IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Histopathological Image Classification using Machine Learning techniques

Paramita Bhattacharya, Nadeem Anwar, Pritha Roy

Abstract: Malignancy identification for diagnosis of cancer is crucial in medical science. Histopathology refers to the microscopic examination of tissue in order to study the manifestations of disease. The analysis of these complex histopathological images (HIs) is done manually by pathologies. The varied result leads to subjectivity of the pathologists and largely depends on the expertise of the examiner. Different features like colour, size, shape, structure of cells and tissues in biopsy samples are examined to classify these samples as benign or malignant. This is time consuming and prone to the subjectivity of the pathologist. To overcome this, computer assisted analysis is needed. In this project, different machine learning techniques are used to analyse their performance on BreaKHis dataset, which is composed of microscopic images of breast tumour tissue collected using different magnifying factors. Machine learning can work with limited computer resources. In this report, a performance analysis has been carried out on different machine learning techniques like K- nearest neighbour (KNN), support vector machine (SVM), decision tree, Gaussian naïve bayes to classify those HIs of breast cancer into benign or malignant. Decision tree works better with 87.98% accuracy while KNN achieves 92.62% accuracy when the data are scaled and normalised.

1. INTRODUCTION

Histology is the study of tissues and pathology is the study of disease. Hence, histopathology means study of tissues related to disease. Traditionally it is analysed manually by pathologists by observing under microscope. It requires expertise & experience of a pathologist. Thus, the result very much depends on the pathologist, which produces varied results. They study different features of those tissues. This includes size, structure, shape, density of cancerous cells in the tissue. Histopathological diagnosis is considered the gold standard in diagnosing cancer. Histopathological images are informative and contain diagnostic information as means of effective analysis.

Computer aided diagnosis overcomes the pathologist subjectivity. Machine learning models can be trained on the HIs using different features of these images like colour and texture features. These features are extracted from the images using different algorithms. Thus, these trained models can diagnose the disease. Here this work is performed to classify HIs into benign and malignant class. Machine learning is a subfield of Artificial Intelligence that gives systems the ability to learn themselves from given features. Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, agriculture, and computer vision. Machine learning techniques can work with limited resource environments. Machine Learning techniques can be classified into (a) supervised learning, (b) unsupervised learning and (c) reinforcement learning. In supervised learning, the system is presented with sample inputs

and their desired outputs so that it could learn the rule to map input to output, as shown in figure 1. Whereas in unsupervised learning, desired labels are not given. It leaves on its own to find the hidden pattern. System interacts with a dynamic environment in reinforcement learning to achieve a certain goal. Here the system is dynamically provided with feedback which is analogous to reward.

For the classification, this work has applied supervised learning techniques only. The project has applied the K-nearest neighbour (KNN or K-NN), Support Vector Machine (SVM), Decision Tree and Naïve Bayes algorithm. KNN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. Usually, this k is chosen to be odd. SVM is a supervised machine learning algorithm used for both classification and regression. The objective of the SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. A Decision Tree is a supervised learning algorithm. It is a graphical representation of all the possible solutions. All the decisions were made based on some conditions. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. Naive Bayes is a simple supervised machine learning algorithm that uses the Bayes' theorem with strong independence assumptions between the features to procure results. That means that the algorithm assumes that each input variable is independent.

In this project, we have analysed breast histopathology images by extracting features like colour, texture, etc. Further it is classified by traditional machine learning models like SVM, KNN, etc.

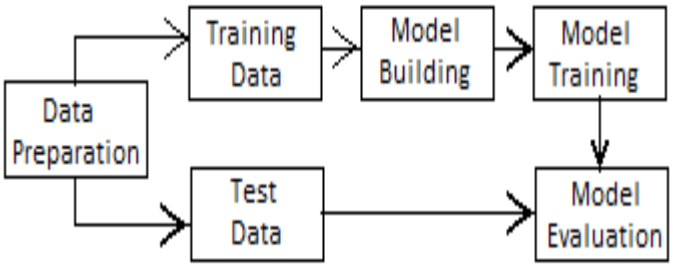

Figure 1: Process of supervised learning model

2. OBJECTIVES:

The objective of this project may be summarised as follows:

a. Extraction of colour, texture, etc. features from breast histopathology images.

b. Classification of breast histopathology image by traditional supervised machine learning approaches.

