HISTOPATHOLOGIC CANCER DETECTION

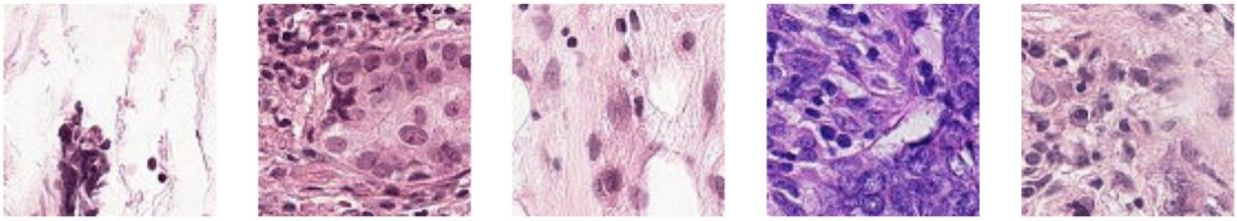
Abstract – This paper reports our experience with Histopathologic Cancer Detection.In the dataset, we are provided with a large number of small pathology images to classify.There are many well-developed tools and techniques for performing binary image classification, and these could hopefully be used to assist with the task of Histopathologic Cancer Detection.We use various classifiers and compare their results in this report..

INTRODUCTION

The goal of this project is to use image classification techniques to attempt to identify the presence of "Histopathologic Cancer" in small image patches taken from larger digital pathology scans.

