Histopathological Cancer Detection

Jeet Shah
Master of Engineering in Electrical and Computer Engineering,
Western University,
London, Ontario, Canada
jshah72@uwo.ca

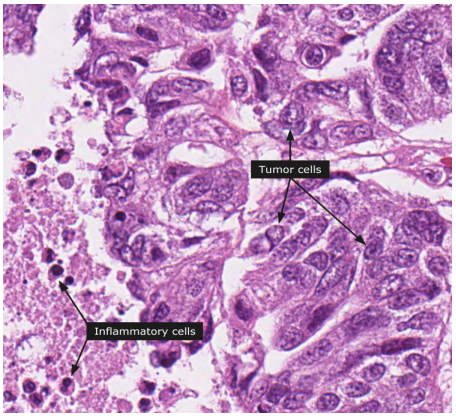
Khushali Patel
Master of Engineering in Electrical And Computer Engineering,
Western University,
London, Ontario, Canada
kpate372@uwo.ca

Divyam Bhatt
Master of Engineering in Electrical And Computer Engineering,,
Western University,
London, Ontario, Canada
dbhatt6@uwo.ca

Abstract*— Histopathological Cancer Detection is to detect metastatic cancer from a plethora of small image patches which were withdrawn from large digital pathology scans. The dataset was endowed by Kaggle website which contains label and ID columns- labels are two classes 0 and 1 in which images will be categorized and ID is unique number given to a particular image. Dataset consists of 3,27,680 color images (96 x 96px) extracted from histopathologic scans of lymph node sections. Blood samples are represented in the form of images in this dataset. In this project, we are going to build a model which will classify whether the blood sample has a cancer or not. Model will be trained and tested using one of the most popular machine learning algorithms, Convolution Neural Network (CNN). CNN is specifically designed to handle pictorial data such as images. The outcome of this project will be helpful to the pathology field where doctors will be able to detect the cancer rapidly using an image of a blood sample. This report gives detailed explanation on CNN and its working with appropriate illustrations. It also includes step by step tutorial on how to use CNN algorithm to detect cancer from an image.

Keywords*— Cancer Detection, CNN, histopathology, metastatic tissue, Biopsy

I. Introduction



Cancer is one of the most fatal diseases in the world. Cancer is the second leading cause for death. Globally, one out of six deaths are due to cancer. Biopsy is the method to remove a sample of cells from body in order to analyze it in a laboratory. If based on a symptom in body, doctor has found an area of concern one may undergo a biopsy in order to determine whether one has cancer or not. For major concern, the best way to undergo a diagnosis is to perform a biopsy to collect cells for examination. Histopathological Cancer Detection (HCD) is used to determine whether the blood sample has a cancer or not. Machine Learning algorithms will be applied to obtain the results. It will be useful for the pathology doctors to analyze the blood samples and rapidly generate the outcomes of the analysis. This process will be helpful to descry cancer rapidly. Convolutional neural networks are used for image classification where input images are cataloged in diverse classes. Histopathology Cancer Detection is a binary image classification problem. The main goal is to identify the presence of metastases from 96 x 96 pixels histopathology image.

II. Related Work

The research paper on Arxiv.org (Deep Neural Network Improve Radiologists Performance in Breast Cancer Screening) describes a deep Convolutional Neural Networks (CNN) – for classifying image that has the area under Receiver Operating Characteristics (ROC) curve of 0.895 to predict the cancerous tissues. They insisted that the probability of positive cells predicted by a radiologist from Artificial Intelligence system result, higher Area Under the Curve (AUC) than any other method.

The researchers Herin Zhaa, Tony of Google Brain are working on the Histopathological Cancer Detection techniques. The researchers achieved an excellent accuracy of 99.12% but they are still trying diverse algorithms to reduce the complexity.

III. Methodology

In this section, we are going to describe the methods which were used to pre-process the data, model the train and test data, and measure the accuracy.

A. Dataset Description

The dataset was bestowed publicly by Kaggle for its competition purposes. Dataset includes around 3.20 lakhs images of different blood sample. There are three files- train_labels.csv, train.zip, test.zip- given by Kaggle. Train_labels.csv includes two columns labels and ID. The dataset contains images annotated with binary label showing presence of metastatic tissue.