3. PROPOSED METHODOLOGY

3.1 Dataset Description

Our proposed supervised model is experimented on Breast Cancer Histopathological Image (BreaKHis) dataset, which is made of 9,109 microscopic images of breast tumour tissue collected from 82 patients using different magnifying factors (40X, 100X, 200X, and 400X). It has a total 2480 benign and 5,429 malignant samples (700 X 460 pixels, 3-channel RGB, 8-bit depth in each channel, PNG format). This database has been built in collaboration with the P&D Laboratory – Pathological Anatomy and Cytopathology, Parana, Brazil [1].

The BreaKHis dataset is divided into two classes: benign and malignant. In the current version, samples present in the dataset were collected by SOB method, also named partial mastectomy or excisional biopsy.

Various types/subtypes of breast tumours can have different prognoses and treatment implications. The dataset currently contains four histological distinct types of benign breast tumours: adenosis (A), fibroadenoma (F), phyllodes tumour (PT), and tubular adenoma (TA); and four malignant tumours (breast cancer): carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC) and papillary carcinoma (PC). Each image filename stores information about the image itself: method of procedure biopsy, tumour class, tumour type, patient identification, and magnification factor, for example, SOB_B_TA-14-4659-40-001.png.

Current work is focused on 1820 images of 400X magnifying factor only which has 588 benign and 1232 malignant images. Sample benign images are given in figure 2 and those of malignant are in figure 3.

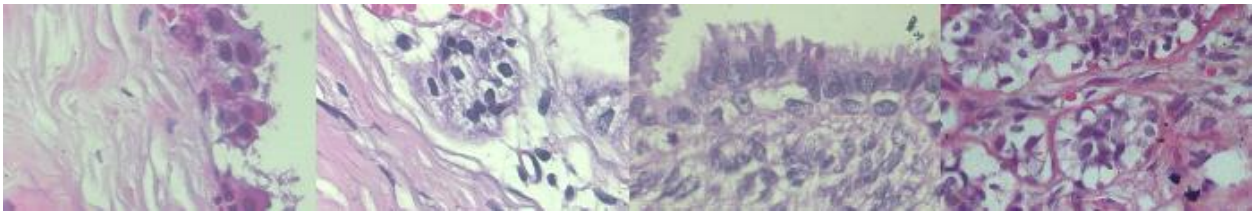

Figure 2 Sample benign images

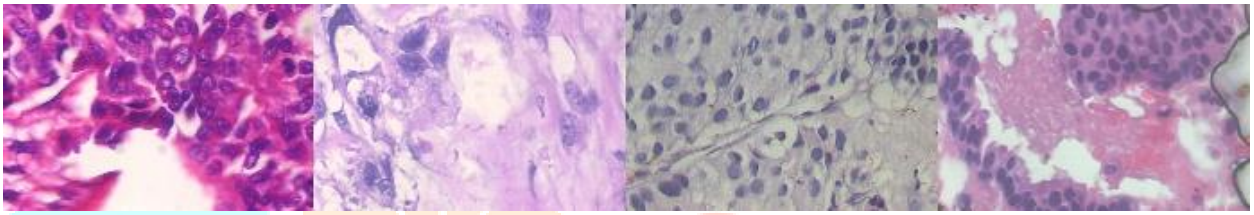

Figure 3 Sample malignant images

The dataset is split into train and test in 4:1 ratio as given in table 1 and saved to respective folders.

