JOURNAL OF MECHANICS OF CONTINUA AND MATHEMATICAL SCIENCES
www.journalimcms.org

ISSN (Online) : 2454 -7190 Vol.-17, No.-11, November (2022) pp 26-45 ISSN (Print) 0973-8975

BREAST CANCER HISTOLOGICAL IMAGES CLASSIFICATION AND PERFORMANCE EVALUATION OF DIFFERENT CLASSIFIERS

Md. Rakibul Islam[1], Shariful Islam[2], Md. Shahadot Hosen (Rony) [3] Md. Nur Alam[4]

[1,2]Department of Electrical and Electronic Engineering, Pundra University of Science and Technology, Bogura-5800, Bangladesh

[3] Department of Electrical and Electronic Engineering, Hajee Mohammad Danesh Science and Technology University, Dinajpur-5200, Bangladesh

[4]Department of Mathematics, Pabna University of Science and Technology, Pabna-6600, Bangladesh

Corresponding author: **Md Rakibul Islam**

Email: rakibhstu15@gmail.com[1], sharifulislamsharif560@gmail.com[2]

(Received: August 30, 2022; Accepted: November 8, 2022)

https://doi.org/10.26782/jmcms.2022.11.00002

Abstract

Breast cancer is a serious trouble and one of the greatest causes of death for women throughout the world. Computer-aided diagnosis (CAD) techniques can help the doctor make more credible decisions. We have determined the possibility of knowledge transfer from natural to histopathological [IX][XII] images by employing a pre-trained network ResNet-50.This pre-trained network has been utilized as a feature generator and extracted features are used to train support vector machine (SVM), random forest, decision tree, and K nearest nneighbour(KNN) classifiers[X]. We altered the softmax layer to support the vector machine classifier, random forest classifier, decision tree classifier, and k-nearest neighbour classifier, to evaluate the classifier performance of each algorithm. These approaches are applied for breast cancer classification and evaluate the performance and behavior of different classifiers on a publicly available dataset named Bthe reak-HIS dataset. In order to increase the efficiency of the ResNet[III] model, we preprocessed the data before feeding it to the network. Here we have applied to sharpen filter and data augmentation techniques, which are very popular and effective image pre-processing techniques used in deep models.

Keywords: Machine learning, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), RESNET (Residual Network) model, Random Forest.[VII]

I. Introduction

Breast cancer is the most common and dangerous cancer spreading significantly in women both in the developed and developing world in the last few

decades. According to a study by the world health organization (WHO), breast cancer is impacting 2.1 million women each year and also causes the greatest number of cancer-related death among women. In 2018 it is estimated that 627,000 women died from breast cancer that is approximately 15% of all cancer deaths among women.[XV]

Breast Cancer Rates

Recently, multi-classification of breast cancer from histopathological images was presented using a structured deep-learning model called CSDCNN. This new DL architecture shows the superior performance when compared to different machine learning and deep learning-based approaches on the BreakHIS dataset [VIII]. This model shows state-of-the-art performance for both image-level and patient-level classification. An average of 93.2% accuracy for patient-level breast cancer classification has been reported. In 2017, different SMV-based techniques were applied for breast cancer recognition, an accuracy of 94.97% for data with a 40X magnification factor was achieved using an Adaptive Sparse SVM (ASSVM).

Belgium had the highest rate of breast cancer in women, followed by Luxembourg.[IV]

Table : 1

Rank	Country	Age-standardized rate per 100000	Rank	Country	Age-standardized rate per 100000
1	Belgium	113.2	14	Denmark	88.8
2	Luxembourg	109.3	15	Switzerland	88.1
3	Netherlands	105.9	16	Montenegro	87.8
4	France (metropolitan)	99.1	17	Malta	87.6
5	New Caledonia (France)	98.0	18	Norway	87.5
6	Lebanon	97.6	19	Hungary	85.5
7	Australia	94.5	20	Germany	85.4
8	UK	93.6	21	Iceland	85.2
9	Italy	92.8	22	US	84.9
10	New Zealand	92.6	23	Canada	83.8
11	Ireland	90.3	24	Cyprus	81.7
12	Sweden	89.8	25	Samoa	80.1
13	Finland	89.5			

A breast tissue biopsy allows the pathologist to histologically access the microscopic level structures and components of the breast tissue. These histological images allow the pathologist to distinguish between normal tissue, non-malignant (benign) tissue, and malignant lesions.[XIX] Benign lesions refer to changes in normal tissue of breast parenchyma and are not related to the progression of malignancy. There are two different carcinoma tissue types including in-situ and invasive. The in-situ tissue type refers to tissue contained inside the mammary ductal-lobular[XV] On the other hand, the invasive carcinoma cells spread beyond the mammary ductal-lobular structure. The tissue samples that are collected during the biopsy are commonly stained with Hematoxylin[XXI] and Eosin (H & E) before the visual analysis performed by the specialist.

During the diagnosis process, the affected region is determined from whole-slide tissue scams. In addition, the pathologist analyses microscopic images of the tissue samples from the biopsy with different magnification factors. Nowadays, to produce the correct diagnosis, the pathologist considers different characteristics within the images including patterns, textures, and different morphological properties. Analyzing images with different magnification factors requires panning, zooming, and focusing scanning of each image in its entirety. This process is very time consuming and tiresome, as a result this manual process sometimes leads to inaccurate diagnoses for breast cancer identification.