Labels are the values 0 and 1 which will determine whether the image has cancer or not. ID are the unique number given to an image.

B. Exploratory Analysis

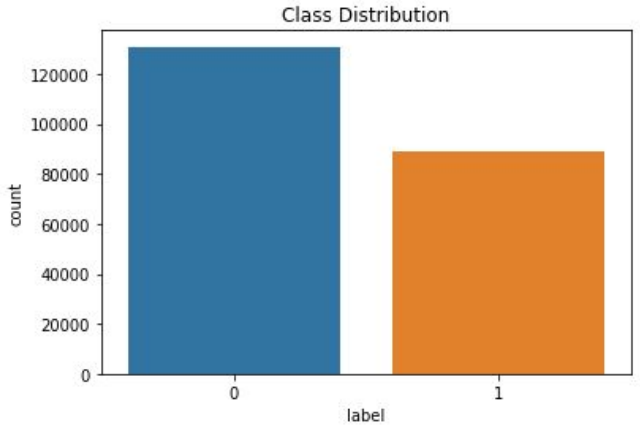


figure 1.1 Class Distribution

Train data consist of 220k images and test has 100K images. All images are in color form with Tagged Image File with 96 x 96px extracted from histopathologic scans of lymph node sections.

Above figure illustrates the distribution of non-cancerous cell (Label 0) and cancerous cell (Label 1) in the dataset. There are 130K and 90K cells which are labelled 0 and 1 respectively.

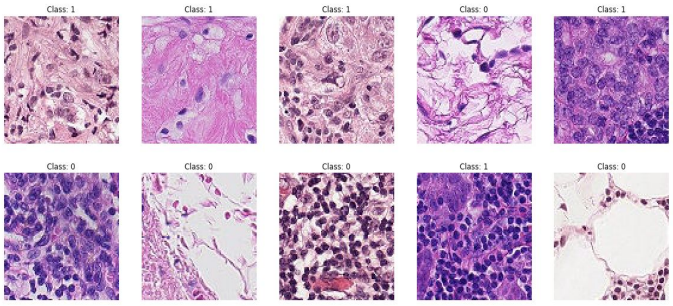
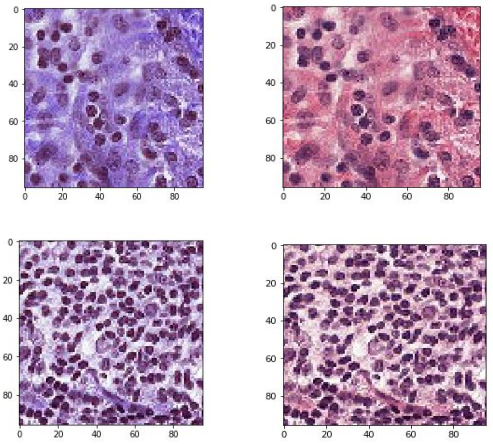


figure 1.2 Histopathology Images with and without cancerous tissues

C. Data Pre-Processing

The main aim of pre-processing is to determine the focal portions present in the image. Due to the considerable proportion of noise present in the image, it becomes necessary to reduce noise prior to focal area identification. In feature engineering, noise reduction is followed by the segmentation method to define the indexing of biological cells. The cell segmentation encompasses two methods- Region-based and Boundary-based. Data cleaning is implemented due to the occurrence of completely dark and bright images in the dataset.

- **Color Augmentation:**



A B

figure 1.3 Converting of Histopathologic images to similar pixels(bgr format). A represents Original Images and B represents the outputs of the original images

The images given to us as a form of data, varies a lot in terms of the color. So we need to perform the color augmentation and normalization to adjust the color values of images on pixel by pixel basis. We cannot convert these images in grayscale as some of the crucial properties may get lost.

- **Data Augmentation:**

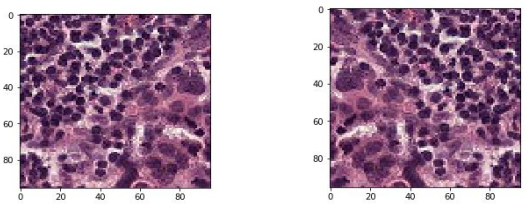
Data augmentation is used to creating new data based on existing data, without changing high level content of the original image. Augmentation techniques used on existing data are;

1) Cropping original image of size 96 X 96 px to 90 X 90 px size.

