Breast Cancer Multi-classification through Deep Neural Network and Hierarchical Classification Approach

Ghulam Murtaza[1, 2*], Liyana Shuib[1*], Ghulam Mujtaba[1, 2], Ghulam Raza[3]
[1]Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, 50603, Kuala Lumpur, Malaysia
[2]Department of Computer Science, Sukkur IBA University, Sukkur, Pakistan.
[3]St. James's Hospital, James's Street, Ushers, Dublin 8, Ireland.
Email: gmurtaza@iba-suk.edu.pk, liyanashuib@um.edu.my, mujtaba@iba-suk.edu.pk, ghulam.raza@ymail.com
*Corresponding Authors: gmurtaza@iba-suk.edu.pk, liyanashuib@um.edu.my

Abstract

Breast cancer (BC) is the third leading cause of deaths in women globally. In general, histopathology images are recommended for early diagnosis and detailed analysis for BC. Thus, state-of-the-art classification models are required for the early prediction of BC using histopathology images. This study aims to develop an accurate and computationally feasible classification model named Biopsy Microscopic Image Cancer Network (BMIC_Net) to classify BC into eight distinct subtypes through deep learning (DL) and hierarchical classification approach. For experiments, the publicly available dataset BreakHis is used and splitted into training and testing set. Furthermore, data augmentation was performed on training set only and 4096 result-oriented features were extracted through DL. In order to improve the classification performance, feature reduction schemes were experimented to elicit the most discriminative feature subset. Finally, six machine-learning algorithms were analyzed to acquire the best results. The experimental results revealed that BMIC_Net outperformed existing baseline models by obtaining the highest accuracy of 95.48% for first-level classifier and 94.62% and 92.45% for second-level classifiers. Thus, this model can be deployed on a normal desktop machine in any healthcare center of less privileged areas in under-developing countries to serve as second opinion for breast cancer classification.

Keywords: Breast Cancer, Deep Learning and Transfer Learning, Hierarchical Classification, Histopathology Images, Image Classification, Convolutional Neural Network

1. Introduction

Cancer-related mortality has drastically increased in recent years. As per World Health Organization (WHO) report (WHO, 2018), cancer is the leading cause of death, and approximately 8.8 million people have died globally in 2015 due to this disease. In addition, the number of new cancer cases is expected to increase by 70% in the next two decades. Among the various types of cancer, breast cancer (BC) is the most common among women and is the third leading cause of cancer-related deaths (1.7 million, 11.3%) (WHO, 2018). Early and precise diagnosis is important to improve the prognosis and increase the survival rate of patients with BC by 30% to 50%. Histopathological imaging (HI) is more commonly used for detection of BC compared with other imaging technologies, such as computed tomography, magnetic resonance imaging (MRI), and mammography (Kasban, El-Bendary, & Salama, 2015). Nonetheless, manual histopathological image analysis has three major limitations (Gurcan et al., 2009). First, expert pathologists are rare in healthcare organizations in several developing countries. Second, the procedure is cumbersome and time consuming for pathologists. Therefore, pathologists may experience fatigue and reduced attention during the image analysis. Finally, a reliable analysis is highly dependent on the professional experience, expertise, and domain knowledge of pathologists. Thus, the aforementioned limitations may cause misdiagnosis of histopathology image analysis for BC and may lead to unreliable and incorrect treatment. Hence, to address the above-mentioned limitations, computer-aided diagnostic systems can be used as a second opinion to analyze the histopathology images for BC.

Recent studies have developed BC classification models based on histopathology images (Al-masni et al., 2018; Chougrad, Zouaki, & Alheyane, 2018; Ribli, Horvath, Unger, Pollner, & Csabai, 2018). In general, these models were developed using either traditional machine learning (ML) or deep learning (DL) approaches. In traditional ML-based classification approaches, initially the collected labeled images are preprocessed. Afterwards, the result-oriented handcrafted features (HF) are extracted in order to form a master feature vector (MFV). Finally, the MFV is fed as an input to traditional ML algorithms to construct the BC classification model. However, the constructed model is evaluated by using various performance measures and by comparing its performance with existing state-of-the-art traditional ML approaches (Aggarwal & Zhai, 2012; Domingos, 2012; Witten, Frank, Hall, & Pal, 2016; David H Wolpert & Macready, 1995). For instance, Spanhol, Oliveira, Petitjean, and Heutte (2016) introduced a BC histopathology image dataset called BreakHis for research community. Authors extracted

Preprint Version

several image features to form MFV provided it as input to four classification algorithms like *k*NN, quadratic linear analysis, support vector machines, and random forests. The experimental results showed the accuracy is ranging from 80% to 85%. Nonetheless, traditional ML-based approaches have been successfully employed to develop BC classification models; the major limitation in traditional ML-based approaches is the extraction of result-oriented features.

Thus, DL approaches overhaul traditional ML approaches where the HF engineering step is eliminated. Moreover, it requires minor preprocessing (if needed) and then finds relevant information in a self-taught manner. The primary advantage of DL is that it incorporates the feature engineering step into its learning process (Dimitropoulos et al., 2017; Kowal, Filipczuk, Obuchowicz, Korbicz, & Monczak, 2013; Loukas et al., 2013; Spanhol et al., 2016; Y.-D. Zhang, Pan, Chen, & Wang, 2018; Yudong Zhang et al., 2016). Thus, the burden of the human expert is shifted toward the machine. In addition, the least domain knowledge is required in DL-based process as compared with traditional ML-based process. In deep learning, CNN is mostly used for image segmentation and classification(Shen, Wu, & Suk, 2017). It works with structured data, such as 2D or 3D images, where each pixel possesses embedded characteristics of its neighboring pixels, which can be a great source of information for CNNs (Shen et al., 2017). Given this relationship between structured data and CNNs, medical image analysis has highly benefitted from CNNs in recent years (Bayramoglu, Kannala, & Heikkilä, 2016; Cruz-Roa et al., 2017; Han et al., 2017; Litjens et al., 2016; Spanhol, Cavalin, Oliveira, Petitjean, & Heutte, 2017; Wan, Cao, Chen, & Qin, 2017). However, training a CNN from scratch requires a large amount of labelled data to avoid overfitting issue (Krizhevsky, Sutskever, & Hinton, 2012) and consumes high computational resources, such as graphics processing unit (GPU). These requirements may create hurdles for researchers. Specially, the availability of labelled BC images in a large quantity is merely possible. To overcome these issues, researchers introduced some new pathways. First, the volume of data must be increased artificially through image augmentation to avoid the CNN overfitting issue. Image augmentation (such as scaling, translation, rotation, padding, flipping, adding noise, lighting conditions, and perspective transform) generates enough data to train CNNs (Krizhevsky et al., 2012). The second pathway suggests to retrain a DL model (e.g., AlexNet (Krizhevsky et al., 2012), GoogLeNet, ResNet50, VGG16 and etc.) called transfer learning (TL) (Chougrad et al., 2018) and fine tune the network options for a target data such as medical images. It enables CNNs to become computationally cost effective and accelerates the training process.

In recent years, a few researchers have proposed DL-based classification models for BC classification using histopathology BC images. Han et al. (2017) presented a CSDCNN end-to-end model to predict eight subclasses of BreakHis BC histopathology images by using CNN and obtained 93.2% overall accuracy. Bayramoglu et al. (2016) developed a seven-layered CNN model to predict benign or malignant classification using BreakHis dataset and reported 81% accuracy. Spanhol et al. (2017) proposed a TL-based classification model using BreakHis dataset. Here, authors extracted deep convolutional activation features (DeCAFs) by using AlexNet. The extracted DeCAFs were served as an input to traditional ML algorithms to construct the classification model. Author's experimental results showed 93.5% accuracy. Khosravi, Kazemi, Imielinski, Elemento, and Hajirasouliha (2018) proposed the DL-based pipeline CNN_Smoothie to classify numerous cancer tissues, subtypes, and their relative staining markers and scores. The author trained two CNNs; the first was trained using TL, and the other network was trained from scratch. The author used multiple datasets, and the reported accuracy is 92%. However, this method is limited to binary classification and requires high computational resources. Moreover, it is highly complex and has a wide scope (can classify lung, breast, and bladder cancer). It does not focus on multiclass classification of BC, and its overall classification accuracy is lower than those of many existing models possess. Thus, it is unfeasible to use as a real-time system. Apart from BrC image classification other type of images are also used for classification like face recognition (Lu, Chen, Zhang, Chen, & Xiong, 2018), classification of noisy images (Jiang, Ma, Wang, Chen, & Liu, 2018) and hyperspectral image classification (Ma, Li, Li, Mei, & Ma, 2018).

Thus, to overcome the limitations of aforementioned studies, this study proposed an accurate and computationally feasible classification model called Biopsy Microscopic Image Cancer Network (BMIC_Net) to classify the BC images into eight distinct classes. BMIC_Net is a hierarchical-based multiclass model constructed by customizing the existing AlexNet architecture. The proposed BMIC_Net is composed of three classifiers (namely, BC_1, B_2, and M_2) arranged in two levels in a cascade manner. In the first-level, the BC_1 classifier is placed to classify images into two distinct classes (benign and malignant). In the second-level classification, the B_2 and M_2 classifiers are deployed whereby the B_2 classifier is responsible for predicting four subtypes of Benign BC [adenosis (A), fibroadenoma (F), tubular adenoma (TA), and phyllodes tumor (PT)]. Furthermore, the M_2 classifier is responsible for classifying the images into four distinct subtypes of malignant BC [ductal carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC), and papillary carcinoma (PC)]. Extensive experiments of the BMIC_Net along with supervised machine learning algorithms are performed to achieve high-performance

Preprint Version

classification of eight BC classes from BreakHis images. Six supervised machine learning algorithms, namely, naïve Bayes (NB), support vector machine (SVM), *k*-nearest neighbor (*k*NN), decision tree (DT), linear discriminant analysis (LDA), and linear regression (LR), are compared to evaluate their performances using three performance measures (namely, overall accuracy, sensitivity, and area under the receiver-operating characteristic. The major contributions of this study are listed below.

1. This study proposes a deep learning-based end to end and computationally effective BC classification model named BMIC_Net. It uses the hierarchical classification model to classify BC into eight distinct classes belonging to two major types of BC (i.e. Benign and Malignant).
2. The proposed BMIC_Net employs convolutional neural network to discover discriminative features from BreakHis images. In addition, it also elicits the most discriminative subset of features without compromising the classification performance by employing information gain feature reduction scheme.
3. To demonstrate the significance of proposed BMIC_Net model, its performance is compared with five existing state-of-the-art baseline classification models (Han et al., 2017; Nahid, Mehrabi, & Kong, 2018; Samah, Fauzi, & Mansor, 2017; Spanhol et al., 2017; Spanhol et al., 2016) for BC classification. The experimental results show that the proposed BMIC_Net model outperforms other baseline classification models for BC classification.

The rest of the paper is structured as follows. Section 2 presents the literature review. The proposed model and experimental settings are discussed in Section 3. The experimental results are reported in Section 4. The significance of the observed experimental results is discussed in Section 5. Finally, section 6 concludes this work.

Preprint Version

2. Literature Review

In recent years, most classification models are developed through traditional ML approaches by two "gold standard" imaging techniques, namely, gray-scale mammography and colored HI. For instance, Rabidas, Midya, and Chakraborty (2018), proposed a mammogram image-based classification model. The authors used two feature extraction methods to perform binary classification (benign or malignant) after finding the ROI in the BC image. In addition, the authors extracted features by exploiting the similarity between neighboring regions of masses to capture two global similarity features at different scales along with uniform local binary patterns to support classification accuracy. The authors' experimental results showed 94.57% classification accuracy and 0.98 AUC by using the Fisher linear discriminant analysis classifier. Kowal et al. (2013) performed nuclear segmentation by using *k*-means, fuzzy C-means, competitive learning neural networks, or Gaussian mixture models to verify segmentation results. In the experiments, the authors used 500 biopsy microscopic images of 50 patients. Furthermore, 42 morphological, topological, and texture features were extracted from the segmented ROI to prepare an MFV. Finally, *k*NN, DT, and NB were used on the MFV to construct a classification model. Experimental results showed 96% to 100% accuracy. Wang, Hu, Li, Liu, and Zhu (2016) proposed a classification model through the use of segmentation for classifying BC images. Initially, the ROIs were identified by combining wavelet decomposition and multiscale region-growing. Afterward, corner detection methods were used to split overlapping cells for increased segmentation accuracy. Finally, classification was performed on four shape-based features and 18 textural-based features on color space. In addition, SVM was used to optimize the number of features for classification. The accuracy observed is 96.19% for binary classes i.e., either BC tumor is benign or malignant. Antropova, Huynh, and Giger (2017) proposed a BC binary classification model to work robustly on different medical imaging modalities, such as mammography, ultrasound, and MRI. In addition, six types of features were extracted by two methods, such as TL using VGG19 and handcrafted features. In the experimental results, the maximum AUC value achieved was 0.90.

