

Breast Cancer Detection via Global and Local Features using Digital Histology Images

Ghulam Murtaza¹, Ainuddin Wahid Abdul Wahab¹, Ghulam Raza², Liyana Shuib¹

Abstract:

Globally, breast cancer (BC) is the prevailing cause of unusual deaths in women. Breast tumor (BT) is a primary symptom and may lead to BC. Digital histology (DH) image modality is a gold standard medical test for a definite diagnosis of BC. Traditionally, DH images are visually examined by two or more pathologists to come up with a consensus for authentic BC detection which may cause a high error rate. Therefore, researchers had developed automated BC detection models using a machine learning (ML) based approach. Thus, this study aims to develop a BC detection model through ten feature extraction methods which extract both local and global type features from publicly available breast histology dataset. The extracted features are sorted by their weights, which are computed by the neighborhood component analysis method. A feature selection algorithm is developed to find the minimum number of discriminating features, evaluated through seven heterogeneous traditional ML classifiers. The proposed ML-based BC detection model acquired 90% accuracy for the initial testing set using 51 Harris features. Whereas, for the extended testing set, only three Harris features is shown 93% accuracy. The proposed BC detection model can assist the doctor in giving a second opinion.

Keywords: breast cancer detection, histology image, cancer classification, machine learning, deep learning.

1. Introduction

Breast cancer is one of the most frequent causes of death in women than in other types of cancer. According to the American Cancer Society (2018) report, around 316120 cases were diagnosed with BC, and 40610 women died unusually. Usually, BC is initiated when breast cells propagate abnormally i.e., out of control. Consequently, such abnormal cell growth forms a lump, also known as a tumor. Breast tumors (BT) are of two types, non-carcinoma and carcinoma. Moreover, non-carcinoma BT has two basic subtypes,

normal and benign. Benign tumors are categorized as non-cancerous and usually do not lead to any severe health problems. Whereas in situ and invasive are subtypes of carcinoma BT, causes BC and create severe health issues. In situ is non-invasive and does not spread to the other organs of a woman's body, and remains in the mammary ductal-lobular system. In contrast, invasive BT tends to invade the surrounding breast tissues. However, if BC is timely diagnosed and pursued with proper treatment, it can be cured by 85% at the initial stage and decreased by up to 10% for later stages (Wang, Khosla, Gargeya, Irshad, & Beck,

¹ Faculty of Computer Science and Information Technology, University of Malaya, 50603, Kuala Lumpur, Malaysia

² Our Lady of Lourdes hospital Drogheda Ireland.

Corresponding Authors: gmurtaza@siswa.um.edu.my, ainuddin@um.edu.my

2016). In routine checkups, BC can be detected by screening examination even before any physical symptoms like a lump. Initially, detection and diagnosis of BC can be made via radiology image modalities such as mammograms (MGs), magnetic resonance imaging (MRI), computed tomography (CT), positron emission tomography (PET), and thermography. However, MGs are commonly adopted for the last three decades for BC screening examination (Tabar et al., 2011). In the case of abnormal tissue growth found in breast screening examination, biopsy techniques are recommended to ensure the presence of BC. The most popular biopsy techniques are fine needle aspiration biopsy, core needle biopsy, vacuum-assisted biopsy, incisional biopsy, and excisional biopsy (BREASTCANCER.ORG, 2015). In a biopsy procedure, the sample of cells or tissues is collected from the affected area of the breast and placed over a glass microscope slide for dyeing and further microscopic analysis. The visual microscopic study of the histology samples allows both cell and tissue level examination of the breast, which enables a pathologist to detect BC presence assertively. Thus, histology samples are used as the gold standard for BC detection as well as for classification (Rubin, Strayer, & Rubin, 2008). Recently, due to the advent of digital pathology labs, the microscopic histology slides are scanned and transformed into digital images called whole slide images (WSI). The cell and tissue level examination in WSI enabled pathologists to mark regions of interest (ROI) (i.e., cancerous and non-cancerous areas). The ROI is marked based on interpretations made for overall tissues' architectures by examining cell distribution, nuclei organization, density, shape (i.e., regular or irregular), and variations in stained tissues (Gurcan et al., 2009). For a detailed analysis of each ROI, various image patches are captured from marked ROI of WSI using a camera with various zooming factors (i.e., 40x, 100x, 200x and 400x) (Spanhol, Oliveira, Petitjean, & Heutte, 2016b) named as digital histology/histopathological (DH) images. The final decision of BC detection and classification is made with the consensus of interpretations made by two or more expert

pathologists after adopting rigorous facial/visual examination of DH images. However, the manual facial analysis of DH images is an inefficient time task. Moreover, the interpretation accuracy of DH image diagnosis depends upon the domain knowledge (Bige et al., 2011), training, experience, and differences in circumstances and personal interest of learning (Desjardins, 1960) of a pathologist. However, a minor misinterpretation among BC types may lead to significant health and financial loss. Therefore, there is an urge for a machine learning (ML) based automated computerized system to assist the pathologist for better interpretation of DH images of BC.

The ML detection from BC histology tissues is characterized by two types of features, namely local and global features. The global features are also called structural features, include contour, edge density, direction, nuclei shape, and glandular unit shape representations like histogram-oriented gradients (HOG) features. Whereas local features (also named texture features) represent the spatial distribution of intensity (rough, smooth, silky, or bumpy) in a region of an image like local binary patterns (LBP). Harris and Stephens (1988) proposed a combined corner and edge detection algorithm based on the local auto-correlation function. The auto-correlation function had eliminated the limitation of a shifting window by a small amount of variation proposed by Moravec (1980). The correlation function not only detects the corners and edges but also measures the edge quality by selecting isolated corner pixels for thinning the edge pixels. It is achieved by taking three measures; first, all small window shifts are made through analytic expansion about the shift origin. Second, a smooth circular window was adopted like Gaussian instead of a binary and rectangular window. Third, the corner measure can use both edges with direction-shift instead of depending on minimum edge value. Thus, the Harris method based features are more accurate than the Moravec method due to less number of false positives to detect edges and corners. Jianbo and Tomasi (1994) developed the eigenvalue algorithm to detect the corners of an object. Unlike Harris features using the response function score,

the author used minimum eigenvalue (MinEigen) to decide corners. It can be mathematically expressed by $T(v) = \lambda v$, where λ eigenvalue associated with eigenvector v . Geometrically eigenvalue is a point stretched by a transformation in a direction by some non-zero factor called eigenvector. Ojala, Pietikäinen, and Mäenpää (2002) presented the LBP method to recognize certain binary patterns named as uniform. The uniform patterns are basic characteristics of local image texture, and their histogram is an authenticated source of dominant texture features. The proposed method computes a generalized gray-scale and rotation invariant operator to detect uniform patterns. The extracted uniform patterns are resilient to any quantization of the angular space and for any spatial resolution. Moreover, it combines multiple operators for multi-resolution analysis. Thus, the developed approach is highly robust to grayscale variation and rotation. Moreover, this approach is computationally simple, efficient because operators can be realized with few comparisons in a small neighbourhood and a lookup table. Such capabilities make LBP method more applicable to time critical applications. Matas, Chum, Urban, and Pajdla (2004) proposed an efficient detection algorithm for an affinely invariant stable subset of external regions termed as maximally stable extremal regions (MSER). The external regions are set of image elements when put into correspondence possess two vital properties; for instance, these sets are closed under the continuous transformation of image coordinates and monotonic transformation of image intensities. MSER algorithm extracts features from brighter or darker regions distinct from their surroundings. Whereas, these regions are stable by their relevant range of thresholds of the intensity function. Therefore, MSER can extract regions more efficiently, like crosswalk regions or regions with distinct illumination variations. Dalal and Triggs (2005) presented work on HOG features to predict pedestrian detection in still images. The basic idea is the distribution of intensity gradients or edge directions, which can identify the appearance and shape of an object in an image. Therefore, HOG features are also categorized as global or structural

features. It is achieved by splitting an image into cells i.e., small overlapping regions. Afterward, a histogram for gradient direction of each cell is computed to form a single HOG feature. Ultimately, a combination of all cells HOG features represents the object present in the entire image. Moreover, the intensity of an image block (group of cells) is calculated for histogram contrast normalization of each cell present in respective block. This normalization enhances the HOG features accuracy and robustness to alterations in illumination and shadowing. Rosten and Drummond (2005) introduced real-time tracking system by collective use of both points and edges properties present in an image. For real time feature tracking the author proposed features from accelerated segment test (FAST) corner feature detector technique. FAST techniques use 16 pixels to detect target pixel p belongs to a corner pixel. Thus, each pixel which surrounds p in a circle is labelled in clockwise fashion by values starting from 1 to 16, see Fig 1. If a set of continuous surrounded pixels N in a circle are brighter than the intensity of p plus threshold value t or all darker than the intensity of p minus threshold value t , then p is detected as corner pixel. Therefore, FAST technique will show good results if three or more N are available around p . Otherwise, there is a need to add blur with normal distribution to get better corner detection using FAST technique. Bay, Ess, Tuytelaars, and Van Gool (2008) coined rotation invariant detectors and features named as speed-up robust feature (SURF). These can be computed faster and outperformed (by scale variant feature transform - SIFT) due to distinctiveness and robustness. This is achieved by using integral images for image convolutions, Hessian matrix approximation for the detection, sums of Haar wavelet components for feature extraction, along with simplifying these methods as needed. SURF is a faster method due to the use of integral images and Laplacian indexing for the matching step. Some of the effective applications are object recognition, image classification, reconstruction, content-based image retrieval, and registration. Leutenegger, Chli, and Siegwart (2011) proposed a binary robust invariant scalable keypoint (BRISK) method for keypoint detection, description,

and matching. This method had proved adaptive, better quality algorithm on benchmark datasets by showing low power consumptions compared to SURF. The success lies in the use of unique scale-space derived from FAST detection algorithm along with building a bit-string descriptor by comparing the intensities of the keypoint neighbourhood. Thus, BRISK uses an easily computed circular sampling pattern for brightness comparison to create a binary descriptor string. Due to its unique properties, it is useful for many real-time applications where limited time with low power consumption matters like a content-based image revival. Rublee, Rabaud, Konolige, and Bradski (2011) proposed oriented FAST and Rotated BRIEF(ORB) as a cost-effective and faster than SURF extraction method based on BRIEF. Moreover, this method is also rotation invariant and robust to noise. To achieve this, the first keypoint extracted by FAST. Afterwards, the top point is selected by Harris corner detection. Moreover, an efficient orientation component is added to the FAST feature algorithm. This makes ORB as rotation invariant compared to FAST. KAZE features are named to tribute the father of scale-space analysis (Weickert, Ishikawa, & Imiya, 1999). Alcantarilla, Bartoli, and Davison (2012) proposed the KAZE feature extraction method, which involves various steps. For instance, initially, nonlinear scale space is computed using an efficient Additive Operator Splitting (AOS) technique with variable conductance diffusion. Next, using nonlinear scale-space the response of scale-normalized determinant of Hessian at multiple levels is computed to detect 2D features of interest. Finally, rotation-invariant features were extracted by approximating the dominant orientation found in local neighbourhood centred at the keypoint location for first order image derivatives.

The main contributions of this study are given below.

1. This study develops an efficient (consumes less training and detection time), cost-effective (can be trained on a normal desktop computer), and feasible (requires a smaller number of labelled images to show

good results) model for BC detection using DH images.

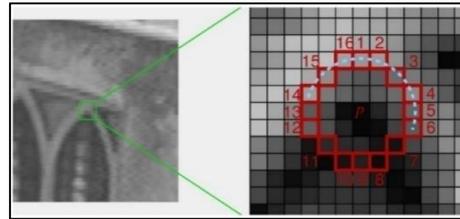


Fig 1: FAST feature extraction technique (Rosten & Drummond, 2005)

2. Ten (local and global) feature extraction methods (i.e., HOG, MinEigen, Harris, LBP, KAZE, MSER, SURF, FAST, BRISK, and ORB) are evaluated to get the most distinct features.

3. A feature selection algorithm is developed to extract the minimum number of discriminative feature subsets to complement classification results.

4. The selected reduced number of features are evaluated via six performance evaluation metrics, namely accuracy, sensitivity, specificity, precision, F-measure, and Matthews correlation coefficient using seven heterogeneous traditional ML classifiers, namely SVM, NB, LR, LDA, kNN, GK, and DT.

5. The results of the proposed BC detection model (complemented with feature reduction algorithm) are compared with ten state-of-the-art baseline models.

Further sections of this study are arranged as follows. Section 2 gives an overview of current literature, limitations of existing BC detection and classification models, and the main contributions of this study. The overall research methodology employed is discussed in section 3. Whereas Section 4 discusses the experimental setups and subsequent results. Next, section 5 discussion covers the significance of experimental results, limitations of this research with future directions. Finally, section 6 conclusion recaps the overall research made in this study.

2. Literature Review

In the past three decades, most of the research had been focused on radiology images (Tabar et al., 2011) especially

mammogram (breast x-rays) images using traditional machine learning (ML) based approaches for binary classification (Hassanien & Ali, 2006; Kontos & Maragoudakis, 2013; Biswas, Nath, & Roy, 2016; Nusantara, Purwanti, & Soelistiono, 2016; Ponraj, Poongodi, & Mercy, 2017). For instance, Ponraj et al. (2017) performed BC detection using MG images of a publicly available dataset. The author extracted textural features using LBP and local gradient pattern (LGP) along with related histograms. The extracted feature set is classified through SVM, where LGP outperformed the LBP by showing better accuracy, sensitivity, and specificity as 95%, 92%, and 94%, respectively. Moreover, Nusantara et al. (2016) used digital MGs to classify them into normal and abnormal categories. Where ROI from MGs was cropped to extract features like energy, mean, and standard deviation from the third level of wavelet decomposition coefficients. The highest reported accuracy, sensitivity, and specificity were 96.8%, 100%, and 95%, respectively. Similarly, Biswas et al. (2016) utilized the MGs of a public dataset for normal and abnormal tissue classification. The author extracted texture features from image ROIs using a gray level co-occurrence matrix (GLCM) followed by many pre-processing steps like removal of artifacts, noise reduction, and contrast enhancement. The author achieved a higher 95% accuracy, 100% sensitivity, and 90% specificity through SVM compared to kNN and artificial neural networks. Apart from a binary classification using radiology images, many researchers used DH images to solve the BT multiclassification problem via deep learning (DL) based approaches (Araujo et al., 2017; Han et al., 2017; Wan, Cao, Chen, & Qin, 2017; Zheng et al., 2017; Murtaza, Shuib, Mujtaba, & Raza, 2019; Murtaza, Shuib, Wahab, et al., 2019; Murtaza, Shuib, Abdul Wahab, et al., 2020; Murtaza, Shuib, Wahab, et al., 2020). For instance, Han et al. (2017) proposed a class structure-based deep convolutional neural network (CSDCNN) which embeds the non-linear representation learning model instead of handcrafted features. CSDCNN was initially trained on imageNet dataset to show better results for the BreakHis dataset. The author reported accuracy for eight subtypes of BT was

93.2%. Moreover, Zheng et al. (2017) proposed a nucleus guided feature extraction method. Initially, the proposed deep convolution network (DCNN) was trained using nucleus features without labels; afterward, the trained DCNN was fine tuned and retrained using labelled DH images of an exclusive dataset. The reported accuracy for the 15-classes dataset was 96.4%. A summary of proposed methodology and limitations of aforementioned studies given in Table I.