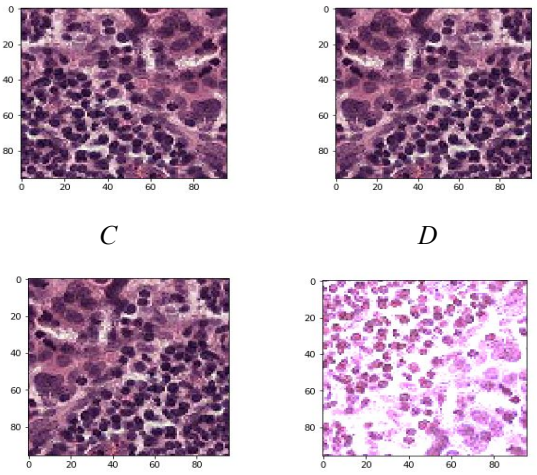
2) Rotating random images of 96x96px at right angles will preserve size of image

3) Lighting: To let the model clearly learn about metastatic tissue, scale the brightness component and generate new images.

4) Zoom: Taking a subsection of the original image, and intelligently increase the subsection so that it still has the exact same dimension in pixels as the original image.



A *B*



C D

E F

figure 1.4 A) Original Image B) Flip Left Right C) Flip Bottom Up D) Rotation 90 E) Rotation 180 F) Brightness

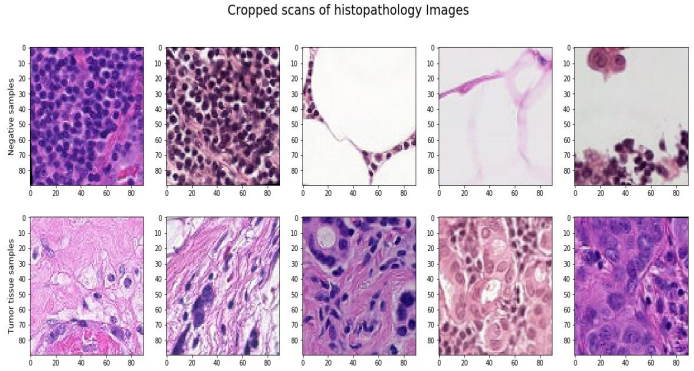
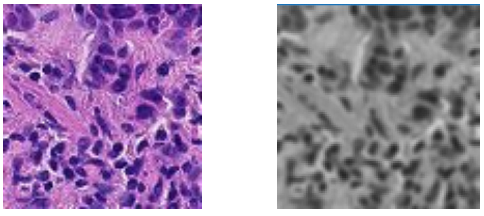


figure 1.5 Cropped Images

We cropped the images from 96 X 96 pixel to 90 X 90 pixel.

- **Gaussian Blur**

  Image smoothing is done by combining images with a low pass filter kernel. It removes high frequency content from the image result in edges being blurred. Here we applied gaussian kernel using Gaussian Blur. Width and height of the kernel are specified as positive and odd. Specifying standard deviation in X and Y directions, sigma X and sigma Y respectively. If they are given as zeros, they are calculated from kernel size.

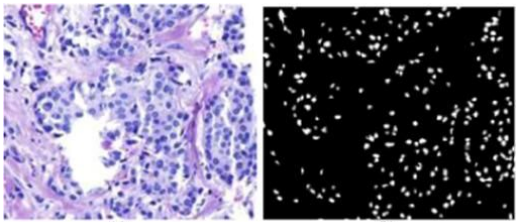


a. Original Image b. Gaussian Blur

figure 1.6 Applying gaussian blur to Original Image

- **Cell Segmentation**

  Nuclei cell segmentation is the most important step to diagnosis cancer cell using machine learning models. Data preprocessing and segmentation stages help enhancing the image quality and extracting the nuclei regions. Segmentation is an important step towards automatic image analysis. Segmentation helps to discriminate between foreground and background of the image. The main objective is to extract cell nuclei from the image. We applied two techniques to perform cell segmentation. 1.Otsu thresholding 2.Watershed Algorithm



a b

figure 1.7 Applying Cell Segmentation to Original Image

  - **Otsu's Thresholding technique (OTT)**:

    It is segmentation method in terms of grey-level image histogram. It differentiates foreground and background of image by selecting an adequate threshold value. The aim of Otsu's method is to determine the optimal threshold that minimizes the intraclass variance [1]. Algorithm steps are:

    1) Find probability and histogram for all levels.