Nonetheless, the four aforementioned studies showed improved classification performance to classify BC images. However, these studies have two major limitations. First, the dataset consisting of a low number of training images was used. Furthermore, these images were collected only from one organization. Thus, the proposed models of aforementioned studies cannot be used at a large scale due to lack of generalization. Second, binary classification problems where the images were classified as benign or malignant were considered. Nevertheless, the BC images can be classified into more specific classes of BC. Thus, to address the aforementioned issues, Spanhol et al. (2016) has coined a large dataset, BreakHis, for researchers working on BC classification using medical images. This dataset is composed of 7079 images belonging to 82 patients. These images were generated from breast tissue biopsy slides stained with hematoxylin and eosin. Moreover, the dataset contained images at magnifications of 40×, 100×, 200×, and 400×. Furthermore, the BC images in this dataset were classified into eight types of BC. The samples were collected by cancer signs of patient method, prepared for histological study, and labeled by pathologists. Spanhol et al. (2016) performed feature extraction by five textural descriptor methods [local binary patterns (LBP), completed LBP, local phase quantization, gray-level co-occurrence matrix (GLCM), and threshold adjacency statistics (TAS)] and one key point descriptor (oriented FAST and rotated BRIEF) method to analyze classification accuracy. The key point descriptor obtained 32-D vector at feasible 500 key points. Three well-known classifiers, *k*NN, SVM, Random Forest, and quadratic linear analysis, were used to classify an image into benign or malignant BC class. Experimental results showed a maximum accuracy of 85.1% on 200× magnification images by using the PFTAS feature extraction method and SVM classier. Samah et al. (2017) proposed an automated classification program using the BreakHis dataset for two classes. They used different types of feature extractors, such as GLCM, LBP, pyramid-structured wavelet transforms (PWT), and tree-structured wavelet transform. Experimental results showed a classification performance ranging from 83% to 86% using the *k*NN classifier. In addition, the maximum accuracy was obtained through PWT feature extraction. Nahid et al. (2018) proposed a binary classification method using the BreakHis dataset. They used DNN after deriving statistical information from images using *k*-means and mean-shift clustering. For feature extraction, authors employed CNN, Long-Short-Term Memory (LSTM), and a combination of both CNN and LSTM. Finally, *softmax* and SVM were used to classify BC images, and an accuracy of 91% was obtained. Nonetheless, the aforementioned studies used the standard dataset for BC classification. However, the aforementioned studies proposed binary classification models to classify BC in benign and malignant classes only. In addition, the classification accuracy can still be improved for use in real-time applications.

To overcome the aforementioned weaknesses of existing studies, Spanhol et al. (2017) proposed a multiclass classification model using a CNN TL model and the BreakHis dataset in experiments. Furthermore, the images were split into patches using sliding window and random approaches. Furthermore, pretrained networks LeNet and AlexNet were retrained on the collected images. Experimental results showed that LeNet yielded 72% accuracy and AlexNet showed 90% accuracy on 40× magnification images. Han et al. (2017) proposed a multiclass BC CNN model, CSDCNN, using the BreakHis dataset and obtained an average accuracy of 93.2%. Nonetheless, the experimental results showed reasonable classification accuracy as compared to aforementioned studies. However, the BC image classification model proposed by Han et al. (2017) has two major limitations. First, in the experiments, the images were augmented in a large quantity to increase the classification performance. Second, the trained classification model was computationally highly expensive and required an enormous amount of time and computational resources to develop the BC classification model.

To address the aforementioned issues, this study proposes and develops a hierarchical multiclass model that classifies BC images into eight distinct classes. The BreakHis dataset was used to construct the classification model. Furthermore, in the proposed method, features were extracted through DL architecture. Moreover, feature reduction schemes were applied to obtain the most informative and discriminative extracted features. Finally, the obtained informative and discriminative feature subset was fed to the proposed hierarchical classification model to predict the class of the BC images. Compared with existing BC classification models, the proposed BC classification model yielded better accuracy. Moreover, it required minimal resources to classify the BC into eight classes. The detailed functionality of the proposed model is discussed in Section 3.0.

3. Proposed Method

This section discusses the detailed research methodology (Figure 1) to construct the proposed cancer prediction model from BC images. The overall research methodology comprises five main phases, namely, data collection, image preprocessing, feature extraction through the proposed DL method, classification through traditional ML algorithms, feature reduction, and classification using the best accuracy-proving classifier. In data collection, publicly available BreakHis dataset is used. In the image preprocessing phase, some essential image preprocessing tasks are used on the collected image corpus to remove non-informative features and to realize efficient image processing and classification. This phase also discusses the image augmentation technique to avoid CNN overfitting and class imbalance in order to improve the classification performance. In the feature extraction phase, several discriminative and informative features are extracted from the preprocessed images through the DL method. The fourth phase discusses the hierarchical classification model to classify the collected images into their respective cancer types. Finally, the fifth phase discusses the feature reduction and classification phase that is used to extract the most differential subset of features in order to improve the classification accuracy. However, feature reduction is performed to reduce the classifier training time and computational cost as well as to improve the overall classification performance. All these phases are described in detail in subsequent subsections.

Preprint Version

3.1. Data Collection

This study uses the publicly available BreakHis (Breast Cancer Histopathology Image Classification) corpus (Spanhol et al., 2016) for the construction of the proposed cancer prediction model. This dataset was prepared by joint collaboration of P&D Laboratory and Pathological Anatomy and Cytopathology, Parana Brazil. BreakHis was collected by excisional biopsy from 82 patients who were diagnosed with breast tumor tissues. The dataset comprised 7909 microscopic images at 40×, 100×, 200×, and 400× magnifications (Figure 2). Each image was 700×460 pixels in size with three-channel RGB, 8-bit depth in each color, and stored in png format. All images were classified as benign or malignant. Benign lesions grow gradually, localized, and non-invasive. Conversely, malignant lesions are invasive, destroy neighboring structures, spread to faraway sites, and can cause death. The BreakHis dataset contained 2480 benign images and 5429 malignant images (Table-1). The BC images were further categorized into four distinct cancer subtypes (Table-2 and Table-3). Benign BC tumor subtypes included A, F, TA, and PT, whereas malignant BC tumor subtypes included DC, LC, MC, and PC. As mentioned earlier, the BreakHis corpus comprised 40×, 100×, 200×, and 400× images. Nonetheless, this study used only images of 58 patients at 40× magnification level because the highest classification performance was achieved at this level (Han et al., 2017). Noticeably, DC class contains 38 patients' images that is 46% of overall dataset (Table-3). Thus to avoid class imbalance issue we randomly selected nine patients after excluding the borderline cases from the majority class i.e DC. The selected 58 patients' images are splitted into training set (70%) and testing set (30%) by using random sampling method (Table-4).

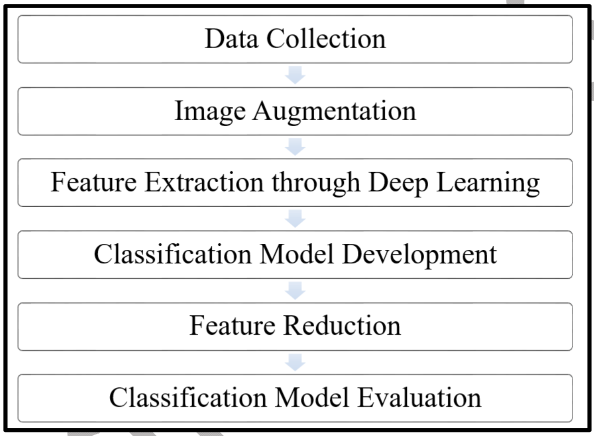
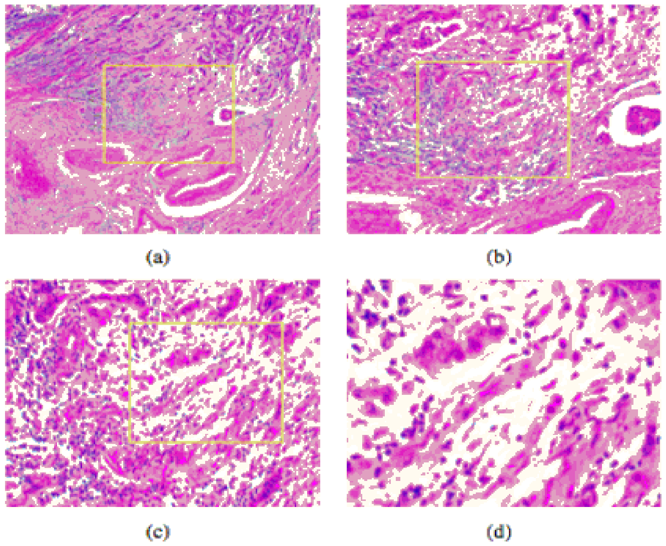

Data Collection
Image Augmentation
Feature Extraction through Deep Learning
Classification Model Development
Feature Reduction
Classification Model Evaluation

Figure 1: Research methodology flowchart



A slide of breast malignant tumor (stained with HE) seen in different magnification factors: (a) 40X, (b) 100X, (c) 200X, and (d) 400X.

Preprint version

Figure 2: BreakHis dataset image magnification factors

Table-1: Image division in benign and malignant types with magnifications

Magnification	*40×*	*100×*	*200×*	*400×*	*Total*	*No. of Patients*
Benign	625	644	623	588	2480	24
Malignant	1370	1437	1390	1232	5429	58
Total	1995	2081	2013	1820	7909	82

Table-2: Image division in benign subtypes with magnifications

Magnification	*40×*	*100×*	*200×*	*400×*	*Total*	*No. of Patients*
A	114	113	111	106	444	4
F	253	260	264	237	1014	10
PT	149	150	140	130	569	3
TA	109	121	108	115	453	7
Total	625	644	623	588	2480	24

Table-3: Image Division of malignant subtypes with magnifications

Magnification	*40×*	*100×*	*200×*	*400×*	*Total*	*No. of Patients*
DC	864	903	896	788	3451	38
LC	156	170	163	137	626	5
MC	205	222	196	169	792	9
PC	145	142	135	138	560	6
Total	1370	1437	1390	1232	5429	58

3.2. Image Augmentation

CNNs require a large amount of training data to achieve good performance. Therefore, image augmentation can be used to improve the training accuracy by using a small amount of image data. Image augmentation synthetically creates images (over-sampling) by various image-processing techniques (e.g., image rotation, shift, shear, flip, and padding) and their random combinations. In this study, the augmentation is only applied on training set only by using five techniques. Each image was augmented 25 times using rotation 90°, flip in vertical and horizontal directions, translation by the fifth part of image size, shear image using four affine transforms, and image padding (Algorithm 1). In addition, the original images were used along with 50% of the minimum quantity available in any class out of 25 times augmented images. The original number of images taken in each class is shown in Table-4.

3.3. Feature Extraction through the Proposed BMIC-Net Deep Learning Model

The present study used the DL model to extract informative and discriminative features from the preprocessed images. Pre-trained (a.k.a. AlexNet) DL architecture (Krizhevsky et al., 2012) was used to extract useful features from the preprocessed images. AlexNet was trained on one million object images for 1000 classification labels, such as keyboard, pen, pencil, coffee mug, and many animals. In brief, AlexNet was composed of eight layers divided into two main parts. The first part consisted of five convolution layers, and the second part comprised three fully connected layers. However, the part of the overall convolution layers was responsible for extracting low-level features, such as edges, blobs, and colors. Moreover, in the second part, the first two fully connected layers were dedicated to extracting high-level features at individual-pixel level. The final fully connected layer was the *softmax* layer, which is an output layer of AlexNet architecture. This layer predicts the probability of each class label, and a class label for each object. Moreover, a nonlinear activation function *ReLU* was inducted after each convolution layer and fully connected layer to avoid complicated computation for gradient descent function in backpropagation [$ReLU(x) = max(0,x)$ *i.e if* $x < 0$, $ReLU\ (x) = 0$ and *if* $x >= 0$, $ReLU\ (x) = x$]. In the backpropagation process, the loss function maps values of one or more variables onto a real number that represents some cost associated with the event. In the training process of BMIC_Net (based on AlexNet) to minimize the cross entropy the following loss function has been used.

Preprint version

$$L(w) = \sum_{i=1}^{N}\sum_{c=1}^{n} -y_{ic}\log f_c(x_i) + \mathcal{E}\|w\|_2^2 \quad (1)$$

Where y is the ground truth label of ith training instance and x is the prediction result of classifier for the ith training instance, the c represent number of class label, value of n will be 2 for BC_1 classifier and 4 for both B_2 and M_2 classifier, while Ɛ represents the learning rate. During training L(w) has been minimized in backpropagation process by using chain-rule. Furthermore, the *maxpool* and *dropout* functions were applied in CNN layers to reduce deep neural network overfitting issue. A *maxpool* function was added after the first, second, and last convolution layers, whereas a *dropout* function was placed in the first two fully connected layers. The first convolution layer, also called an input layer, accepts RGB images of size 227×227 (Figure 3).