It has been observed from the aforementioned studies that the traditional ML-based BC detection models are better than DL based models due to five reasons. First, DL based BC detection models require high computational resources like GPU, RAM with longer training time compared to traditional ML models. Second, DL models need a large number of images to avoid overfitting issue. Whereas, traditional ML model can be trained efficiently using a small number of images to show comparable results. Third, it is highly difficult to optimize the hyper-parameters of deep layered networks before initiating the training process. Fourth, DL based models get very smaller size input images like AlexNet(Krizhevsky, Sutskever, & Hinton, 2012) to get 227x227 pixels, whereas DH images usually are of very high resolution like BCBH DH image size is 2040 x 1536 pixels. Thus, rescaling is mandatory before feeding into a deep neural network, which causes loss of information (Komura & Ishikawa, 2018). Hence, DL based models are more suitable to solve the multiclassification problem where a larger number of labelled images are available with ample computational resources. However, the traditional ML-based BC detection model is a better and worthwhile choice for BC detection using DH images.

In this study, DH images are utilized for BC detection confidently compared to radiology images. Moreover, ten aforementioned feature extraction methods (FEMs) namely HOG, MinEigen, Harris, LBP, KAZE, MSER, SURF, FAST, BRISK, and ORB are used to extract local and global features from DH images. Both, local and global types of features are helping to choose the most distinct features using various ML

classifiers to detect BC accurately from DH images. Moreover, many studies used a similar type of FEMs, thus provides direct results comparison of the proposed classification approach with state-of-the-art baseline models. Next, the extracted features from each FEMs are stored in the master feature table (MFT). Where, a row in MFT represents many features of each image present in the training set, and two (initial and extended) testing sets. Thus, ten MFTs are created to represent each of the FEM. Finally, MFTs are classified through seven

heterogeneous traditional ML learning classifiers namely support vector machine (SVM), naïve Bayes (NB), linear regression (LR), linear discriminant analysis (LDA), k nearest neighbor (kNN), Gaussian Kernel (GK) and decision tree (DT) using six performance evaluation metrics such as accuracy, sensitivity, specificity, precision, F-measure and Matthews correlation coefficient.

Table I: A summary of the methodology and their limitations of existing models for breast cancer detection

Reference	Methodology	Limitations
Nusantara et al. (2016)	ROI-based wavelet decomposition is used for feature extraction and classification is made by kNN.	The result can be compromised if ROI is not extracted properly. Thus, have a higher dependency of mammogram expert.
Biswas et al. (2016)	Gray Level Co-occurrence Matrix (GLCM) used on ROI of a mammogram of MIAS dataset. kNN(k=3) is used for classification.	The result can be compromised if ROI is not extracted properly. Thus, have a higher dependency of mammogram expert.
Ponraj et al. (2017)	Features extracted via local binary pattern (LBP) and local gradient pattern (LGP) and classified through SVM.	Uses a small dataset thus model training can be biased. Needs to show other PEMs like F-measure or MCC to show class level biasness.
Han et al. (2017)	Proposed CNN-based model for classification of BreakHis dataset for 8 subtypes of BrT using histopathology images.	The model requires a large number of images for pretraining like ImageNet. Thus, needs GPU and longer training time.
Zheng et al. (2017)	A CNN-based model trained on histopathology images where the nucleus of cancer lesion is already marked.	Needs a histopathology image expert to mark the nucleus of the cancer lesion.

3. Proposed Method

This section demonstrates the employed overall research methodology to develop the BC classification model for carcinoma and non-carcinoma types. Four stages (see Fig 2) are involved namely data collection, classification models training, and classification with evaluation. The tasks performed in each phase are discussed in detail in the subsequent sections.

3.1. Data Collection

In the first stage of the research methodology (see Fig 2) a publicly available bioimaging challenge 2015 breast histology (BCBH) dataset (Teresa Araújo, 2015) is utilized to develop the BC cancer detection model. This dataset is hosted by the Institute for Systems and Computer Engineering, Technology, and Science (INESC TEC) associated laboratory. BCBH dataset possesses breast histology, annotated, uncompressed images with 2040 x 1536 pixels' resolution. Whereas, all the images

were acquired with the magnification of 200x with the same protocol. Moreover, the dataset contains H&E stained images divided into four types of BT, namely normal, benign, in situ, and invasive. Normal and benign subjects are assumed as non-cancerous (i.e., non-carcinoma), whereas the rest of the two subtypes (i.e., in situ and invasive) are cancerous (i.e., carcinoma). Thus, in this study BC detection is made as carcinoma and non-carcinoma basic types of a tumor. BCBH corpus possesses overall 285 DH images (see Table

2). Where 249 out of 285 images are provided for model training purposes, and the rest of the images (i.e., 36) are designated for model testing. Moreover, the testing set is divided into two parts, named as an initial testing set and an extended testing set. The initial testing set possesses 20, whereas the extended testing set consists of 16 DH images. The images in both training and testing sets are almost equally distributed for each class label. BCBH dataset is publicly accessible at URL³.

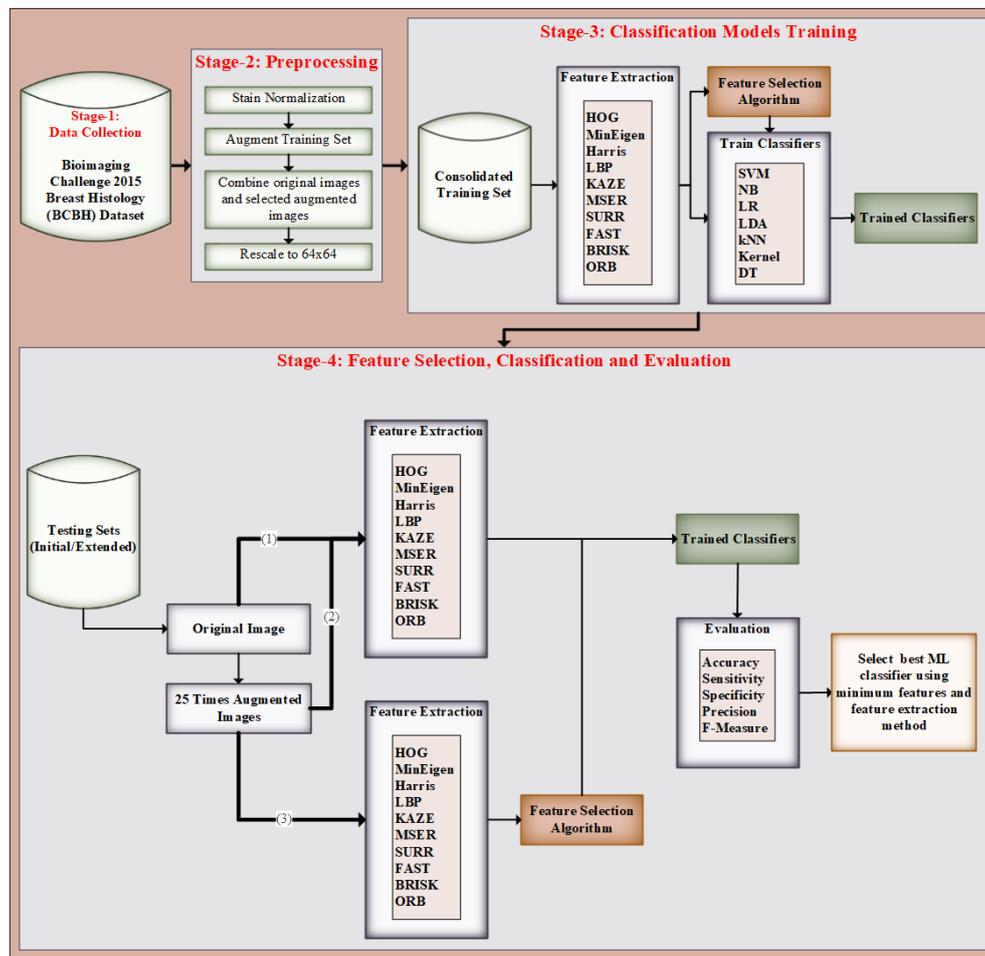


Fig 2: Overall research methodology. Feature selection strategies (1) Original image, (2) Augmented images, and (3) Augmented images with feature selection algorithm.

³ <https://rdm.inesctec.pt/dataset/nis-2017-003>.

Table II. Distribution of images in bioimaging challenge 2015 breast histology dataset

		Non-carcinoma	Carcinoma	Total
		Normal + Benign	In situ + Invasive	
Training Set		55+69	63+62	249
Testing Sets	Initial	5+5	5+5	20
	Extended	4+4	4+4	16
Total		132	143	285

3.2. Image pre-processing

In the second stage of research methodology (see Fig 2), the image pre-processing consists of many necessary tasks required to reduce the misclassification rate by cleaning the raw data before initiating classifier training as well as testing processes. In this study, the adopted pre-processing tasks are stain normalization, augmentation, and the rescaling of images.

3.2.1. Stain normalization

DH images inherently possess higher inconsistencies due to two major reasons. First, the cumbersome process of making biopsy slides for microscopic facial analysis. Second, the scanning process of biopsy slides into digital images to create WSI. The preparation of biopsy slides involves pathology lab protocols, human skills, embedding, sectioning, and coloring (McCann, Ozolek, Castro, Parvin, & Kovacevic, 2015). Moreover, the preparation of WSI in digital pathological labs involves various scanners (to convert biopsy microscopic slides into digital WSI) of many vendors, which possess different scanning protocols. Thus, inconsistencies (i.e., visible variability) are found due to different resolution, color, brightness, and contrast settings of scanners. Therefore, to avoid such dominant variabilities which may largely distract the overall classification process, need to be eliminated before initiating the training and testing process for

DH images. In this study, Reinhard, Adhikhmin, Gooch, and Shirley (2001) method is adopted to harmonize the stain of DH image. However, there are many stain normalization methods (Macenko et al., 2009; Khan, Rajpoot, Treanor, & Magee, 2014), but these methods are weaker to preserve the BC lesion structures. Thus, the use of such methods may cause a loss of important information to detect carcinoma lesions.

3.2.2. Image augmentation

The image augmentation is performed to enhance the performance of traditional ML classifiers for the collected BCBH dataset. In this study, the training set has been augmented by twenty-five times using basic image processing techniques like rotation, flip, transform, and padding. An algorithm (see Fig 3) is developed to generate twenty-five synthetic images from a single original image by applying the aforementioned basic image processing techniques. Each image of the training set is augmented by twenty-five times and overall, 6225 (i.e., 249x25) augmented training images are generated. However, out of 6225, 3100 belong to non-carcinoma, whereas 3125 are carcinoma augmented images. Whereas, the original images of the training set are also used along with augmented images for training purposes. The overall training set possesses 3224 (i.e., 3100+55+69) non-carcinoma and 3250 (i.e., 3125+63+62) carcinoma images, which are utilized for training of ML classifiers. Moreover, the developed augmentation algorithm is also used for testing sets to get further analysis of each augmented image to classify the original testing set image.

3.2.3. Image rescaling

BCBH dataset DH images are larger like 2040 x 1536 pixels. Thus, it needs lots of computational time to extract the features. In this study, it has been experimentally examined that the larger image size does not make a major effect on the extraction of meaningful features. Besides, the translation of images in the image augmentation algorithm returns an arbitrary size. Thus, rescaling is required to make all images of the same size before applying any FEM. This pre-processing step makes the entire classification process resources efficient

without affecting the overall performance of the BC classification.

3.3. Classification models training

In stage three of the research methodology (see Fig 2), two tasks are carried out using the augmented training set. First, feature extraction using various feature extraction techniques. Second, training of multiple ML classifiers using two approaches. In the first approach, the training of ML classifiers is performed by using all extracted features. Whereas, the second approach reduces the number of features by implementing the proposed feature selection algorithm. The details of each task are given in the subsequent sections.

3.3.1. Feature Extraction

The preprocessed training set is utilized to extract features by ten FEMs, namely HOG, MinEigen, Harris, LBP, KAZE, MSER, SURF, FAST, BRISK, and ORB. Many studies have used a similar type of FEMs; thus, it provides an opportunity for direct comparison of results for the proposed classification model. The extracted features of each FEM method are stored in MFT. Where a row in MFT represents many features of each image of training and two testing sets. Thus, ten MFTs are created to represent each of the aforementioned FEM.

3.3.2. ML classifiers training

In this study, seven heterogeneous traditional ML classifiers, namely SVM, NB, LR, LDA, kNN, GK, and DT are trained through aforesaid ten MFTs. Because most of the classifiers are used by baseline studies, that allows direct comparison of results of the proposed BC detection model. Here, the ML classifiers' training task is based on two approaches. First, to train above mentioned seven classifiers using all MFTs. Second, the use of the proposed feature selection algorithm (see **Error! Reference source not found.**) to reduce the features for the training of the above-given ML classifiers.

```

Input: OrgImg
Output: AugImg(25)
Function ImageAugmenter
1   define array AugImg(25)
2   AugImg(1) = flip(Img,TopToBottom)
3   AugImg(2) = flip(Img,LeftToRight)
4   AugImg(3) = rotate(AugImg(1),90)
5   AugImg(4) = rotate(AugImg(2),90)
6
7   AugImg(5) = translate(AugImg(1),H/2,-W/2)
8   AugImg(6) = translate(AugImg(1),H/2, W/2)
9   AugImg(7) = translate(AugImg(1),-H/2,-W/2)
10  AugImg(8) = translate(AugImg(1),-H/2, W/2)
11
12  AugImg(9) = translate(AugImg(2),H/2,-W/2)
13  AugImg(10) = translate(AugImg(2),H/2, W/2)
14  AugImg(11) = translate(AugImg(2),-H/2,-W/2)
15  AugImg(12) = translate(AugImg(2),-H/2, W/2)
16
17  AugImg(13) = transform(OrgImg, affine([1 0 0; 1 1 0; 0 0 1]))
18  AugImg(14) = transform(OrgImg, affine([1 0.33 0; 0 1 0; 0 0 1]))
19  AugImg(15) = transform(OrgImg, affine([1 0.9 0; 0 1 0; 0 0 1]))
20  AugImg(16) = transform(OrgImg, affine([1 0 0; .5 1 0; 0 0 1]))
21
22  AugImg(17) = transform(AugImg(1), affine([1 0 0; 1 1 0; 0 0 1]))
23  AugImg(18) = transform(AugImg(1), affine([1 0.33 0; 0 1 0; 0 0 1]))
24  AugImg(19) = transform(AugImg(1), affine([1 0.9 0; 0 1 0; 0 0 1]))
25  AugImg(20) = transform(AugImg(1), affine([1 0 0; .5 1 0; 0 0 1]))
26
27  AugImg(21) = transform(AugImg(2), affine([1 0 0; 1 1 0; 0 0 1]))
28  AugImg(22) = transform(AugImg(2), affine([1 0.33 0; 0 1 0; 0 0 1]))
29  AugImg(23) = transform(AugImg(2), affine([1 0.9 0; 0 1 0; 0 0 1]))
30  AugImg(24) = transform(AugImg(2), affine([1 0 0; .5 1 0; 0 0 1]))
31
32  AugImg(25) = Padding(OrgImg,affine[1 0; 1 1; 0 0]/2))
end
    
```

