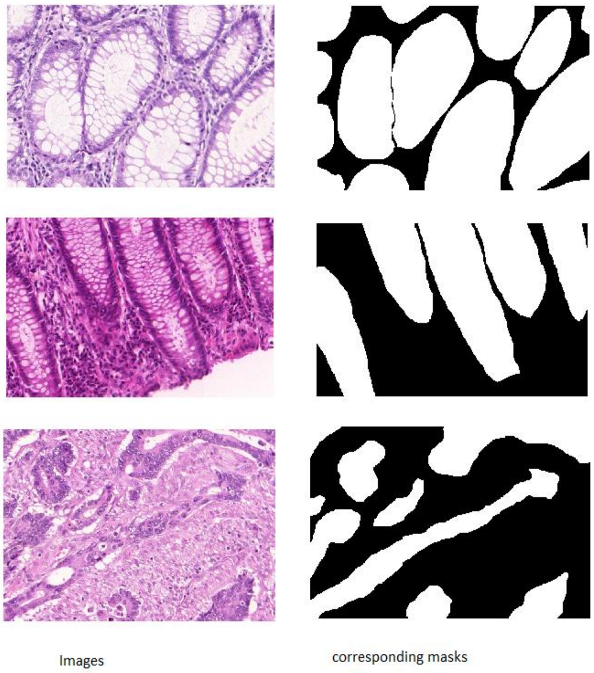


Fig. 16: Histology images of gland and their corresponding masks

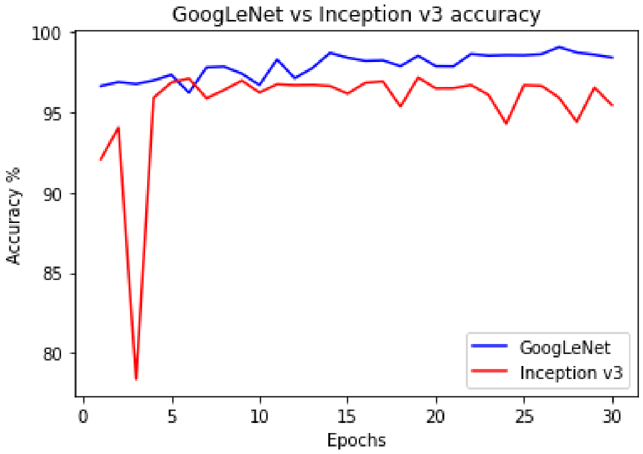


Fig. 19: GoogLeNet vs Inception v3 validation accuracy

E. *Semantic segmentation of glands on the histopathology images.*

The scope of this experiment was limited by the unavailability of sufficient images and masks to train the UNet model. As seen in Fig 17. there are 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.

F. *Comparison of performance with other researches.*

Below table Fig 20. demonstrates the accuracy obtained by various researchers who have aimed to solve similar problems. It is observed that the Xception and GoogLeNet models used by me has provided better results than that of the other researchers.

