Signal, Image and Video Processing (2020) 14:665–673
https://doi.org/10.1007/s11760-019-01584-4

ORIGINAL PAPER

Microscopic images classification for cancer diagnosis

Yashwant Kurmi[1] · Vijayshri Chaurasia[1] · Narayanan Ganesh[2] · Abhimanyu Kesharwani[3]

Received: 20 June 2019 / Revised: 6 September 2019 / Accepted: 10 October 2019 / Published online: 16 November 2019
© Springer-Verlag London Ltd., part of Springer Nature 2019

Abstract
Computer aided diagnosis of cancer is a field of substantial worth in current scenario since approximately 38% population of the world is suffering from the disease. The detection of cancer is based on the observation of deformation in nuclei structure using histopathology slides/images. The proposed technique utilizes nuclei localization prior to classification of histopathology images as benign and malignant. The features used for classification are an ensemble of 150 bag of visual word features, extracted from preprocessed image and 20 handcrafted features, extracted from the internal parts of nuclei using localized histopathology images. The simulation results confirm the superiority of proposed localization based cancer classification method as compared to existing methods of the domain. It has reported average classification accuracy of 95.03% on BreakHis dataset.

Keywords Medical imaging · Histopathology · Histopathology image · Feature extraction · Image classification

1 Introduction

This microscopic image analysis in the medical field is a very helpful and emerging field. The hematoxylin and eosin (H&E)-stained histopathology images (HIs) are segmented to get the specific region characteristics for the detailed analysis. The statistical features are utilized in image segmentation to mark the object of interest [1–3]. Wang et al. presented a multi-scale region growing and curvature scaling for automatic breast cell nuclei segmentation and classification (ANSC) [2]. Wang in [4], has proposed a semi-automatic method (SAM) for cell segmentation. Various statistical features have been studied to enhance different regions of interest [4–6]. The HI segmentation finds applications in identification of diverse objects like tissue, gland, etc. [7–9]. Vu et al. [8] proposed class specific features learning based technique to separate the interclass difference named as discriminative feature-oriented dictionary learning method (DFDLM). Naylor et al. [9] presented a nuclei segmentation using deep regression (NSDR) approach in order to target the touching nuclei regions. In most of the HI segmentation methods, firstly the basic marking is performed, followed by stain decomposition [10] as per the dataset requisites. The Laplacian-of-Gaussian filtering and the Gaussian mixture model (GMM)-based pixel clustering have been investigated for seed point extraction for nuclei segmentation in [11] and [12], respectively.

Many researchers have been investigating the use of localization using segmentation as preprocessing of feature extraction. Support vector machine (SVM) and convolutional neural network based classifiers are the classifiers based on hyperplane selection [13–17]. Yan et al. [18] presented a hybrid technique for breast cancer HI classification using convolutional and recurrent deep neural network (CRDNN). Yang et al. [19] worked on the feature selection for high-dimensional data mining using the nearest neighbor-based feature weighting. Klein et al. [20] developed a fast Bayesian optimization technique of machine learning hyperparameters on large datasets. Most of the above discussed methods faces problem in extraction of features due to the overlapped nuclei which in turn leads to the reduction in the dependability of classification.

This paper suggests a classification of HIs for cancer detection using nuclei localization as a preprocessing step. A significant number of relevant features have been extracted using a combination of bag of visual words and the handcrafted features form segmented HI. The significance

✉ Yashwant Kurmi
yashwantkurmi18@gmail.com

[1] Maulana Azad National Institute of Technology, Bhopal 462003, India

[2] Jawaharlal Nehru Cancer Hospital and Research Center, Bhopal, India

[3] All India Institute of Medical Sciences Bhopal, Bhopal, India

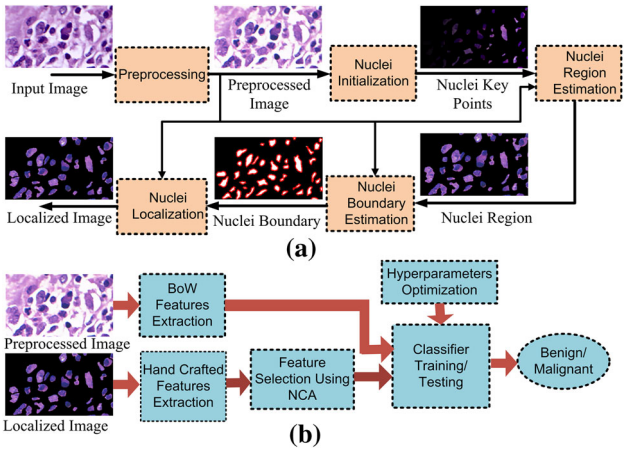

Fig. 1 Block diagram of the proposed algorithm: **a** nuclei localization in HI and **b** HI classification in benign and malignant