Fig 3: Image augmentation algorithm

The proposed feature selection algorithm performs two tasks in a cascade manner. First, it sorts (in descending order) the features of MFT of each feature extraction method by computing feature weights through neighborhood component analysis (NCA), see Fig 5, shows the sample graphs of HOG features only, whereas the same approach is implemented for all FEMs to get rid of those features which have a very low contribution (i.e., weight) to classify carcinoma and non-carcinoma distinctly. Thus, NCA weights are computed for the MFT of each feature extraction method. Fig 5 first column, shows the sample graphs of HOG unsorted features. Next, all MFTs' columns are sorted in descending order by NCA weights, (see Fig 5 second column) the sample graphs of HOG sorted features. The sorted MFT is referred to hereafter as SMFT. Finally, the top 64 features from all SMFTs are used for further analysis. Because it has been experimentally observed, the classification results are at their maximum level while using the top 59

```

Input: TrainingSet,TestingSetInitial,TestingSetExtended
Output: BestClassifier
Function FeatureSelection
1   Augment TrainingSet,TestingSetInitial,TestingSetExtended by ImageAugmenter
2   Extract Features TrainingSet,TestingSetInitial,TestingSetExtended using ten methods
3   Create MFT_Training,MFT_Testing_Initial,MFT_Testing_Extended for ten feature extraction methods
   Comments: compute weights by neighbourhood component analysis (NCA)
4   FeatureWeights = nca(MFT_Training,Labels_Training)
5   Ftr_Idx = sort(FeatureWeights,'desc')
   Comments: training set features by NCA weight
6   SMFT_Training = MFT_Training(:,Ftr_Idx)
   Comments: initial testing set features by NCA weight
7   SMFT_Testing_Initial = MFT_Testing_Initial(:,Ftr_Idx)
   Comments: sort extended testing set features by NCA weight
8   SMFT_Testing_Extended = MFT_Testing_Extended(:,Ftr_Idx)
   Comments: classify using reduced features by 1,3,5,7,.....
9   for Ftr = 1:TotalNoFeatures step by 2
10      Train SVM, NB, LR, LDA, kNN, GK and DT using SMFT_Training(:,Ftr)
11      Classify SVM, NB, LR, LDA, kNN, GK and DT using SMFT_Testing_Initial(:,Ftr)
12      Classify SVM, NB, LR, LDA, kNN, GK and DT using SMFT_Testing_Extended(:,Ftr)
13      Evaluate SVM, NB, LR, LDA, kNN, GK and DT using accuracy, sensitivity and F-measure
14   end
15   Select best feature extraction method, best classifier with minimum feature subset
end

```

Fig 4: Feature selection algorithm

features of each SMFT. Fig 5 third column, shows the sample graphs of HOG 64 features of the highest NCA weights. Moreover, the proposed feature selection algorithm takes features subset (i.e., 1,3,5, 7...) incrementally from SMFT to train the aforementioned seven ML classifiers. The goal of the proposed feature selection algorithm is to get a minimum feature subset to train the classifier without compromising the overall classification performance. In summary, Fig 5 first column graphs display the NCA weights (unsorted) and the second column represents the NCA weights (sorted in descending order) of overall HOG features. However, the third column graphs of Fig 5 show the selection of the minimum number of features based on higher weights. It has been experimentally observed that the classification results are at their maximum level while using the top 59 out of 1764 NCA weighted HOG features.

3.4. Feature selection, classification, and evaluation

In the fourth stage of the overall research methodology (see Fig 2), the classification and evaluation of two testing sets (i.e., initial and extended) are made by following three strategies of feature selection. In the first feature selection strategy, the original images of both testing sets are used to extract features using the aforementioned

ten FEMs. Here, all features are used for classification and evaluation purposes. Whereas, in the second feature selection strategy, the features are extracted from ten methods followed by applying twenty-five augmentation methods individually on all original images of both testing sets. For classification and evaluation, the overall extracted features of each image augmentation method are used for results analysis. However, in the third strategy of feature selection, the proposed feature selection algorithm is implemented, followed by all steps mentioned in the second strategy. In contrast, to both first and second feature selection strategies, the classification and evaluation are performed using selected reduced features subset (i.e., 1,3,5, 7...). After adopting the three feature selection strategies, the selected features are used for classification using aforesaid seven trained ML classifiers. Whereas, the classification results are evaluated through six performance evaluation metrics (PEMs) like accuracy, sensitivity, specificity, precision, F-measure, and Matthews correlation coefficient. Finally, the top-performing ML classifier using a minimum number of features is selected for each of the two testing sets.

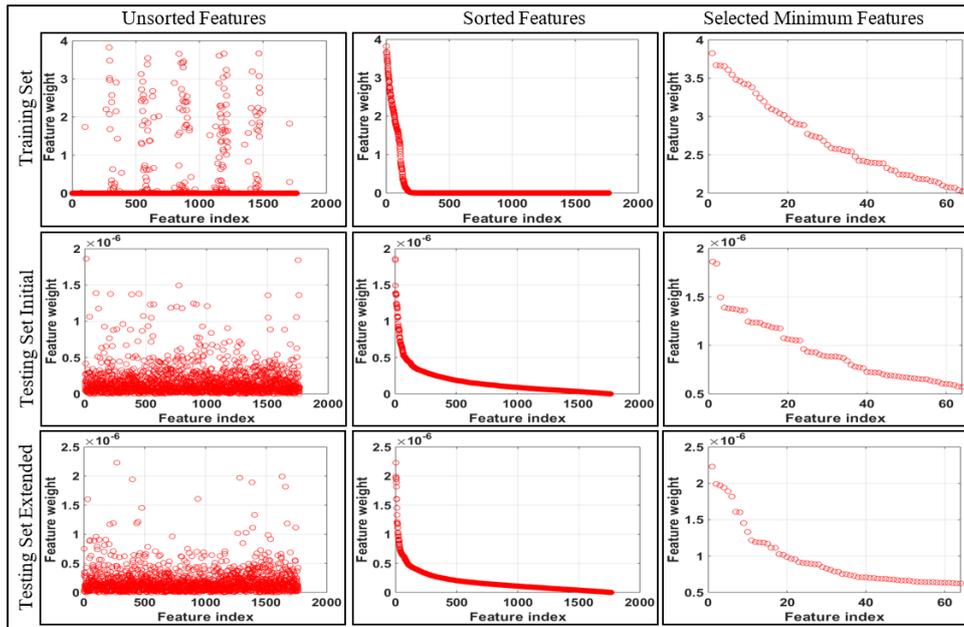


Fig 5. NCA weights of extracted HOG features for the training set and two testing sets

4. Experimental setup and analysis

The experiments are organized to assess the performance of ten FEMs namely HOG, MinEigen, Harris, LBP, KAZE, MSER, SURF, FAST, BRISK, and ORB through seven traditional ML classifiers namely SVM, NB, LR, LDA, kNN, GK, and DT. This research involves three experimental settings, see Fig 6

- I. Experiments using original images:** In the first experimental setting, the performance of ten FEMs through seven ML classifiers is analysed using two testing sets namely initial and extended. Finally, all the results are reported in this study.
- II. Experiments using augmented images:** In the second experimental setting, each original image is augmented twenty-five times. Afterward, each augmented image is used to evaluate the performance of ten FEMs through seven ML classifiers using two testing sets. Finally, the best results among twenty-five augmentation methods are shown in this study.
- III. Experiments using augmented images with feature selection algorithm:** In the third experimental

setting, thirty-two feature subset (1,3, 5...63) with twenty-five augmentation methods are evaluated by showing the performance of ten FEMs through seven ML classifiers using two testing sets. Finally, the best results with reduced features are shown for further analysis.

All classification model development steps like data splitting, image pre-processing, ML classifiers training, testing, and evaluation are performed in MATLAB R2019a. Furthermore, all experiments are carried out using default parameters except those which are specifically mentioned in this study.

4.1. Classification Performance evaluation metrics

In this study, the overall experimental results analysis is made through three main PEMs namely accuracy (A_c), sensitivity (S_n), F-measure (F_m), and Matthews Correlation Coefficient (MCC) see Fig 7. Because A_c (Fig 7, equation 1) is the most

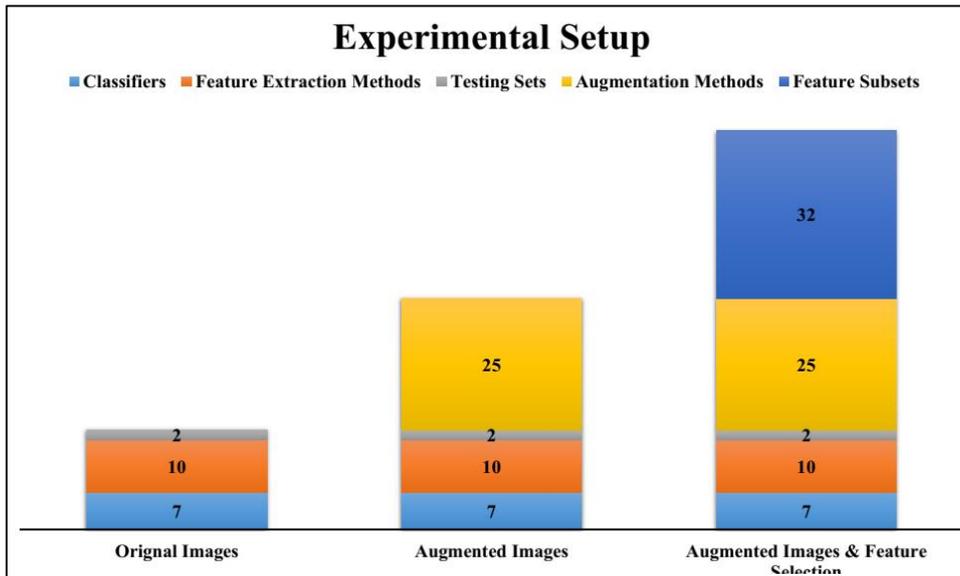


Fig 6. Experimental analysis made for each experimental setup

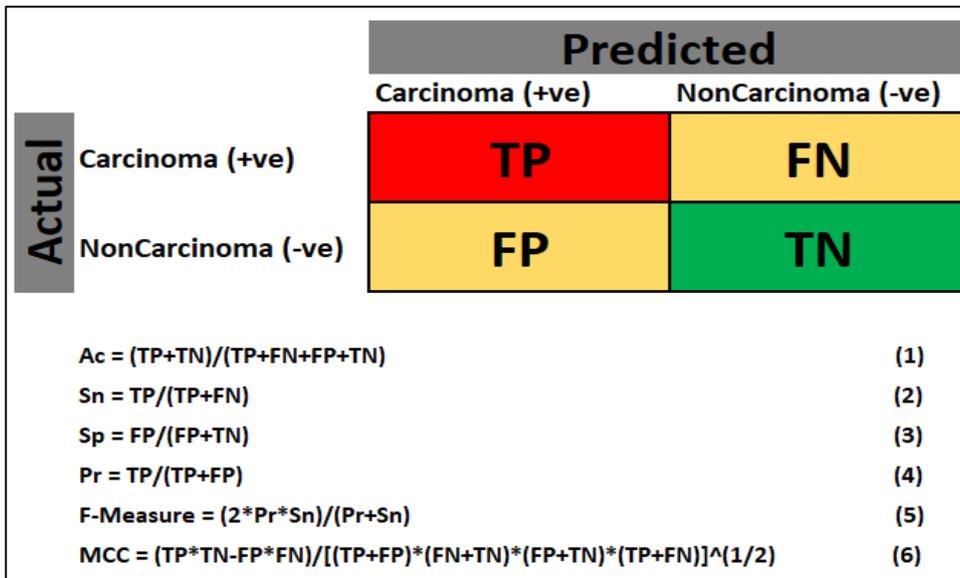


Fig 7. PEMs (like Ac, Sn, Sp, Pr, F-Measure and MCC) equations adopted for experimental results analysis

common PEMs needed to compare baseline studies, while Sn (Fig 7, equation 2) is highly important in medical science diagnosis of cancer (i.e., cancer positive cases). However, the other PEMs like specificity (Sp) (Fig 7, equation 3), and precision (Pr) (Fig 7, equation 4) are also reported along with four outcomes of the confusion matrix (see Fig 7) namely true

positive (TP), false negative (FN), false positive (FP), and true negative (TN) in the appendix-A (see Table 1, Table 2, and Table 3). Whereas, Fm (Fig 7, equation 5) and MCC (Fig 7, equation 6) can evaluate the quality (i.e., biasedness) of classifier predictions. MCC involves the overall outcomes of the confusion matrix. It returns a coefficient value between +1 and -1.

Where, coefficient value +1, 0, and -1 represents the perfect, average, and reverse predictions of a classifier respectively. Thus, in this study, MCC is adopted to select the top-performing classifier with unbiased predictions.

4.2. Experimental Results Analysis

In this section, the results of above mentioned three experimental setups are reported for analysis. The BC classification results are discussed in terms of four (i.e., Ac, Sn, Fm, and MCC) PEMs.

4.2.1. Experimental results using original images

Table 3 that the kNN (k=1,3,5) classifier using ORB features has outperformed by showing the highest 75.00% Ac, 78.26% Fm, and 0.5241 MCC with the second-highest Sn of 90.00%. Whereas the 100.00% Sn is achieved by kNN using HOG features with lower Ac, Sn, and MCC as 55.00%, 68.00%, and 0.2294.

Table that the better Ac and MCC like 81.25% and 0.6742 are gained by LDA+MSER but with a lower Sn of 62.50%. Whereas, the better Sn (i.e., 100.00%) is attended by both SVM+HOG and kNN+FAST. Moreover, the top Fm value of 87.50% is shown by LR for BRISK features with compromised Ac, Sn, and MCC. Conversely, LR produces the worst

This section reports the results of seven (i.e., SVM, NB, LR, LDA, kNN, GK, and DT) ML classifiers using ten (i.e., HOG, MinEigen, Harris, LBP, KAZE, MSER, SURF, FAST, BRISK, and ORB) FEMs for two (i.e., Initial and Extended) testing sets. In the first experimental setup, overall, 140 (7 classifiers x 10 FEMs x 2 testing sets) analyses are reported using four PEMs like Ac, Sn, Fm, and MCC see Table 3. In Table 3, the experimental results are split into two parts based on initial and extended testing sets. While using the initial testing set it can be observed from

Furthermore, the highest accuracy of 75.00% is also achieved through kNN using LBP features but with compromised MCC as 0.5025. In contrast, NB+SURF has shown the lowest performance (i.e., Ac=35.00% and Sn=00.00%) using the initial testing set. On the other hand, while using the extended testing set, it can be seen in

results while using both SURF and HOG features i.e., Ac=50.00%, Sn=00.00%. Similarly, NB+HOG has shown poor results (i.e., Ac=50.00%, Sn=00.00%) using the extended testing set. Summarizing, ORB and MSER features have outperformed among all FEMs using initial and extended testing sets by acquiring better MCC value for BC detection.

