

## A REVIEW ON FEATURE EXTRACTION TECHNIQUES FOR IMPROVED BIOPSY IMAGE CLASSIFICATION

Bindu S<sup>1</sup>, Devika G S<sup>2</sup>, Dr. Mohammad Raf<sup>3</sup>

<sup>1,2,3</sup>Computer Science & Engineering, University BDT College Of Engineering, Hadadi Road, Davanagere,  
Karnataka, India – 577004.

### ABSTRACT

In this paper, Improving biopsy image classification often involves employing advanced feature extraction techniques to extract relevant information from the images. These techniques help in capturing important patterns and structures that can aid in accurate classification. Here are some feature extraction techniques commonly used in medical image analysis, including biopsy image classification. Breast cancer (BC) classification has become a point of concern within the field of biomedical informatics in the health care sector in recent years.

This is because it is the second-largest cause of cancer-related fatalities among women. The medical field has attracted the attention of researchers in applying machine learning techniques to the detection, and monitoring of life-threatening diseases such as breast cancer (BC). Automatic detection of BC based on pathological images and the use of a Computer-Aided Diagnosis (CAD) system allow doctors to make a more reliable decision. we propose a colon biopsy image classification system called CBIC, which benefits from discriminatory capabilities of information rich hybrid feature spaces, and performance enhancement based on ensemble classification methodology.

In this work, handcrafted feature extraction techniques (Hu moment, Haralick textures, and color histogram) and Deep Neural Network (DNN) are employed for breast cancer multi-classification using histopathological images on the BreakHis dataset. The features extracted using the handcrafted techniques are used to train the DNN classifiers with four dense layers and Softmax.

Further, the data augmentation method was employed to address the issue of overfitting. The results obtained reveal that the use of handcrafted approach as feature extractors and DNN classifiers had a better performance in breast cancer multi classification than other approaches in the literature.

### 1. INTRODUCTION

Cancer detection has always been a major issue for the pathologists and medical practitioners for diagnosis and treatment planning. The manual identification of cancer from microscopic biopsy images is subjective in nature and may vary from expert to expert depending on their expertise and other factors which include lack of specific and accurate quantitative measures to classify the biopsy images as normal or cancerous one. The automated identification of cancerous cells from microscopic biopsy images helps in alleviating the abovementioned issues and provides better results if the biologically interpretable and clinically significant feature based approaches are used for the identification of disease. The chances of curing from cancer are primarily in its detection and diagnosis. The selection of the treatment of cancer totally depends on its level of malignancy. Medical professionals use several techniques for detection of cancer. These techniques may include various imaging modalities such as X-ray, Computer Tomography (CT) Scan, Positron Emission Tomography (PET), Ultrasound, and Magnetic Resonance Imaging (MRI) and pathological tests such as urine test and blood test. For the detection and diagnosis of cancer from microscopic biopsy images, the histopathologists normally look at the specific features in the cells and tissue structures. The various common features used for the detection and diagnosis of cancer from the microscopic biopsy images include shape and size of cells, shape and size of cell nuclei, and distribution of the cells. The brief descriptions of these features are given as follows.

(A) Shape and Size of the Cells. It has been observed that the overall shape and size of cells in the tissues are mostly normal. The cellular structures of the cancerous cells might be either larger or shorter than normal cells. The normal cells have even shapes and functionality. Cancer cells usually do not function in a useful way and their shapes are often not even.

(B) Size and Shape of the Cell's Nucleus. The shape and size of the nucleus of a cancer cell are often not normal. The nucleus is decentralized in the cancer cells. The image of the cell looks like an omelet, in which the central yolk is the nucleus and the surrounding white is the cytoplasm. The nuclei of cancer cells are larger than the normal cells and deviated from the centre of the mass. The nucleus of cancer cell is darker. The segmentation step mainly focuses on separation of regions of interests (cells) from background tissues as well as separation of nuclei from cytoplasm.