Table-4: Images selected for training and testing

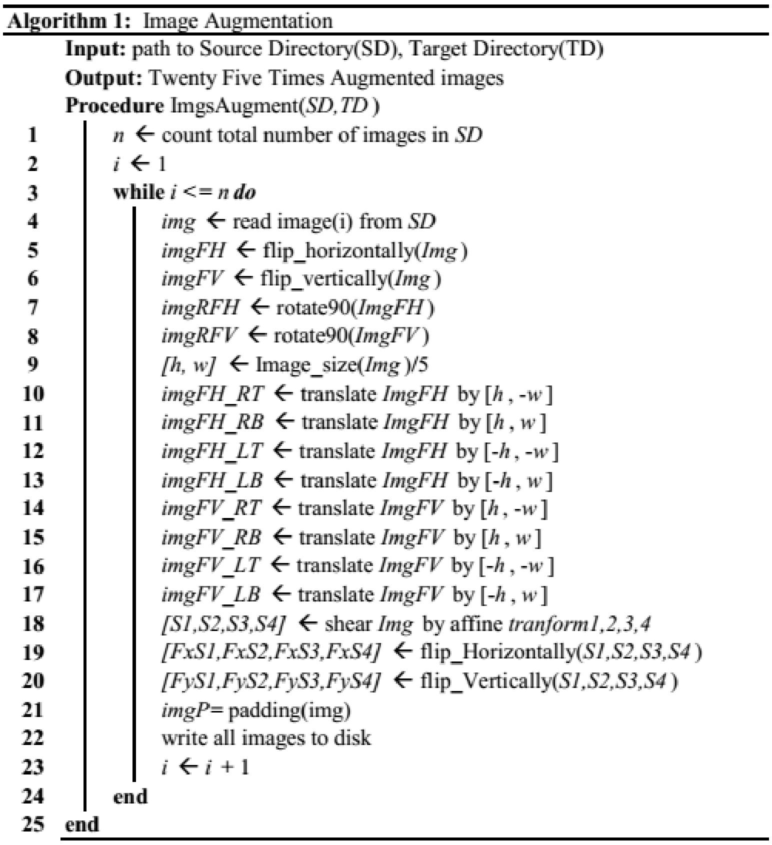
Main Types	*Subtypes*	*40× (Total)*	*40× (Selected)*	*No. of Patients*	*Training*	*Testing*
					70%	*30%*
Benign	***A***	114	114	4	80	34
	F	253	253	10	177	76
	PT	109	109	3	76	33
	TA	149	149	7	104	45
Total (Benign)		625	625	24	438	187
Malignant	***DC***	864	208	14	146	62
	LC	156	156	5	109	47
	MC	205	205	9	144	61
	PC	145	145	6	102	43
Total (Malignant)		1370	714	34	500	214
Grand Total		1995	1339	58	937	402

As aforementioned in Section 3.1, the collected corpus comprised only 7909 images belonging to eight cancer types. Thus, this small number of images may not be highly effective for any DL algorithm to train from scratch. In such cases, the pretrained DL architecture plays a decisive role in the classification of new types of images, such as medical images. Moreover, it can be easily and quickly retrained on new images to obtain a reasonable classification performance using a personal computer. Thus, one of the objectives of this research was to train a CNN on a small number of medical images by using least computation resources with a fast training process. Hence, this study used AlexNet architecture because it can be trained on new image data and because it consists of less layer number compared with most of the pretrained CNN architectures. To achieve this objective, this study used TL by fine-tuning AlexNet to construct a BMIC-Net model. In TL, the first part of BMIC-Net behaves like a trained AlexNet to extract the best low-level features. Only the second part that consists of fully connected layers was fine-tuned and retrained for specific tasks, such as medical biopsy images. Thus, BMIC-Net can predict various BC types. Finally, the output layer of the BMIC-Net model is limited according to the number of cancer subtypes to be predicted.

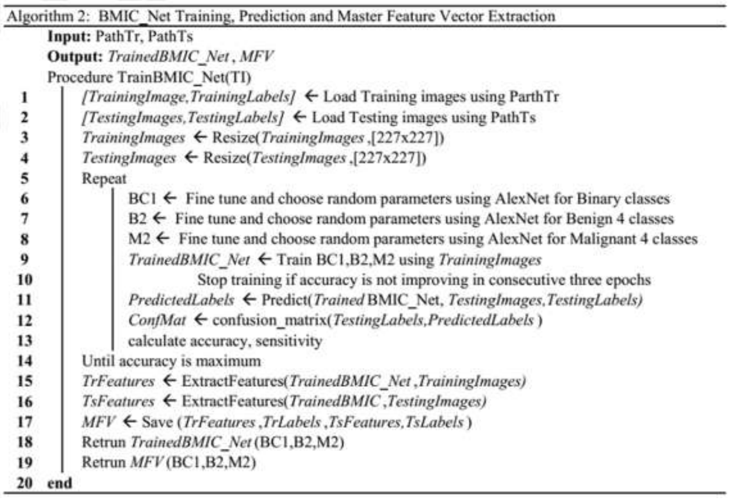
In the construction of a fine-tuned BMIC-Net, several experiments were run recursively to obtain optimum training results by adjusting a few training options. The BMIC-Net architecture was trained (Algorithm 2) using gradient descent with momentum. Some of the parameter adjustments are as follows: momentum of 0.9, maximum epochs of 30, mini-batch size of 50, initial learning rate of 1e-4, and learning rate drop factor of 0.5. The CNN was fine-tuned with stochastic gradient descent with a learning rate adjusted to be lower than the initial learning rate. Hence, the features previously learned from the larger dataset were guaranteed to be not entirely ignored during retraining. Furthermore, the network was compelled to stop training if validation accuracy was not improving or was decreasing constantly in the last three validation predictions. Ultimately, BMIC-Net was devised in such a way to avoid overfitting and underfitting. Moreover, the best feasible validation accuracy of BMIC-Net was achieved. BMIC-Net was trained to obtain the best features. It should be noted that random sub-sampling approach was used for training and testing set where 70% used for training and 30% for testing. Furthermore, by default, the *softmax* classifier was used in the validation process for training. Finally, the trained network was applied to predict the class labels for testing data. Moreover, the training graph, accuracy, sensitivity, and AUROC

Preprint Version

were calculated for analysis. In ROC function, x-coordinates for the performance curve, returned as a vector. X values are the false positive rate, FPR (1- specificity). It computes the pointwise confidence bounds, by using vertical averaging. Similarly, the Y-coordinates for the performance curve, returned as a vector. Y values are the true positive rate, TPR (recall or sensitivity). ROC computes the confidence bounds using vertical averaging; AUC is a 3-by-1 vector. The first column of AUC contains the mean value. The second and third columns contain the lower bound and the upper bound, respectively, of the confidence bound. For a perfect classifier, AUC value will be 1 and for a classifier that randomly assigns observations to classes, then AUC will be 0.5. Finally, the features that were finalized by BMIC-Net were extracted before the output layer to form MFV, which was composed of 4096 unique features for classification and served as an input to traditional ML algorithms to evaluate the classification predictive performance.



```
Algorithm 1: Image Augmentation
Input: path to Source Directory(SD), Target Directory(TD)
Output: Twenty Five Times Augmented images
Procedure ImgsAugment(SD, TD)
    n ← count total number of images in SD
    i ← 1
    while i <= n do
        img ← read image(i) from SD
        imgFH ← flip_horizontally(Img)
        imgFV ← flip_vertically(Img)
        imgRFH ← rotate90(ImgFH)
        imgRFV ← rotate90(ImgFV)
        [h, w] ← Image_size(Img)/5
        imgFH_RT ← translate ImgFH by [h , -w]
        imgFH_RB ← translate ImgFH by [h , w]
        imgFH_LT ← translate ImgFH by [-h , -w]
        imgFH_LB ← translate ImgFH by [-h , w]
        imgFV_RT ← translate ImgFV by [h , -w]
        imgFV_RB ← translate ImgFV by [h , w]
        imgFV_LT ← translate ImgFV by [-h , -w]
        imgFV_LB ← translate ImgFV by [-h , w]
        [S1,S2,S3,S4] ← shear Img by affine tranform1,2,3,4
        [FxS1,FxS2,FxS3,FxS4] ← flip_Horizontally(S1,S2,S3,S4)
        [FyS1,FyS2,FyS3,FyS4] ← flip_Vertically(S1,S2,S3,S4)
        imgP= padding(img)
        write all images to disk
        i ← i + 1
    end
end
```



```
Algorithm 2: BMIC_Net Training, Prediction and Master Feature Vector Extraction
Input: PathTr, PathTs
Output: TrainedBMIC_Net , MFV
Procedure TrainBMIC_Net(TI)
    [TrainingImage,TrainingLabels] ← Load Training images using ParthTr
    [TestingImages,TestingLabels] ← Load Testing images using PathTs
    TrainingImages ← Resize(TrainingImages ,[227x227])
    TestingImages ← Resize(TestingImages ,[227x227])
    Repeat
        BC1 ← Fine tune and choose random parameters using AlexNet for Binary classes
        B2 ← Fine tune and choose random parameters using AlexNet for Benign 4 classes
        M2 ← Fine tune and choose random parameters using AlexNet for Malignant 4 classes
        TrainedBMIC_Net ← Train BC1,B2,M2 using TrainingImages
                Stop training if accuracy is not improving in consecutive three epochs
        PredictedLabels ← Predict(Trained BMIC_Net, TestingImages,TestingLabels)
        ConfMat ← confusion_matrix(TestingLabels,PredictedLabels )
        calculate accuracy, sensitivity
    Until accuracy is maximum
    TrFeatures ← ExtractFeatures(TrainedBMIC_Net ,TrainingImages)
    TsFeatures ← ExtractFeatures(TrainedBMIC ,TestingImages)
    MFV ← Save (TrFeatures ,TrLabels ,TsFeatures,TsLabels )
    Retrun TrainedBMIC_Net (BC1,B2,M2)
    Retrun MFV (BC1,B2,M2)
end
```