Table III. Experimental results using original images

		Initial Testing Set						Extended Testing Set							
		SVM	NB	LR	LDA	kNN	GK	DT	SVM	NB	LR	LDA	kNN	GK	DT
HOG	Ac	75.00	60.00	50.00	70.00	55.00	45.00	60.00	68.75	50.00	50.00	62.50	56.25	62.50	43.75
	Sn	80.00	100.00	0.00	80.00	100.00	50.00	50.00	75.00	87.50	0.00	62.50	100.00	62.50	50.00
	Fm	76.19	71.43	NaN	72.73	68.97	47.62	55.56	70.59	63.64	NaN	62.50	69.57	62.50	47.06
	MCC	0.5025	0.3333	NaN	0.4082	0.2294	-0.1005	0.2041	0.3780	0.0000	NaN	0.2500	0.2582	0.2500	0.1260
MinEigen	Ac	60.00	55.00	40.00	40.00	50.00	70.00	35.00	31.25	43.75	50.00	56.25	43.75	37.50	68.75
	Sn	40.00	20.00	30.00	20.00	50.00	70.00	50.00	25.00	25.00	25.00	50.00	50.00	12.50	62.50
	Fm	50.00	30.77	33.33	25.00	50.00	70.00	43.48	26.67	30.77	33.33	53.33	47.06	16.67	66.67
	MCC	0.2182	0.1400	0.2041	-0.2182	0.0000	0.4000	0.3145	-0.3780	-0.1348	0.0000	0.1260	0.1260	-0.2887	0.3780
Harris	Ac	40.00	60.00	60.00	60.00	55.00	55.00	45.00	62.50	50.00	31.25	37.50	43.75	62.50	43.75
	Sn	40.00	30.00	40.00	40.00	50.00	70.00	50.00	75.00	50.00	25.00	37.50	37.50	50.00	25.00
	Fm	40.00	42.86	50.00	50.00	52.63	60.87	47.62	66.67	50.00	26.67	37.50	40.00	57.14	30.77

		Initial Testing Set						Extended Testing Set							
		SVM	NB	LR	LDA	kNN	GK	DT	SVM	NB	LR	LDA	kNN	GK	DT
MC C		-	0.2500	0.218	0.2182	0.1005	0.1048	0.100	0.2582	0.0000	0.378	0.2500	0.126	0.2582	0.134
		0.2000		2				5			0		0		8
	Ac	65.00	60.00	70.00	60.00	75.00	60.00	65.00	62.50	68.75	62.50	68.75	62.50	68.75	37.50
	Sn	80.00	70.00	90.00	80.00	70.00	80.00	80.00	50.00	50.00	50.00	75.00	62.50	50.00	25.00
LBP MC C	Fm	69.57	63.64	75.00	66.67	73.68	66.67	69.57	57.14	61.54	57.14	70.59	62.50	61.54	28.57
		0.3145	0.2041	0.436	0.2182	0.5025	0.2182	0.314	0.2582	0.4045	0.258	0.3780	0.250	0.4045	0.258
				4				5			2		0		2
	Ac	55.00	55.00	55.00	50.00	65.00	45.00	50.00	56.25	56.25	56.25	43.75	43.75	56.25	43.75
KAZE MC C	Sn	20.00	20.00	20.00	50.00	50.00	20.00	50.00	12.50	25.00	12.50	25.00	37.50	12.50	50.00
	Fm	30.77	30.77	30.77	50.00	58.82	26.67	50.00	22.22	36.36	22.22	30.77	40.00	22.22	47.06
		0.1400	0.1400	0.140	0.0000	0.3145	-	0.000	0.2582	0.1601	0.258	-	0.126	0.2582	0.126
				0			0.1155	0			2	0.1348	0		0
MSER MC C	Ac	55.00	70.00	65.00	60.00	55.00	70.00	60.00	62.50	68.75	62.50	81.25	75.00	56.25	56.25
	Sn	20.00	90.00	40.00	30.00	30.00	40.00	40.00	37.50	62.50	37.50	62.50	75.00	25.00	25.00
	Fm	30.77	75.00	53.33	42.86	40.00	57.14	50.00	50.00	66.67	50.00	76.92	75.00	36.36	36.36
		0.1400	0.4364	0.346	0.2500	0.1155	0.5000	0.218	0.2887	0.3780	0.288	0.6742	0.500	0.1601	0.160
			4				2			7		0		1	
SURF MC C	Ac	60.00	35.00	40.00	45.00	55.00	65.00	50.00	56.25	56.25	50.00	43.75	56.25	56.25	56.25
	Sn	80.00	0.00	0.00	0.00	10.00	80.00	10.00	50.00	12.50	0.00	12.50	37.50	50.00	37.50
	Fm	66.67	NaN	NaN	NaN	18.18	69.57	16.67	53.33	22.22	NaN	18.18	46.15	53.33	46.15
		0.2182	-	-	-	0.2294	0.3144	0.000	0.1259	0.2581	NaN	-	0.134	0.1259	0.134
		0.4200	0.333	0.2294	0.16	0.85	0	88	99		0.1601	84	88	84	
		8	3	2							3				
FAST MC C	Ac	70.00	60.00	60.00	60.00	50.00	45.00	50.00	62.50	62.50	50.00	50.00	62.50	56.25	43.75
	Sn	90.00	20.00	20.00	20.00	10.00	70.00	20.00	100.00	62.50	25.00	37.50	62.50	62.50	25.00
	Fm	75.00	33.33	33.33	33.33	16.67	56.00	28.57	72.73	62.50	33.33	42.86	62.50	58.82	30.77
		0.4364	0.3333	0.333	0.3333	0.0000	-	0.000	0.3780	0.2500	0.000	0.0000	0.250	0.1260	0.134
			3			0.1155	0			0		0		8	
BRISK MC C	Ac	65.00	50.00	50.00	50.00	55.00	60.00	45.00	50.00	62.50	56.25	68.75	62.50	43.75	68.75
	Sn	80.00	10.00	20.00	20.00	20.00	60.00	10.00	87.50	75.00	25.00	62.50	62.50	50.00	62.50
	Fm	69.57	16.67	28.57	28.57	30.77	60.00	15.39	12.50	50.00	87.50	75.00	62.50	37.50	75.00
		0.3145	0.0000	0.000	0.0000	0.1400	0.2000	0.140	0.0000	0.2582	0.160	0.3780	0.250	-	0.378
			0				0			1		0	0.1260	0	
ORB MC C	Ac	50.00	50.00	55.00	50.00	75.00	55.00	40.00	62.50	43.75	50.00	50.00	50.00	37.50	50.00
	Sn	50.00	60.00	40.00	40.00	90.00	50.00	50.00	75.00	37.50	50.00	50.00	50.00	62.50	37.50
	Fm	50.00	54.55	47.06	44.44	78.26	52.63	45.46	66.67	40.00	50.00	50.00	50.00	50.00	42.86
		0.0000	0.0000	0.104	0.0000	0.5241	0.1005	0.204	0.2582	-	0.000	0.0000	0.000	-	0.000
			8				1			0		0	0.2887	0	

4.2.2. Experimental results analysis using image augmentation algorithm

This section represents the results of the aforementioned seven ML classifiers for ten FEMs using twenty-five image augmentation methods for two (i.e., Initial and Extended) testing sets. In the second

experimental setup, overall, 140 (7 classifiers x 10 FEMs x 2 testing sets x 1 Augmentation method) analyses of top-performing augmentation method is reported using three PEMs like Ac, Sn, and Fm, see Table 4.

Table , by analysing the results of the initial testing set, it can be observed that the

highest 0.7035 MCC with comparable Sn (i.e.; 90.00%) is shown by DT+MSER. However, similar MCC is also achieved by NB+MSER but with a compromised 80.00%. Whereas, the highest Sn (i.e., 100.00%) is produced by SVM in most of the FEMs with lower MCC. Contrary, NB, and LR have shown the lowest performance (i.e., Ac=50.00%, Sn=00.00%) using BRISK features. On the flip side, using extended testing set the better Ac (i.e., 90.00%), Sn (i.e., 100.00%), and MCC (i.e.,

0.8165) are achieved by SVM+BRISK with a lower Fm value of 80.00%. Whereas, a better Fm of 100.00% is observed using both LDA and DT with BRISK features. In contrast, the lowest Ac (i.e., 50.00%) and Sn (i.e., 00.00%) are noticed with both SURF and FAST features when classified through LR. In summary, the MSER and BRISK feature outperformed among all FEMs using initial and extended testing sets by acquiring better MCC values for BC detection.

Table IV. Experimental results analysis using augmented images

		Initial Testing Set							Extended Testing Set						
		SVM	NB	LR	LDA	kNN	GK	DT	SVM	NB	LR	LDA	kNN	GK	DT
HOG	Ac	65.00	75.00	70.00	80.00	75.00	70.00	75.00	81.25	62.50	56.25	75.00	75.00	75.00	43.75
	Sn	90.00	90.00	100.00	70.00	90.00	90.00	80.00	87.50	100.00	37.50	87.50	75.00	62.50	50.00
	Fm	72.00	78.26	76.92	77.78	78.26	75.00	76.19	82.35	72.73	46.15	77.78	75.00	71.43	47.06
	MC C	0.346 4	0.524 1	0.500 0	0.612 4	0.524 1	0.436 4	0.502 5	0.629 9	0.378 0	0.1348	0.516 4	0.500 0	0.516 4	- 0.1260
MinDigen	Ac	65.00	70.00	80.00	80.00	85.00	65.00	70.00	50.00	62.50	50.00	87.50	81.25	75.00	62.50
	Sn	100.00	100.00	70.00	90.00	70.00	70.00	80.00	100.00	75.00	87.50	87.50	75.00	87.50	50.00
	Fm	74.07	76.92	77.78	81.82	82.35	66.67	72.73	66.67	66.67	63.64	87.50	80.00	77.78	57.14
	MC C	0.420 1	0.500 0	0.612 4	0.612 4	0.733 8	0.301 5	0.408 2	NaN	0.258 2	0.0000	0.750 0	0.629 9	0.516 4	0.2582
Harris	Ac	75.00	60.00	65.00	65.00	65.00	75.00	70.00	68.75	81.25	50.00	75.00	81.25	62.50	37.50
	Sn	100.00	90.00	80.00	80.00	90.00	90.00	70.00	100.00	100.00	62.50	62.50	87.50	87.50	37.50
	Fm	80.00	69.23	69.57	69.57	72.00	78.26	70.00	76.19	84.21	55.56	71.43	82.35	70.00	37.50
	MC C	0.577 4	0.250 0	0.314 5	0.314 5	0.346 4	0.524 1	0.400 0	0.480 4	0.674 2	0.0000	0.516 4	0.629 9	0.288 7	- 0.2500
LBP	Ac	65.00	70.00	70.00	70.00	80.00	70.00	75.00	68.75	62.50	43.75	81.25	81.25	68.75	62.50
	Sn	100.00	90.00	90.00	90.00	90.00	90.00	100.00	100.00	87.50	50.00	87.50	100.00	100.00	75.00
	Fm	74.07	75.00	75.00	75.00	81.82	75.00	80.00	76.19	70.00	47.06	82.35	84.21	76.19	66.67
	MC C	0.420 1	0.436 4	0.436 4	0.436 4	0.612 4	0.436 4	0.577 4	0.480 4	0.288 7	- 0.1260	0.629 9	0.674 2	0.480 4	0.2582
KAZE	Ac	70.00	75.00	80.00	85.00	80.00	80.00	75.00	75.00	68.75	50.00	75.00	68.75	68.75	50.00
	Sn	100.00	80.00	80.00	100.00	70.00	70.00	80.00	75.00	75.00	100.00	75.00	100.00	75.00	100.00
	Fm	76.92	76.19	80.00	86.96	77.78	77.78	76.19	75.00	70.59	66.67	75.00	76.19	70.59	66.67
	MC C	0.500 0	0.502 5	0.600 0	0.733 8	0.612 4	0.612 4	0.502 5	0.500 0	0.378 0	NaN	0.500 0	0.480 4	0.378 0	NaN
MSER	Ac	65.00	85.00	70.00	75.00	70.00	80.00	85.00	68.75	68.75	50.00	81.25	87.50	75.00	18.75
	Sn	80.00	80.00	70.00	70.00	70.00	70.00	90.00	62.50	87.50	37.50	62.50	87.50	87.50	25.00
	Fm	69.57	84.21	70.00	73.68	70.00	77.78	85.71	66.67	73.68	42.86	76.92	87.50	77.78	23.53
	MC C	0.314 5	0.703 5	0.400 0	0.502 5	0.400 0	0.612 4	0.703 5	0.378 0	0.404 5	0.0000	0.674 2	0.750 0	0.516 4	- 0.6299
SURF	Ac	75.00	75.00	75.00	75.00	70.00	80.00	70.00	56.25	62.50	50.00	68.75	75.00	62.50	50.00
	Sn	90.00	90.00	80.00	90.00	80.00	90.00	60.00	87.50	87.50	0.00	75.00	87.50	87.50	0.00
	Fm	78.26	78.26	76.19	78.26	72.73	81.82	66.67	66.67	70.00	NaN	70.59	77.78	70.00	NaN
	MC C	0.524 1	0.524 1	0.502 5	0.524 1	0.408 2	0.612 4	0.408 2	0.160 1	0.288 7	NaN	0.378 0	0.516 4	0.288 7	NaN
FA	Ac	80.00	65.00	60.00	60.00	65.00	75.00	70.00	75.00	81.25	50.00	68.75	75.00	68.75	50.00

		Initial Testing Set							Extended Testing Set						
		SVM	NB	LR	LDA	kNN	GK	DT	SVM	NB	LR	LDA	kNN	GK	DT
	Sn	100.0 0	30.00	20.00	20.00	40.00	80.00	40.00	100.0 0	87.50	0.00	50.00	75.00	87.50	0.00
	Fm	83.33	46.15	33.33	33.33	53.33	76.19	57.14	80.00	82.35	NaN	61.54	75.00	73.68	NaN
	MC C	0.654 7	0.420 1	0.333 3	0.333 3	0.346 4	0.502 5	0.500 0	0.577 4	0.629 9	NaN	0.404 5	0.500 0	0.404 5	NaN
BRISK	Ac	75.00	75.00	50.00	81.25	75.00	68.75	50.00	90.00	50.00	60.00	65.00	55.00	80.00	65.00
	Sn	100.0 0	75.00	0.00	75.00	62.50	87.50	0.00	100.0 0	40.00	30.00	30.00	40.00	100.0 0	30.00
	Fm	80.00	75.00	NaN	80.00	71.43	73.68	NaN	80.00	60.00	90.00	100.0 0	70.00	60.00	100.00
ORB	MC C	0.577 4	0.500 0	NaN	0.629 9	0.516 4	0.404 5	NaN	0.816 5	0.000 0	0.2500	0.420 1	0.104 8	0.654 7	0.4201
	Ac	85.00	65.00	60.00	60.00	70.00	70.00	65.00	62.50	68.75	43.75	81.25	68.75	68.75	43.75
	Sn	80.00	70.00	80.00	80.00	90.00	80.00	80.00	62.50	75.00	37.50	75.00	75.00	75.00	50.00
	Fm	84.21	66.67	66.67	66.67	75.00	72.73	69.57	62.50	70.59	40.00	80.00	70.59	70.59	47.06
	MC C	0.703 5	0.301 5	0.218 2	0.218 2	0.436 4	0.408 2	0.314 5	0.250 0	0.378 0	0.1260	0.629 9	0.378 0	0.378 0	- 0.1260

4.2.3. Experimental results using image augmentation algorithm with feature selection algorithm

This section demonstrates the results of the aforementioned seven traditional ML classifiers using ten FEMs and twenty-five image augmentation methods followed by minimum features selection for two (i.e., Initial and Extended) testing sets. In the third experimental setup, overall, 140 (7 classifiers x10 FEMs x 2 testing sets x 1 Augmentation method x 1 reduced features subset) analyses of top-performing classifiers using reduced features are reported using four PEMs like Ac, Sn, Fm, and MCC.