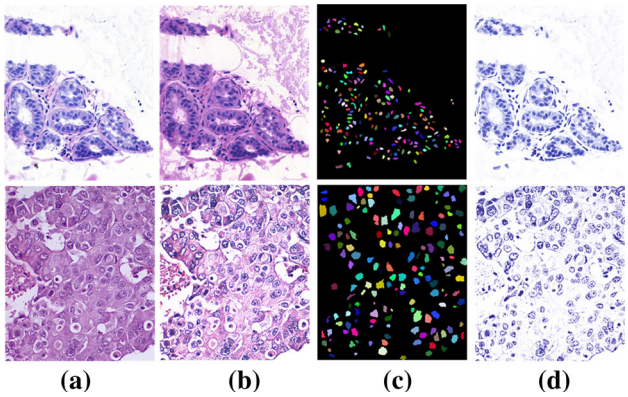

Fig. 2 Hematoxylin & Eosin stained images illustration. **a** original image, **b** preprocessed image, **c** ground truth and **d** the decomposed hematoxylin stain component for benign (upper) and malignant (lower) images

of handcrafted features has been tested using neighborhood component analysis (NCA) [19]. Further, the SVM [15] and multilayer perceptron (MLP) [16] classifiers have been applied along with optimized hyperparameters using Bayesian optimization [20] for the benign and malignant HI classification.

In remaining manuscript, the second section presents nuclei localization method, Sect. 3 explains the selection of appropriate features, and details of classification models are given in Sect. 4. Experimental setup is presented in Sect. 5, followed by result analysis in Sect. 6. Finally, Sect. 7 concludes the research findings.

2 Localization method

The proposed method is presented through the block diagram in Fig. 1a shows the localization part and Fig. 1b shows the classification part. In localization, input HI (f) is preprocessed followed by the identification of nuclei region and nuclei boundaries. The final outcome of the described HI processing provides complete nuclei segmented (localized) image (f_L).

Fig. 1b is presenting the combination of proposed feature extraction and classification. Details of proposed feature extraction and classification are explained in Sect. 3 and 4, respectively.

For preprocessing, firstly the stain decomposition is applied on f. It focuses on stains co-occurrence in association with the circular mixture model and soft-clustering of pixels [10]. The pixel level clustering is done through periodicity of hue signals on the unit for decomposition. It results a preprocessed image f_{p}. The nucleus contains euchromatic (active region comparatively brighter than other region) and heterochromatic (inactive region comparatively darker than other region). The nucleus is always a darker region and called as key points. Figure 2 illustrates the benign and malignant images with their ground truth and stain decomposed counterparts.

2.1 Nuclei initialization

The nuclei initialization is performed on preprocessed HI to initiate the segmentation. Firstly, f_{p} is converted to respective grayscale image f_{g} in the range $\{0, 1\}$ and further enhanced using normalization factor α in order to transform the range from [0.15–0.4] to [0–0.4]. The pixel values of enhanced grayscale HI (f_{r}) are shown in Eq. 1

$$f_{\mathrm{r}}(x,y) = \begin{cases} 0; & \text{if } f_{\mathrm{g}}(x,y) \leq 0.15 \\ \alpha \times f_{\mathrm{g}}(x,y); & \text{if } 0.15 < f_{\mathrm{g}}(x,y) \leq 0.4 \\ f_{\mathrm{g}}(x,y); & \text{else} \end{cases}. \quad (1)$$

The difference of Gaussian (DoG) is applied on f_{r} and returns f_{DoG}. Similarly, Hessian of Laplacian of Gaussian (HLoG) operators is applied on f_r and produces f_{HLoG}. All three images (f_{DoG}, f_{HLoG} and f_{r}) are segmented through Otsu thresholding [21], and three segmented images are combined to identify the nuclei key points. The nuclei region is processed with morphological erosion and overlapped nuclei are separated by considering nuclei radius $r \leq R$ as constraints to get the nuclei seeds. The large regions are taken as multiple nuclei, while considering the nuclei shape and size. The ultimate result of this step is nuclei center marked image (f_{Mark}).

2.2 The nuclei region estimation

The nuclei regions are identified through application of the normalized graph cut method [22] on f_{p} followed by the application of key point-contour link creation algorithm [5]

by considering marked key points of f_{Mark}. It connects the key points with the outcome of the normalized graph cut method. The link length is chosen as 3–7 pixels based on the size of nuclei. Image f_{p} is represented as a graph $G = (V, E)$, where V defines set of vertices $\{\text{v}_1, \text{v}_2, \ldots\}$ and E defines set of edges $\{\varepsilon_1, \varepsilon_2, \ldots\}$. The links are categorized in two sections: object O and background B.

$$N_{\text{cut}}(O, B) = \frac{\text{cut}(O, B)}{\text{assoc}(O, V)} + \frac{\text{cut}(B, O)}{\text{assoc}(B, V)} \tag{2}$$

where

$$\text{cut}(O, B) = \sum_{u \in O, v \in B} w(u, v)$$

$$\text{assoc}(O, V) = \sum_{u \in O, v \in V} w(u, v) \tag{3}$$

$$\text{Here } \text{w}_{\text{ij}} = \begin{cases} 1 \text{ if } \text{v}_i \text{v}_{\text{j}} \in \varepsilon, \forall(\text{v}_i, \text{v}_{\text{j}}) \in V \\ 0 \qquad \text{otherwise} \end{cases} \tag{4}$$

The normalized graph cut method is applied in a recursive manner and separates strong and weak links. Strong links signify the nuclei connects and other objects are signified by weak links. The overall shape of the nuclei in HI is defined by strong links and images containing the extracted nuclei region is defined as f_{RE}. The image f_{RE} may suffer from the boundary region problem. In the proposed method, the solution of this problem is addressed by nuclei boundary estimation.