    2) Finds the class mean and probability of image.

    3) Move to all possible maximum thresholds value

    4) Improve value of mean and probability

    5) Selects maximum value of class variances.

    The below figures demonstrates the OTSU thresholding technique where we convert the image into grayscale.Using OTSU thresholding, cells in the original image from the right will be extracted by making the background brighter.

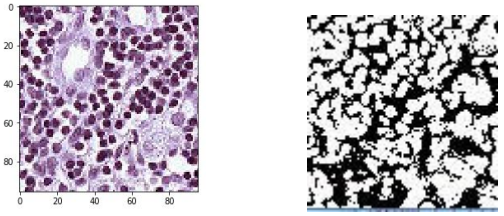


figure 1.8 OTSU Thresholding

- **Watershed Algorithm:**

  Watershed Algorithm is used for segmentation through which we can separate the objects in an image. Watershed algorithm manages an image as a topographic map.

1) Convert an image into grayscale image where image can be seen with black and white regions.
2) Extract the area where the objects are placed i.e. tissues/cells.
3) OTSU Thresholding is applied on the image to differentiate between background and foreground
4) Unknown regions and known regions are differentiated using colors.

From the below figure we can see the Watershed Algorithm on the original image. In the original image some cells are overlapping which cannot be seen through the naked eye. So from the resultant image we can see the region where the cells gets overlapped.

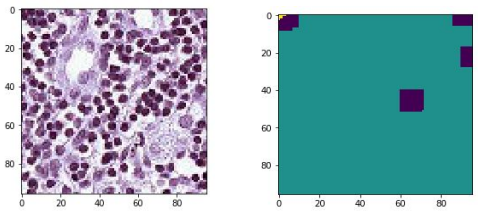


figure 1.9 Watershed Algorithm

D. *Workflow Overview*

Our overall workflow can be understood as three top-down multi-classification stages. We describe the steps as follows:

Training stage: the purpose of the training stage is to learn to represent the feature. After importing histopathological images, the model first learns the hierarchical feature representation during training and share the same parameters of weights and biases. The high-level feature maps then enter into ℓ2 normalizations. The outputs of the four branches are transferred to maximize and minimize the Euclidean distance of inter-class and intra-class respectively. Finally, a stochastic gradient descent algorithm optimizes the losses.

Validation Stage: The main goal of this stage is to avoid overfitting, tune the parameters, and select the optimal model based on the epochs for testing.
Testing Stage: The goal of this stage is to evaluate the performance of the CNN. After the first step of the input layer, the former layers can learn low-level features that include colors, edges, and boundaries via repeated iterations of high-level layers, discriminative semantic features can be extricated and inserted into a trainable classifier.

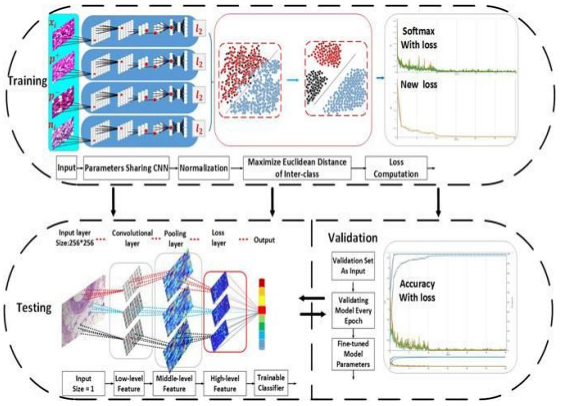


figure 1.10 Workflow

IV. CLASSIFICATION ALGORITHMS AND EVALUATION

A. *Convolutional Neural Networks (CNN)*

Neural Networks (NN) was influenced by biological processes. There are many types of NN algorithm. In this report, we are going to discuss about Convolutional Neural Network (CNN). CNN are most commonly applied to the pictographic data such as images, video and so forth. Architectural view of CNN includes Input Layer, Feature Learning Layer and Classification Layer. Major Application of CNN are Image classification, Object Detection from Images and Videos, Image Analysis.