(C) Distribution of the Cells in Tissue. The function of each tissue depends on the distribution and arrangements of the normal cells. The numbers of healthy cells per unit area are less in the cancerous tissues. These adjectives of microscopic biopsy images have been included in shape and morphology based features, texture features, color based features, Color Gray Level Cooccurrence Matrix (GLCM), Law's Texture Energy (LTE), Tamura's features, and wavelet features which are more biologically interpretable and clinically significant.

The main aim of this paper is to design and develop a framework and a software tool for automated detection and classification of cancer from microscopic biopsy images using the abovementioned clinically significant and biologically interpretable features. This paper focuses on selecting an appropriate method for each design stage of the

framework after making a comparative analysis of the various commonly used methods in each category. The various stages involved in the proposed methodology include enhancement of microscopic images, segmentation of background cells, features extraction, and finally the classification.

## 2. RELATED WORKS

The investigation and analysis of the breast cancer histopathological images by specialists is a delicate and rigorous procedure demanding time and great requirement. Classification of breast cancer method was proposed. The diagnoses of the histopathological images using computer-aided tools require utilization of machine learning techniques. In the past, classification of such images would require feature engineering techniques to extract features that were supplied to a classical machine learning classifier. By Vijayarajeswari et al. (2019) using Hough transform and support vector machine (SVM).. An incremental boosting strategy was also proposed to achieve a better result by combining weak classifiers with strong ones. Most suggested classification systems in the state-of-the-artwork use “texture” as a feature of the images, which is based on a variety of feature descriptions such as gray-level co-occurrence matrix (GLCM), histogram of gradient (HOG), and local binary pattern (LBP) (Alhindi et al., 2018), and graph run-length matrix (GRLM) (Belsareetal., 2016). Local phase quantization (LPQ) methods outperformed scale-invariant feature transform (SIFT) and speeded-up resilient features (SURF) features due to their good efficiency, reliability to noise, and low computing power environments (Rublee et al., 2011). Zhangetal. (2014) introduced the KPCA model as a single-class kernel theory component analysis model.

## 3. DNN MODEL

Deep learning approaches are very powerful because they have a mechanism that allows attributes to be extracted from data without preprocessing. Classical methods such as machine learning are tedious in tasks like pre-processing, feature extraction, and segmentation. These tasks degrade the efficiency and accuracy performance of the system (Khan et al., 2019). However, deep learning models yield more accurate results automatically. The working mechanism of a human brain is imitated by profound learning strategies (Bengio, 2012; LeCun et al., 2015). The DNN is made up of many simple structures which form a stack. Almost everybody undertakes non-linear operations in this basic framework, which changes the scale of the data in another area and helps to expose hidden features in the data (Bengio, 2012; LeCunetal., 2015). However, in this study dense layer is used because of the hardware requirement and time-consuming nature of CNN models like VGG16, GoogleNet. The proposed model used handcrafted features and dense layers because of its strong gradient flow, computational efficiency, and maintains low complexity features.