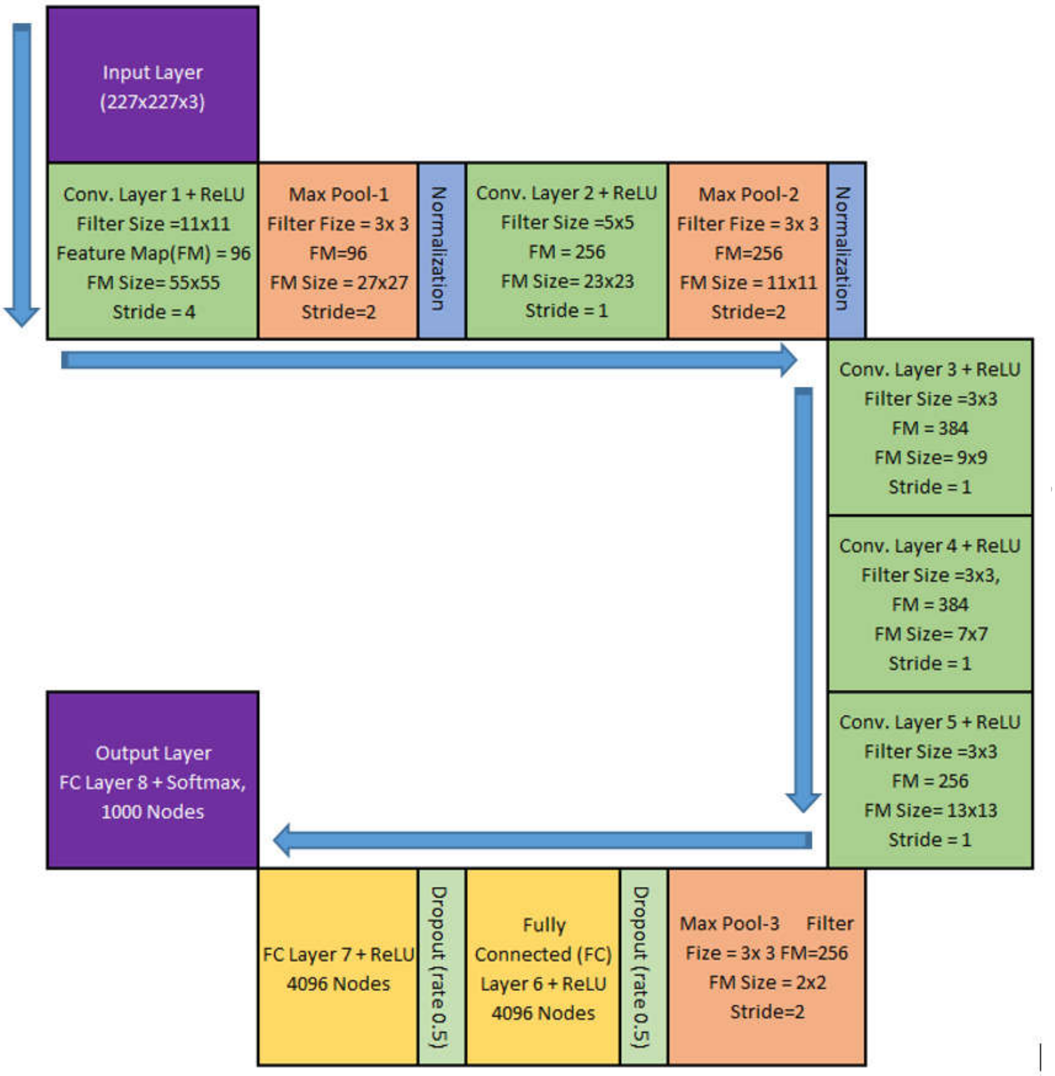


Figure 3 AlexNet basic architecture

Prepri

3.4. Development of Classification Model

A hierarchal classification model was used to classify the BC images into their respective cancer subtypes. Two-level classifications were used (Figure 4). In the first level, the classifier predicts whether the image belongs to benign or malignant cancer. For instance, if the classifier predicted that the image belongs to benign class, then the second level classifier is responsible to predict further the subtype of benign cancer for the given image. Thus, three classification models were prepared, namely, first-level classification model called Breast Cancer Classifier (BC_1), second-level classification model called Benign classifier (B_2), and second-level classification model called Malignant Classifier (M_2). For the construction of BC_1, all training images were divided into benign and malignant classes. For the construction of B_2, all images belonging to benign subtypes were labelled into specific benign cancer types. For the construction of M_2, all images belonging to malignant subtypes were labelled into specific malignant cancer types. To summarize, BC_1 is a binary model that classifies an image into BC main types, such as benign or malignant, formally represented by equation 2. B_2 is a multiclass model used to classify an input image into four distinct benign cancer subtypes, namely, A, F, TA, and PT, formally expressed by equation 3. Finally, M_2 is also a multiclass model used to predict further four malignant cancer subtypes, namely, DC, LC, MC, and PC, mathematically denoted by equation 4. Thus, two-level classifiers were organized in two-level cascade architecture. A biopsy microscopic image was initially processed by first-level classifier BC_1; for instance, it was predicted as malignant. At the second level of hierarchy, the M_2 classifier was trained to predict its four types; for instance, M_2 was predicted as the DC subtype.

For each classification level, six traditional ML algorithms were applied, namely, the *k*NN where *k*=1, SVM, NB, DT, LDA classifier, and LR. These six traditional ML algorithms were applied to see which one produces better classification accuracy on the BreakHis dataset. Furthermore, according to (D. H. Wolpert & Macready, 1997), no single traditional ML algorithm can perform consistently better on all types of data. Thus, the performance of various algorithms on the collected dataset must be evaluated to investigate which one produces better classification results on the collected dataset. Hence, this study selected the six aforementioned traditional ML algorithms to evaluate their performances on the extracted features.

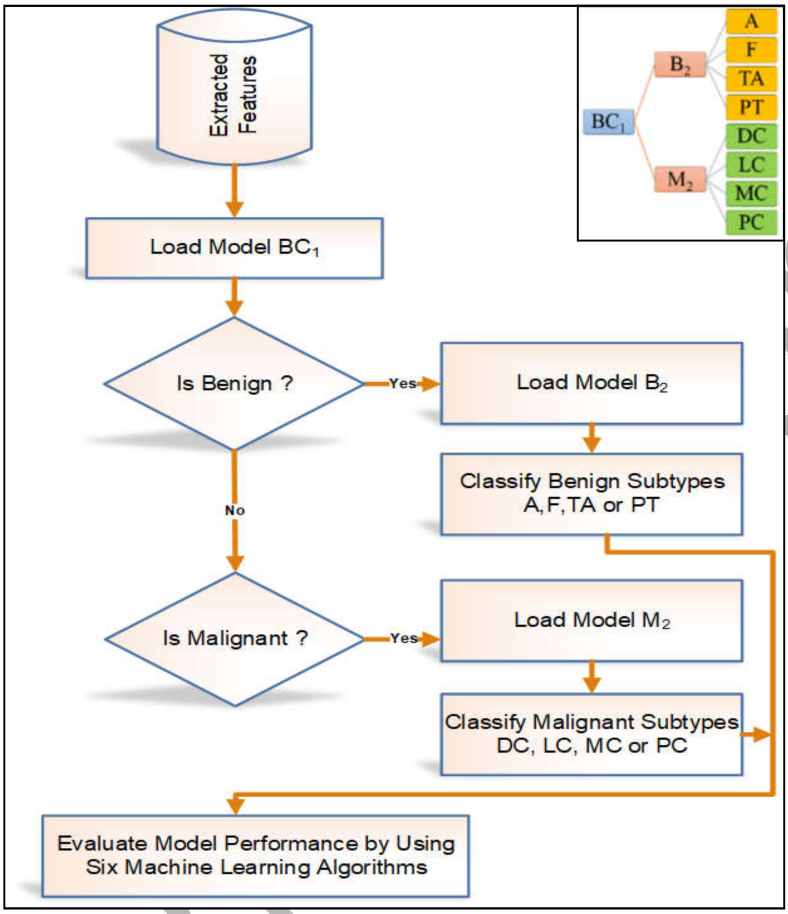

Extracted Features
Load Model BC$_1$
Is Benign ?
Yes
Load Model B$_2$
Classify Benign Subtypes A,F,TA or PT
No
Is Malignant ?
Yes
Load Model M$_2$
Classify Malignant Subtypes DC, LC, MC or PC
Evaluate Model Performance by Using Six Machine Learning Algorithms
BC$_1$
B$_2$
M$_2$
A
F
TA
PT
DC
LC
MC
PC

Figure 4: Proposed hierarchical model

$$Y_{BC_1}(I) = Softmax(I.W + b) \quad (2)$$

$$Y_{B_2}\left(I|Y_{BC_1}\right) = Softmax(I.W + b) \ \ , \ if \ Y_{BC_1} is\ Benign \quad (3)$$

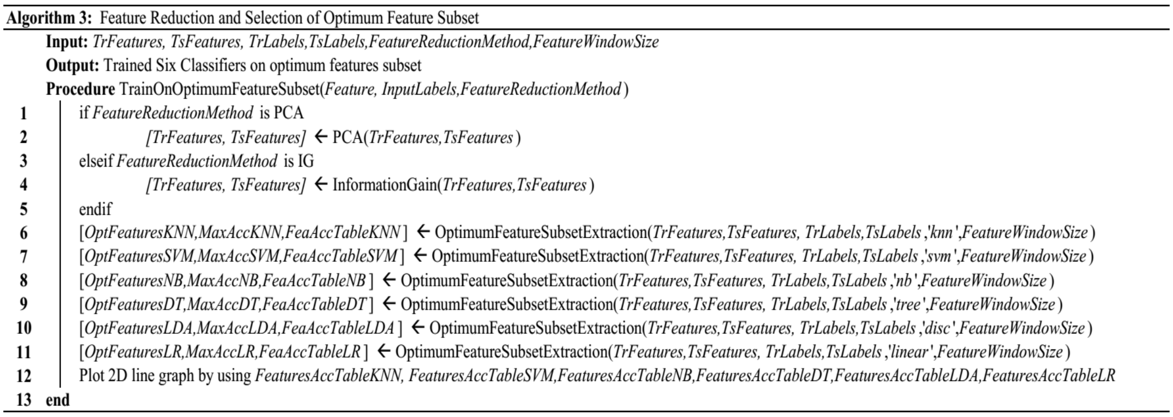
$$Y_{M_2}\left(I|Y_{BC_1}\right) = Softmax(I.W + b) \ \ , \ if \ Y_{BC_1} is\ Malignant \quad (4)$$

Where, I, W, and b represent input image, weights and biases respectively computed by BC$_1$, B$_2$, and M$_2$ classifiers. Y_{BC_1} represents the classification probabilities predicted for benign and malignant classes. Y_{B_2} denotes the classification probabilities predicted for four subclasses of benign BrC. Similarly, Y_{M_2} denotes the classification probabilities predicted for four subclasses of malignant BrC.

3.5. Feature Reduction

A large number of features in MFV may not be feasible for effective traditional ML algorithms to obtain the highest accuracy within limited computational time and resources. Thus, the extracted features were analyzed and optimized using two well-known feature reduction schemes, namely, IG and PCA, to obtain the most informative and discriminative feature subset from the MFV. The feature reduction process is based on three steps. In the first step, a feature score table (FST) was created using MFV. In the second step, a feature accuracy table was generated using the FST. Finally, the highest accuracy was achieved when the least number of feature subsets was used (Algorithms 3 and 4). There were three major reasons behind the selection of IG and PCA feature reduction methods. First, in several studies, these method have shown the promising results compared to other methods

(Babu, Sukumar, & Anandan, 2013; Bovis, Singh, Fieldsend, & Pinder, 2000; Buciu & Gacsadi, 2009; Buciu & Gacsadi, 2011; Kozegar, Soryani, Minaei, & Domingues, 2013; Naik et al., 2008; Surendiran & Vadivel, 2010; Swiniarski, Lim, Shin, & Skowron, 2006; Yu Zhang, Tomuro, Furst, & Raicu, 2012). Second, images mostly have highly correlated features due to similarity among neighboring pixels. However, real life histopathology images usually possess some noise/inconsistencies due to different color, intensity and lightning effects because of image acquisition protocols and different standards followed in digital pathology labs. Thus entropy based feature selection (like IG) method helps to find out the purity of contribution for each dimension towards intended class label (Kent, 1983). Third, for high dimensional data like histopathology images, PCA mostly used in order handle the curse of dimensionality without losing the important information. Moreover, variant information in the data needs to be preserved. Thus, PCA is a well-established mathematical technique for reducing the dimensionality of images and keeps the embedded information variations as its maximum (Abdi & Williams, 2010).



Algorithm 3: Feature Reduction and Selection of Optimum Feature Subset

```
Input: TrFeatures, TsFeatures, TrLabels,TsLabels,FeatureReductionMethod,FeatureWindowSize
Output: Trained Six Classifiers on optimum features subset
Procedure TrainOnOptimumFeatureSubset(Feature, InputLabels,FeatureReductionMethod)
    if FeatureReductionMethod is PCA
            [TrFeatures, TsFeatures] ← PCA(TrFeatures,TsFeatures)
    elseif FeatureReductionMethod is IG
            [TrFeatures, TsFeatures] ← InformationGain(TrFeatures,TsFeatures)
    endif
    [OptFeaturesKNN,MaxAccKNN,FeaAccTableKNN] ← OptimumFeatureSubsetExtraction(TrFeatures,TsFeatures, TrLabels,TsLabels,'knn',FeatureWindowSize)
    [OptFeaturesSVM,MaxAccSVM,FeaAccTableSVM] ← OptimumFeatureSubsetExtraction(TrFeatures,TsFeatures, TrLabels,TsLabels,'svm',FeatureWindowSize)
    [OptFeaturesNB,MaxAccNB,FeaAccTableNB] ← OptimumFeatureSubsetExtraction(TrFeatures,TsFeatures, TrLabels,TsLabels,'nb',FeatureWindowSize)
    [OptFeaturesDT,MaxAccDT,FeaAccTableDT] ← OptimumFeatureSubsetExtraction(TrFeatures,TsFeatures, TrLabels,TsLabels,'tree',FeatureWindowSize)
    [OptFeaturesLDA,MaxAccLDA,FeaAccTableLDA] ← OptimumFeatureSubsetExtraction(TrFeatures,TsFeatures, TrLabels,TsLabels,'disc',FeatureWindowSize)
    [OptFeaturesLR,MaxAccLR,FeaAccTableLR] ← OptimumFeatureSubsetExtraction(TrFeatures,TsFeatures, TrLabels,TsLabels,'linear',FeatureWindowSize)
    Plot 2D line graph by using FeaturesAccTableKNN, FeaturesAccTableSVM,FeaturesAccTableNB,FeaturesAccTableDT,FeaturesAccTableLDA,FeaturesAccTableLR
end
```

3.6. Experimental Setup and Implementation Details:

To evaluate the performance of the proposed model, this study designed three experimental settings.

I- In the first setting, the performance of extracted features through BMIC-Net was evaluated using six traditional ML algorithms, namely, *k*NN, SVM, NB, DT, LDA, and LR. Hence, in this setting, 18 analyses were run (6 traditional ML algorithms × 3 classifiers: BC_1, B_2, and M_2) (Figure 5).