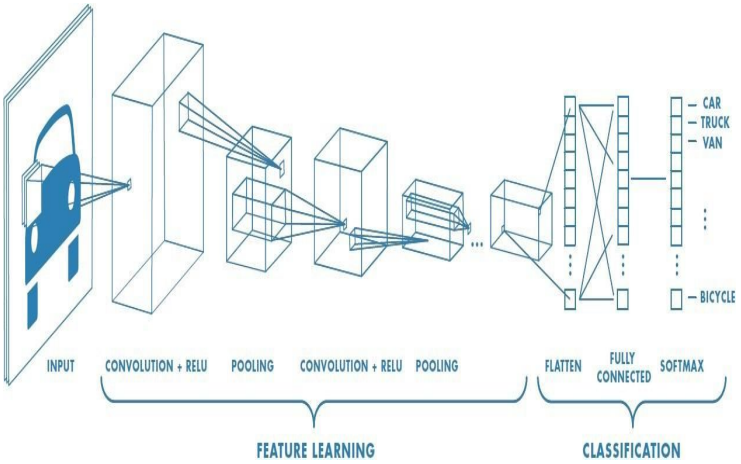


figure 2.1 Working of CNN

- **Input Layer:**

  Input Layer holds the raw pixels of an image.

- **Feature Learning Layer:**

Feature Learning Layer is used to extract features from an image using Convolutional Layer, Rectified Linear Unit (ReLU) and Pooling Layer

- **Convolutional Layer:**
  Convolutional Layer is a layer that takes an input image and generates the activation map with the help of filters. Filters are the set of weights in the convolutional layer. When input image is convolved with the filter, the outcome is activation map. Filter count is a hyper parameter which controls the production of activation maps. Activation Maps are the result of dot product of an input image and filters.

- **Rectified Linear Unit (ReLU) Layer:**
  ReLU Layer is commonly used with convolutional layer. ReLU Layer is used to remove non-linearity between images. ReLU Layers uses rectified function. It implements an elementwise activation function over the input data thresholding.

- **Pooling Layer:**
  Pooling Layer is used to decrease the spatial size of the convolved feature which helps in decreasing the computational power required to process the data with the help of dimension reduction. There are two type of Pooling – Max Pooling and Average Pooling. Max Pooling is used as Noise Suppressant. Along with Dimensionality reduction, Max Pooling also eliminates noisy activations and accomplishes de-noising of an image. On the contrary, Average polling achieves only dimensionality reductions Following Image is the best example of Pooling Layer.

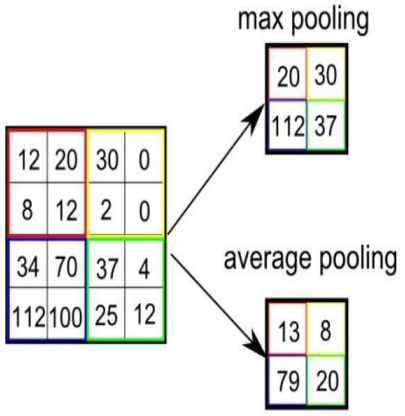


figure 2.2 Pooling Layer

- **Classification Layer:**

  Classification Layer categorizes an image into respective classes using Fully Connected Layer. Fully Connected Layer makes use of SoftMax Function, Sigmoid Cross-entropy and Euclidean Functions. SoftMax Function predicts a single class of K mutually exclusive classes. Sigmoid Cross-Entropy envisages K independent probability values in [0,1].

B. Evaluation:

Confusion Matrix: We divided the data using the Holdout method where we partitioned 70% of the data to train the model and 30% to test the model. We will use the Confusion Matrix and measure the True Positive rate to calculate the accuracy of all the models.

		Prediction	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

figure 2.3 Confusion Matrix

True Positive (TP): Number of instances labeled as 'true' and classified as 'true'.
True Negative (TN): Number of instances labeled as 'false' and classified as 'false'.
False Positive (FP): Number of instances labeled as 'false' but classified as 'true'.
False Negative (FN): Number of instances labeled as 'true' but classified as 'false'.