Datasets

The train_labels.csv file provides the ground truth for the images in the train folder. *A positive label indicates that the centre 32x32px region of a patch contains at least one pixel of tumor tissue.*
The train data we have here contains 220,025 images and the test set contains 57,468 images.
Labels column : ***Class label - 0 for malignant and 1 for non-malignant***
Visualisation of sone train images



Images without Tumor Tissue

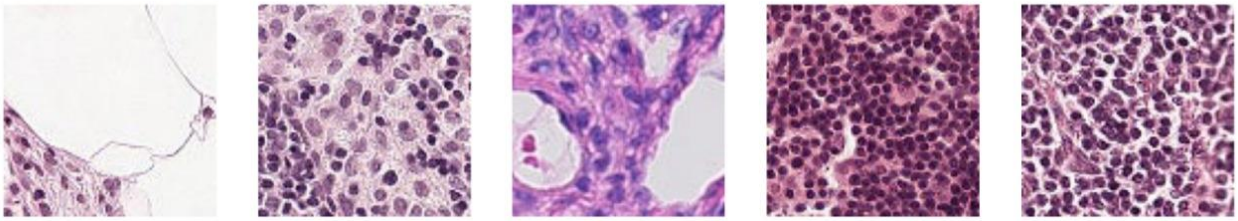
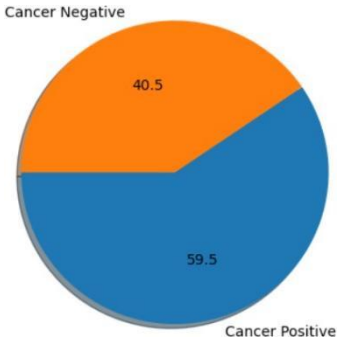
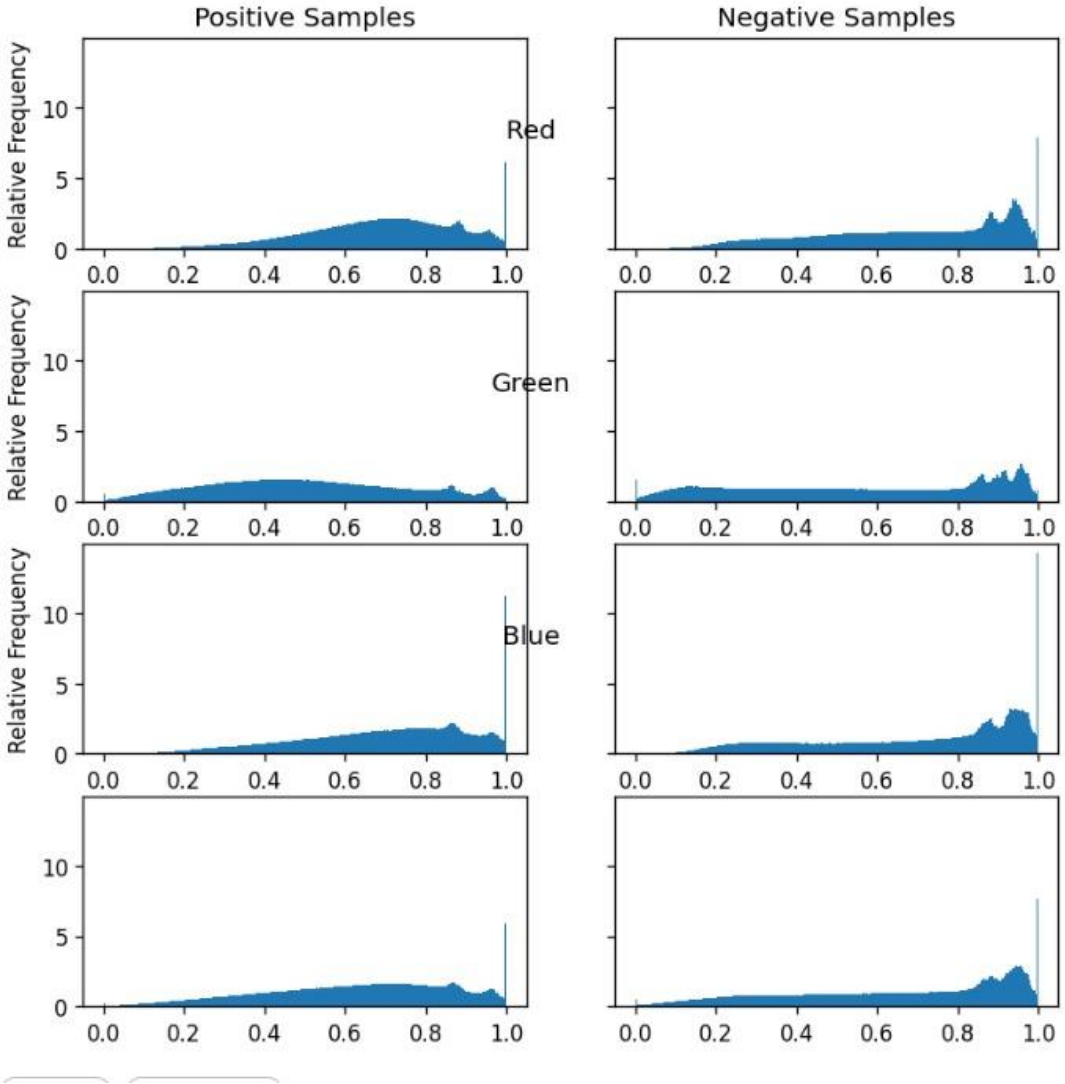