2.3 Nuclei boundary estimation

The nuclei boundary estimation corresponding to estimated nuclei region points is graphically illustrated in Fig. 3a, b. The contours are extracted by nuclei edges with an optimum boundary estimation as displayed at bottom in Fig. 3b. Nuclei boundary extraction by the combination of receptive field (CORF) model is based on the edge detection which are extracted unit wise. The combination of small edge sections is taken as the receptive field (RF) unit. The response R_{S} of a CORF operator is defined as the weighted geometric mean of the responses of all edges sections, for more detail please refer [11].

The segmentation outcome has boundaries that are not aligned to each other are removed using the nucleus center to boundary association. CORF followed by modified gredient at discontinuity [23] results a clear demarcation of nuclei and indicates the nuclei boundary in HI and returns f_{BE}. Further, the complete nuclei localization is furnished by transforming the inside region of estimated boundary by 1 s and the rest of the area of image f_{BE} as 0 s which results as compete nuclei mask f_{Mask}. The application of f_{Mask} on f_{p} produces final localized image f_{L}.

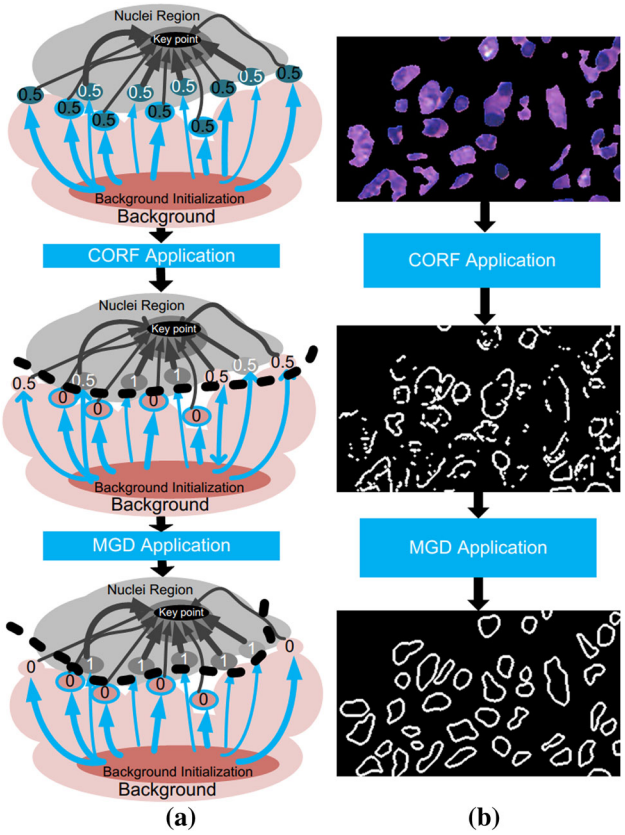

Fig. 3 Nuclei boundary estimation by combination of receptive field (CORF) and improved by modified gradient at discontinuity (MGD), **a** working principle of the CORF model and MGD model, **b** the visual illustration of nuclei boundary estimation by CORF model and boundary refinement by MGD model

3 Feature extraction and selection

Appropriate class prediction is the prime focus of the proposed HI analysis. Basically, a classifier needs a set of features to classify the data into their suitable classes.

3.1 Bag of visual words

A set of hundred and fifty shape features is extracted using BoW model from f_{p} [24]. The codebook containing a certain number of code words (or visual words) is constructed with their local descriptors or features.

3.2 Handcrafted features

Here, the handcrafted (HC) features based on the internal structure of the nuclei are proposed. The f_{L} is utilized for the extraction of HC features. The f_{L} is separated into two regions: heterochromatic region (HCR) [25] and euchromatic region (ECR) [25], through the application of stain color differentiation. A modified threshold is utilized to separate the

HCR and ECR of the nuclei which is twice of the Otsu threshold. The nuclei component above the threshold value is the ECR and the rest is HCR. As per the histopathology analysis, HIs have the ECR and HCR constituents about to be equivocal for benign tumor and for malignant tumors the HCR is dominating as that of ECR in all the tissue structures. The size of the nuclei increases in the presence of the tumor along with the increases in shape irregularity and heterochromaticity. The set of 31 HC features (F_1–F_{31}) is defined in Table 1.

3.3 Neighborhood components analysis

Neighborhood components analysis (NCA) is a supervised learning method for classifying multivariate data into distinct classes according to the significance parameter in data [19].