Similarly, we would also make use of the precision, recall and F1 score to measure the accuracy of the semantic analysis model. The mathematical equation to calculate the above-mentioned methods are:

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Precision: It is the proportion of the true prediction and all true prediction in the test sets.

Recall: It is the proportion of the true prediction and all true test sets.

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

F1 score: f1 score is the harmonic mean of the precision and recall.

ROC-AUC Curve: This curve measures the performance for classification problem at thresholds settings. ROC is a probability curve and AUC are a degree or measure of separability that describes the capability of the model in distinguishing between its classes. The model will predict well if there is higher AUC that means it will distinguish between patients with cancer and no cancer.

The Roc curve is plotted with the TP against FP where TP will be on the Y-axis and FP will be on the X-axis.

Cross-entropy Loss: It is used to measure the performance of a classification model whose output is a probability value between 0 and 1. The cross entropy value increases as the predicted probability diverges from the actual label.

V. EXPERIMENTS

Using Number of Epochs: Initially we used 25 epochs to train the model. The model got overfit after 10 epoch.So we tried tweaking the learning rates.But the model did not improve. We also tried with 10 epochs but we faced the same problem. So we tried with 15 epochs and achieved some good results.

Activation Function: For model training, Relu, tanh, Softmax, Sigmoid are the four activation functions which were implemented in the algorithm. But the best fitted was ReLU function.

VI. RESULTS:

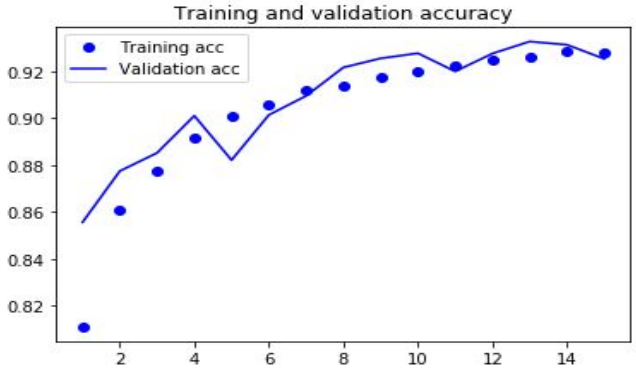


figure 3.1 Training and Validation Accuracy

The figure describes the training and validation accuracy against the number of epochs. The training accuracy for the first epoch is 81% and the last epoch (15) is 93% whereas the validation accuracy starts at 86% for the first epoch. We can observe that the validation accuracy does not improve for the last three epochs. So here, we used a regularization parameter called Early Stopping. It is clearly visible that the model is not getting overfit.

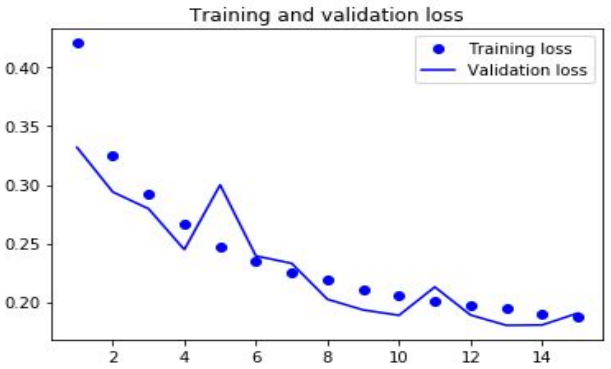


figure 3.2 Training and Validation Accuracy

The above figure demonstrates the training and validation loss. As the epochs increase both the training and the validation error decreases. We can observe some fluctuations in the validation loss at some particular time interval. Similarly, the validation loss is not increasing than the training loss, hence the model is not getting overfit.

.

