# Histopathologic Cancer Detection

Apoorv Singh, Vinayak
2017027, 2017122
IIIT Delhi
[apoorv17027, vinayak17122]@iiitd.ac.in

Yatin Kumar Arora
2017124
IIIT Delhi
yatin17124@iiitd.ac.in

## Abstract

*Metastatic Cancer is the presence of fast-spreading cancer cells. Our objective is to classify images of lymph nodes on the basis of metastatic cancer cells. Machine Learning in the field of medicine is becoming incredibly important, especially medical imaging which makes it possible to predict the occurrence and potential spreading of cancer in lymph nodes. A binary classification model helps to predict whether a particular sample of images contains patches of metastatic cancer or not.*

## Introduction

We have taken PatchCamelyon (PCam) data from Kaggle dataset, which is a slight modification of the original PCam benchmarking dataset, to avoid duplicates present in the original dataset. The images are microscope images of lymph nodes that are stained with hematoxylin and eosin (HE). It consists of 327,680 color images (96 x 96px) extracted from histopathologic scans of lymph node sections. Each image is annoted with a binary label indicating presence of metastatic tissue.

## Related Work

This work used a pre-trained Resnet50 ImageNet model for training a neural network as the model. Furthermore, data augmentation techniques like resizing, random cropping, flipping, etc. were used. Optimization techniques like batch normalization, max-pooling, weight decay, ReLU activation were also used.

## Methodology

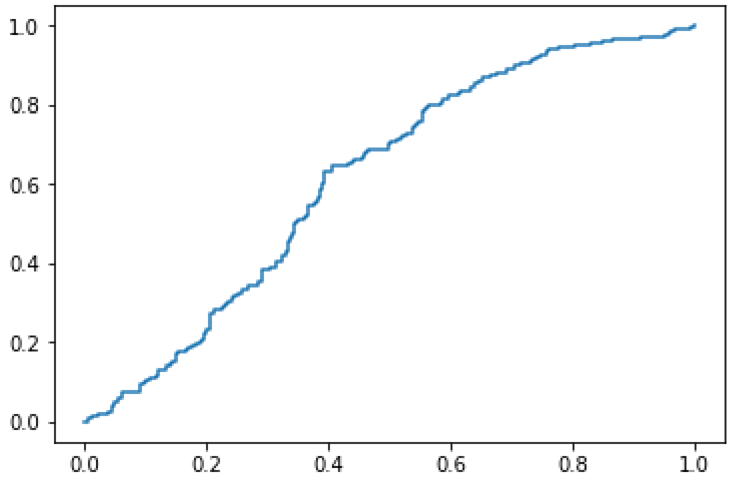
We decided to split the data set 80-20 into training and validation sets respectively. The original data set was extremely large and contained 220,000 samples. We decided to randomly sample 50,000 data points, which contained an equal number of samples belonging to the positive and negative classes. A similar problem occurred with the test set. We tested against a randomly-sampled test set containing 57,500 samples. The problem statement requires that we concentrate on the central 32X32 pixel region. The sample is labelled as positive if it contains at least one pixel of tumour tissue. Tumour tissue in the outer region of the patch does not influence the label.

We decided to evaluate using Area under ROC curve (which is the desired evaluation metric for the abovementioned Kaggle tournament). We also decided to include, AUC, Confusion Matrices and Accuracy metrics.

## Results

## Logistic Regression

We started with a base model which is convenient for binary classification. The model was implemented using the inbuilt Sklearn Library. The optimal learning rate was 0.001. The optimum regularization method was Lasso or L1 regularisation. The optimal results were:



ROC Curve

| | |
|---|---|
| 2483 | 2733 |
| 1334 | 3450 |

Confusion Matrix

## Support Vector Machine

We tried a vanilla SVM and predicted loss using linear, polynomial and RBF kernels. For hyperparameter tuning of the parameters “C” and “gamma”, we used sklearn’s GridSearchCV routine. In an attempt the show the influence of the blue pigment and its dominance over all the colors, we observed the accuracy we get with respect to every color, for each pair of “C” and “gamma” we took under consideration. The blue color was dominant around 80% of the times.

RBF

| **Gamma** | **C** | **Red** | **Green** | **Blue** | **Dominant Colour** |
|---|---|---|---|---|---|
| 0.0000027 | 100 | 0.6875 | 0.67625 | 0.705 | Blue |
| 0.000001 | 1000 | 0.7235 | 0.69375 | **0.74125** | Blue |
| 0.0000001 | 10000 | 0.73875 | 0.70875 | 0.725 | Red |