To understand NCA, let us consider a set of N number of training samples $T = \{(x_1, y_1), \ldots, (x_i, y_i), \ldots (x_N, y_N)\}$. where vector x is d-dimensional feature space with class label y. The weighting vector w is determined in such a way to select a feature subset by optimizing the nearest neighbor classification. The weighted distance D_{W} between two samples x_i and x_j in terms of the weighting vector w is given as:

$$D_{\mathrm{w}}(x_i, x_j) = \sum_{l=1}^{d} w_l^2 \left| x_{il} - x_{jl} \right| \tag{5}$$

where w_l represents the associated weight of lth feature. The probability distribution based effective approximation of reference point is determined first using 1-nearest neighbor classification by maximizing its leave-one-out classification accuracy on the training set T. The related probability of x_i to pick x_j as its reference point is given as:

$$p_{ij} = \begin{cases} \frac{\kappa(D_w(x_i, x_j))}{\sum_{k \neq i} \kappa(D_w(x_i, x_k))} & \text{if } i \neq j \\ 0, & \text{if } i = j \end{cases} \tag{6}$$

where the kernel $\kappa(z) = \exp(-z/\sigma)$ is used with a kernel width σ. The kernel width is taken as input variable that plays a key role in deciding the reference point. Thus, the probability of correct classification of the query point x_i is given by:

$$p_i = \sum\nolimits_j y_{i\,j}\, p_{i\,j} \tag{7}$$

where $y_{ij} = 1$ only for $y_i = y_j$ and $y_{ij} = 0$ else. As a result, the leave-one-out classification accuracy can be approximated as:

$$\rho(w) = \frac{1}{N} \sum_i p_i = \frac{1}{N} \sum_i \sum_j y_{i\,j}\, p_{i\,j}. \tag{8}$$

Table 1 Features description of nuclei and image level

F_1—Average nuclei size: The nuclei size will be determined based on the average number of pixels
F_2—Nuclei count: Number of nuclei present in the image
F_3—Perimeter: The nuclei perimeter is calculated through which we determine the irregular border
F_4—Irregular border: The nuclei shape is matched with a circular cum elliptical shaped nucleus and provided a score in between 0 and 1. For complete match, it is 1 for no match, it is 0. The average value is taken for the whole image
F_5—Streaks: The streaks are counted as darker region appear due inappropriate staining and not the part of nuclei
F_6—Max-diameter: Average of major axis of elliptical shape
F_7—Projections: Represents the major axis projection with respect to minimum axis.
F_8—Min-diameter: Average of minor axis of elliptical shape
F_9—Aspect ratio: It is equal to the difference between maximum column value and the minimum column value of the nuclei divided by the difference between the maximum row value and minimum row value of the nuclei
F_{10}—Entropy1: Average nuclei entropy
F_{11}—Entropy2: Image entropy
F_{12}—Number of colors: The dominated color components in the nucleus region 1, 2 or 3 corresponding to R, G, or B
$F_{13–14}$—Color ratios-1&2: The ratio of the dominant color components to non-dominant color component
F_{15}—Euler number: The parameter defines the number of nuclei minus the number of holes in the image
F_{16}—Thinness ratio: It is the ratio of the area and the perimeter of the object
F_{17}—Crowdedness or nuclei density: Median of the distances between nuclei centroids from each individual to all others
F_{18}—Nuclei mean intensity: It is clear from its name itself, the individual nuclei intensity is calculated and the median is taken from that
F_{19}—Image mean intensity: The mean image intensity of the original image
$F_{20–21}$—Heterochromatic area and Euchromatic area: The internal part of the nucleus
F_{22}—Center of area of heterochromatic area
F_{23}—Axis of least second moment of heterochromatic area
F_{24}—Center of area of euchromatic area
F_{25}—Axis of least second moment of euchromatic area
F_{26}—Total nuclei area: It is the area taken of nuclei regions
F_{27}—Center of area of nuclei regions
F_{28}—Axis of least second moment of nuclei regions
F_{29}—HTE ratio: It is the ratio of the heterochromatic and euchromatic area
F_{30}—HTT ratio: It is the ratio of the heterochromatic and the total area
F_{31}—ETT ratio: It is the ratio of the Euchromatic and the total area

As σ tends to zero, $\rho(w)$ becomes the true leave-one-out classification accuracy. A regularization term ($\lambda > 0$) is further introduced to perform feature selection and alleviate overfitting, hence the object function modified as:

$$\rho(w) = \frac{1}{N}\sum_{i}\sum_{j} y_{i\,j}\, p_{i\,j} - \lambda \sum_{l=1}^{d} w_l^2 \tag{9}$$

The regularization parameter is tuned using cross validation. To update weights, the object function with regularization $\rho(w)$ is differentiated with respect to w_l as follows:

$$\frac{\partial \rho(w)}{\partial w_l} = \sum_{i}\sum_{j} y_{i\,j}\left[\frac{2}{\sigma}\, p_{i\,j}\left(\sum_{k\neq i} p_{i\,k}\,|x_{il} - x_{kl}| - |x_{il} - x_{kl}|\right) w_l\right] - 2\lambda w_l \tag{10}$$

Using the above derivative that leads to the corresponding gradient ascent update equation, features optimization is performed on 31 extracted HC features. The NCA is applied here to reduce redundant features. Which results that one-third of the features are not carrying useful information and removed. This optimization provides 20 significant HC features; hence, further processing is performed using significant HC features.

4 Classification models

The block diagram of the general classification model with features extraction and selection, along with model hyperparameter optimization is shown in Fig. 1b. In the proposed computer aided cancer diagnosis method, two classifiers SVM and MLP model are employed for classification.