Initial testing set results in Table shows that GK outperformed using the least number of Harris features (i.e., 517 out of 64) with the highest Ac (i.e., 90%), Sn (i.e., 90%), Fm (i.e., 95.24%) and MCC (i.e., 0.80). However, the same results are produced via SVM by using 57 out of 1764 HOG features. Besides, HOG and LBP features

have shown 100% Sn using most of the classifiers. Other classifiers like NB (i.e., 3 out of 64 FAST features) and LDA (i.e., 19 out of 64 FAST features) have shown the worst performance by achieving the lowest Sn and Fm like 30% and 46.15%. Contrastingly, for extended testing, set the results in Table illustrate that kNN got better results (i.e., Ac=93.75, Sn=100%, Fm=94.12%, and MCC=0.8819) using 3 out of 1764 Harris features. Moreover, 100% Sn is observed using HOG, Harris, and ORB reduced features using almost all classifiers. Moreover, approximately the same results are obtained via kNN+MinEigen by using 32 out of 64 features. In contrast, LR using (13 out 64) BRISK features has shown the lowest Sn (i.e., 37.50%) and compromised Ac (i.e., 68.75%) using the extended testing set. In a nutshell, Harris and FAST selective features have outperformed among all FEMs using initial and extended testing set by acquiring better MCC using the least number of features for BC detectio

Table V. Experimental results using augmented images with the feature selection algorithm

		Initial Testing Set							Extended Testing Set						
		SVM	NB	LR	LDA	kNN	GK	DT	SVM	NB	LR	LDA	kNN	GK	DT
HOG	Ac	90.00	75.00	85.00	85.00	85.00	80.00	85.00	75.00	81.25	81.25	81.25	93.75	87.50	87.50
	Sn	90.00	100.0 0	100.0 0	100.0 0	100.0 0	90.00	80.00	100.0 0	100.0 0	100.0 0	87.50	87.50	100.0 0	87.50
	Fm	90.00	80.00	86.96	86.96	86.96	81.82	84.21	80.00	84.21	84.21	82.35	93.33	88.89	87.50

	Initial Testing Set							Extended Testing Set							
	SVM	NB	LR	LDA	kNN	GK	DT	SVM	NB	LR	LDA	kNN	GK	DT	
MCC	<u>0.800</u> ₀	0.577 ₄	0.733 ₈	0.733 ₈	0.733 ₈	0.612 ₄	0.703 ₅	0.577 ₄	0.674 ₂	0.674 ₂	0.629 ₉	0.881 ₉	0.774 ₆	0.750 ₀	
	Features	<u>57</u>	3	61	63	5	11	29	23	1	61	1	27	3	1
MinEigen	Ac	80.00	75.00	80.00	80.00	90.00	85.00	80.00	81.25	75.00	87.50	87.50	<u>93.75</u>	93.75	87.50
	Sn	<u>100.0</u> ₀	<u>100.0</u> ₀	70.00	90.00	80.00	80.00	80.00	<u>100.0</u> ₀	87.50	<u>100.0</u> ₀	75.00	<u>87.5</u>	87.50	87.50
	Fm	83.33	80.00	77.78	81.82	88.89	84.21	80.00	84.21	77.78	88.89	85.71	<u>93.33</u>	93.33	87.50
	MCC	0.654 ₇	0.577 ₄	0.612 ₄	0.612 ₄	0.816 ₅	0.703 ₅	0.600 ₀	0.674 ₂	0.516 ₄	0.774 ₆	0.774 ₆	<u>0.881</u> ₉	0.881 ₉	0.750 ₀
	Features	49	29	7	63	11	13	39	19	13	17	13	<u>23</u>	19	55
Harris	Ac	85.00	70.00	80.00	75.00	80.00	90.00	80.00	87.50	81.25	81.25	81.25	<u>93.75</u>	87.50	87.50
	Sn	<u>100.0</u> ₀	<u>100.0</u> ₀	90.00	90.00	<u>100.0</u> ₀	90.00	90.00	<u>100.0</u> ₀	87.50					
	Fm	86.96	76.92	81.82	78.26	83.33	90.00	81.82	88.89	84.21	84.21	84.21	<u>94.12</u>	88.89	87.50
	MCC	0.733 ₈	0.500 ₀	0.612 ₄	0.524 ₁	0.654 ₇	0.800 ₀	0.612 ₄	0.774 ₆	0.674 ₂	0.674 ₂	0.674 ₂	<u>0.881</u> ₉	0.774 ₆	0.750 ₀
	Features	55	7	41	19	29	51	3	39	3	3	3	<u>3</u>	9	4
LBP	Ac	75.00	70.00	80.00	75.00	85.00	75.00	80.00	75.00	75.00	75.00	87.50	81.25	81.25	87.50
	Sn	<u>100.0</u> ₀	<u>100.0</u> ₀	<u>100.0</u> ₀	90.00	80.00	<u>100.0</u> ₀	<u>100.0</u> ₀	<u>100.0</u> ₀	87.50	87.50	87.50	<u>100.0</u> ₀	<u>100.0</u> ₀	75.00
	Fm	80.00	76.92	83.33	78.26	84.21	80.00	83.33	80.00	77.78	77.78	87.50	84.21	84.21	85.71
	MCC	0.577 ₄	0.500 ₀	0.654 ₇	0.524 ₁	0.703 ₅	0.577 ₄	0.654 ₇	0.577 ₄	0.516 ₄	0.516 ₄	0.750 ₀	0.674 ₂	0.674 ₂	0.774 ₆
	Features	3	1	5	4	7	3	10	6	9	3	3	3	6	7
KAZE	Ac	80.00	85.00	90.00	90.00	90.00	85.00	85.00	81.25	81.25	81.25	81.25	87.50	75.00	87.50
	Sn	<u>100.0</u> ₀	90.00	90.00	<u>100.0</u> ₀	90.00	<u>100.0</u> ₀	90.00	87.50	75.00	87.50	<u>100.0</u> ₀	<u>100.0</u> ₀	75.00	<u>100.0</u> ₀
	Fm	83.33	85.71	90.00	90.91	90.00	86.96	85.71	82.35	80.00	82.35	84.21	88.89	75.00	88.89
	MCC	0.654 ₇	0.703 ₅	0.800 ₀	0.816 ₅	0.800 ₀	0.733 ₈	0.703 ₅	0.629 ₉	0.629 ₉	0.629 ₉	0.674 ₂	0.774 ₆	0.500 ₀	0.774 ₆
	Features	3	21	7	51	11	3	61	53	13	39	43	17	11	9
MSER	Ac	80.00	85.00	80.00	80.00	80.00	90.00	85.00	68.75	68.75	75.00	81.25	87.50	75.00	87.50
	Sn	90.00	90.00	90.00	80.00	80.00	80.00	90.00	75.00	87.50	62.50	75.00	87.50	<u>100.0</u> ₀	<u>100.0</u> ₀
	Fm	81.82	85.71	81.82	80.00	80.00	88.89	85.71	70.59	73.68	71.43	80.00	87.50	80.00	88.89
	MCC	0.612 ₄	0.703 ₅	0.612 ₄	0.600 ₀	0.600 ₀	0.816 ₅	0.703 ₅	0.378 ₀	0.404 ₅	0.516 ₄	0.629 ₉	0.750 ₀	0.577 ₄	0.774 ₆
	Features	9	43	9	13	5	19	47	5	29	21	25	17	9	11
SURF	Ac	75.00	75.00	80.00	85.00	80.00	75.00	80.00	75.00	68.75	68.75	68.75	87.50	68.75	81.25
	Sn	<u>100.0</u> ₀	<u>100.0</u> ₀	90.00	90.00	90.00	90.00	80.00	87.50	87.50	87.50	87.50	75.00	87.50	87.50
	Fm	80.00	80.00	81.82	85.71	81.82	78.26	80.00	77.78	73.68	73.68	73.68	85.71	73.68	82.35
	MCC	0.577 ₄	0.577 ₄	0.612 ₄	0.703 ₅	0.612 ₄	0.524 ₁	0.600 ₀	0.516 ₄	0.404 ₅	0.404 ₅	0.404 ₅	0.774 ₆	0.404 ₅	0.629 ₉
	Features	17	49	5	17	3	5	15	3	7	3	9	11	7	41
FAST	Ac	85.00	60.00	65.00	65.00	70.00	80.00	70.00	75.00	87.50	75.00	81.25	81.25	87.50	81.25
	Sn	<u>100.0</u> ₀	40.00	30.00	30.00	40.00	100.0 ₀	40.00	<u>100.0</u> ₀	87.50	50.00	62.50	62.50	<u>100.0</u> ₀	75.00
	Fm	86.96	50.00	46.15	46.15	57.14	83.33	57.14	80.00	87.50	66.67	76.92	76.92	88.89	80.00
	MCC	0.733 ₈	0.218 ₂	0.420 ₁	0.420 ₁	0.500 ₀	0.654 ₇	0.500 ₀	0.577 ₄	0.750 ₀	0.577 ₄	0.674 ₂	0.674 ₂	0.774 ₆	0.629 ₉
	Features	45	3	3	19	5	7	63	5	9	27	37	7	15	31

		Initial Testing Set							Extended Testing Set						
		SVM	NB	LR	LDA	kNN	GK	DT	SVM	NB	LR	LDA	kNN	GK	DT
BRISK	Ac	90.00	65.00	65.00	65.00	70.00	85.00	65.00	75.00	75.00	68.75	81.25	87.50	81.25	87.50
	Sn	100.00	40.00	30.00	30.00	40.00	100.00	40.00	100.00	75.00	37.50	75.00	75.00	100.00	87.50
	Fm	90.91	53.33	46.15	46.15	57.14	86.96	53.33	50.00	75.00	100.00	87.50	100.00	62.50	87.50
	MCC	0.8165	0.3464	0.4201	0.4201	0.5000	0.7338	0.3464	0.5774	0.5000	0.4804	0.6299	0.7746	0.6742	0.7500
	Features	25	17	27	27	5	25	31	57	13	13	21	17	31	47
ORB	Ac	80.00	75	85	85	85	80	80	81.25	87.5	81.25	75	81.25	87.5	87.5
	Sn	90.00	90.00	90.00	100.00	90.00	80.00	80.00	100.00	100.00	100.00	100.00	100.00	75	75
	Fm	81.82	78.26	85.71	86.96	85.71	80.00	80.00	84.21	88.88	84.21	80	84.21	85.71	85.71
	MCC	0.6124	0.5241	0.7035	0.7338	0.7035	0.6000	0.6000	0.6742	0.7746	0.6742	0.5774	0.6742	0.7746	0.7746
	Features	1	7	11	11	31	7	19	37	33	59	9	5	29	21

In general, the BC classification results are drastically improved by proposed algorithms like image augmentation and feature selection algorithms, see Fig 8. Fig 8 (a), bar-graph represents the results for the initial testing set, whereas the Fig 8 (b) bar-graph shows the results for the extended testing set. While using the initial testing set, it can be observed from Fig 8(a) that PEMs are improved radically. Such as Ac is increased by 15% (from 75% to 90%), Fm is raised by 11.74% (from 78.26 to 90%), and MCC elevated from 0.5241 to 0.8 when augmented images are used with feature selection algorithm. Here, Sn remains stagnant like 90%, during all experiments. However, features are reduced by 20.31% i.e., only 51 out of 64 Harris features are

utilized to get the highest results like 90% for each Ac, Sn, Fm. Here, the highest MCC obtained is 0.80 using the initial testing set to detect BC.

Similarly, while using an extended testing set the PEMs trend is exponentially increased, see Fig 8 (b). For instance, Ac is increased by 11.75% (from 81.25% to 93%), Sn is enhanced by 37.5% (from 62.5% to 100%), Fm value is improved by 17.2% (from 76.92% to 94.12%) when the image augmentation is applied by using feature selection algorithm. Moreover, the features are reduced by 95.31%, i.e., only 3 out of 64 Harris features are used to get maximum results like 93% Ac, 100% Sn, 94.12% Fm, and 0.8819 MCC.

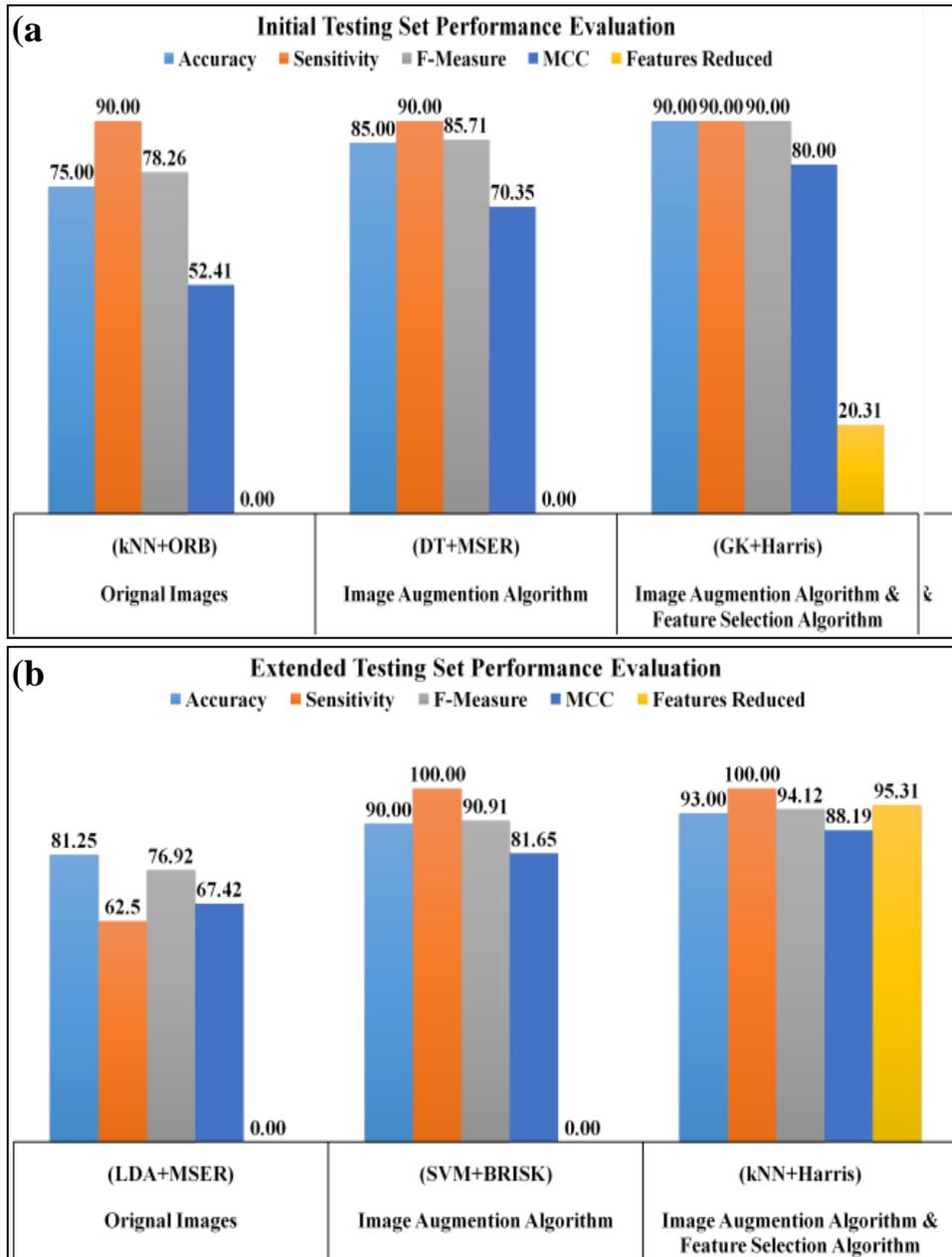


Fig 8. Three experimental setup results comparison among features of original images, augmented images and augmented images with selected features. a). Initial dataset: Three types of features ORB, MSER and Harris are used with kNN, DT and GK classifiers. b). Extended dataset: Three types of features MSER, BRISK and Harris are used with LDA, SVM and kNN classifiers.