Table 1 Class wise data distribution

Image Class	Training Set	Test Set
Benign	470	992
Malignant	118	248

3.2 Pre-Processing and Features Extractions

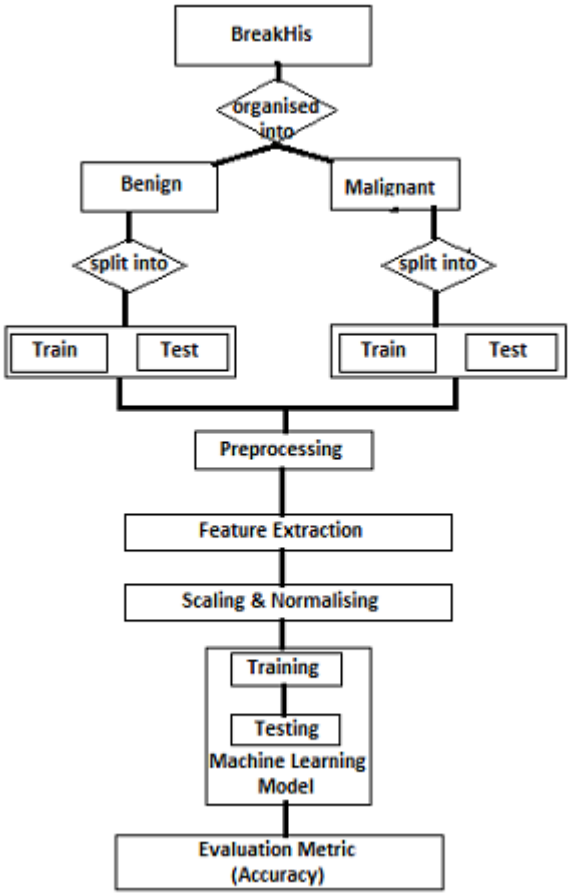

Figure 4 Proposed methodology

These images are pre-processed using histogram equalisation, also known as contrast enhancement and then grayscale image is binarized using Otsu thresholding. Sample binarized images are given in figure 5. Various colour features have been extracted from HIs. Those colour features include (Red-Green-Blue), HSV (Hue-Saturation-Value), LAB (Lightness-A-B), YCrCb extracted from original HIs and Gray from those binarized HIs. Similarly, GLCM texture features have been extracted from the binarized HIs. GLCM, stands for Gray-level co-occurrence matrix, is a method of examining texture that considers the spatial relationship of pixels. GLCM includes properties like contrast, dissimilarity, homogeneity, energy, correlation and ASM. Now both of these features have been saved in csv files separately for each set. Proposed methodology of the project has been figured out in figure 4.

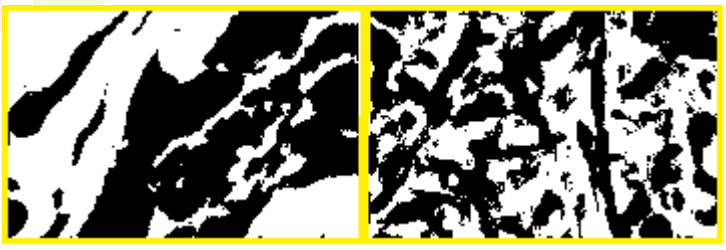

Figure 5 Sample pre-processed binarized images

3.3 Classification:

3.3.1 Experimental Setup

This work is carried out on Google Colaboratory updated up to 06.12.2022[2]. Firstly, the dataset has been uploaded on Google Drive and then mounted on Google Colaboratory. SKLEARN is used to split the dataset and design machine learning models. SKIMAGE.FEATURE is used to extract texture features. CSV has been used to work with csv files to save and read extracted data. To work on images, this project applied CV2 and MATPLOTLIB.

3.3.2 Training of ML classifiers

ML classifiers are separately trained on the actual extracted values. Furthermore, these are trained on scaled and normalised data. KNN model is built with various values of k (that is no of neighbours) which is usually taken to be odd e.g., 1,3,5 etc. SVM model is constructed using various kernel parameters like linear, poly, rbf and sigmoid. Similarly, the decision tree model is based on distinct values for criterion parameters e.g., Entropy and Gini. The Gaussian naive bayes model does not take any parameter.

4. RESULT ANALYSIS & DISCUSSION

Detailed discussion on experimental setup carries out experiment on classifying the dataset using multiple machine learning techniques has been described in Table 2. Performance of the techniques is measured with standard evaluation metrices which includes accuracy, precision, recall and F1 score. For better insight of these metrices, confusion matrix is also given in same table.

Table 2 Standard Evaluation Metrics and Confusion Matrix Score (in Percent)

ML Classifier			**Standard Evaluation Metrics**			
Classifier	***Parameter***	***Value***	***Accuracy***	***Precision***	***Recall***	***F1 Score***
KNN	**K**	**7**	**93**	**92**	**91**	**91**
SVM	**Kernel**	**Rbf**	**90**	**90**	**87**	**89**
Decision Tree	**Criterion**	**Gini**	**88**	**88**	**85**	**86**
Gaussian Naïve Bayes	**--**	**--**	**81**	**78**	**81**	**79**