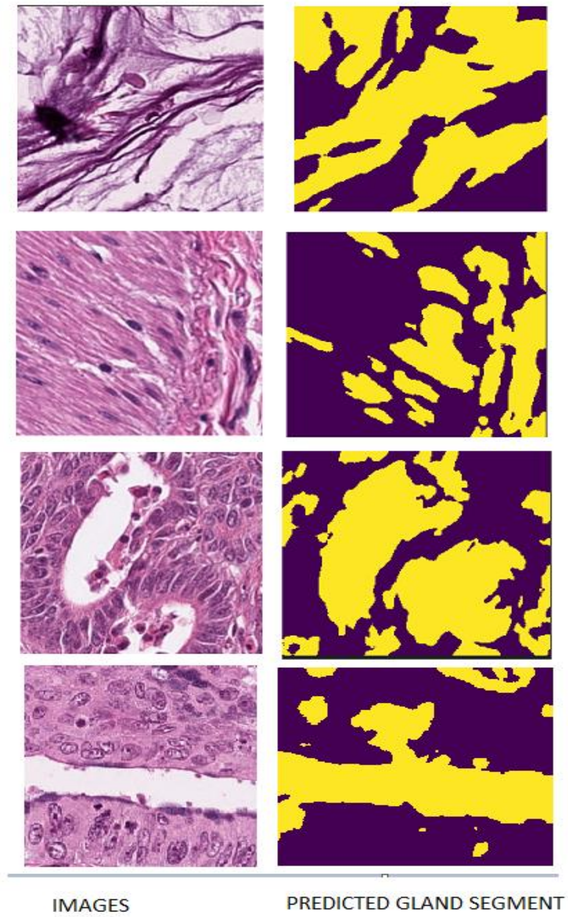


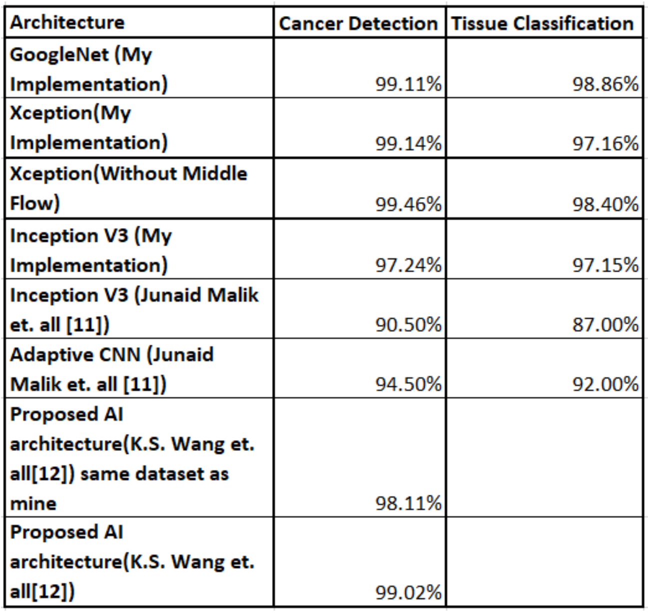
Fig. 18: Gland Segmentation results on the colorectal cancer images

The primary reason for better performance of my models could be due to the fact that huge amount of data (almost 30,000) was used by me to train the models, thus the models had very efficient learnable parameters. Another aspect of my experiments are all datasets used by me have no class imbalance. I have used equal number of images for each class in the dataset.

G. Future Scope

I have tested all my models primarily on the dataset: ‘NCT-CRC-HE-100K’ [5].

- 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 i.e., patients.
- I have trained GoogLeNet on Lung Cancer images [24] which gave me an accuracy of 99.47% . The tests can include histology images for breast cancer, skin cancer, brain tumors etc.
- Train and test models on different types of pathological images like biopsy, CT scan, X-rays.
- 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 my work. Also, the performance of models on images where blur and noises have been introduced can be tested.



Architecture	Cancer Detection	Tissue Classification
GoogleNet (My Implementation)	99.11%	98.86%
Xception(My Implementation)	99.14%	97.16%
Xception(Without Middle Flow)	99.46%	98.40%
Inception V3 (My Implementation)	97.24%	97.15%
Inception V3 (Junaid Malik et. all [11])	90.50%	87.00%
Adaptive CNN (Junaid Malik et. all [11])	94.50%	92.00%
Proposed AI architecture(K.S. Wang et. all[12]) same dataset as mine	98.11%	
Proposed AI architecture(K.S. Wang et. all[12])	99.02%	

Fig. 20: Comparison of similar works

- 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.
- Fine tune models and study the effect on efficiency.
- Semantic segmentation of glands in histopathology images can be improved with the proper usage of images and 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.

VIII. Conclusion

In this project I have done a comparative study of the different CNN based deep-learning architectures on primarily solving two problems Cancer Detection or diagnosis and tissue classification using histopathology images of colorectal cancer. It has been experimentally determined that deep learning-based algorithms are better at cancer diagnosis than the traditional machine learning approaches. My results show that GoogLeNet has fared better than the more advanced and deeper neural networks in classifying the tissues. Whereas Xception performed best in Cancer detection. Omitting few layers from the Xception architecture increases its overall efficiency. I have further studied the effects of using pre-trained models in Cancer diagnosis and results show Transfer Learning technique is effective with performance slightly better than the normal deep learning approach. This technique comes handy when core data is scarce and there is availability of similar data. I have also performed semantic segmentation of the glands in the histology images of colorectal cancer. However, due to unavailability of appropriate images and masks the results of segmentation is not very accurate. In future there is a huge scope of improvement in area.

IX. declaration

Declaration of Originality. I am aware of and understand that this assignment is my own work, except where indicated by referencing, and that I have followed the good academic practices

Declaration of Ethical Concerns. This work does not raise any ethical issues. No human or animal subjects are involved neither has personal data of human subjects been processed. Also, no security or safety critical activities have been carried out

X. REFERENCES

[1] World Cancer Research Fund/American Institute for Cancer Research. "Diet, Nutrition, Physical Activity and Cancer: a Global Perspective". Continuous Update Project Expert Report 2018.
Available at: https://www.wcrf.org/diet-activity-and-cancer/

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Adv. Neural Inf. Process. Syst., 2012.