5. Discussion

This section reveals the significance of the proposed BC detection model, developed using the traditional ML-based approach for

DH images. As mentioned in the introduction section breast DH images are used for confident detection and classification of BC. For automatic detection, most of the research is made for

BC detection using DH images via DL based models and very few studies have employed traditional ML-based models. Both types of BC detection models have their pros and cons. For instance, DL based models usually require high computation power, RAM, and longer training time. Such type of deeply layered models also requires a large number of annotated images for proper training (Han et al., 2017). Moreover, it is a cumbersome task (based on the trial-and-error method) to adjust hyperparameters before initiating the training process to get the desired results. However, these models have shown good results to solve multiclassification (up to eight classes) problem (Han et al., 2017; Murtaza, Shuib, Mujtaba, et al., 2019). In contrast, the traditional ML-based approach had shown compromised results to solve the multiclassification problem. Moreover, for BC detection, the ML-based models achieved comparable results to DL based models. Apart from comparable performance, the traditional ML-based BC detection models can be trained efficiently using a normal desktop computer and do not expect a large number of labelled images to better results. Thus, it can be concluded that, apart from DL based multiclassification model, the traditional ML-based model is a feasible solution for BC detection, where we have limited computational resources and smaller datasets like medical image datasets.

Most of the existing state-of-the-art studies used DL based approach (Janowczyk & Madabhushi, 2016; Spanhol, Oliveira, Petitjean, & Heutte, 2016a; Araujo et al., 2017; Song, Zou, Chang, & Cai, 2017; Spanhol, Oliveira, Cavalin, Petitjean, & Heutte, 2017; Gupta & Bhavsar, 2018; Huang & Chung, 2018) for BC detection using DH images. For instance, Janowczyk and Madabhushi (2016) used AlexNet configuration to create a DL based model to detect invasive BC. The proposed model was trained on GPU for 22 hours using an exclusive dataset of DH images for BC detection. The reported Ac and Fm are 84.68% and 76.48%. Spanhol et al. (2016a) proposed a transferred learning-based DL

model trained through the BreakHis dataset. The author acquired Ac was $85.6 \pm 4.8\%$ using a GPU for 3 hours. Araujo et al. (2017) hosted the BCBH dataset for BT classification using DH images. The author proposed a DL based model for feature extraction and classification is made through softmax and SVM. SVM outperformed softmax by showing accuracies of 90% and 81.3% for initial and extended testing sets. However, there is room to improve the performance for BC detection mentioned in the aforementioned studies. Huang and Chung (2018) developed a CNN based spatial fused residual network using the BCBH dataset. The DH images were split into non-overlapping patches to avoid image rescaling. Moreover, a spatial relationship among patches was utilized to enhance the BC detection using high-resolution DH images. The author reported a high 98.5% accuracy and 99.6% AUC. However, the model training time was very high (i.e., 30 minutes/iteration) using GPU. Moreover, the classification duration reported was 80 minutes/image. It can be concluded from the aforementioned studies that most of the DL-based models consumed high computational resources, longer training time, and required a large number of labeled images for training, which is not a cost-effective solution for BC detection using DH images. Moreover, most of the studies only reported Ac and it can be biased towards a particular class. Thus, any other PEM like Sn, Sp, Pr or Fm is needed to be measured to show unbiased/reliable classification results. Where, Sn is highly important in medical science to identify the misclassification of true positives.

A few of the existing state-of-the-art studies (Wan, Liu, Chen, & Qin, 2014; Belsare, Mushrif, Pangarkar, & Meshram, 2015; Spanhol et al., 2016b; Gupta, Bhavsar, & Ieee, 2017) have implemented a traditional ML-based approach for BC detection using DH images. As proof, Belsare et al. (2015) used an exclusive DH image dataset for BC detection as non-malignant and malignant tissues. The author extracted GLCM, Graph Run Length Matrix (GRLM) features, and

Euler number followed by the spatio-color-texture graph segmentation method. The LDA, kNN and SVM traditional ML classifiers were employed for classification. Where LDA outperformed the others by showing an accuracy of 80% for malignancy detection. However, Belsare’s BC detection model needs to be tested over any standard dataset for fair comparison of results. Spanhol et al. (2016b) hosted BreakHis, a bit larger, multifaceted, publicly available standard dataset, which carries 7909 DH images of 82 patients classified into eight subtypes of BT. The author detected BC as benign or malignant using various FEMs namely LBP, completed LBP (CLBP), local phase quantization (LPQ), GLCM, threshold adjacency statistics (TAS), and ORB. Next, the extracted features were classified via

kNN, quadratic linear analysis (QDA), SVM, and random forest (RF) classifiers. Where, SVM outperformed the others by getting 85% accuracy and 86.1% AUC for binary classification. Moreover, Gupta et al. (2017) performed BC detection using the BreakHis dataset. The author created a fusion of six texture features, such as normalized color space representation, multilayer coordinate clusters representation, Gabor features on Gaussian color model, Gabor chromatic features, complex wavelet features, and chromatic features, and opponent color local binary pattern (OCLBP). Afterward, classification is made based on voting through various traditional ML classifiers, namely SVM, kNN, DT, DA, and ensemble classifiers. The reported accuracy is 88.69%.

Table , that most traditional ML-based models are unable to get better accuracy compared to DL based models. Nonetheless, DL based models consumed high computational resources, longer training time, and need a large number of annotated images to avoid overfitting issues. It is a cumbersome task to collect a large number of labelled medical images. Conversely, traditional ML-based models Table , that the proposed model achieved better Sn, and MCC, even higher than DL based models. The proposed model obtained 90% results for each Ac and Fm with MCC value 0.8 using 51 (20.31% reduced) features for the initial testing set. However, for extended testing set, the proposed model acquired 93% accuracy, 94.12% Fm, and 0.8819 MCC value using 3 (95.31% reduced) features. Further results like Pr and confusion matrix can be seen in

are efficient (in terms of computational resources and time) and usually do not require a large number of labelled images, therefore suitable for medical image data. Thus, in this study, we aim to achieve better accuracy while using the least resources (normal desktop computer), training time, and a small number of labelled images. It can be seen from

Table 5 and Table 9. Thus, the proposed model is an efficient, feasible (can show better results on a small number of images), and cost-effective solution for BC automatic detection using a desktop computer. Moreover, it can be used for other applications, where time for detection matters like content-based image retrieval

Table VI. Performance evaluation metrics comparison of the proposed model to the state-of-the-art exiting models

Referenc e	Classific ation approach	Ac (%)	Sn (%)	Class labels	Dataset type, name	Limitations
Wan et al. (2014)	Tradition al Machine Learning	87.97	86.8	Mitotic/ Non- mitotic	Public, MITOS	• Needs to improve performance.

Reference	Classification approach	Ac (%)	Sn (%)	Class labels	Dataset type, name	Limitations
Belsare et al. (2015)	Traditional Machine Learning	80	100	Non-malignant/Malignant	Exclusive, Not given	<ul style="list-style-type: none"> Needs to improve performance of Malignancy, The use of a public dataset can show different results.
Gupta et al. (2017)	Traditional Machine Learning	88.69	---	Benign/Malignant	Public, BreakHis	<ul style="list-style-type: none"> Needs to improve performance, Sn, Fm, and MCC need to be reported
Spanhol et al. (2016b)	Traditional Machine Learning	85	---	Benign/Malignant	Public, BreakHis	<ul style="list-style-type: none"> Needs to improve performance, Sn, Fm, and MCC need to be reported.
Song et al. (2017)	Deep Learning	86.2 ± 3.7	---	Benign/Malignant	Public, BreakHis	<ul style="list-style-type: none"> Needs to improve performance, Sn, Fm, and MCC need to be reported, It requires high computational resources and training time.
Spanhol et al. (2017)	Deep Learning	84.2 ± 1.7	---	Benign/Malignant	Public, BreakHis	<ul style="list-style-type: none"> Needs to improve performance, Sn, Fm, and MCC need to be reported, It requires high computational resources and training time.
Gupta and Bhavsar (2018)	Deep Learning	95.9 ± 4.2	---	Benign/Malignant	Public, BreakHis	<ul style="list-style-type: none"> Only Ac is reported which can be biased.

Reference	Classification approach	Ac (%)	Sn (%)	Class labels	Dataset type, name	Limitations
						<ul style="list-style-type: none"> • Sn, Fm, and MCC need to be reported, • Requires very high computational resources and training time,
Spanhol et al. (2016a)	Deep Learning	85.6 ± 4.8	---	Benign/Malignant	Public, BreakHis	<ul style="list-style-type: none"> • Needs to improve performance, • Sn, Fm, and MCC need to be reported, • It requires very high computational resources and training time.
Janowczyk and Madabhushi (2016)	Deep Learning	84.68	---	Invasive/ In situ	Exclusive, Not given	<ul style="list-style-type: none"> • Needs to improve performance, • Sn, Fm, and MCC need to be reported, • Requires very high computational resources and training time, • The use of a public dataset can show different results.
Huang and Chung (2018)	Deep Learning	98.5	---	Carcinoma/ Non-carcinoma	Public, BCBH	<ul style="list-style-type: none"> • Sn, Fm, and MCC need to be reported, • Requires very high computational resources and training time, • Need to be tested using a larger size dataset.

Reference	Classification approach	Ac (%)	Sn (%)	Class labels	Dataset type, name	Limitations
Araujo et al. (2017)	Deep Learning	Initial = 90, Extended = 81.3	Initial = (In situ = 80, Invasive = 100), Extended = (In situ = 50, Invasive = 100)	In situ/Invasive	Public, BCBH	<ul style="list-style-type: none"> Needs to improve performance, Requires high computational resources and training time, Need to be tested using a larger size dataset.
Proposed	Traditional Machine Learning (GK, kNN)	Initial = 90, Extended = 93	Initial = (In situ = 90, Invasive = 90), Extended = (In situ = 100, Invasive = 87.50)	In situ (i.e., Non-carcinoma) / Invasive (i.e., Carcinoma)	Public, BCBH	<ul style="list-style-type: none"> Need to be tested using a larger size dataset.

6. Conclusion

An automated breast cancer detection is performed by classifying histology images in carcinoma and non-carcinoma cancer types. A publicly available BCBH dataset of DH images was utilized to develop the BC detection model. Next, a few images pre-processing tasks like stain normalization, training set augmentation, and rescaling is

performed to enhance the BC detection performance. Where stain normalization harmonizes the inherent color inconsistencies of DH images and image augmentation is required to create synthetic sample data to improve classifiers' detection performance. Moreover, rescaling is adopted to extract features efficiently from smaller size images. The preprocessed images are used to extract both local and

global types of features using ten feature extraction methods. An MFT is created for each feature extraction method using an augmented training set, initial testing set, and extended testing set. Afterwards, the features of all MFTs are sorted by their weights calculated through the NCA method to create SMFTs. An algorithm is developed to select the minimum number of features from SMFTs to train seven heterogeneous ML classifiers. It has been observed from numerous experiments that the performance of ML classifiers using SMFTs is fluctuating up to 59 features and then started degrading as the weight is decreasing. Furthermore, the goal of the feature selection algorithm is to look for the best performing feature extraction method (out of ten), the best ML classifier (out of seven) using the minimum number of features. After performing many experiments, it has been found that for the initial testing set GK outperformed the other ML classifiers using only 51 (out of 64) Harris features by showing 90% results for each like Ac, Sn, and Fm with 0.8 MCC value. On the other hand, while using extended testing kNN outperformed the rest of the classifiers using only 3 (out of 64) Harris features by achieving 93% Ac, 94% Fm, and 0.8819 MCC value. Thus, it can be concluded that Harris features outperformed all feature extraction methods. The performance of the proposed BC detection model shows that the model is efficient (consumes less training and detection time), cost-effective (can be trained on a normal desktop computer), and feasible (requires fewer number of labelled images to show good results) solution for BC detection using DH images. Thus, the model can be implemented efficiently in any health care centre to assist doctors as a second opinion for BC early detection using DH images, especially in the low privileged area of a country. As future work, the proposed model can be tested by using other datasets with a large number of BC DH images. Moreover, other types of texture features and the wavelet-based statistical features can be extracted to make further analysis.

Acknowledgment: This work was supported by University Malaya Research Grant Program – [grant number GPF016D-2019]

AUTHOR CONTRIBUTION: Ghulam Murtaza(GM), Ainuddin Wahid Abdul Wahab(AWAW), Ghulam Raza(GR), Liyana Shuib(LS)

Data curation: GM, LS.

Methodology: GM, AWAW.

Software: GM.

Supervision: AWAW, LS.

Writing – original draft: GM.

Writing – review & editing: GM, AWAW, GR, LS.

Funding: AWAW.

DATA AVAILABILITY STATEMENT: A publicly available dataset is used for this research work.

URL: <https://rdm.inesctec.pt/dataset/nis-2017-003>.

CONFLICT OF INTEREST: None

FUNDING: This work was supported by University Malaya Research Grant Program – [grant number GPF016D-2019]

References

- [1] Alcantarilla, P. F., Bartoli, A., & Davison, A. J. (2012). KAZE features. Paper presented at the European Conference on Computer Vision.
- [2] American Cancer Society. (2018). Breast Cancer Facts & Figures 2017-2018. Atlanta: American Cancer Society, Inc. 2017. Retrieved 03/08/2019
- [3] Araujo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., . . . Campilho, A. (2017). Classification of breast cancer histology images using Convolutional Neural Networks. *Plos One*, 12(6), 14. doi: 10.1371/journal.pone.0177544
- [4] Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-up robust features (SURF). *Computer vision and image understanding*, 110(3), 346-359.