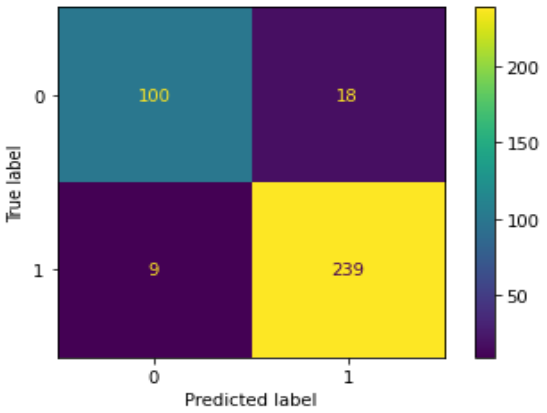

Figure 6

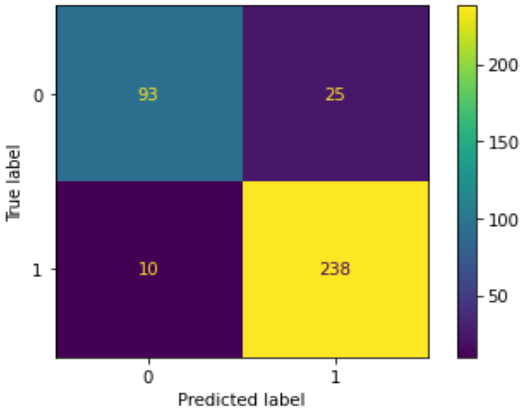

Figure 7

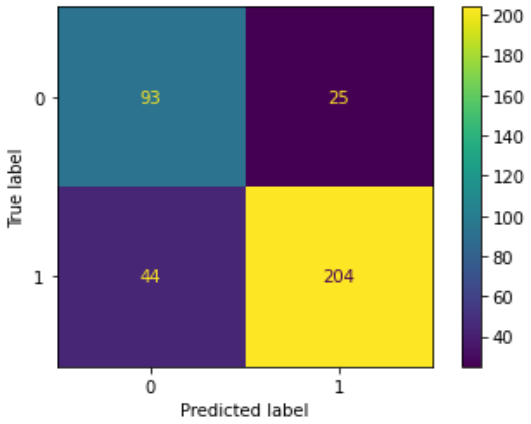

Figure 8

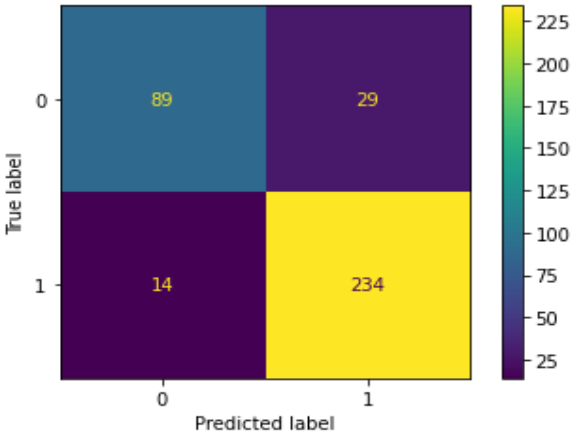

Figure 9

Confusion matrix for KNN (figure 6), SVM(figure7), GAUSIAN NAIVE BAYES (figure 8) and DECISION TREE(figure 9) classifier respectively.

The resulting decision tree of our project:

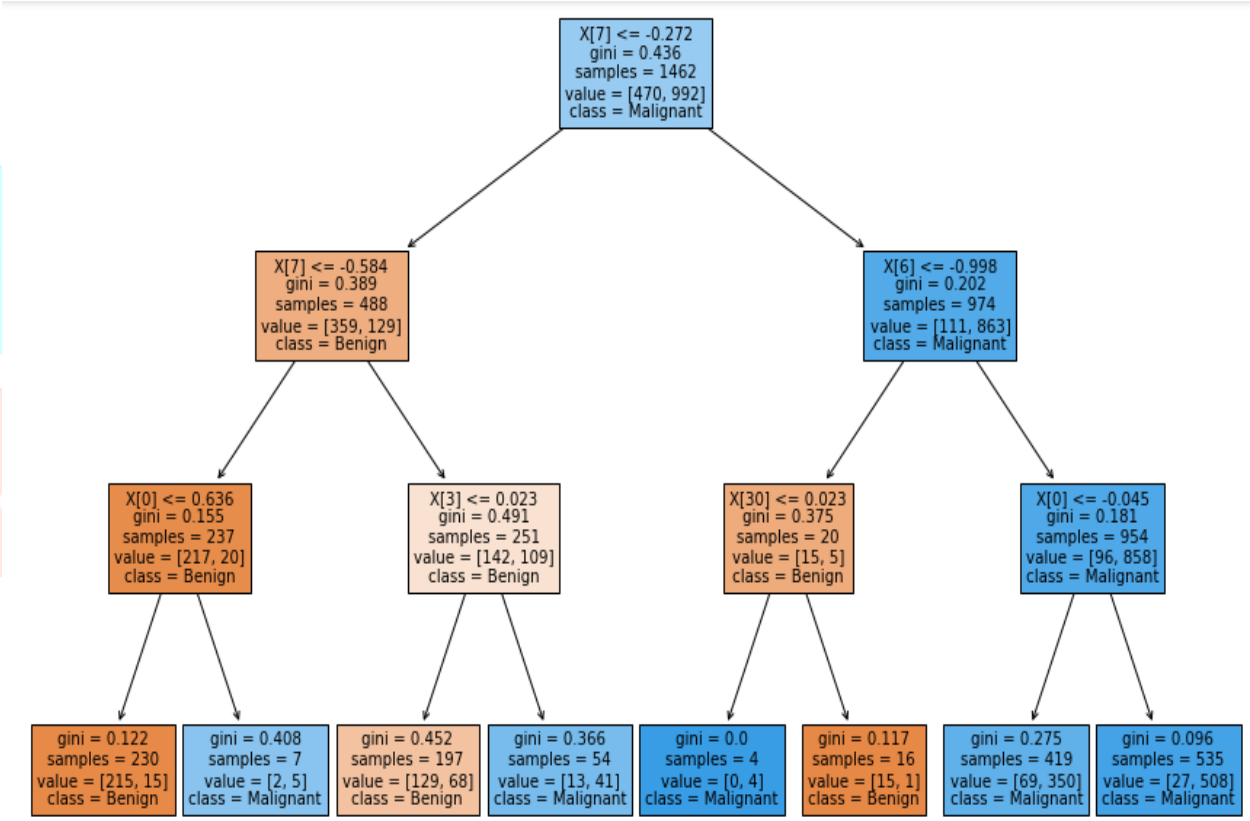

Figure 10

Machine learning techniques (KNN, SVM, Decision Tree and Naïve Bayes) have been trained and tested on scaled and normalised data. Each technique is trained with varied parameters. Best result of those parameter options is given in table 2. In scaled and normalised data, KNN results best at 93% for K=7 followed by SVM with 90% precision, recall and F1 score of KNN is better than all with more than 90% score followed by SVM, decision tree and then Gaussian naïve bayes. However, all results are good and are above 70%. Confusion matrices are also given where TN stands for true negative, FN for false negative, FP for false positive and TP for true positive. TN and TP are correct predictions while others are not.

5. CONCLUSION

This work is carried on to classify BreaKHis dataset by different supervised machine learning techniques. Scaling and normalising have improved results as it's expected. It gives a fair idea to the general reader about the performance of different techniques of machine learning. Current work may be extended to include unsupervised and reinforcement machine learning techniques, and deep learning as well. There are multiple standard datasets of HIs, which could be included in the future project apart from breast cancer HIs. This project uses only two features viz. colour and texture. For better analysis other features like morphological etc could be extracted.

6. REFERENCES

1. http://www.prevencaoediagnose.com.br
2. Colaboratory Release Notes - Colaboratory (google.com)
3. Breast Cancer Histopathological Database (BreaKHis)- Laboratório Visão Robótica e Imagem (ufpr.br)
4. Machine Learning Methods for Histopathological Image Analysis: A Review, J.D. Matos et al, arXiv:2102.03889v1 [cs.CV] 7 Feb 2021
5. https://medium.com/mlearning-ai/color-shape-and-texture-feature-extraction-using-opencv-cb1feb2dbd73
6. https://medium.com/analytics-vidhya/image-equalization-contrast-enhancing-in-python-82600d3b371c
7. Image Feature Extraction | Feature Extraction Using Python (analyticsvidhya.com)
8. Investigation on Feature Extraction and Classification of Medical Images, P. Gnanasekar et. al., World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering, Vol:5, No:12, 2011
9. Machine Learning for Medical Imaging , Erickson, Bradley J et al., Radiographics : a review publication of the Radiological Society of North America, Inc vol. 37,2 (2017): 505-515. doi:10.1148/rg.2017160130