II- In the second setting, the best feature subset was obtained using IG and PCA. The performance of 50 to all 4096 features was evaluated with the increment of 50 from all three MFVs (BC_1, B_2, and M_2). Thus, 82 sub-MFVs (50 features, 100 features, 150 features, … , and 4096 features) were prepared overall from each of the three super-MFVs. Moreover, the same six traditional ML algorithms, which were used in setting-I, were adopted to evaluate the performance of the prepared 82×3 MFVs. Thus, in this setting, 2916 analyses were run (81 sub-MFVs × 3 super-MFVs × 2 feature reduction schemes × 6 traditional ML algorithms) to perform feature reduction (Figure 6).

III- In the third setting, a non-hierarchical model was evaluated using all same traditional ML algorithms. Here, the aims is to compare the performance of proposed hierarchical classifier with a non-hierarchical classifier, see Figure 10. Therefore, in this experiment 6 analysis were executed (6 traditional ML algorithms × 1 non-hierarchical classifier) and the results are shown in Figure 11.

The image preprocessing, construction of proposed BMIC-Net, all classification experiments, and feature reduction were performed in MATALB R2017b. All classification experiments were solely executed on default parameters.

Preprint v

Algorithm 4: Optimum Feature Subset Extraction

```
Input: TrainingFeatures, TestingFeatures, TrainingLabels, TestingLabels, ClassifierName, FeaturesWindowSize
Output: OptimumFeaturesSubset,MaxAccuracy, FeaturesSubsetAccuracyTabl
Function OptimumFeatureSubsetExtraction(TrainingFeatures, TestingFeatures,TrainingLabels,TestingLabels,
                                        ClassifierName,FeaturesWindowSize)
    NoPredictions = NumberOfFeatures(TestingFeatures )/FeaturesWindowSize
    i ← 1
    k ← 0
    while i <= NoPredictions do
        k ← k+FeaturesWindowSize
        TrainingFeaturePart ← take k Features from TrainingFeatures
        TestingFeaturePart ← take k Feature from TestingFeatures
        TrainedClassifier ← Train(ClassifierName,TrainingFeaturePart, TrainingLabels)
        PredictedLabels ← Predict(TrainedClassifier,TestingFeaturePart, TestingLabels)
        ConfMatrix ← confusion_matrix (TestingLabels,PredictedLabels )
        Accuracy ← Calculate Accuracy by using ConfMatrix
        FeaturesSubsetAccuracyTable[i] ← table [k,Accurracy ]
        i ← i + 1
    end
    MaxAccuracy ← Maximum_Accuracy(FeaturesSubsetAccuracyTable )
    OptimumFeaturesSubset ← FeatureSubset(Maximum_Accuracy(FeaturesSubsetAccuracyTable ))
    return OptimumFeaturesSubset ,MaxAccuracy, FeaturesSubsetAccuracyTable
end
```

3.6.1. Performance Metrics

Accuracy, sensitivity and AUC were used as performance metrics for all 2940 analyses (18 from setting-I and 2916 from setting-II). A brief description of the performance metrics is discussed in subsequent subsections.

3.6.1.1. Overall Predictive Accuracy

The prediction accuracy of a classification model is the ratio of summation of correctly predicted class labels to the summation of overall predicted labels of BC images. Mathematically, it can be defined as

$$Accuracy_{Avg} = \frac{\sum_{i=1}^{C} \frac{TP_i + TN_i}{TP_i + FN_i + TN_i + FP_i}}{C} \quad (5)$$

3.6.1.2. AUC using receiver-operating characteristic

The receiver operating characteristic (ROC) curve is a recognized performance metric used in traditional ML predictive analysis. The main advantage of using this measure is that it is not influenced by the class imbalance problem of a dataset (Provost & Fawcett, 1997; Provost, Fawcett, & Kohavi, 1998). Moreover, the ROC curve is used to measure the AUC values of each predicted class label based on the true and false positive rates of a confusion matrix. Computationally, AUC values lie between 0 and 1. Usually, an AUC value of 1 determines a perfect test and a value of 0.5 or below characterizes poor performance of the predictive model (Fawcett, 2006; Provost et al., 1998). Equation 2 is the mathematical representation of AUC (Hand & Till, 2001).

$$AUC = \frac{S_0 - n_0(n_0+1)/2}{n_0 n_1} \quad (6)$$

3.6.1.3. Sensitivity

Sensitivity (Sn) is the ratio of true positive (TP) predictions with the sum of TP) and false negatives (FN). It is alternatively known as recall (Rec) or true positive rate (TPR). The best Sn is consider as 1.0, whereas the worst is 0.0.

$$Sensitivity\ (Sn) = \frac{TP}{TP+FN} \quad (7)$$

4. Results and Discussion

This section reports and discusses the experimental results of three experimental settings (as mentioned in Section 3.6) in terms of overall predictive accuracy, sensitivity and area under an ROC (AUROC).

4.1. Setting-I Experimental Results

Preprint version

This section presents all 18 analysis [6 traditional ML algorithms × 3 models (BC_1, B_2, and M_2)] results in terms of overall predictive accuracy and AUROC using overall 4096 features. The overall predictive accuracies of the three proposed models (BC_1, B_2, and M_2) using the six traditional ML algorithms are shown in Figure 5. In the BC_1 model, *k*NN outperformed the five other traditional ML algorithms by obtaining an overall predictive accuracy of 94.55% (Sn = 91.94, 97.17). Nonetheless, the overall predictive accuracies and sensitivity of the LDA and SVM traditional ML algorithms were slightly less than that of *k*NN in the BC_1 model. The DT and NB algorithms showed the lowest overall predictive accuracies in the BC_1 model. In the B_2 and M_2 models, the *k*NN traditional ML algorithm outperformed the five other traditional ML algorithms by obtaining overall predictive accuracies of 92.13% (Sn = 96.97, 96.05, 84.38, 91.11) and 91.28% (Sn = 91.80, 95.74, 86.89, 90.69), respectively, followed by the SVM and linear regression algorithms. In addition, the lowest overall predictive accuracies were observed in the NB algorithm. To summarize, in setting-I, the best performance in all models was observed in the *k*NN algorithm followed by SVM. However, in the BC_1 model, the best performance was shown by the *k*NN algorithm followed by the LDA algorithm. Thus, the AUROC is also shown for best-performing algorithms in all three models in Figure 6(a), Figure 6(b), and Figure 6(c).

Figure 6 shows the AUROC of all three models using the *k*NN algorithm. The highest AUC values of 0.9750, 0.9484, 0.9121, and 0.9555 were obtained by A, F, PT, and TA classes, respectively, in the B_2 model [Figure 6(b)]. However, the BC_1 model depicted slightly lower AUC values of 0.9455 and 0.9455 for benign and malignant classes, respectively, [Figure 6(a)]. Conversely, the M_2 model showed the lowest AUC values of 0.9358, 0.9544, 0.9245, and 0.9505 for the DC, LC, MC, and PC classes, respectively [Figure 6(c)]. As can be witnessed from Figure 6, the intraclass performances of each class across all three models are reasonable. Moreover, the AUROC figure of each model shows that all three models are good enough to predict BC across the eight classes. Furthermore, the AUROC shows that the proposed models are not either over-fitted or under-fitted or biased towards any particular class or classes. Thus, the performance of all three proposed models is satisfactory for deployment in real-time applications.

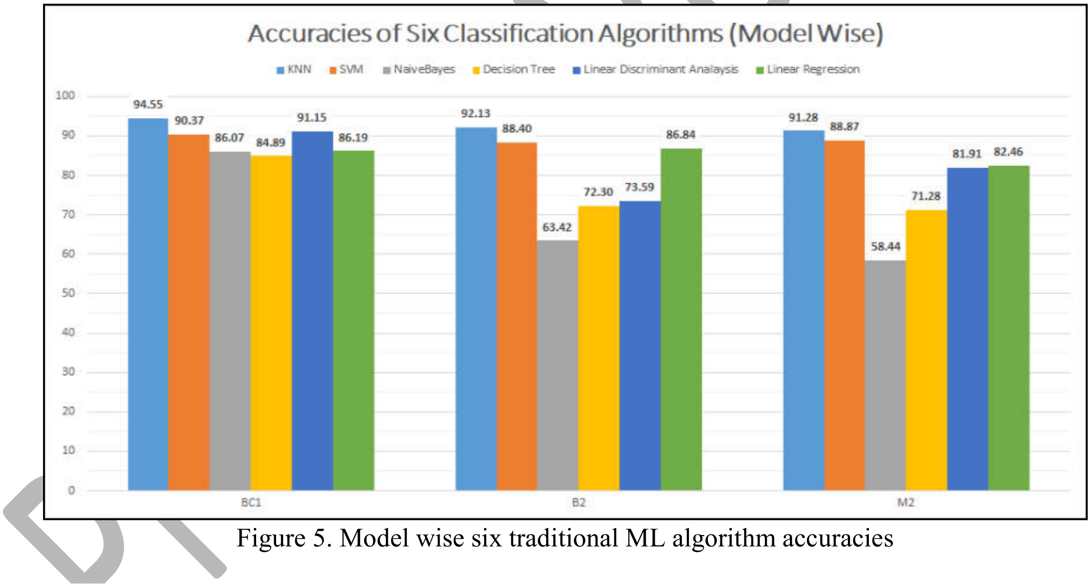
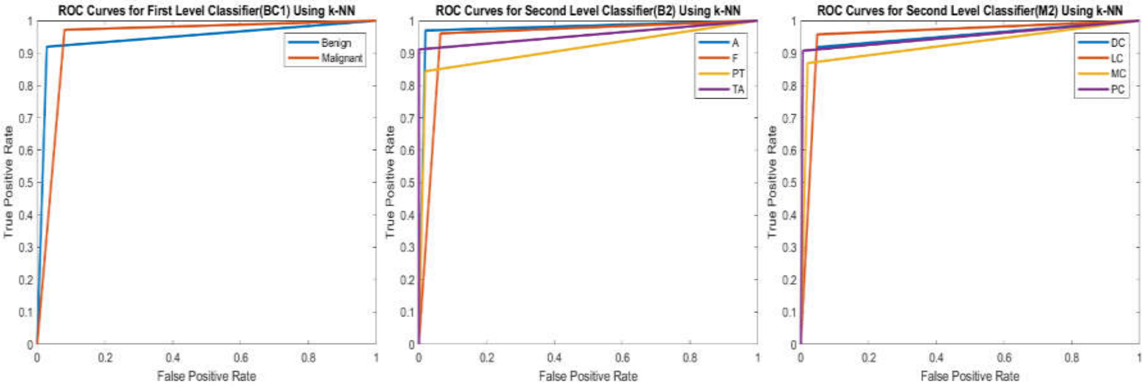


Figure 5. Model wise six traditional ML algorithm accuracies



(a) (b) (c)

Figure 6: ROCs for BC_1, B_2, and M_2 classifier

Draft Version

4.2. Setting-II Experimental Results

This section reports and discusses the results of 2916 analyses (82 sub-MFVs × 3 super-MFVs × 2 feature reduction schemes × 6 traditional ML algorithms) in terms of overall predictive accuracy and AUROC. To recapitalize, the aim of this setting-II was to obtain the best feature subset for high classification accuracy and to reduce the computational time. Thus, in this setting, various feature subsets were tested with the six aforementioned traditional ML algorithms across all three models to observe their classification performance. Moreover, two feature reduction schemes (PCA and IG) were compared to see which one elicits the best subset of features for classification. The experimental results of these 2916 analyses are shown in Figures 7(a) to 7(f). Figure 7(a) and 7(b) show the overall predictive accuracies of the BC_1 model across the six traditional ML algorithms, 82 feature subsets (50, 100, 150, … 4096), and two feature reduction schemes. IG outperformed PCA by obtaining the highest accuracy of 95.48% using 900 features and *k*NN. In addition, Figures 7(c) and 7(d) show the overall predictive accuracies of the B_2 model across the six traditional ML algorithms, 82 feature subsets (50, 100, 150, … 4096), and two feature reduction schemes. PCA performed slightly better than IG by obtaining a higher accuracy of 94.62% using only 50 features through *k*NN. Finally, Figures 7(e) and 7(f) show the overall predictive accuracies of the M_2 model across the six traditional ML algorithms, 82 feature subsets (50, 100, 150, … 4096), and two feature reduction schemes. IG marginally performed better than PCA by obtaining a higher accuracy of 92.45% using only 350 features through *k*NN.