- [5] Belsare, A. D., Mushrif, M. M., Pangarkar, M. A., & Meshram, N. (2015). Classification of Breast Cancer Histopathology Images using Texture Feature Analysis Tencon 2015 - 2015 Ieee Region 10 Conference. New York: Ieee.
- [6] Bige, O., Demir, A., Saygili, U., Gode, F., Uslu, T., & Koyuncuoglu, M. (2011). Frozen section diagnoses of 578 ovarian tumors made by pathologists with and without expertise on gynecologic pathology. *Gynecologic Oncology*, 123(1), 43-46. doi: <https://doi.org/10.1016/j.ygyno.2011.06.030>
- [7] Biswas, R., Nath, A., & Roy, S. (2016). Mammogram Classification using Gray-Level Cooccurrence Matrix for Diagnosis of Breast Cancer. New York: Ieee.
- [8] BREASTCANCER.ORG. (2015, October 7, 2015). Biopsy techniques. from <https://www.breastcancer.org/symptoms/testing/types/biopsy>
- [9] Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection.
- [10] Desjardins, A. U. (1960). Is the Pathologist Infallible? *JAMA Internal Medicine*, 106(5), 596-602. doi: 10.1001/archinte.1960.03820050008003
- [11] Gupta, V., & Bhavsar, A. (2018). Sequential modeling of deep features for breast cancer histopathological image classification. Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.
- [12] Gupta, V., Bhavsar, A., & Ieee. (2017). Breast Cancer Histopathological Image Classification: Is Magnification Important? 2017 Ieee Conference on Computer Vision and Pattern Recognition Workshops (pp. 769-776). New York: Ieee.
- [13] Gurcan, M. N., Boucheron, L. E., Can, A., Madabhushi, A., Rajpoot, N. M., & Yener, B. (2009). Histopathological image analysis: a review. *IEEE reviews in biomedical engineering*, 2, 147-171. doi: 10.1109/RBME.2009.2034865
- [14] Han, Z., Wei, B., Zheng, Y., Yin, Y., Li, K., & Li, S. (2017). Breast Cancer Multi-classification from Histopathological Images with Structured Deep Learning Model. *Scientific Reports*, 7(1). doi: 10.1038/s41598-017-04075-z
- [15] Harris, C. G., & Stephens, M. (1988). A combined corner and edge detector. Paper presented at the Alvey vision conference.
- [16] Hassanien, A. E., & Ali, J. M. (2006). Rough set approach for classification of breast cancer mammogram images. In V. DiGesù, F. Masulli & A. Petrosino (Eds.), *Fuzzy Logic and Applications* (Vol. 2955, pp. 224-231). Berlin: Springer-Verlag Berlin.
- [17] Huang, Y., & Chung, A. C.-S. (2018, 2018//). Improving High Resolution Histology Image Classification with Deep Spatial Fusion Network. Paper presented at the Computational Pathology and Ophthalmic Medical Image Analysis, Cham.
- [18] Janowczyk, A., & Madabhushi, A. (2016). Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of Pathology Informatics*, 7(1), 29-29. doi: 10.4103/2153-3539.186902
- [19] Jianbo, S., & Tomasi, C. (1994). Good features to track. Paper presented at the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.
- [20] Khan, A. M., Rajpoot, N., Treanor, D., & Magee, D. (2014). A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. *IEEE Transactions on Biomedical Engineering*, 61(6), 1729-1738.
- [21] Komura, D., & Ishikawa, S. (2018). Machine Learning Methods for

- Histopathological Image Analysis. Computational and Structural Biotechnology Journal, 16, 34-42. doi: <https://doi.org/10.1016/j.csbj.2018.01.001>
- [22] Kontos, K., & Maragoudakis, M. (2013). Breast Cancer Detection in Mammogram Medical Images with Data Mining Techniques. In H. Papadopoulos, A. S. Andreou, L. Iliadis & I. Maglogiannis (Eds.), Artificial Intelligence Applications and Innovations, Aiai 2013 (Vol. 412, pp. 336-347). Berlin: Springer-Verlag Berlin.
- [23] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Paper presented at the Advances in neural information processing systems.
- [24] Leutenegger, S., Chli, M., & Siegwart, R. (2011). BRISK: Binary robust invariant scalable keypoints. Paper presented at the 2011 IEEE international conference on computer vision (ICCV).
- [25] Macenko, M., Niethammer, M., Marron, J. S., Borland, D., Woosley, J. T., Guan, X., . . . Thomas, N. E. (2009). A method for normalizing histology slides for quantitative analysis. Paper presented at the Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on.
- [26] Matas, J., Chum, O., Urban, M., & Pajdla, T. (2004). Robust wide-baseline stereo from maximally stable extremal regions. Image and Vision Computing, 22(10), 761-767. doi: <https://doi.org/10.1016/j.imavis.2004.02.006>
- [27] McCann, M. T., Ozolek, J. A., Castro, C. A., Parvin, B., & Kovacevic, J. (2015). Automated histology analysis: Opportunities for signal processing. IEEE Signal Processing Magazine, 32(1), 78-87.
- [28] Moravec, H. P. (1980). Obstacle avoidance and navigation in the real world by a seeing robot rover: Stanford Univ CA Dept of Computer Science.
- [29] Murtaza, G., Shuib, L., Abdul Wahab, A. W., Mujtaba, G., Mujtaba, G., Nweke, H. F., . . . Azmi, N. A. (2020). Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges. Artificial Intelligence Review, 53(3), 1655-1720. doi: 10.1007/s10462-019-09716-5
- [30] Murtaza, G., Shuib, L., Mujtaba, G., & Raza, G. (2019). Breast Cancer Multi-classification through Deep Neural Network and Hierarchical Classification Approach. Multimedia Tools and Applications. doi: 10.1007/s11042-019-7525-4
- [31] Murtaza, G., Shuib, L., Wahab, A. W. A., Mujtaba, G., Raza, G., & Azmi, N. A. (2019). Breast cancer classification using digital biopsy histopathology images through transfer learning. Paper presented at the Journal of Physics: Conference Series.
- [32] Murtaza, G., Shuib, L., Wahab, A. W. A., Mujtaba, G., Raza, G. J. M. T., & Applications. (2020). Ensembled deep convolution neural network-based breast cancer classification with misclassification reduction algorithms. 1-33.
- [33] Nusantara, A. C., Purwanti, E., & Soelistiono, S. (2016). CLASSIFICATION OF DIGITAL MAMMOGRAM BASED ON NEAREST-NEIGHBOR METHOD FOR BREAST CANCER DETECTION. International Journal of Technology, 7(1), 71-77. doi: 10.14716/ijtech.v7i1.1393
- [34] Ojala, T., Pietikäinen, M., & Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis & Machine Intelligence(7), 971-987.
- [35] Ponraj, N., Poongodi, & Mercy, M. (2017). Texture Analysis of Mammogram for the Detection of

- Breast Cancer using LBP and LGP: A Comparison 2016 Eighth International Conference on Advanced Computing (pp. 182-185). New York: Ieee.
- [36] Reinhard, E., Adhikhmin, M., Gooch, B., & Shirley, P. (2001). Color transfer between images. *IEEE Computer graphics and applications*, 21(5), 34-41.
- [37] Rosten, E., & Drummond, T. (2005). Fusing points and lines for high performance tracking. Paper presented at the ICCV.
- [38] Rubin, R., Strayer, D. S., & Rubin, E. (2008). *Rubin's pathology: clinicopathologic foundations of medicine*: Lippincott Williams & Wilkins.
- [39] Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. R. (2011). ORB: An efficient alternative to SIFT or SURF. Paper presented at the ICCV.
- [40] Song, Y., Zou, J. J., Chang, H., & Cai, W. (2017). Adapting fisher vectors for histopathology image classification. Paper presented at the 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017).
- [41] Spanhol, F. A., Oliveira, L. S., Cavalin, P. R., Petitjean, C., & Heutte, L. (2017). Deep features for breast cancer histopathological image classification. Paper presented at the Systems, Man, and Cybernetics (SMC), 2017 IEEE International Conference on.
- [42] Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2016a). Breast cancer histopathological image classification using convolutional neural networks. Paper presented at the Neural Networks (IJCNN), 2016 International Joint Conference on.
- [43] Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2016b). A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7), 1455-1462.
- [44] Tabar, L., Vitak, B., Chen, T. H., Yen, A. M., Cohen, A., Tot, T., . . . Duffy, S. W. (2011). Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades. *Radiology*, 260(3), 658-663. doi: 10.1148/radiol.11110469
- [45] Teresa Araújo, G. A., Eduardo Castro, José Rouco, Paulo Aguiar, Catarina Eloy, António Polónia, Aurélio Campilho. (2015, November 15, 2018, 7:19 PM (UTC+08:00)). Bioimaging Challenge 2015 Breast Histology Dataset. from <https://rdm.inesctec.pt/dataset/nis-2017-003>
- [46] Wan, T., Cao, J., Chen, J., & Qin, Z. (2017). Automated grading of breast cancer histopathology using cascaded ensemble with combination of multi-level image features. *Neurocomputing*, 229, 34-44. doi: <https://doi.org/10.1016/j.neucom.2016.05.084>
- [47] Wan, T., Liu, X., Chen, J., & Qin, Z. (2014, 27-30 Oct. 2014). Wavelet-based statistical features for distinguishing mitotic and non-mitotic cells in breast cancer histopathology. Paper presented at the 2014 IEEE International Conference on Image Processing (ICIP).
- [48] Wang, D., Khosla, A., Gargeya, R., Irshad, H., & Beck, A. H. (2016). Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*.
- [49] Weickert, J., Ishikawa, S., & Imiya, A. (1999). Linear scale-space has first been proposed in Japan. *Journal of Mathematical Imaging and Vision*, 10(3), 237-252.
- [50] Zheng, Y., Jiang, Z., Xie, F., Zhang, H., Ma, Y., Shi, H., & Zhao, Y. (2017). Feature extraction from histopathological images based on nucleus-guided convolutional neural network for breast lesion classification. *Pattern Recognition*, 71, 14-25. doi: 10.1016/j.patcog.2017.05.010

Appendix-A

Table 1: Other Performance Evaluation Metrics of using the original image

		Initial Testing Set							Extended Testing Set							
		SVM	NB	LR	LDA	kNN	GK	DT	SVM	NB	LR	LDA	kN N	GK	DT	
HOG	Sp	70.00	20.00	100.00	60.00	10.00	40.00	70.00	62.50	12.50	100.00	62.50	12.50	62.50	37.50	
	Pr	72.73	55.56	NaN	66.67	52.63	45.46	62.50	66.67	50.00	NaN	62.50	53.33	62.50	44.44	
	Confusion Matrix	T P	8	10	0	8	10	5	5	6	7	0	5	8	5	4
		F N	2	0	10	2	0	5	5	2	1	8	3	0	3	4
		F P	3	8	0	4	9	6	3	3	7	0	3	7	3	5
		T N	7	2	10	6	1	4	7	5	1	8	5	1	5	3
MinEigen	Sp	80.00	90.00	50.00	60.00	50.00	70.00	20.00	37.50	62.50	75.00	62.50	37.50	62.50	75.00	
	Pr	66.67	66.67	37.50	33.33	50.00	70.00	38.46	28.57	40.00	50.00	57.14	44.44	25.00	71.43	
	Confusion Matrix	T P	4	2	3	2	5	7	5	2	2	2	4	4	1	5
		F N	6	8	7	8	5	3	5	6	6	6	4	4	7	3
		F P	2	1	5	4	5	3	8	5	3	2	3	5	3	2
		T N	8	9	5	6	5	7	2	3	5	6	5	3	5	6
Harris	Sp	40.00	90.00	80.00	80.00	60.00	40.00	40.00	50.00	50.00	37.50	37.50	50.00	75.00	62.50	
	Pr	40.00	75.00	66.67	66.67	55.56	53.85	45.46	60.00	50.00	28.57	37.50	42.86	66.67	40.00	
	Confusion Matrix	T P	4	3	4	4	5	7	5	6	4	2	3	3	4	2
		F N	6	7	6	6	5	3	5	2	4	6	5	5	4	6
		F P	6	1	2	2	4	6	6	4	4	5	5	4	2	3
		T N	4	9	8	8	6	4	4	4	4	3	3	4	6	5
LBP	Sp	50.00	50.00	50.00	40.00	80.00	40.00	50.00	75.00	87.50	75.00	62.50	62.50	87.50	50.00	
	Pr	61.54	58.33	64.29	57.14	77.78	57.14	61.54	66.67	80.00	66.67	66.67	62.50	80.00	33.33	
	Confusion Matrix	T P	8	7	9	8	7	8	8	4	4	4	6	5	4	2
		F N	2	3	1	2	3	2	2	4	4	4	2	3	4	6
		F P	5	5	5	6	2	6	5	2	1	2	3	3	1	4
		T N	5	5	5	4	8	4	5	6	7	6	5	5	7	4
KAZ	Sp	90.00	90.00	90.00	50.00	80.00	70.00	50.00	100.00	87.50	100.00	62.50	50.00	100.00	37.50	

		Initial Testing Set							Extended Testing Set						
		SVM	NB	LR	LDA	kNN	GK	DT	SVM	NB	LR	LDA	kN N	GK	DT
Confusion Matrix	Pr	66.67	66.67	66.67	50.00	71.43	40.00	50.00	100.00	66.67	100.00	40.00	42.86	100.00	44.44
	TP	2	2	2	5	5	2	5	1	2	1	2	3	1	4
	FN	8	8	8	5	5	8	5	7	6	7	6	5	7	4
	FP	1	1	1	5	2	3	5	0	1	0	3	4	0	5
	TN	9	9	9	5	8	7	5	8	7	8	5	4	8	3
MSER	Sp	90.00	50.00	90.00	90.00	80.00	100.00	80.00	87.50	75.00	87.50	100.00	75.00	87.50	87.50
	Pr	66.67	64.29	80.00	75.00	60.00	100.00	66.67	75.00	71.43	75.00	100.00	75.00	66.67	66.67
	TP	2	9	4	3	3	4	4	3	5	3	5	6	2	2
	FN	8	1	6	7	7	6	6	5	3	5	3	2	6	6
	FP	1	5	1	1	2	0	2	1	2	1	0	2	1	1
SURF	Sp	40.00	70.00	80.00	90.00	100.00	50.00	90.00	62.50	100.00	100.00	75.00	75.00	62.50	75.00
	Pr	57.14	0.00	0.00	0.00	100.00	61.54	50.00	57.14	100.00	NaN	33.33	60.00	57.14	60.00
	TP	8	0	0	0	1	8	1	4	1	0	1	3	4	3
	FN	2	10	10	10	9	2	9	4	7	8	7	5	4	5
	FP	6	3	2	1	0	5	1	3	0	0	2	2	3	2
FAST	Sp	50.00	100.00	100.00	100.00	90.00	20.00	80.00	25.00	62.50	75.00	62.50	62.50	50.00	62.50
	Pr	64.28	100.00	100.00	100.00	50.00	46.67	50.00	57.14	62.50	50.00	50.00	62.50	55.56	40.00
	TP	9	2	2	2	1	7	2	8	5	2	3	5	5	2
	FN	1	8	8	8	9	3	8	0	3	6	5	3	3	6
	FP	5	0	0	0	1	8	2	6	3	2	3	3	4	3
BRISK	Sp	50.00	90.00	80.00	80.00	90.00	60.00	80.00	12.50	50.00	87.50	75.00	62.50	37.50	75.00
	Pr	61.54	50.00	50.00	50.00	66.67	60.00	33.33	50.00	60.00	66.67	71.43	62.50	44.44	71.43
	TP	8	1	2	2	2	6	1	7	6	2	5	5	4	5
	FN	2	9	8	8	8	4	9	1	2	6	3	3	4	3

		Initial Testing Set							Extended Testing Set						
		SVM	NB	LR	LDA	kNN	GK	DT	SVM	NB	LR	LDA	kN N	GK	DT
		FP	5	1	2	2	1	4	2	7	4	1	2	3	5
TN	5	9	8	8	9	6	8	1	4	7	6	5	3	6	
ORB	Sp	50.00	40.00	70.00	60.00	60.00	60.00	30.00	50.00	50.00	50.00	50.00	50.00	12.50	62.50
	Pr	50.00	50.00	57.14	50.00	69.23	55.56	41.67	60.00	42.86	50.00	50.00	50.00	41.67	50.00
	TP	5	6	4	4	9	5	5	6	3	4	4	4	5	3
	FN	5	4	6	6	1	5	5	2	5	4	4	4	3	5
	FP	5	6	3	4	4	4	7	4	4	4	4	4	7	3
	TN	5	4	7	6	6	6	3	4	4	4	4	4	1	5