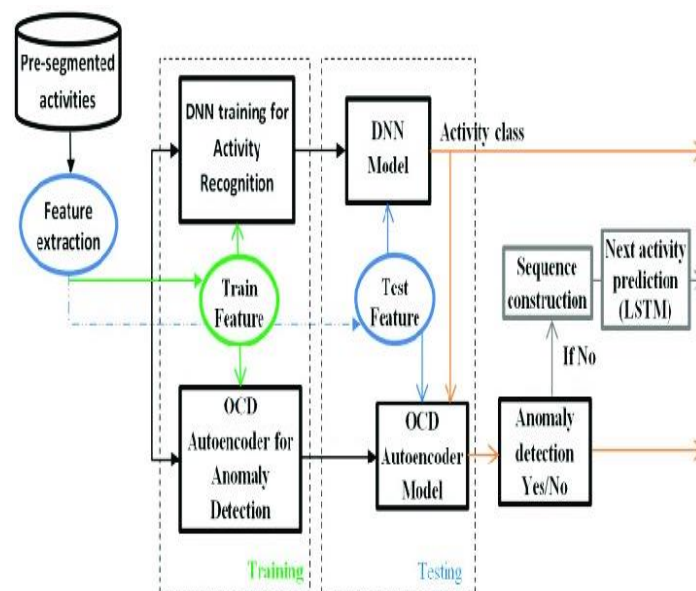


Fig 3.1 DNN Model Architecture

## 4. PROPOSED WORK

The proposed model for multi-classification of breast cancer into either benign (B) or malignant (M) cells in breast histopathological images are explained, which is based on Handcrafted features and the DNN classification model. Different levels of features are extracted separately in the proposed model using three wellknown techniques for extraction of color, shape, and texture features from images: Colored Histogram, Hu-Moment, and Haralick. Fig. 4.1 shows a block diagram of the proposed work. The paragraphs that follow provide more information on each phase of the proposed work.

### 4.1. Handcrafted feature

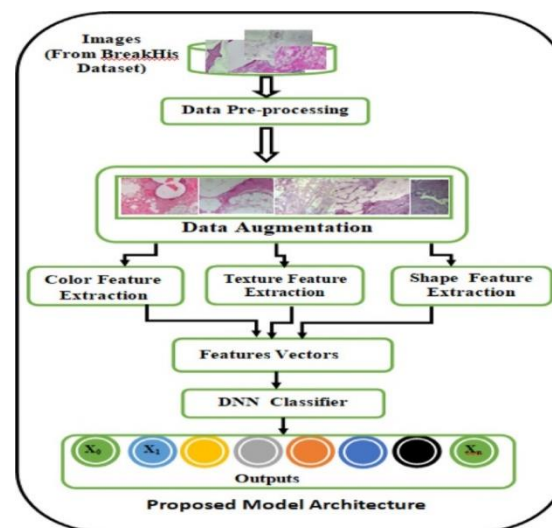
Handcrafted features are attributes that are derived from existing information in a histopathological image using a computational technique. The classification system using handcrafted features is typically divided into two stages: the

feature extraction phase, in which the three most important properties color, shape, and texture are considered to measure the breast cancer histopathological images, and the classification phase, in which we considered the use of Deep Neural Network as the classifier in this study.

#### 4.1.1. Colored histogram (color)

In digital pathology, before digitization of tissue samples, the staining procedure provides a thorough insight into the structures (nuclei, stroma, and cytoplasm) of the tissue. Blue or violet nuclei with hematoxylin and other pink structures with eosin were coloured in a normal dyeing procedure in which hematoxyline and eosin (H&E) colouring are used (Nahid & Kong, 2018). As a result, color is an important factor to consider when classifying histopathological images. A color histogram is a technique for determining the distribution of colours in an image (Nahid et al., 2018). The color histogram reflects the amount of pixel value within every bin that is the same color for the color spectrum list. The additive frequencies of the occurrence of each color give the feature vector. For example, red, green, and blue are considered coloured images. The length of vector features obtained will be  $8 \times 8 \times 8 = 512$ , if, for example, we used a histogram of 8 bins for each of the channels. Owing to the presence of inconsistency, the feature obtains using color histogram only is not sufficient to measure the breast cancer histopathological images. Variations in fixation and staining protocols being used in various laboratories, differences in light sources or detectors used in the scanner, use of different reagents, long waits in fixation, and lack of consistency in dyeing conditions are the causes of inconsistency (Veta et al., 2014). So, to develop a reliable classification model, other attributes need to be looked into in addition to the color histogram.

4.1 shows a block diagram of the proposed work. The paragraphs that follow provide more information on each phase of the proposed work.



#### Haralick textures (texture)

The texture of an image is a quality that describes the surface and appearance of the object in the image.. In the Haralick Texture Framework, the Gray Level Co-occurrence matrix is calculated by using grayscale pixel I and j, which are expressed as the number of co-occurrence matrix in different directions as represented below

$$p(i, j | d, \theta) = p(i, j | d, \theta) \div (\sum_i \sum_j p(i, j | d, \theta))$$

where  $p(i, j)$  is the matrix of relative frequencies,  $d$  is the pixel distance that occurs within  $(x1, y1)$  and  $\theta$  is the direction of pixel  $(x1, y1)$ . In  $p(i, j)$  gives the statistical probability values for changes between gray-levels  $i$  and  $j$  at a given distance  $d$  and angle.