In sum, compared with all 4096 features, 900 feature subsets showed the best performance (95.48% overall accuracy, Sn = 93.55, 97.16) using IG and *k*NN in the BC_1 model. In addition, compared with all 4096 features, only 50 feature subsets showed the best performance (94.62% overall accuracy, Sn = 96.97, 96.05, 93.65, 91.11) using PCA and *k*NN in the B_2 model. Finally, in the M_2 model, 350 out of 4096 feature subsets showed the best performance (92.45% overall accuracy, Sn = 88.52, 97.87, 91.80, 93.02) using IG and *k*NN. Thus, the AUROC was also calculated for all these three best analyses in Figure 8(a), Figure 8(b), and Figure *8*(c) to observe the intraclass performance across the BC_1, B_2, and M_2 models, respectively. The AUC values of the BC_1 model [see Figure 8(a)] using 900 feature subsets extracted by IG and *k*NN were 0.9536 and 0.9536 for the benign and malignant classes, respectively. In addition, the AUC values of the B_2 model [see Figure 8(b)] using 50 feature subsets extracted by PCA and *k*NN were 0.9718, 0.9621, 0.9623, and 0.9556 for the A, F, PT, and TA classes, respectively. Finally, the AUC values of the M_2 model [Figure 8(c)] using 350 feature subsets extracted by IG and *k*NN were 0.9294, 0.9651, 0.9524, and 0.9529 for the DC, LC, MC, and PC classes, respectively. As shown in Figure 8, the intraclass performance of each class across all three models was satisfactory. Moreover, the confusion matrices (see Figure 9) and AUROC figures of each model show that all three models can predict BC across the eight classes using the reduced feature subset. Furthermore, the proposed models with reduced features are neither over-fitted nor under-fitted or biased toward any particular class or classes. Thus, the performance of all three proposed models with reduced feature subset is better and more accurate compared with all 4096 features for deployment in real-time applications.

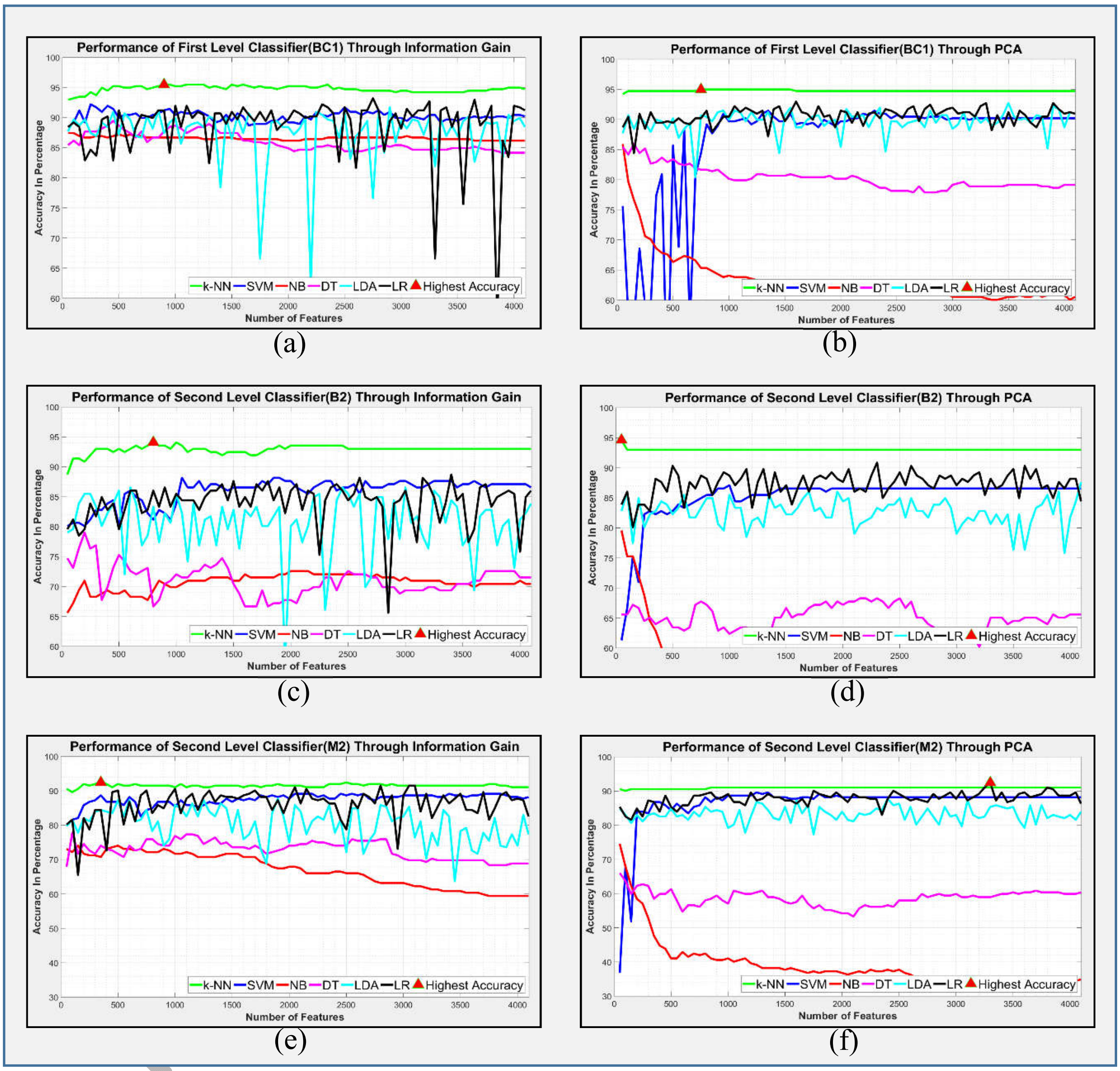
Moreover, the top 900 features extracted through IG from the 4096 master features should be consumed as an input to *k*NN when constructing the top-level BC_1 classification model. In addition, the top 50 features extracted through PCA from the 4096 master features should be given as an input to *k*NN when constructing the second-level B_2 classification model. Moreover, the top 350 features extracted through IG from the 4096 master features should serve as an input to *k*NN when constructing the second-level M_2 classification model. Ultimately, the constructed models should be deployed in a cascading manner to predict the eight BC types.

Preprint version

4.3. Setting-III Experimental Results

This section in particular presents 6 analyses using non-hierarchical classifier to predict eight classes of breast cancer. The performance shown here is solely based on all 4096 features. The overall predicted accuracies of six traditional ML algorithms for eight classes using a non-hierarchical model are shown in Figure 11. Here it can be easily observed that, SVM outperformed (see Figure 11) the rest of five traditional ML algorithms by gaining the highest accuracy of 86.97% (Sn = 93.94, 85.53, 81.25, 84.44, 86.89, 93.62, 77.05, 93.02, AUC = 0.9972, 0.9779, 0.9849, 0.9914, 0.9877, 0.9937, 0.9715, 0.9879). However, the accuracy (86.84%), sensitivity (93.94, 88.16, 62.5, 91.11, 90.17, 93.62, 86.89, 88.37) and AUC (0.9615, 0.9175, 0.8043, 0.9542, 0.9404, 0.9624, 0.9226, 0.9390) of *k*NN is hairless than the SVM. In contrast, DT proven the lowest performing traditional ML algorithm by getting accuracy of 38.39% (Sn = 75.76, 57.89, 56.25, 64.44, 60.66, 68.09, 44.26, 72.09, AUC = 0.9264, 0.8706, 0.7998, 0.8546, 0.8465, 0.8908, 0.7568, 0.9051). Precisely, in six analysis the best performance has been observed by SVM while *k*NN is slightly lower. However, the rest of the four traditional ML algorithms were unable to give better results at all.

The experimental results showed that the proposed hierarchical classification model outperformed as compare to non-hierarchical classification model using biopsy images due many reasons. First, there is a great challenge in classification due to broad inconsistency in high-resolution of image appearance. Second, there is a greater similarity of cancerous tissues between two borderline types of cancers like in BreakHis dataset one of the subject (ID: 13412) was a borderline case. Subject possess characteristics of two cancer types like ductal and lobular carcinoma. Third, due to inhomogeneous staining the color distribution in image slides varies among patients. Due to these inherent biopsy image issues a classifier can be jumbled and leads to a higher misclassification rate. Therefore, it is simpler and easier to classify among four subtypes of cancer instead of eight simultaneously. This is the main notion of choosing hierarchical classification model.

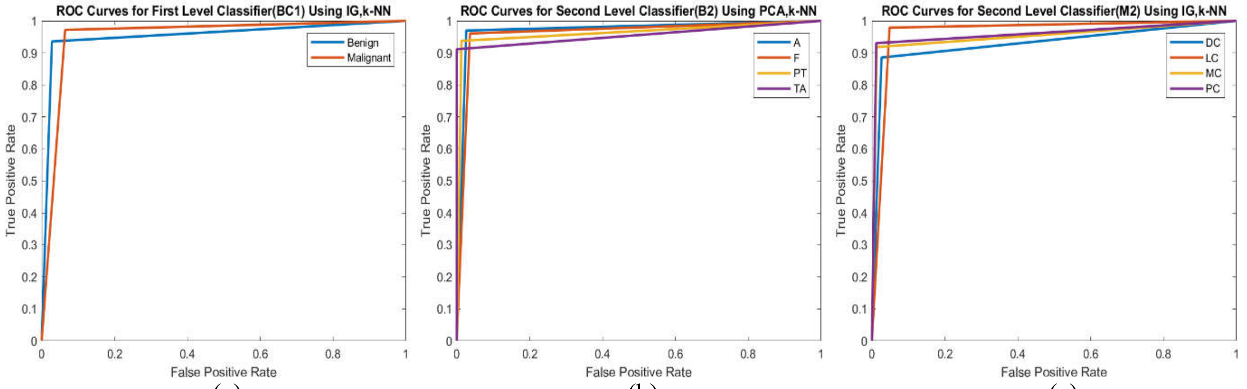


(a) (b)

(c) (d)

(e) (f)

Figure 7. Feature reduction and overall accuracies



(a) (b) (c)

Figure 8: ROCs after feature reduction

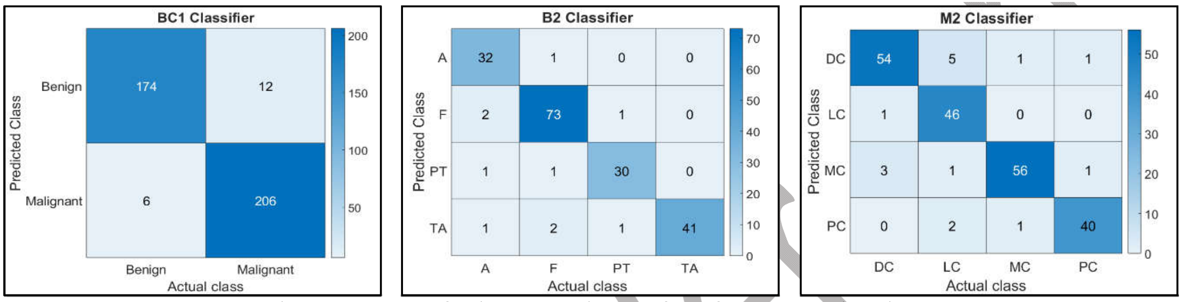


Figure 9: Confusion matrices after feature reduction

5. Discussion

This section presents the hypothetical analysis and significant results of the proposed BMIC_Net classification model using magnified biopsy medical images. The proposed BMIC_Net classification model obtained reliable and improved classification performance. The rigorous experimental evaluation on complex, challenging, and standard publicly available dataset proved that the proposed BMIC_Net model is less complex, computationally effective, reliable, and more accurate compared with existing baseline classification models for BC classification. Several studies have proposed classification models that claim high accuracies for the early prediction of BC. However, such models suffer from three major limitations. First, these models are mostly capable of predicting only two classes of BC, namely, benign and malignant. Second, several studies have evaluated the performances of those classification models on exclusive datasets containing a low amount of training images. Thus, the reported results in existing studies may not be applicable on a wider scale. Finally, most of the existing classification models have been developed using traditional ML approaches, whereby the handcrafted feature extraction and selection process were performed with the help of domain experts. Thus, extracting and selecting the features manually are tiring and time-consuming tasks. In addition, traditional ML approach-based models for BC images usually use segmentation and edge smoothing techniques to determine the exact ROIs before applying any feature extraction technique. However, all these traditional ML preprocessing steps make the classification model complex, computationally expensive, and vulnerable to high error rates. In specific, segmentation is not a trivial task and is prone to errors.