4.1 Support vector machine

The SVM [15] classification model provides high flexibility to classify distinct classes. The nonlinearity can be introduced in SVM using a soft margin parameter C. The formulations of soft margin linear SVM are given as:

$$\text{Minimize}\left[\frac{1}{2}\sum_{i=1}^{n} w_i^2 + C\sum_{i=1}^{N}\xi_i\right] \tag{11}$$

subjected to $y_i(\vec{w}.\vec{x} + b) \geq 1 - \xi_i$ for $i = 1, \ldots, N$.

The additional separation distance can be introduced by nonlinear projection in the high-dimensions using Gaussian kernel, defined as:

$$K\left(\vec{x}_i, \vec{x}_j\right) = \exp\left(-\gamma \left\|\vec{x}_i - \vec{x}_j\right\|^2\right) \tag{12}$$

The SVM model's training is performed using k-fold cross validation. The training and testing process are repeated k times, by tracking the performance of the model in predicting the holdout set using a performance metric such as accuracy. The tenfold cross validation with $C = 1$, and $\gamma = 1$ is used for SVM model training and testing. A set of 40 objective evolutions provide best feasible point box constraint 0.0012 and kernel scale 0.0192.

4.2 Multilayer perceptron model

The neural network based classifier performance depends on the model selected and appropriate training of the model. We have trained a MLP [16] for the binary classification with some nonlinearity, described for input feature vector x as:

$$O^0 = x, \quad O^l = F^l\left(W^l \hat{o}^{(l-1)}\right) \quad \text{for} \quad l = 1, \ldots, L. \tag{13}$$

The input vector x is taken as "output of the zeroth layer". A hat notation $\hat{o}^{(l-1)}$ represents an operation where a number 1. prepended to a vector to increase its dimension. Hence, the bias terms of the layer l can be written as the first column of matrix W^l. The notation F^l represents the application of an activation function (sigmoid) on all components of a vector. The softmax function is also used as the activation function in the output layer of the five layered (Input + 3hidden + output) MLP model. Number of neurons in 3 hidden layers are 6, 10, and 8, respectively. The MLP feed forward fully connected model is implemented with a sigmoid activation. The model is trained using backpropogation with learning rate $\eta = 0.12$ by considering input features as training parameters and image labels as target variable. For the parameters update, the gradient is computed using the stochastic gradient descent algorithm, termed as weight error. The parameters are updated in such a way as they move the MLP, one step closer to the error minimum. We have taken the batch size of 5 samples and 1000 epochs for optimum result.

4.3 Hyperparameter tuning framework

Further, hyperparameter tuning is done with the objective to maximize the validation accuracy as:

$$\mathrm{X}* = \arg\max_{\mathrm{X}\in\chi} f(\mathrm{X}) \tag{14}$$

where $\mathrm{X} \subseteq R^D$ and f(x) represent the model performance on validation data for a set of hyperparameters X. Let the hyperparameters search space is bounded between l and u are the D-dimensional vectors denoting the lower and upper ranges, respectively. The ultimate goal is to optimize the hyperparameters on whole training data. We start by taking a small subset of the training data to identify the optimal hyperparam-

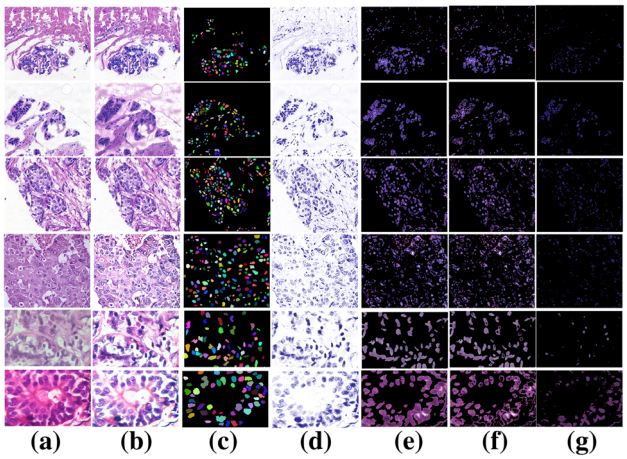

Fig. 4 **a** Original image, **b** preprocessed image, **c** ground truth, **d** the hematoxylin component of stain decomposed image, **e** nuclei segmented image, **f** heterochromatic component, **g** euchromatic component of the nuclei

eters using Bayesian optimization [20], which is repeatedly applied to a number of different smaller subsets. The mean of the optimal hyperparameters is determined to find the robust estimate. The parameters C, γ and kernel are optimized for SVM and the hidden layers, activation, solver and learning rate are optimized for MLP classifier.

5 Experimental setup

The experimental setup covers the datasets used in the localization and classification and the evaluation parameters calculated for performance analysis. The dataset name Bisque is an acronym for Bio-Image Semantic Query User Environment. Which provides a cloud based system to store, organize, visualize and analyze the various dataset images, and breast cancer (BC) dataset is one of the collections of Bisque. The Bisque dataset contains 32 benign and 26 malignant cases [26]. The dataset Breast Cancer Histopathological Image Classification (BreakHis) contains 9109 microscopic images (from 82 patients) of breast tumor tissue using 40×, 100×, 200× and 400× magnifying factors [7].