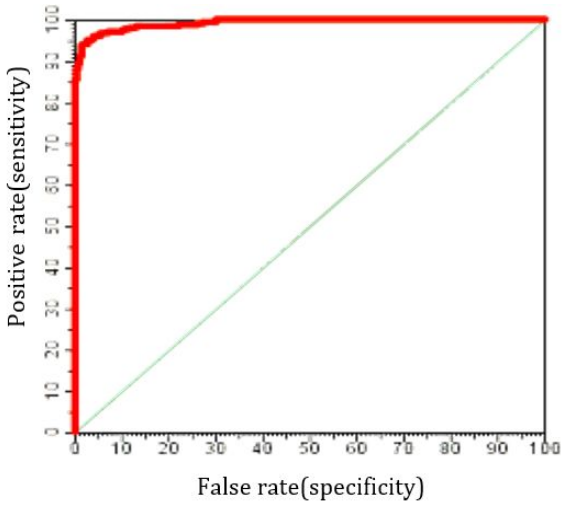


figure 3.4 ROC Curve

The above plot shows the ROC curve against the true positive rate(sensitivity) and the false positive rate(specificity). The central line shows the threshold value. From the plot we can say that the model performs well on the true positive values (sensitivity) which means the labels are correctly predicted. The curve starts in the lower left corner or when sensitivity is 0 and specificity is 1 corresponding to the cutoff of one which means classifying all the case as not defaults.The other end of the curve in the upper right corner corresponds to a cutoff equal to 0 where sensitivity is equal to 1 and specificity equal to 0. The curve is closer to the top left corner so it means it is a good plotting. The ROC curve has a value of 94.88%.

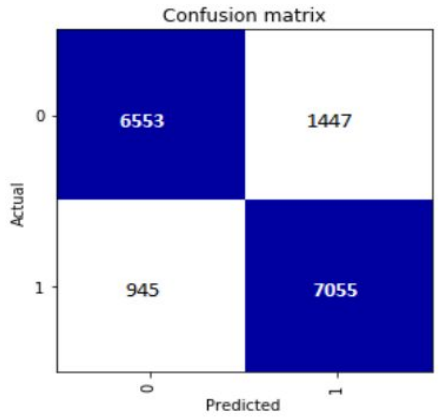


figure 3.5 Confusion Matrix

The above figure of confusion matrix helps us to understand the model performance and ratio of true positive and negative. This is the matrix table showing the count of True label and Predicted label. Considering 80% of data, model was able to predict 6553 images as True Positive i.e this number of samples are predicted to have tumor cells. On the other side, about 2000 samples were predicted as False Positive and False Negative. These samples were actually classified as a samples with cancer or not but prediction was completely opposite. The remaining images are considered to be True Negative which means that they were samples with no metastatic cells and prediction supported the fact.

	Precision	Recall	F1_Score	Support
no_tumor	0.87	0.82	0.85	8000
has_tumor	0.83	0.88	0.86	8000

Here, are the results for our prediction set. The F1 score of the image having the tumor cells is 0.86 whereas not having tumor cells is 0.85.

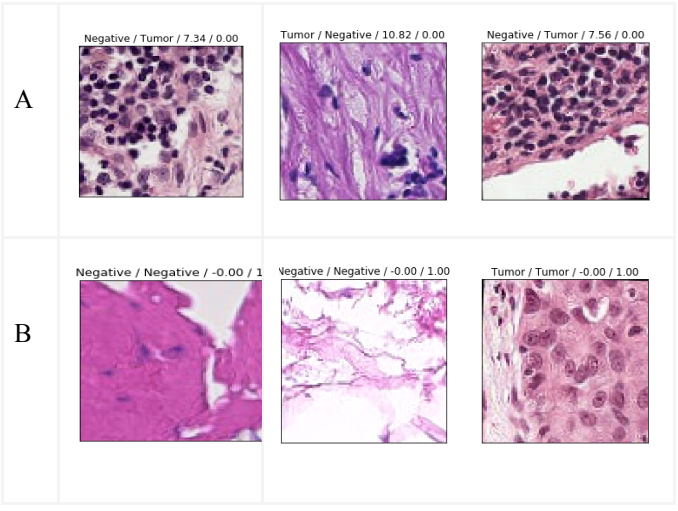


figure 3.6 A: represents the incorrect samples
B: represents the correct samples
Predicted and Actual Label with Probability

The above figure pictorizes the probability of the samples with the actual property of the labels.We can classify the actual difference between the samples that are correctly predicted and incorrectly predicted.