To overcome the issues of classification models developed using traditional ML approaches, recent studies have developed several classification models through DL approaches to produce accurate predictions by involving an autofeature extraction step through CNN training. However, the current studies of DL for BC image classification use two approaches to develop a classification model. First is to create a new CNN model and perform training from scratch. Second is to retrain a pretrained CNN model after fine-tuning the last layers, generally known as TL. However, the training CNN model from scratch has many limitations. First, it requires a large amount of labelled data and may need hours to train it on a computationally expensive machine, such as GPU. Hence, it may not be a feasible solution for medical image analysis. By contrast, TL was adopted to obtain a faster, reliable, and computationally feasible solution. However, the classification performance of TL is insufficient for it to be deployed as a classification model in health sectors. Hence, TL can be used instead as a source for autofeature extraction and selection. Finally, the extracted features can be reduced and optimized for reliable and computationally effective classification. Therefore, TL was used as a fast, accurate, and reliable feature extraction method in the proposed model. The experimental results of this study show that the hierarchical classification

Preprint Version

model based on TL can produce accurate and reliable results by using least computation resources in a feasible time duration. The overall multiclass prediction accuracy of hierarchical BMIC_Net ranged from 92.45% to 95.48%, and the AUC values ranged from 0.92 to 0.97. In addition, the performance of proposed hierarchical classification model is compared with non-hierarchical model (Figure-10). It can be clearly seen from Figure-10, the non-hierarchical classification model is unable to show better performance (90.45% accuracy) compared to hierarchical BMIC_Net model. The reason behind the success of hierarchical classification approach lies in the inherent characteristics of BreakHis dataset. The BreakHis dataset contains the images belonging to eight classes. Of these eight classes, four belongs to Benign and remaining four belong to Malignant. Thus, to take an advantage of this dataset characteristic, we came up with the idea of hierarchical classification approach. Hence, the proposed model produced promising results and showed better accuracy with reliability, and it can be deployed as a classification model for the early diagnosis and prognosis of lethal lesions in BC. In specific, the proposed classification model can be used as a second opinion where expert pathologists are unavailable, such as in rural areas of under-developing countries. Han et al. (2017) classified BC using the BreakHis dataset through a CSDCNN model and obtained a classification accuracy of 93.2%. Nonetheless, the accuracy obtained through CSDCNN was very high; this model is computationally very expensive and needs extensive resources in training and extracting useful features. This is because; initially this model was trained on ImageNet dataset to construct a pre-trained CSDCNN. Afterwards, BreakHis images were trained on pre-trained CSDCNN model to classify BC images. Thus, CSDCNN might have taken extensive time and computational resources to achieve the accuracy of 93.2%. Conversely, our proposed model is less complex, consumes fewer resources, and obtains comparable accuracy to classify BC using the BreakHis dataset. This is because; our proposed BMIC_Net model was directly trained on BreakHis images and not on ImageNet like CSDCNN. Moreover, it uses the hierarchical classification model and this resulted in reduction of classification time. Table-5 shows the comparison of the proposed model results with the existing baseline models results using the BreakHis dataset. As can be seen here, the proposed model obtained the highest classification accuracy and AUC values across existing baselines. Noticeably, we are unable to compared any other performance metric because, in the given baseline models none of the study have used any other common evaluation measure except accuracy. However, some of the baseline studies like Spanhol et al. (2016) shown 0.848 value of AUC, Spanhol et al. (2017) reported 88.0 score of F1 measure. Similarly, Nahid et al. (2018) presented F-measure by 96%. All aforementioned baseline studies had shown the overall performance instead of subset classification performance measure. Finally, we have decided to compare commonly used performance measure (i.e. accuracy) used by all baseline studies. Apart from accuracy comparison, most of the baseline studies did not reported the training time except Han et al. (2017). In Han et al. (2017) the author first created a model and trained from scratch by using ImageNet dataset, referred as pretraining. Afterward, pretrained model has been fine-tuned and retrained by using BreakHis dataset. However, the author reported only retraining time i.e. 10 hours and 13 Minutes, but did not reported the pretraining time of newly created model.

The classification accuracy of the model generally relies on the quality of features used. Insufficient, superfluous, and irrelevant features may lead to poor and inconvincible results. Therefore, extracting relevant and discriminatory features from MFV by different feature selection and reduction schemes before classification is crucial Fukunaga (2013). The intention behind selecting result-oriented features is to ensure that the classifier will only focus on result-oriented features and is not distracted or misguided by unwanted attributes. Moreover, it enables the proposed model to utilize least computational resources and processing time to achieve the primary goal of this research. In this study, numerous subsets of features were analyzed to evaluate the performance of the model in terms of overall accuracy and AUROC values. Thus, the best feature subset size for the classification of BC must be determined, and the classifier overall efficiency must be enhanced. A group of 50 features was selected using the above-mentioned feature selection schemes to assess the behavior of all six classifiers and thus optimize the feature subset size. Subsequently, accuracy was computed after increasing the feature subset size by 50 in each iteration (total of 82 iterations) to cover the 4096 overall extracted features. The performance accuracy of the model increased by up to 2.5% when the size of the feature subset was increased from 50 to 3600. Moreover, a slightly lower accuracy was observed on only 50 feature subsets. Therefore, the feature subset can be selected in terms of the highest accuracy on a large number of feature sets (e.g., 95.48% using 900 features) or the least number of feature subsets by slightly compromising accuracy (e.g., 94.22% using 250 features), see Figure 7 (a). Hence, a large number of attributes do not contribute to the enhancement of classification predictive accuracy. Therefore, to mine the optimum size of features from MFV, data analysists are advised to perform sensitivity analysis to evaluate a range of feature sizes, starting from 50 features until no more improvement in accuracy is observed.

Preprint Version

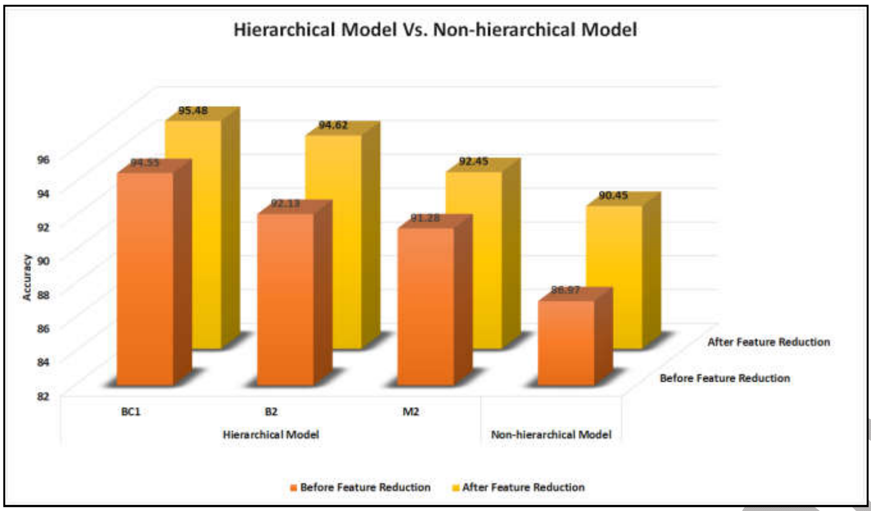


Figure 10: Performance Comparison of Proposed Hierarchical Model with Non-Hierarchical Model, Before and After Feature Reduction

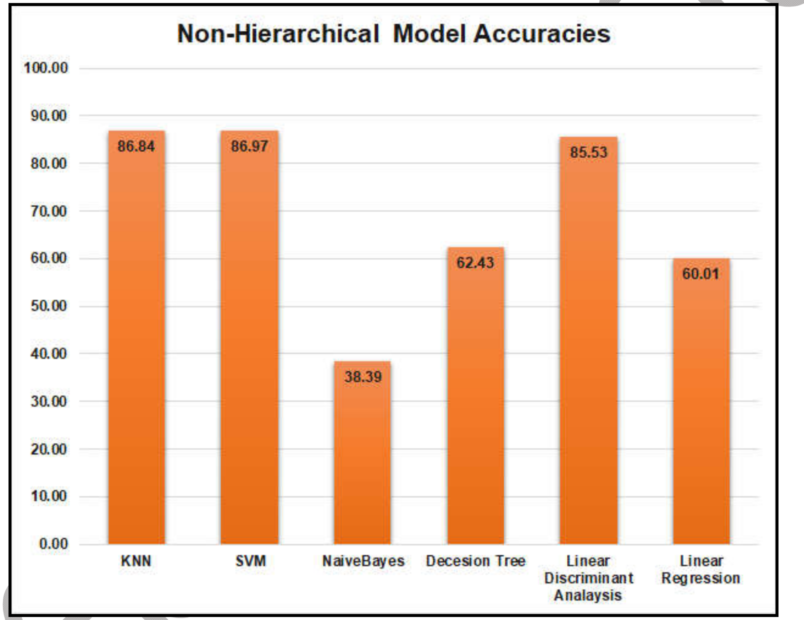


Figure 11: Non-hierarchical model accuracies using six traditional ML algorithms.

Table-5: Comparison of proposed models with baseline models using the BreakHis dataset

Study	Approach	Preeminent Classifier	No. of Classes	Overall Accuracy	Training Time Duration	Limitations
(Spanhol et al., 2016)	Traditional machine learning with feature optimization	SVM	2	85%	Not reported	It only considers binary classification for breast cancer prediction. Overall accuracy needs improvement
(Samah et al., 2017)	Traditional machine learning with	*k*NN	2	86%	Not reported	It only considers binary classification for BC prediction.

Study	Approach	Preeminent Classifier	No. of Classes	Overall Accuracy	Training Time Duration	Limitations
	feature optimization					Overall accuracy needs improvement
(Spanhol et al., 2017)	Transfer learning	softmax	2	90%	Not reported	It only considers binary classification for BC prediction. Overall accuracy needs improvement
(Nahid et al., 2018)	Feature fusion extracted from transfer learning and proposed deep learning model	softmax, SVM	2	91%	Not reported	It only considers binary classification for BC prediction. Overall accuracy needs improvement
(Han et al., 2017)	Proposed CSDCNN deep learning model	softmax	8	93%	10 Hrs 13 Mins (Pretraining time duration is not reported)	Computationally expensive and requires high computational resources. Model was pretrained on ImageNet dataset before using BreakHis images for target training. Proposed model is too complex.
Proposed Work	Feature extraction through transfer learning. Feature selection and classification through traditional machine learning. Proposed hierarchical classification approach	*k*NN	8	BC_1 = 95.48%, B_2 = 94.62%, M_2 = 92.45%	20 Hrs Average	Training time is extensive because all experiments are performed on an ordinary workstation with Corei7 CPU and 8 GB RAM.

In several analyses, IG was observed to perform better than PCA because of certain limitations. First, PCA assumes that principal components are linear combinations of the original data and such components are rarely found in real-life data, especially in multidimensional data, such as medical images. Therefore, PCA may not provide satisfactory results. Second, PCA creates high variance axes for its principal components; thus, low

Preprint Version

variance axes are treated as noise. Third, PCA assumes that the PC are orthogonal; thus, if data is either in the shape of sphere or sin wave or sparse on a plane, then PCA may not be an effective choice for feature reduction (Abdi & Williams, 2010). Conversely, IG is the amount of information that is obtained by calculating the entropy (purity or impurity in an arbitrary collection of examples) of an attribute. Moreover, the largest IG is equivalent to the smallest entropy. In addition, IG determines the contribution (purity) of an attribute in its class prediction. Therefore, features that perfectly contribute toward classification task should yield maximal IG value. Conversely, nondiscriminative features should yield least IG values. Finally, a value (shows its purity) assigned to each attribute is called IG, and on the basis of IG values, features can be ranked and selected best features with high prediction probability. IG is easy to implement thus may select reliable, result-oriented, and discriminative feature subsets (Kent, 1983).

According to the "no free lunch" theorem (David H Wolpert & Macready, 1995), no single traditional ML algorithm performs best in all domains for data mining. Hence, the results of multiple traditional ML algorithms warrant verification. Therefore, the performances of the six traditional ML algorithms (*k*NN, SVM, NB, DT, LDA classifier, and LR) were evaluated to classify BC images. In top-level classification, *k*NN outperformed all five other traditional ML algorithms. Nonetheless, the performance of LR and SVM was slightly less than that of *k*NN in top-level classification. Justifiable reasons exist for the outperformance of *k*NN in top-level classification. First, *k*NN performs better if the number of samples is large enough. Second, it can be applied to the data of any distribution; even if the data is large enough and inseparable with a linear boundary, it shows remarkable results. Third, it is robust to noisy data to reduce the probability of misclassification. Fourth, by nature, it is an extremely flexible algorithm for feature or distance choices and well suited for multi-modal classes. Fifth, building a model is simple and computationally cheap and can sometimes be the best method (Kuramochi & Karypis, 2005).

In the experimental results of this research, LR and SVM outperformed in many analyses because of some reasons. First, LR can determine the relative influence of one or more predictor variables to the criterion value. For instance, adjacent pixels usually have a high correlation in images. The second reason can be the ability to identify outliers or anomalies that can help reduce misclassification (Kotsiantis, Zaharakis, & Pintelas, 2007). However, the SVM classification model is suitable for both linear and nonlinear data, such as medical images. In addition, it is highly effective and performs a nonlinear mapping to transform the original training data into a higher dimension. Furthermore, it is memory efficient and uses a subset of training points (support vectors) as decisive factors for classification. Finally, it is a versatile algorithm that creates a novel kernel for a specific decision function until it provides correct results. Therefore, one can define customized kernels based on specific requirements (Kotsiantis et al., 2007).