5.1 Proposed multi-organ dataset

A set of microscopic images from multiple organs is prepared and analyzed with 10 cases for each. The images are taken at different magnification as 40×, 100×, 400× and 1000× . The images at 100× are utilized for the gross analysis of cancer detection based on localized nuclei characteristics and 1000× are utilized for analysis of nuclei localization along with ECR and HCR segmentation.

Table 2 Average FS, JI and HD for Bisque and BreakHis

Parameters	Jaccard Index			Hausdorff Distance			F_1-Score			Accuracy			Complexity
Datasets/methods	Bisque	BreakHis	Avg	Bisque	BreakHis	Avg	Bisque	BreakHis	Avg	Bisque	BreakHis	Avg	
ANSC [2]	0.389	0.391	0.39	9.51	9.46	9.49	0.702	0.71	0.706	0.784	0.793	0.7885	$O(N^3)$
SAM [3]	0.451	0.456	0.4535	12.02	11.97	12.00	0.673	0.707	0.69	0.791	0.755	0.773	$O(N^3)$
DFDLM [8]	0.508	0.499	0.5035	11.31	11.39	11.35	0.708	0.699	0.7035	0.838	0.826	0.832	$O(N^4)$
NSDR [9]	0.671	0.589	0.63	8.28	8.33	8.30	0.696	0.727	0.7115	0.821	0.819	0.82	$O(N^3)$
Proposed	***0.704***	***0.721***	***0.7125***	***7.00***	***6.96***	***6.98***	***0.861***	***0.874***	***0.8675***	***0.914***	***0.923***	***0.9185***	$O(N^2)$

The bold face values represent the best performance in between baseline methods, while bold and italic show the overall best

Table 3 The FS, JI and HD of the proposed multi-organ (Breast = Bst, Cervix = Cvx, Tongue = Ton) image dataset at 1000X magnification

Methods	Jaccard Index				Hausdorff Distance				F_1-Score				Accuracy			
Organ	Bst	Cvx	Ton	Avg	Bst	Cvx	Ton	Avg	Bst	Cvx	Ton	Avg	Bst	Cvx	Ton	Avg
ANSC [2]	0.395	0.389	0.363	0.382	9.422	10.498	9.35	9.76	0.716	0.699	0.771	0.729	0.859	0.814	0.704	0.792
SAM [3]	0.514	0.497	0.466	0.492	11.352	12.394	10.34	11.36	0.72	0.818	0.779	0.772	0.863	0.833	0.712	0.803
DFDLM [8]	0.457	0.454	0.424	0.445	11.93	11.296	14.51	12.58	0.683	0.761	0.836	0.760	0.826	0.876	0.769	0.824
NSDR [9]	0.677	0.587	0.536	0.600	8.287	10.405	9.43	9.37	0.727	0.681	0.869	0.759	0.87	0.89	0.858	0.873
Proposed	**0.71**	***0.719***	**0.597**	**0.675**	**6.919**	***7.54***	**8.24**	**7.57**	**0.847**	**0.853**	***0.901***	**0.867**	***0.925***	**0.915**	**0.915**	***0.918***

The bold face values represent the best performance in between baseline methods, while bold and italic show the overall best

5.2 Evaluation parameters

To measure the performance of the proposed localization method, the parameters used are: $F1$-score (FS) [27, 28], Jaccard index (JI), [29] and Hausdorff distance (HD) [30] for segmentation. The accuracy and area under curve (AUC) [31] are used for the classification performance measurement along with the receiver operating characteristic (ROC) curve [31].

Table 4 Dataset description for classification experiments five times the original samples images

Experiment	Operation	Benign samples	Malignant samples	Total
Exp.1	Total	160	130	290
Exp.2	Total	2890	6160	9050
Exp.3	Total	2890	2960	5850

Exp.1 Bisque dataset the only available combination
Exp.2 BreakHis dataset with given image combination 1
Exp.3 BreakHis dataset with given image combination 2

6 Results and discussion

The evaluation of localization and classification is analyzed qualitative and quantitative.

6.1 Localization work evaluation

The image localization is visually illustrated in Fig. 4. The images in the upper three rows are of Bisque dataset, fourth row image is from proposed dataset, and lower two rows are from the BreakHis dataset. Figure 4a shows the H&E stained original image, and Fig. 4b depicts the preprocessed image followed by their corresponding ground truth in Fig. 4c. The hematoxylin component of the stain decomposed image is shown in Fig. 4d. Figure 4e–g visualizes the nuclei segmented image, HCR and ECR components, respectively. First and fifth rows are the benign cases, and rests are the malignant cases. For the malignancy, the HCR increases inside the nuclei.

The quantitative performance of the proposed localization method is presented in terms of average FS, JI and HD is shown in Table 2, for Bisque and BreakHis (400× magnification images) dataset. The proposed method provides an average FS of 0.861, which is 23% and 21% better than the best performing baseline methods NSDR and DFDLM, respectively. The proposed method provides a JI value for BreakHis dataset images about 0.721 with overall average is 0.713. The average accuracy is 0.919 which is 10% greater than the NSDR method.