A set of textural features are extracted from the gray-tone spatial dependence such as Contrast, Entropy, Correlation, and Homogeneity Energy.

**Contrast:** Measure pixel-neighbor intensity throughout the full image is deemed as zero for the constant image as well as variance and moment of inertia.

$$\text{Contrast} = \sum_{i,j} (i - j)^2 p(i, j)$$

**Correlation:** measures how the pixel is correlated to its neighbor over the entire image

$$\text{Correlation} = \sum_{i,j} (i - \mu_i)(j - \mu_j)p(i, j) \div (\delta_i \delta_j)$$

where  $\delta_i$ ,  $\delta_j$  and  $\mu_i$ ,  $\mu_j$  are the standard deviations and means of  $p_i$ , and  $p_j$ .

**Entropy:** gives the intricacy and complexity of the image metrics tends to increase entropy.

$$\text{Entropy} = - \sum_{i,j} p(i, j) \log p(i, j)$$

Energy is the sum of squared elements in the GLCM and it is a permanent image by default.

$$\text{Energy} = \sum_{i,j} p(i, j)^2$$

## 5. CBIC SYSTEM

In this paper, we propose a colon biopsy image classification (CBIC) system, which performs ensemble classification of samples based on discriminatory capabilities of hybrid feature spaces. In order to exploit the color information present in colon biopsy images, variants of traditional statistical moments and Haralick features have been proposed. Further, traditional histogram of oriented gradients (HOG) based features have been used. These features have been combined to form various hybrid feature sets. The minimum Redundancy Maximum Relevance (mRMR) method has been employed to select discerning feature sets from individual as well as hybrid feature sets. Samples are then classified into normal and malignant classes by employing ensemble classification through majority voting. The experimental results in this work have been obtained from various aspects. First, the performance of individual as well as hybrid feature types has been investigated. Second, the performance of original feature sets and the feature sets selected by mRMR method has been examined. Third, the performance of individual as well as ensemble classifier has been studied. The experimental results verify that the proposed system is quite suitable for the classification of colon biopsy images. Further, an analysis on computational efficiency of feature extraction and classification stages has been presented in order to validate the suitability of the proposed CBIC system to serve in real-time scenarios where histopathologists receive many images per day.

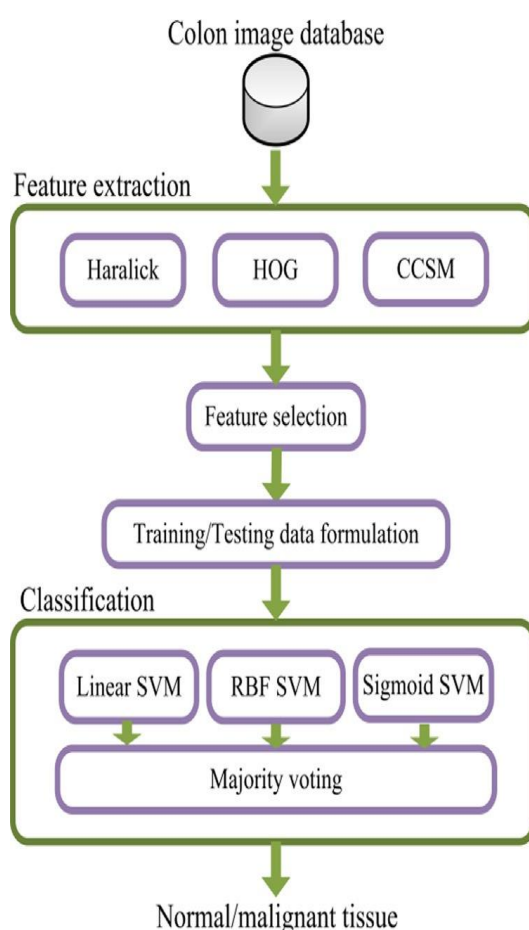


Fig.5.1 Top-level architecture of the proposed CBIC system.