[3] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do & Kaori Togashi , "Convolutional neural networks: an overview and application in radiology" ,22 June 2018.

[4] Teresa Araújo ,Guilherme Aresta ,Eduardo Castro ,José Rouco,Paulo Aguiar,Catarina Eloy,António Polónia,Aurélio Campilho., "Classification of breast cancer histology images using convolutional neural networks," PLoS One, 2017.

[5] Kather, Jakob Nikolas; Halama, Niels; Marx, Alexander, "100,000 histological images of human colorectal cancer and healthy tissue", 7 April 2018.

[6] Sari CT, Gunduz-Demir C. , "Unsupervised feature extraction via deep learning for histopathological classification of colon tissue images. IEEE Trans Med Imaging". May 2019.

[7] Korsuk Sirinukunwattana, Shan E Ahmed Raza, Yee-Wah Tsang, David R J Snead, Ian A Cree, Nasir M Rajpoot, "Locality sensitive deep learning for detection and classification of nuclei in routine Colon Cancer histology images". May 2016.

[8] Ahmad Chaddadand Camel Tanougast, "Texture Analysis of Abnormal Cell Images for Predicting the Continuum of Colorectal Cancer", 17 January 2017.

[9] Jakob Nikolas Kather et. all ,"Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study", 24 January 2019.

[10] Pavel Kr´al, Ladislav Lenc,"LBP FEATURES FOR BREAST CANCER DETECTION", September 2016, DOI:10.1109/ICIP.2016.7532838.

[11] Junaid Malik, Serkan Kiranyaz, Suchitra Kunhoth, Turker Ince, Somaya Al-Maadeed, Ridha Hamila, Moncef Gabbouj, "Colorectal cancer diagnosis from histology images: A comparative study", 28 March 2019.

[12] K. S. Wang, G. YuWang et. all, "Accurate diagnosis of colorectal cancer based on histopathology images using artificial intelligence", 23 March 2021.

[13] K. Sirinukunwattana, J. P. W. Pluim, H. Chen, X Qi, P. Heng, Y. Guo, L. Wang, B. J. Matuszewski, E. Bruni, U. Sanchez, A. Böhm, O. Ronneberger, B. Ben Cheikh, D. Racoceanu, P. Kainz, M. Pfeiffer, M. Urschler, D. R. J. Snead, N. M. Rajpoot, "Gland Segmentation in Colon Histology Images: The GlaS Challenge Contest", 1 March 2016.

[14] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, "ImageNet classification with deep convolutional neural networks", December 2012.

[15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, "Going Deeper with Convolutions", ,17 September 2014.

[16] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, Zbigniew Wojna, "Rethinking the Inception Architecture for Computer Vision", 11 December 2015.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Deep Residual Learning for Image Recognition" ,10 December 2015.

[18] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam, April 17, 2017, 'MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications'

[19] François Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions", 4 April 2017.

[20] Gao Huang; Zhuang Liu; Laurens Van Der Maaten; Kilian Q. Weinberger, "Densely Connected Convolutional Networks", 26 July 2017.

[21] Saining Xie; Ross Girshick; Piotr Dollár; Zhuowen Tu; Kaiming He, "Aggregated Residual Transformations for Deep Neural Networks" , 26 July 26 2017

[22] Paszke, A. et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Advances in Neural Information Processing Systems 32", 2019, Curran Associates, Inc., pp. 8024–8035

[23] Martín Abadi, et. all,"TensorFlow: Large-scale machine learning on heterogeneous systems",2015. Software available from tensorflow.org.

[24] Andrew A. Borkowski, Marilyn M. Bui, L. Brannon Thomas, Catherine P. Wilson, Lauren A. DeLand, Stephen M. Mastorides, "Lung and Colon Cancer Histopathological Image Dataset (LC25000)", 16 Dec 2019.

[25] Olaf Ronneberger, Philipp Fischer, Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", 18 May 2015.

[26] Coelho, L.P., "Mahotas: Open source software for scriptable computer vision", 2013. Journal of Open Research Software, 1(1), p.e3. DOI: http://doi.org/10.5334/jors.ac.

[27] T. Ojala; M. Pietikainen; T. Maenpaa st. all, "Multiresolution Grayscale and Rotation Invariant Texture Classification with Local Binary Patterns", 7 July 2002.

[28] Alice Porebski; Nicolas Vandenbroucke; Ludovic Macaire et. all, "Haralick feature extraction from LBP images for color texture classification", 9 January 2009.

[29]https://github.com/Mayurji/Image-Classification-PyTorch

[30] https://github.com/hawrot/image-classification-pytorch