The performance of the proposed localization method is also validated by the proposed multi-organ dataset, which comprises 10 images from each organ, including breast, cervix and tongue with equal counts of benign and malignant cases, shown in Table 3. The accuracy of the proposed method is 0.918 which is at par to the standard datasets result.

6.2 Experimental setup for classification

The experimental setup is divided into three categories of different dataset and image augmented combinations. The small dataset size is increased by flipping and shearing operation at 10 degrees along horizontal (or vertical) axis gave two set of images. The random cropping and 10% random noise addition provided other two sets of images. The image dataset size becomes (four regenerated sets + original) five times as shown in Table 4. Exp. 1 is designed using Bisque dataset with 160 benign and 130 malignant HIs. Exp. 2 is designed with BreakHis data that has total 9050 images with 2890 benign and 6160 malignant cases as an imbalance dataset. In Exp. 3, the dataset imbalance problem is taken care. For balancing purpose, five–five selected images are taken from the ductal carcinoma in situ (DCiS) [7]. It provides a set of 5850 images at 400× from BreakHis dataset. The Exp.3 comprises of 2890 benign image and 2960 malignant HIs.

Table 5 Average accuracy (Ac%) and area under the curve (AUC) for Bisque and BreakHis datasets classification

Features	Measures	Features by Bag of words (BoW)		BoW with handcrafted features	
Datasets	Methods	SVM	MLP	SVM	MLP
Bisque Exp1	Ac (%)	82.83	78.27	90.06	94.48
	AUC	0.75	0.77	0.86	0.91
BreakHis Exp2	Ac (%)	81.83	79.71	93.47	93.86
	AUC	0.9	0.79	0.91	0.92
BreakHis Exp3	Ac (%)	85.35	83.34	93.73	**96.75**
	AUC	0.86	0.86	0.92	**0.94**
Average	Ac (%)	83.34	80.44	92.42	**95.03**
	AUC	0.84	0.81	0.90	0.92

Table 6 Confusion matrix Exp. 1–Exp. 3(in %) data into benign (B) and malignant (M) classes

Class	Exp.1		Exp.2		Exp.3		Method
	B	M	B	M	B	M	
B	89.38	10.62	93.11	6.89	93.63	6.37	SVM
	93.75	6.25	93.63	6.37	**95.85**	4.15	MLP
M	9.23	90.77	6.64	93.36	7.26	92.84	SVM
	6.92	93.08	5.96	94.04	3.58	**96.42**	MLP

Table 7 Comparison of classification performance with previous works on BreakHis dataset

Method	Accuracy (%)	AUC
MDC [17]	88.32	0.84
CRDNN [18]	86.37	0.85
Proposed	**95.03**	**0.92**

The bold face values represent the best performing values

6.3 Evaluation of the classification work

The MLP classifier performed best using the combination of BoW and HC features with an average accuracy of 96.75% for the balanced dataset (Exp.3), while for imbalanced dataset (Exp.2) 93.86% as shown in Table 5.

The MLP provides the highest AUC of 0.94 for balance dataset. The confusion matrix for Exp. 1–Exp.3 using BoW with HC features is shown in Table 6. The average accuracy of the proposed method is 95.03%, which is 10% and 7% higher than the CRDNN and MDC methods, respectively. The average AUC has reported 0.92 in comparison to 0.84 and 0.85 provided by MDC and CRDNN, respectively as shown in Table 7.

7 Conclusion

This paper presents an innovative HI localization based classification method using a combination of BoW features and HC features. The proposed method categorizes the HIs in benign and malignant classes. The extraction of HC features is performed on the basis of the intra-nuclei region separation of localized image in two components: HCRs and ECRs. Total 31 HC features are extracted out of which 20 significant features are selected using neighborhood components analysis. The BoW in association with HC features, is used for classification using MLP and SVM models. The simulation results are obtained using Bisque, BreakHis and proposed datasets. The proposed localization method has attained an average accuracy of 91.85%. The performance of the MLP model using balanced dataset has reported as best with an average accuracy of 95.03%, which is 10% and 7% higher as that of CRDNN and MDC methods.

Acknowledgements Autors are thankful to the supporting team of Jawaharlal Nehru Cancer Hospital & Research Center, (JNCH&RC) Bhopal, India. Specially Smt. Asha Joshi (Chairman), Smt. Divya Parashar (CEO & Research Coordinator), Dr. K. V. Pandya (Director), Dr. Pradeep Kolekar (Medical Director) and Mr. Rakesh Joshi (Additional Director), JNCH&RC Bhopal, India, for facilitating to work with patient data for dataset preparation.