Image dataset : The image dataset has 130908 malignant and 89117 non-malignant data samples.





I. METHODOLOGY

Data Processing:
For our mission, we had to deal with enormous amounts of image data. Since processing this massive quantity of data is a challenge for Artificial Intelligence learning algorithms, we focus on the sampling of data in the big data paradigm. Increasing the amount of data does not inherently increase the amount of information Cancer negative Cancer positive it contains. Subsampling is used to reduce the quantity of data into a reduced data set while maintaining its semantics and structure.

Overview
There are various classification algorithms present out of which we shall implement the following.

- LDA
- Random Forest Classifier
- KNN
- Convolutional Neural Networks
- SVM
- SVM and KNN on LDA transformed data
- Model comparison

II. CLASSIFICATION

Exploring the dataset and pre-processing

On checking for Duplicate Entries in the train dataset , it was found that there are no duplicate ids present. We also refined the data by Removing an image which was causing a training error. Removing an image because it is black.

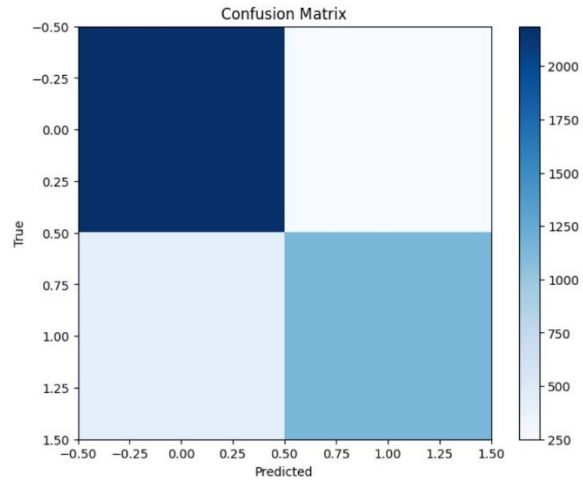
Implementation

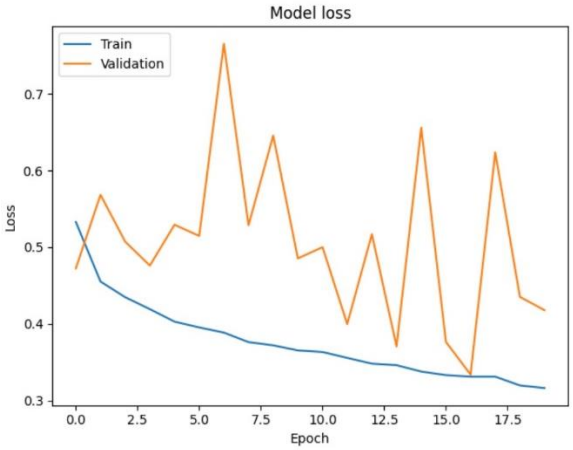
- ***CNN Model using Keras API:*** The CNN model is trained on a large dataset of histopathological images to learn discriminative features that can distinguish between cancerous and non-cancerous tissues.
  val_loss: 0.4177
  val_acc: 0.9440

```
125/125 [==============================] - 6s 50ms/step
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.90      0.87      2437
           1       0.82      0.73      0.77      1563

    accuracy                           0.83      4000
   macro avg       0.83      0.81      0.82      4000
weighted avg       0.83      0.83      0.83      4000
```





The results of CNN on image dataset :

Number of epochs	loss at the end	accuracy at the end	loss on validation set at the end	accuracy on validation set
20	0.3161	0.9356	0.4177	0.9440

Random Forest:

```
RF Model Metrics:
Accuracy: 0.734006734006734
Precision: 0.75625
Recall: 0.5041666666666667
F1 Score: 0.605
Confusion Matrix:
[[315  39]
 [119 121]]
```

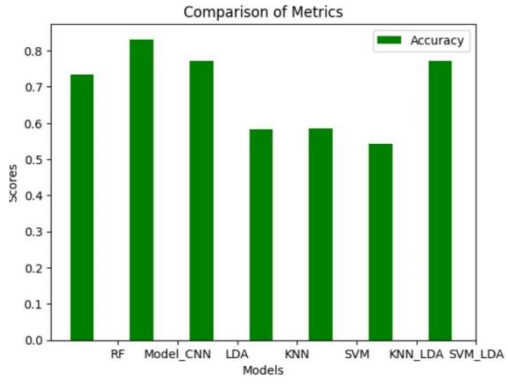
LDA:

```
LDA Model Metrics:
Accuracy: 0.7727272727272727
Precision: 0.7091633466135459
Recall: 0.7416666666666667
F1 Score: 0.7250509164969451
Confusion Matrix:
[[281  73]
 [ 62 178]]
```

KNN:

```
KNN Model Metrics:
Accuracy: 0.5824915824915825
Precision: 0.46923076923076923
Recall: 0.25416666666666665
F1 Score: 0.3297297297297297
Confusion Matrix:
[[285  69]
 [179  61]]
```

Comparison of models:



Applying KNN and SVM on LDA transformed data:

```
KNN accuracy with LDA data : 0.5841750841750841
SVM Accuracy with LDA : 0.5420875420875421
```

III.DISCUSSIONS

- In histopathological cancer detection project, some important discussions were:
- Selecting suitable machine learning algorithms for classification, such as random forest, decision tree, CNN, LDA, KNN, SVM, SVM and KNN on LDA transformed data.
- Evaluating the performance of the model with appropriate metrics, such as accuracy and precision.
- Discussing the trade-offs between different evaluation metrics and selecting the best one for the specific problem.
- Addressing potential ethical concerns, such as data privacy and bias.
- Exploring the potential for future work and improvements to the model.

IV. RESULTS AND ANALYSIS

Training CNN on our sampled dataset gave best validation accuracy of 94.40%, which is better than any of the models we tried previously, hence validating our choice of implementing CNN. Next we trained the same architecture on the original complete dataset for 20 epochs, which reported a testing accuracy of 94.40%. Hyperparameter tuning and model refining could not be attempted due to resource constraints, demanding enormous training runtimes.

CNN undergoes two significant transformations. The first method is convolution, in which pixels are convolved with a filter or kernel. Subsampling is a second significant transformation that can be implemented in a variety of ways (max pooling, min pooling, and average pooling) as required. The pooling layer is responsible for reducing the dimensionality of the data, and it is quite effective at preventing overfitting. After employing a combination of convolution and pooling layers, the output may be sent to a fully connected layer for efficient classification.

REFERENCES

[1] Pattern Classification -Book by David G. Stork, Peter E. Hart, and Richard O. Duda

[2] Link for image dataset: https://www.kaggle.com/competitions/histopathologic-cancer-detection/data

[3] Why CNN are good for image classification https://medium.datadriveninvestor.com/why-are-convolutional-neural-networks-good-for-image-classification-146ec6e865e8