## 6. CONCLUSION

In this paper, we develop a new publicly available Colorectal Histopathological Image Dataset called EBHI with four magnifications of 5532 images: 40x, 100x, 200x and 400x, and five types images of tumor differentiation stages. EBHI has the function to test whether the image classifier has high classification accuracy and good robustness. For classical machine learning methods, this paper focuses on the accuracy differences among classifiers, who perform at best at only 76%, but at worst below 50%. For the deep learning methods, all four models perform excellent classification results, with the highest accuracy rate reaching over 95%. This paper focuses on the analysis of the four models in terms of accuracy, model size, training time, and other metrics.

In this paper, a learning strategy with different magnification levels was used to compare handcrafted features and a DNN classifier for multi-classification of BC histopathology images. Throughout our research, we discovered that the handcrafted approaches as feature extractors perform very well in comparison to other techniques. For all magnification levels (40x, 100x, 200x, and 400x), the combination of handcrafted features with the DNN classifier offers the best result. Due to this fact (Haralick texture, Hu moment, and colored histogram with DNN classifier) considered the current setup to be more stable and stronger. Additionally, data augmentation methods are often used to further improve the precise classification by proper adjustment of the parameters. The influence of magnification on



classification accuracy is depending on the amount of complexity of histopathological pictures, which increases as magnification levels increase. Finally, the model efficiency is measured individually and comparably with other existing methods. The proposed model has been observed to provide excellent accuracy results

In this research study, a classification system (CBIC) has been proposed for predicting cancer in colon tissues. In the proposed system, hybrid feature set comprising CCSM, Haralick-HSV, and HOG is constructed. The mRMR method is employed to select discerning features from the hybrid feature set. The discerning features are then used in different SVM kernels based ensemble classification. Working with colon biopsy images, highest classification accuracy of 98.85% and 96.68% has been observed with hybrid and individual feature set (CCSM), respectively. Results prove that the proposed variants of traditional Haralick features and statistical moments are promising feature types for classification of colon biopsy images.

## 7. REFERENCES

- [1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: Cancer J Clin 2021;1–41. <http://dx.doi.org/10.3322/caac.21660>.
- [2] Pamudurthy V, Lodhia N, Konda VJ. Advances in endoscopy for colorectal polyp detection and classification. In: Baylor university medical center proceedings, vol. 33, no. 1. Taylor & Francis; 2020, p. 28–35.
- [3] Liu W, Li C, Xu N, Jiang T, Rahaman MM, Sun H, et al. CVM-Cervix: A hybrid cervical pap-smear image classification framework using CNN, visual transformer and multilayer perceptron. Pattern Recognit 2022;108829.
- [4] Abdel-Zaher, A. M., & Eldeib, A. M. (2016). Breast cancer classification using deep belief networks. Expert Systems with Applications, 46(November), 139–144. <https://doi.org/10.1016/j.eswa.2015.10.015>.
- [5] Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. Expert Systems with Applications, 36(2 PART 2), 3240– 3247. <https://doi.org/10.1016/j.eswa.2008.01.009>.
- [6] Cancer Facts and Figures, American Cancer Society, (<http://www.cancer.org/research/cancerfactsstatistics>), October 2013.
- [7] Colon Cancer Risk Factors, C.C. Alliance, ([http://www.ccalliance.org/colorectal\\_cancer/riskfactors.html](http://www.ccalliance.org/colorectal_cancer/riskfactors.html)), October 2013.
- [8] D.Myers, Colon Cancer Stages: Basics of Each Colon Cancer Stage, (<http://coloncancer.about.com/od/stagesandsurvivalrate1/a/ColonCancerStag.htm>), October 2013.