id	No_Tumor_Tissue	Has_Tumor_Tissue
0000ec92553fda4ce39889f9226ace43cae3364e	0.931653	0.068347
000c8db3e09f1c0f3652117cf84d78aae100e5a7	0.914055	0.085945
000de14191f3bab4d2d6a7384ca0e5aa5dc0dffe	0.939064	0.060936
000e6341cf18365d35b40f4991002fec8834afc0	0.896984	0.103016
00a2a7d5fbf50f1314a2f35e325c7cb452f4b5c8	0.618088	0.381912

The above table describes the probability of the histopathology images of the test set. There are two classes so if the probability value of any class will be higher, the input sample will be classified as the tumor being malignant or benign. From the table we can say that the initial image id has 93.16% chances that tumor is not present and 6.8% chances of tumor being present.

VII. Summary

On this Histopathologic Cancer detection dataset, we used CNN approach for classification of tumor being malignant or not. We performed random augmentations to remove noise from the image and cell segmentation techniques to make our model easily classify the images. We achieved the accuracy of 83.227%. The false negative rate is somewhat higher than the normal range which cannot be ignored for classifying the tumor. This model is the first step towards the digital histopathology images, which could still improve the patient care. We deploy a model that uses histopathologic and metabolic inspection on biopsy to assist in the diagnosis of the disease.

VIII. Future Work

We inspected the contribution of the biological entities in the detection of mitosis, this could be interesting for future research in the histopathological sector. We could also detect stroma on several datasets. The overall system developed for mitosis detection represents an ability to shift the load of rejection of the false positive rate depending on the magnification setting. The tissue useful for stroma detection is not highly recognized, in order to overcome that the nuclear membrane plays a crucial role.

We can extract some features from the images such as cell size, number of cells and so forth to apply classification algorithms.

IX. References

[1] https://pdfs.semanticscholar.org/6680/a2867a6fef675fee01e2d660aa7a7933443f.pdf

[2] https://wiki.cancer.org.au/oncologyformedicalstudents/Cancer_diagnosis:_Histopathology,_cytology_and_tumour_markers

[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3587978/

[4] https://www.researchgate.net/publication/327065839_Segmentation_of_Nuclei_in_Histopathology_Images_by_deep_regression_of_the_distance_map

[5] https://www.researchgate.net/publication/317723792_Nuclei_segmentation_in_histopathology_images_using_deep_neural_networks

[6] https://www.sciencedirect.com/science/article/pii/S235291481630034X

[7] https://ieeexplore.ieee.org/document/8388338

[8] https://www.kaggle.com/qitvision/a-complete-ml-pipeline-fast-ai

[9] Otsu N. A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics. 1979;9(1):62–66.doi:10.1109/tsmc.1979.4310076

[10] https://www.datacamp.com/community/tutorials/convolutional-neural-networks-python

[11] https://genomemedicine.biomedcentral.com/track/pdf/10.1186/gm332

[12] https://www.who.int/news-room/fact-sheets/detail/cancer

[13] https://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_imgproc/py_watershed/py_watershed.html

[14] https://medium.com/@arindambaidya168/https-medium-com-arindambaidya168-using-keras-imagedatagenerator-b94a87cdefad

[15] https://towardsdatascience.com/histopathologic-cancer-detector-finding-cancer-cells-with-machine-learning-b77ce1ee9b0a

[16] https://www.kaggle.com/vbookshelf/cnn-how-to-use-160-000-images-without-crashing

[17] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5863327/

[18] https://arxiv.org/pdf/1902.06543.pdf

X. Appendix:

Work was divided equally and all members were updated with the work they are doing. Following are the work implemented by team members:

Divyam Bhatt:
Found the dataset on Kaggle and passed it to other group members. Worked on data pre-processing and data augmentation. Build confusion matrix.

Khushali Patel:
Read research papers for different ways to do data preprocessing and discussed in group. build the CNN model. Executed code on google colab with 15 epochs. Performed prediction on train set.

Jeet Shah:
Researched on dataset. Implemented cell segmentation using Otsu Thresholding and Watershed Algorithm. Done prediction on test set. Build ROC and AUC curve.

GITHUB Repository Link

https://github.com/Divyam1102/Histopathologic-Cancer-Detection.git