Experimental results showed that NB, DT, linear regression, and LDA showed a substantial reduction in performance than *k*NN and SVM in most of the experimental analyses. NB yielded the lowest results due to some reasons. First, it assumes that the features are independent of each other to predict a class label. However, in reality, data attributes are correlated. Thus, NB cannot obtain better classification results. Second, a systemic problem is related to the imbalanced training instance per class. Given this reason, NB takes poor weight for decision boundary and may provide biased results. Third, NB estimates a possible likelihood value, between 0 to 1, which can lead to numerically unstable results. Finally, it is specifically related to the continuous feature preprocessing step. In general, one can use binning strategy to convert continuous values into discrete values before using the NB classifier. This process can cause loss of discriminating information (Rennie, Shih, Teevan, & Karger, 2003).

The logic behind the weak predictive performance of the DT classification algorithm may be because it usually produces a weak or noisy classifier. Moreover, it cannot generalize well, and the optimal DT considerably suffers from a small change in the training set. Moreover, DT is unsuitable for numerical data (like images), and complex trees can be created by binary tree splits. Finally, it mostly grows large and needs pruning, which may cause loss of information (Kotsiantis et al., 2007). A possible reason for the low performance of the discriminant analysis classifier may be because it does not work well in data with class imbalance. In addition, it is basically used for binary classification but extended for multi-classification problems. Therefore, it can exhibit less performance than any other multiclass classifiers. Furthermore, it may be unsuitable for nonlinear problems, such as images having nonlinearly spread data. Moreover, it shows unstable results if the classes are well separated. Finally, LDA is sensitive to overfitting; thus, it needs rigorous validation or testing (Kotsiantis et al., 2007).

Preprint Version

6. Conclusion

A multiclass BMIC_Net classification model was developed for early prediction of BC through hierarchical classification. The top-level classifier predicts if the BC is benign or malignant, and the second-level classifier predicts further subtypes of benign or malignant BC. The BreakHis publicly available dataset was used for training and testing the proposed BMIC_Net. Furthermore, the features were extracted through TL after fine-tuning the pre-trained AlexNet architecture and to obtain the MFVs. These vectors contained enormous features. Thus, the most discriminative features were elicited through IG and PCA. Finally, the six traditional ML algorithms were applied on extracted subsets of features to evaluate the classification performance. Results of several analyses showed that IG outperformed PCA in obtaining the most discriminative subset of features. Furthermore, *k*NN outperformed then all other traditional ML algorithms and obtained the highest accuracies of 95.48% (Sn = 93.55, 97.16), 94.62% (Sn = 96.97, 96.05, 93.65, 91.11) and 92.45% (Sn = 88.52, 97.87, 91.80, 93.02) in the BC_1, B_2, and M_2 models, respectively. To show the effectiveness of BMIC_Net, we compared the results obtained using the hierarchical-based classification model with those obtained using a one-level classification model (Han et al., 2017). Results showed that the hierarchical-based classification model outperformed the one-level classification model. The proposed BMIC_Net model produces promising results compared with existing baseline models. Furthermore, it consumes limited computational resources to predict BC from the eight distinct classes. Furthermore, the proposed model can be deployed on any normal desktop computer in any health sector of deprived areas in under-developing countries. In future work, a generic classification model for the early prediction of BC may be constructed. This generic model will be responsible for classifying any type of imaging modality, such as mammography, ultrasound, HI, and MRI, for predicting BC.

Acknowledgement

This research was fully funded by the University Malaya Research Grant – Frontier Science (Grant No: RG380-17AFR).

Preprint Version

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics, 2*(4), 433-459.

Aggarwal, C. C., & Zhai, C. (2012). *Mining text data*: Springer Science & Business Media.

Al-masni, M. A., Al-antari, M. A., Park, J. M., Gi, G., Kim, T. Y., Rivera, P., . . . Kim, T. S. (2018). Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system. *Computer Methods and Programs in Biomedicine, 157*, 85-94. doi:10.1016/j.cmpb.2018.01.017

Antropova, N., Huynh, B. Q., & Giger, M. L. (2017). A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. *Medical Physics, 44*(10), 5162-5171. doi:10.1002/mp.12453

Babu, J. S., Sukumar, L. B., & Anandan, K. (2013). Quantitative Analysis of Digitized Mammograms Using Nonsubsampled Contourlets and Evolutionary Extreme Learning Machine. *Journal of Medical Imaging and Health Informatics, 3*(2), 206-213.

Bayramoglu, N., Kannala, J., & Heikkilä, J. (2016, 4-8 Dec. 2016). *Deep learning for magnification independent breast cancer histopathology image classification.* Paper presented at the 2016 23rd International Conference on Pattern Recognition (ICPR).

Bovis, K., Singh, S., Fieldsend, J., & Pinder, C. (2000). *Identification of masses in digital mammograms with MLP and RBF nets.* Paper presented at the Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on.

Buciu, I., & Gacsadi, A. (2009). *Gabor wavelet based features for medical image analysis and classification.* Paper presented at the Applied Sciences in Biomedical and Communication Technologies, 2009. ISABEL 2009. 2nd International Symposium on.

Buciu, I., & Gacsadi, A. (2011). Directional features for automatic tumor classification of mammogram images. *Biomedical Signal Processing and Control, 6*(4), 370-378.

Chougrad, H., Zouaki, H., & Alheyane, O. (2018). Deep Convolutional Neural Networks for breast cancer screening. *Computer Methods and Programs in Biomedicine, 157*, 19-30. doi:https://doi.org/10.1016/j.cmpb.2018.01.011

Cruz-Roa, A., Gilmore, H., Basavanhally, A., Feldman, M., Ganesan, S., Shih, N. N. C., . . . Madabhushi, A. (2017). Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent. *Scientific Reports, 7*. doi:10.1038/srep46450

Dimitropoulos, K., Barmpoutis, P., Zioga, C., Kamas, A., Patsiaoura, K., & Grammalidis, N. (2017). Grading of invasive breast carcinoma through Grassmannian VLAD encoding. *PLoS ONE, 12*(9), e0185110. doi:10.1371/journal.pone.0185110

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55*(10), 78-87.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters, 27*(8), 861-874.

Fukunaga, K. (2013). *Introduction to statistical pattern recognition*: Elsevier.

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.

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*, 4172. doi:10.1038/s41598-017-04075-z

Preprint Version

Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine learning, 45*(2), 171-186.

Jiang, J., Ma, J., Wang, Z., Chen, C., & Liu, X. (2018). *Hyperspectral Image Classification in the Presence of Noisy Labels* (Vol. PP).

Kasban, H., El-Bendary, M., & Salama, D. (2015). A comparative study of medical imaging techniques. *Int. J. Information Sci. Intelligent System, 4*, 37-58.

Kent, J. T. (1983). Information gain and a general measure of correlation. *Biometrika, 70*(1), 163-173. doi:10.1093/biomet/70.1.163

Khosravi, P., Kazemi, E., Imielinski, M., Elemento, O., & Hajirasouliha, I. (2018). Deep Convolutional Neural Networks Enable Discrimination of Heterogeneous Digital Pathology Images. *Ebiomedicine, 27*, 317-328. doi:10.1016/j.ebiom.2017.12.026

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering, 160*, 3-24.

Kowal, M., Filipczuk, P., Obuchowicz, A., Korbicz, J., & Monczak, R. (2013). Computer-aided diagnosis of breast cancer based on fine needle biopsy microscopic images. *Computers in Biology and Medicine, 43*(10), 1563-1572. doi:https://doi.org/10.1016/j.compbiomed.2013.08.003

Kozegar, E., Soryani, M., Minaei, B., & Domingues, I. (2013). Assessment of a novel mass detection algorithm in mammograms. *Journal of cancer research and therapeutics, 9*(4), 592.

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.

Kuramochi, M., & Karypis, G. (2005). Gene classification using expression profiles: A feasibility study. *International Journal on Artificial Intelligence Tools, 14*(04), 641-660.

Litjens, G., Sanchez, C. I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., . . . van der Laak, J. (2016). Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific Reports, 6*. doi:10.1038/srep26286

Loukas, C., Kostopoulos, S., Tanoglidi, A., Glotsos, D., Sfikas, C., & Cavouras, D. (2013). Breast Cancer Characterization Based on Image Classification of Tissue Sections Visualized under Low Magnification. *Computational and Mathematical Methods in Medicine, 2013*, 829461. doi:10.1155/2013/829461

Lu, T., Chen, X., Zhang, Y., Chen, C., & Xiong, Z. (2018). *SLR: Semi-coupled locality constrained representation for very low resolution face recognition and super resolution* (Vol. PP).

Ma, Y., Li, C., Li, H., Mei, X., & Ma, J. (2018). Hyperspectral Image Classification With Discriminative Kernel Collaborative Representation and Tikhonov Regularization. *IEEE Geoscience and Remote Sensing Letters, 15*(4), 587-591. doi:10.1109/LGRS.2018.2800080

Nahid, A.-A., Mehrabi, M. A., & Kong, Y. (2018). Histopathological Breast Cancer Image Classification by Deep Neural Network Techniques Guided by Local Clustering. *BioMed Research International, 2018*.

Naik, S., Doyle, S., Agner, S., Madabhushi, A., Feldman, M., & Tomaszewski, J. (2008). *Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology.* Paper presented at the Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on.

Preprint Version

Provost, F. J., & Fawcett, T. (1997). *Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions.* Paper presented at the KDD.

Provost, F. J., Fawcett, T., & Kohavi, R. (1998). *The case against accuracy estimation for comparing induction algorithms.* Paper presented at the ICML.

Rabidas, R., Midya, A., & Chakraborty, J. (2018). Neighborhood Structural Similarity Mapping for the Classification of Masses in Mammograms. *IEEE Journal of Biomedical and Health Informatics*, 1-1. doi:10.1109/JBHI.2017.2715021

Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003). *Tackling the poor assumptions of naive bayes text classifiers.* Paper presented at the Proceedings of the 20th international conference on machine learning (ICML-03).

Ribli, D., Horvath, A., Unger, Z., Pollner, P., & Csabai, I. (2018). Detecting and classifying lesions in mammograms with Deep Learning. *Scientific Reports, 8*. doi:10.1038/s41598-018-22437-z

Samah, A. A., Fauzi, M. F. A., & Mansor, S. (2017, 12-14 Sept. 2017). *Classification of benign and malignant tumors in histopathology images.* Paper presented at the 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA).

Shen, D., Wu, G., & Suk, H.-I. (2017). Deep Learning in Medical Image Analysis. *Annual review of biomedical engineering, 19*, 221-248. doi:10.1146/annurev-bioeng-071516-044442

Spanhol, F. A., Cavalin, P. R., Oliveira, L. S., 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.

Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2016). A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering, 63*(7), 1455-1462.

Surendiran, B., & Vadivel, A. (2010). Feature selection using stepwise ANOVA discriminant analysis for mammogram mass classification. *International J. of Recent Trends in Engineering and Technology, 3*(2), 55-57.

Swiniarski, R. W., Lim, H. K., Shin, J. H., & Skowron, A. (2006). *Independent Component Analysis, Princpal Component Analysis and Rough Sets in Hybrid Mammogram Classification.* Paper presented at the IPCV.

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

Wang, P., Hu, X., Li, Y., Liu, Q., & Zhu, X. (2016). Automatic cell nuclei segmentation and classification of breast cancer histopathology images. *Signal Processing, 122*, 1-13. doi:https://doi.org/10.1016/j.sigpro.2015.11.011

WHO, W. H. O. (2018). World Cancer Report.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann.

Wolpert, D. H., & Macready, W. G. (1995). *No free lunch theorems for search*. Retrieved from

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation, 1*(1), 67-82. doi:10.1109/4235.585893

Zhang, Y.-D., Pan, C., Chen, X., & Wang, F. (2018). Abnormal breast identification by nine-layer convolutional neural network with parametric rectified linear unit and rank-based stochastic pooling. *Journal of Computational Science, 27*, 57-68. doi:https://doi.org/10.1016/j.jocs.2018.05.005

Preprint Version

Zhang, Y., Tomuro, N., Furst, J., & Raicu, D. S. (2012). Building an ensemble system for diagnosing masses in mammograms. *International journal of computer assisted radiology and surgery, 7*(2), 323-329.

Zhang, Y., Wu, X., Lu, S., Wang, H., Phillips, P., & Wang, S. (2016). Smart detection on abnormal breasts in digital mammography based on contrast-limited adaptive histogram equalization and chaotic adaptive real-coded biogeography-based optimization. *SIMULATION, 92*(9), 873-885. doi:10.1177/0037549716667834

Preprint Version