Table 2: Other Performance Evaluation Metrics of using image augmentation algorithm

		Initial Testing Set							Extended Testing Set						
		SV M	NB	LR	LDA	kNN	GK	DT	SVM	NB	LR	LDA	kN N	GK	DT
		Sp	40.00	60.00	40.00	90.00	60.00	50.00	70.00	75.00	25.00	75.00	62.50	75.00	87.50
Pr	60.00	69.23	62.50	87.50	69.23	64.29	72.73	77.78	57.14	60.00	70.00	75.00	83.33	44.44	
HOG	TP	9	9	10	7	9	9	8	7	8	3	7	6	5	4
	FN	1	1	0	3	1	1	2	1	0	5	1	2	3	4
	FP	6	4	6	1	4	5	3	2	6	2	3	2	1	5
	TN	4	6	4	9	6	5	7	6	2	6	5	6	7	3
	Sp	30.00	40.00	90.00	70.00	100.00	60.00	60.00	0.00	50.00	12.50	87.50	87.50	62.50	75.00
Pr	58.82	62.50	87.50	75.00	100.00	63.64	66.67	50.00	60.00	50.00	87.50	85.71	70.00	66.67	
MinEigen	TP	10	10	7	9	7	7	8	8	6	7	7	6	7	4
	FN	0	0	3	1	3	3	2	0	2	1	1	2	1	4
	FP	7	6	1	3	0	4	4	8	4	7	1	1	3	2
	TN	3	4	9	7	10	6	6	0	4	1	7	7	5	6
	Sp	50.00	30.00	50.00	50.00	40.00	60.00	70.00	37.50	62.50	37.50	87.50	75.00	37.50	37.50
Pr	66.67	56.25	61.54	61.54	60.00	69.23	70.00	61.54	72.73	50.00	83.33	77.78	58.33	37.50	
Harris	TP	10	9	8	8	9	9	7	8	8	5	5	7	7	3
	FN	0	1	2	2	1	1	3	0	0	3	3	1	1	5

		Initial Testing Set							Extended Testing Set							
		SV M	NB	LR	LDA	kNN	GK	DT	SVM	NB	LR	LDA	kN N	GK	DT	
		FP	T N	FP	T N	FP	T N	FP	T N	FP	T N	FP	T N	FP	T N	
LBP	Sp	30.00	50.00	50.00	50.00	70.00	50.00	50.00	37.50	37.50	37.50	75.00	62.50	37.50	50.00	
	Pr	58.82	64.29	64.29	64.29	75.00	64.29	66.67	61.54	58.33	44.44	77.78	72.73	61.54	60.00	
	Confusion Matrix	TP	10	9	9	9	9	9	10	8	7	4	7	8	8	6
		FN	0	1	1	1	1	1	0	0	1	4	1	0	0	2
		FP	7	5	5	5	3	5	5	5	5	5	2	3	5	4
		TN	3	5	5	5	7	5	5	3	3	3	6	5	3	4
		Sp	40.00	70.00	80.00	70.00	90.00	90.00	70.00	75.00	62.50	0.00	75.00	37.50	62.50	0.00
Pr	62.50	72.73	80.00	76.92	87.50	87.50	72.73	75.00	66.67	50.00	75.00	61.54	66.67	50.00		
Confusion Matrix	TP	10	8	8	10	7	7	8	6	6	8	6	8	6	8	
	FN	0	2	2	0	3	3	2	2	2	0	2	0	2	0	
	FP	6	3	2	3	1	1	3	2	3	8	2	5	3	8	
	TN	4	7	8	7	9	9	7	6	5	0	6	3	5	0	
	MSER	Sp	50.00	90.00	70.00	80.00	70.00	90.00	80.00	75.00	50.00	62.50	100.00	87.50	62.50	12.50
Pr		61.54	88.89	70.00	77.78	70.00	87.50	81.82	71.43	63.64	50.00	100.00	87.50	70.00	22.22	
Confusion Matrix		TP	8	8	7	7	7	7	9	5	7	3	5	7	7	2
		FN	2	2	3	3	3	3	1	3	1	5	3	1	1	6
		FP	5	1	3	2	3	1	2	2	4	3	0	1	3	7
		TN	5	9	7	8	7	9	8	6	4	5	8	7	5	1
		SURF	Sp	60.00	60.00	70.00	60.00	60.00	70.00	80.00	25.00	37.50	100.00	62.50	62.50	37.50
Pr	69.23		69.23	72.73	69.23	66.67	75.00	75.00	53.85	58.33	NaN	66.67	70.00	58.33	NaN	
Confusion Matrix	TP		9	9	8	9	8	9	6	7	7	0	6	7	7	0
	FN		1	1	2	1	2	1	4	1	1	8	2	1	1	8
	FP		4	4	3	4	4	3	2	6	5	0	3	3	5	0
	TN		6	6	7	6	6	7	8	2	3	8	5	5	3	8
	FAST		Sp	60.00	100.00	100.00	100.00	90.00	70.00	100.00	50.00	75.00	100.00	87.50	75.00	50.00
Pr		71.43	100.00	100.00	100.00	80.00	72.73	100.00	66.67	77.78	NaN	80.00	75.00	63.64	NaN	

		Initial Testing Set							Extended Testing Set						
		SV M	NB	LR	LDA	kNN	GK	DT	SVM	NB	LR	LDA	kN N	GK	DT
Confusion Matrix	TP	10	3	2	2	4	8	4	8	7	0	4	6	7	0
	FN	0	7	8	8	6	2	6	0	1	8	4	2	1	8
	FP	4	0	0	0	1	3	0	4	2	0	1	2	4	0
	TN	6	10	10	10	9	7	10	4	6	8	7	6	4	8
BRISK	Sp	50.0 0	75.00	100.0 0	87.50	87.50	50.0 0	100.0 0	80.00	60.0 0	90.00	100.0 0	70.0 0	60.0 0	100.0 0
	Pr	66.6 7	75.00	NaN	85.71	83.33	63.6 4	NaN	83.33	50.0 0	75.00	100.0 0	57.1 4	71.4 3	100.0 0
	TP	8	6	0	6	5	7	0	10	4	3	3	4	10	3
	FN	0	2	8	2	3	1	8	0	6	7	7	6	0	7
	FP	4	2	0	1	1	4	0	2	4	1	0	3	4	0
	TN	4	6	8	7	7	4	8	8	6	9	10	7	6	10
	Sp	90.0 0	60.00	40.00	40.00	50.00	60.0 0	50.00	62.50	62.5 0	50.00	87.50	62.5 0	62.5 0	37.50
	Pr	88.8 9	63.64	57.14	57.14	64.29	66.6 7	61.54	62.50	66.6 7	42.86	85.71	66.6 7	66.6 7	44.44
ORB	TP	8	7	8	8	9	8	8	5	6	3	6	6	6	4
	FN	2	3	2	2	1	2	2	3	2	5	2	2	2	4
	FP	1	4	6	6	5	4	5	3	3	4	1	3	3	5
	TN	9	6	4	4	5	6	5	5	5	4	7	5	5	3

Table 3: Other Performance Evaluation Metrics of using image augmentation with feature selection algorithm

		Initial Testing Set							Extended Testing Set							
		SVM	NB	LR	LDA	kNN	GK	DT	SVM	NB	LR	LDA	kNN	GK	DT	
HOG	Sp	90.00	50.0 0	70.00	70.00	70.00	70.00	90.00	50.00	62.5 0	62.50	75.00	100.0 0	75.00	87.50	
	Pr	90.00	66.6 7	76.92	76.92	76.92	75.00	88.89	66.67	72.7 3	72.73	77.78	100.0 0	80.00	87.50	
	TP	9	10	10	10	10	9	8	8	8	8	7	7	8	7	
	FN	1	0	0	0	0	1	2	0	0	0	1	1	0	1	
	FP	1	5	3	3	3	3	1	4	3	3	2	0	2	1	
	TN	9	5	7	7	7	7	9	4	5	5	6	8	6	7	
	MinFigen	Sp	60.00	50.0 0	90.00	70.00	100.0 0	90.00	80.00	62.50	62.5 0	75.00	100.0 0	100.0 0	100.0 0	87.50
		Pr	71.43	66.6 7	87.50	75.00	100.0 0	88.89	80.00	72.73	70.0 0	80.00	100.0 0	100.0 0	100.0 0	87.50

		Initial Testing Set							Extended Testing Set						
		SVM	NB	LR	LDA	kNN	GK	DT	SVM	NB	LR	LDA	kNN	GK	DT
Confusion Matrix	TP	10	10	7	9	8	8	8	8	7	8	6	7	7	7
	FN	0	0	3	1	2	2	2	0	1	0	2	1	1	1
	FP	4	5	1	3	0	1	2	3	3	2	0	0	0	1
	TN	6	5	9	7	10	9	8	5	5	6	8	8	8	7
Harris	Sp	70.00	40.00	70.00	60.00	60.00	90.00	70.00	75.00	62.50	62.50	62.50	87.50	75.00	87.50
	Pr	76.92	62.50	75.00	69.23	71.43	90.00	75.00	80.00	72.73	72.73	88.89	80.00	87.50	
	TP	10	10	9	9	10	9	9	8	8	8	8	8	8	7
	FN	0	0	1	1	0	1	1	0	0	0	0	0	0	1
	FP	3	6	3	4	4	1	3	2	3	3	3	1	2	1
	TN	7	4	7	6	6	9	7	6	5	5	5	7	6	7
	Sp	50.00	40.00	60.00	60.00	90.00	50.00	60.00	50.00	62.50	62.50	87.50	62.50	62.50	100.00
	Pr	66.67	62.50	71.43	69.23	88.89	66.67	71.43	66.67	70.00	70.00	87.50	72.73	72.73	100.00
	TP	10	10	10	9	8	10	10	8	7	7	7	8	8	6
	FN	0	0	0	1	2	0	0	0	1	1	1	0	0	2
FP	5	6	4	4	1	5	4	4	3	3	1	3	3	0	
TN	5	4	6	6	9	5	6	4	5	5	7	5	5	8	
KAZE	Sp	60.00	80.00	90.00	80.00	90.00	70.00	80.00	75.00	87.50	75.00	62.50	75.00	75.00	75.00
	Pr	71.43	81.82	90.00	83.33	90.00	76.92	81.82	77.78	85.71	77.78	72.73	80.00	75.00	80.00
	TP	10	9	9	10	9	10	9	7	6	7	8	8	6	8
	FN	0	1	1	0	1	0	1	1	2	1	0	0	2	0
	FP	4	2	1	2	1	3	2	2	1	2	3	2	2	2
	TN	6	8	9	8	9	7	8	6	7	6	5	6	6	6
MSER	Sp	70.00	80.00	70.00	80.00	80.00	100.00	80.00	62.50	50.00	87.50	87.50	87.50	50.00	75.00
	Pr	75.00	81.82	75.00	80.00	80.00	100.00	81.82	66.67	63.64	83.33	85.71	87.50	66.67	80.00
	TP	9	9	9	8	8	8	9	6	7	5	6	7	8	8
	FN	1	1	1	2	2	2	1	2	1	3	2	1	0	0
	FP	3	2	3	2	2	0	2	3	4	1	1	1	4	2

		Initial Testing Set						Extended Testing Set							
		SVM	NB	LR	LDA	kNN	GK	DT	SVM	NB	LR	LDA	kNN	GK	DT
SURF	TN	7	8	7	8	8	10	8	5	4	7	7	7	4	6
	Sp	50.00	50.00	70.00	80.00	70.00	60.00	80.00	62.50	50.00	50.00	50.00	100.00	50.00	75.00
	Pr	66.67	66.67	75.00	81.82	75.00	69.23	80.00	70.00	63.64	63.64	63.64	100.00	63.64	77.78
	TP	10	10	9	9	9	9	8	7	7	7	7	6	7	7
	FN	0	0	1	1	1	1	2	1	1	1	1	2	1	1
	FP	5	5	3	2	3	4	2	3	4	4	4	0	4	2
	TN	5	5	7	8	7	6	8	5	4	4	4	8	4	6
	Sp	70.00	80.00	100.00	100.00	100.00	60.00	100.00	50.00	87.50	100.00	100.00	100.00	75.00	87.50
	Pr	76.92	66.67	100.00	100.00	100.00	71.43	100.00	66.66	87.50	100.00	100.00	100.00	80.00	85.71
	TP	10	4	3	3	4	10	4	8	7	4	5	5	8	6
FN	0	6	7	7	6	0	6	0	1	4	3	3	0	2	
FP	3	2	0	0	0	4	0	4	1	0	0	0	2	1	
TN	7	8	10	10	10	6	10	4	7	8	8	8	6	7	
BRISK	Sp	80.00	90.00	100.00	100.00	100.00	70.00	90.00	50.00	75.00	100.00	87.50	100.00	62.50	87.50
	Pr	83.33	80.00	100.00	100.00	100.00	76.92	80.00	66.67	75.00	100.00	85.71	100.00	72.73	87.50
	TP	10	4	3	3	4	10	4	8	6	3	6	6	8	7
	FN	0	6	7	7	6	0	6	0	2	5	2	2	0	1
	FP	2	1	0	0	0	3	1	4	2	0	1	0	3	1
	TN	8	9	10	10	10	7	9	4	6	8	7	8	5	7
ORB	Sp	70.00	60.00	80.00	70.00	80.00	80.00	80.00	62.50	75.00	62.50	50.00	62.50	100.00	100.00
	Pr	75.00	69.23	81.82	76.92	81.82	80.00	80.00	72.73	80.00	72.73	66.67	72.73	100.00	100.00
	TP	9	9	9	10	9	8	8	8	8	8	8	8	6	6
	FN	1	1	1	0	1	2	2	0	0	0	0	0	2	2
	FP	3	4	2	3	2	2	2	3	2	3	4	3	0	0
	TN	7	6	8	7	8	8	8	5	6	5	4	5	8	8

Table 4: List of abbreviations

Abbreviation	Full-Form
Ac	Accuracy
BC	Breast cancer
BCBH	Bioimaging challenge 2015 breast histology
BreakHis	Breast cancer histopathological image classification
BRISK	Binary robust invariant scalable keypoint
BT	Breast tumor
DH	Digital histology
DT	Decision tree
FAST	Features from accelerated segment test
FEMs	Feature extraction methods
Fm	F-measure
HoG	Histogram-oriented gradients
kNN	kNearest Neighbors
LBP	Local binary patterns
LR	Linear regression
MFT	Master feature table
MG	Mammogram
MinEigen	Minimum eigenvalue
ML	Machine learning
MRI	Magnetic resonance imaging
MSER	Maximally stable extremal regions
NB	Naive Bayes
ORB	Oriented fast and rotated brief
PEMs	Performance evaluation metrics
SMFT	Sorted master feature table
Sn	Sensitivity
Sp	Specificity
SURF	Speed-up robust feature
SVM	Support vector machine
US	Ultrasound
WSI	Whole slide image