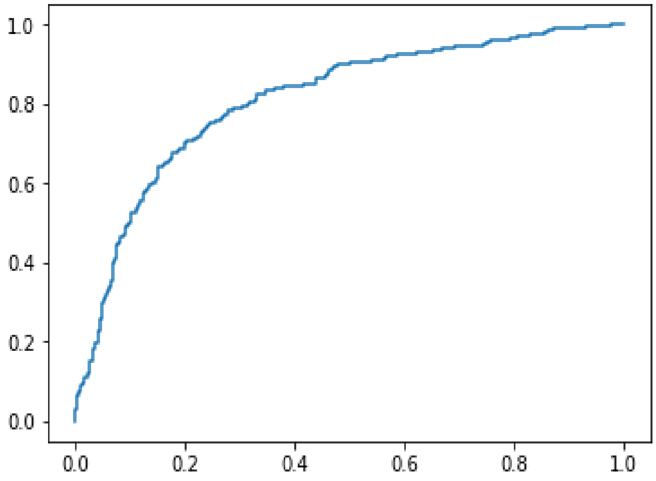
Polynomial

| **Gamma** | **C** | **Degree** | **Red** | **Green** | **Blue** | **Dominant Colour** |
|---|---|---|---|---|---|---|
| 0.0000005 | 100 | 3 | 0.60125 | 0.60625 | 0.61875 | Blue |
| 0.0000007 | 1000 | 2 | 0.62 | 0.5975 | 0.6175 | Blue |
| 0.0000003 | 10000 | 3 | 0.6 | 0.5925 | **0.6275** | Blue |

Linear

| **Gamma** | **C** | **Red** | **Green** | **Blue** | **Dominant Colour** |
|---|---|---|---|---|---|
| 0.000002 | 100 | 0.5945 | 0.6115 | 0.66 | Blue |
| 0.0000005 | 1000 | **0.6940** | 0.663 | 0.6795 | Red |
| 0.00003 | 10000 | 0.5815 | 0.6275 | 0.682 | Blue |

The optimal results were:



ROC Curve

| | |
|---|---|
| 3547 | 1283 |
| 1246 | 3924 |

Confusion Matrix

## Multi-layered Perceptron

A multi-layered perceptron is considered a precursor to CNN solutions. We implemented the model using the inbuilt Sklearn libraries. The parameters we had to determine the optimal number of layers, the number of nodes in each layer and the activation function.

| Layers | Alpha | Logistic | Tanh | Relu | Identity |
|---|---|---|---|---|---|
| 200, 300, 100, 250, 300 | 0.0001 | 0.5725 | 0.71 | **0.73875** | 0.7045 |
| 120, 250, 200, 150 | 0.0001 | 0.5442 | 0.69375 | **0.7325** | 0.665 |
| 300, 250, 200 | 0.0001 | 0.5442 | 0.69375 | **0.7325** | 0.665 |

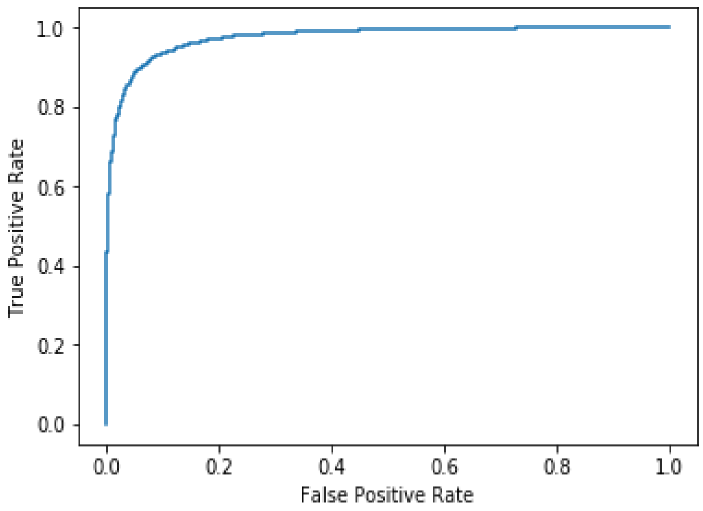
## Convolutional Neural Network

We first created a vanilla CNN model with the following layers in order: Convolutional Layer 1, ReLU, MaxPool, Conv2, ReLU, MaxPool, and three Fully Connected Layers.

It was found that the model was overfitting on the training samples, i.e. the training accuracy was reported to be ~98% and the test accuracy to be 74%. Therefore, we introduced Batch Normalization after the first Convolutional layer so that the extent of variation in the output neurons is decreased. This regularization technique helped achieve an accuracy of 77%.

Next, we applied hyperparameter tuning to set the optimal parameters of learning rate and momentum. After some iterations, we arrived at the optimal learning rate of 0.00006 and momentum equal to 0.9. The accuracy was reportedly increased to 86% after that.

It was noted that the accuracy turns out to be the optimal, amongst all models, for CNNs.



| | |
|---|---|
| 4583 | 388 |
| 414 | 4615 |

Confusion matrix

## Analysis and Conclusion

The property that the blue colour is indeed a good indicator of the presence of metastatic cancer cells in lymph nodes was verified. The accuracy was significantly better in almost all the different models we encountered.

We were faced with the problem of overfitting. We used data augmentation techniques such as rotation of images to counter it with significant results.

CNN’s provided the best results. They gave the best accuracy results as we can generate more complex networks while incurring significantly lesser costs in terms of time complexity. Due to their properties of sparse connectedness and parameter sharing, they are optimum in terms of determining local connectedness and have efficient time complexity.

## References

[1] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, M. Welling. ”Rotation Equivariant CNNs for Digital Pathology”. arXiv:1806.03962

[2] Ehteshami Bejnordi et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA: The Journal of the American Medical Association, 318(22), 2199–2210. doi:jama.2017.14585

[3] PyTorch tutorial https://pytorch.org/tutorials/beginner/blitz/cifar10_tutorial.html