References

1. Jung, C., Kim, C.: Segmenting clustered nuclei using H-minima transform-based marker extraction and contour parameterization. IEEE Trans. Biomed. Eng. **57**(10), 2600–2604 (2010)
2. Wang, P., Hu, X., Li, Y., Liu, Q., Zhu, X.: Automatic cell nuclei segmentation and classification of breast cancer histopathology images. Signal Process. **122**, 1–13 (2016)
3. Wang, Z.: A semi-automatic method for robust and efficient identification of neighboring muscle cells. Pattern Recogn. **53**, 300–312 (2016)
4. Jabeen, A., Riaz, M.M., Iltaf, N., Ghafoor, A.: Image contrast enhancement using weighted transformation function. IEEE Sens. J. **16**(20), 7534–7536 (2016)
5. Nguyen, K., Sarkar, A., Jain, A.K.: Prostate cancer grading: use of graph cut and spatial arrangement of nuclei. IEEE Trans. Med. Imag **33**(12), 2254–2270 (2014)
6. Lu, Z., Carneiro, G., Bradley, A.P., et al.: Evaluation of three algorithms for the segmentation of overlapping cervical cells. IEEE J. Biomed. Health Inf. **21**(2), 441–450 (2017)
7. Spanhol, F.A., Oliveira, L.S., Petitjean, C., Heutte, L.: A dataset for breast cancer histopathological image classification. IEEE Trans. Biomed. Eng. **63**(7), 1455–1462 (2016)
8. Vu, T.H., Mousavi, H.S., et al.: Histopathological image classification using discriminative feature-oriented dictionary learning. IEEE Trans. Med Imag **35**(3), 738–751 (2016)
9. Naylor, P., La, M., Reyal, F., Walter, T.: Segmentation of nuclei in histopathology images by deep regression of the distance map. IEEE Trans. Med. Imag (2018). https://doi.org/10.21227/H26X0H
10. Li, X., Plataniotis, K.N.: Circular mixture modeling of color distribution for blind stain separation in pathology images. IEEE J. Biomed. Health Inf. **21**(1), 150–161 (2017)

11. Wang, W., Ozolek, J.A., Slepcev, D., et al.: An optimal transportation approach for nuclear structure-based pathology. IEEE Trans. Med. Imag. **30**(3), 621–631 (2011)
12. Dundar, M.M., Badve, S., Bilgin, G., et al.: Computerized classification of intraductal breast lesions using histopathological images. IEEE Trans. Biomed. Eng. **58**(7), 1977–1984 (2011)
13. Basavanhally, A.N., et al.: Computerized image-based detection and grading of lymphocytic infiltration in HER2 + breast cancer histopathology. IEEE Trans. Biomed. Eng. **57**(3), 642–653 (2010)
14. Manivannan, S., Li, W., Zhang, J., et al.: Structure prediction for gland segmentation with hand-crafted and deep convolutional features. IEEE Trans. Med. Imag **37**(1), 210–221 (2018)
15. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
16. Raji, C.G., Chandra, S.S.V.: Long-term forecasting the survival in liver transplantation using multilayer perceptron networks. IEEE Trans. Syst. Man Cybern. Syst. **47**(8), 2318–2329 (2017)
17. Li, C., Wang, X., Liu, W., et al.: Weakly supervised mitosis detection in breast histopathology images using concentric loss. Med. Image Anal. **53**, 165–178 (2019)
18. Yan, R., Ren, F., Wang, Z., et al.: Breast cancer histopathological image classification using a hybrid deep neural network. Methods (2019)
19. Yang, W., Wang, K., Zuo, W.: Neighborhood component feature selection for high-dimensional data. JCP **7**, 161–168 (2012)
20. Klein, A., et al.: Fast Bayesian optimization of machine learning hyperparameters on large datasets, 2016. ArXiv:abs/1605.07079
21. Irshad, H., et al.: Methods for nuclei detection, segmentation, and classification in digital histopathology: a review current status and future potential. IEEE Rev. Biomed. Eng. **7**, 97–114 (2014)
22. Azzopardi, G., Petkov, N.: Contour detection by CORF operator. In: ANN and Machine Learning ICANN 2012, Lecture Notes in Computer Science, vol. 7552, pp. 395–402. Springer, Heidelberg (2012)
23. Kurmi, Y., Chaurasia, V., Ganesh, N.: Tumor malignancy detection using histopathology imaging. J. Med. Imaging Radiat. Sci. (2019). https://doi.org/10.1016/j.jmir.2019.07.004
24. E. Mercan, S. Aksoy, L. G. Shapiro, et al.: Localization of diagnostically relevant regions of interest in whole slide images. In: 22nd International Conference on Pattern Recognition, Stockholm, pp. 1179–1184 (2014)
25. Riddle, N.C., et al.: Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin. Genome Res. **21**(2), 147–163 (2011)
26. Kvilekval, K., Fedorov, D., Obara, B., et al.: Bisque: a platform for bioimage analysis and management. Bioinformatics **26**(4), 544–552 (2010)
27. Kurmi, Y., Chaurasia, V.: Multifeature-based medical image segmentation. IET Image Proc. **12**(8), 1491–1498 (2018)
28. Chaurasia, V., Chaurasia, V.: Statistical feature extraction based fast fractal image compression. J. Vis. Commun. Image Represent. **41**, 87–95 (2016)
29. Jaccard, P.: The distribution of the flora in the alpine zone. New Phytol. **11**, 37–50 (1912)
30. Taha, A.A., Hanbury, A.: An efficient algorithm for calculating the exact Hausdorff distance. IEEE Trans. Pattern Anal. Mach. Intell. **37**(11), 2153–2163 (2015)
31. Fawcett, T.: An introduction to ROC analysis. Pattern Recog. Lett. **27**, 861–874 (2006)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.