Due to the advancement of digital imaging techniques [XIV, XV] in the last decade, different computer vision and machine learning techniques have been applied for analyzing pathological images at a microscopic resolution. Recently, deep learning-based approaches were shown to outperform conventional machine learning methods in many image analysis tasks, automating end-to-end processing. In the domain of medical imaging [XXI], Conventional Neural Networks (CNN) [XXVI] have been successfully used for diabetic retinopathy screening, bone disease prediction and age assessment, and other problems. Previous deep learning-based applications in histological microscopic image analysis have demonstrated their potential to provide utility in diagnosing breast cancer. Although these approaches have shown tremendous success in medical imaging, they require a very large amount of label data which is still not available in this domain of applications for several reasons.

In this paper, we propose a Deep Learning (DL)[XXVII],[XXI] based approach for a breast cancer recognition system using a support vector machine (SVM)[XI], random forest[XXVII], decision tree, k-nearest neighbor (KNN) classifiers[XX]. A pre-trained Res-Net 50 model is utilized to extract features from the BreakHIS dataset. More recently, several Res-Net (Residual Network) based architectures are developed, each of which improves previous State-of-the-art performance on the benchmark dataset. The pre-trained 50-layer Res-Net model is very much feasible to handle and also has better performance. To enhance the image quality and reduce noise, it is necessary to preprocess the data. A sharpening filter is one of the effective pre-processing methods [XI]. It sharpens the edges of the images and makes it easier to identify the structure correctly. We also used data augmentation techniques to alleviate the problem of limited data size.

The contribution of this paper are summarized as follows:

- Successful breast cancer classification using Res-Net 50 model as feature generator and classifiers (SVM, Random Forest, Decision Tree, KNN).
- Experiments have been conducted on recently released publicly available datasets for breast cancer histopathology (BreakHIS dataset) where we evaluated images with different magnifying factors (including 40x, 100x, 200x and 400x).
- Transfer learning is applied in breast cancer classification.
- The performance of different classifiers is evaluated and compared.

Breast Cancer

Breast cancer refers to cancers occurring in the breast tissues, developed due to uncontrolled division of different cells of the breast. The risk factor can be genetic but some lifestyle factors such as alcohol intake make it more likely to happen. The first symptoms of breast cancer are usually an area of thickened tissue in the breast or a lump in the breast or armpit. The risk increases with age. At 20 years, the chance of developing breast cancer in the next decade is 0.6 percent. By the age of 70 years, this figure goes up to 3.84 percent. Women who carry the BRCA1 and BRCA2 genes have a higher risk of developing cancer later on. Examples include atypical ductal hyperplasia or lobular carcinoma in situ. Other factors could be obesity, radiation exposure, hormone treatments, cosmetic implants, and breast cancer survival. Early diagnosis and proper treatment can reduce the risk of death.

Types of Breast Cancer

There are two main types of breast cancer: In situ-breast cancers have not spread. Invasive or infiltrating cancers have spread into the surrounding breast tissue.

In Situ Cancer

Ductal carcinoma in situ (DCIS) [XIII] is also called intra-ductal carcinoma and Stage 0 breast cancer. DCIS is a non-invasive or pre-invasive breast cancer. This means the cells that line ducts have changed to cancer cells but they have not spread through the walls of the ducts into the nearby breast tissue. Because DCIS hasn't spread into the breast tissue around it, it can't be spread (metastasize) beyond the breast to other parts of the body. DCIS is considered a pre-cancer because sometimes it can become invasive cancer.

Lobular carcinoma in situ (LCIS) may also be called lobular neoplasia. In this breast change, cells that look like cancer cells are growing in the milk-producing glands of the breast (called the lobules) but they don't grow through the wall of the lobules. LCIS is not considered to be cancer and it typically doesn't spread beyond the lobule (become invasive breast cancer) if it isn't treated. But having LCIS does increase your risk of developing invasive breast cancer in either breast, later on, so close follow-up is important.

Invasive (infiltrating) breast cancer

Breast cancer that has been spread into surrounding breast tissue are known

as invasive breast cancer. There are different kinds of invasive breast cancer. Some kinds are more common than others.

Invasive (infiltrating) ductal carcinoma (IDC)[XXXIII]**:**

This is the most common type of breast cancer. About 8 of 10 invasive breast cancers are invasive (or infiltrating) ductal carcinomas (IDC). IDC starts in the cells that line a milk duct in the breast, breaks through the wall of the duct, and grows into the nearby breast tissues. At this point, it may be able to spread (metastasize) to other parts of the body through the lymph system and bloodstream.

Invasive lobular carcinoma (ILC):

Invasive lobular carcinoma (ILC) starts in the milk-producing glands (lobules). Like IDC, it can spread (metastasize) to other parts of the body. About 1 invasive breast cancer in 10 is an ILC. Invasive lobular carcinoma may be harder to detect on physical exams as well as imaging, like mammograms, than invasive ductal carcinoma. And compared to other kinds of invasive carcinoma, about 1 in 5 women with ILC might have cancer in both breasts.

There are some special types of breast cancer that are sub-types of invasive carcinoma. Some of these may have a better prognosis than standard invasive infiltrating ductal carcinoma. These includes:

- Adenoid cystic or adenocystic carcinoma
- Low-grade adenosquamous carcinoma (this is a type of metaplastic carcinoma)
- Medullary carcinoma
- Mucinous (or colloid) carcinoma
- Papillary carcinoma
- Tubular carcinoma

Some sub-types have the same or maybe worse prognoses than standard invasive infiltrating ductal carcinoma. These include:

- Metaplastic carcinoma (most types, including spindle cell and squamous, except low-grade adenosquamous carcinoma)
- Micropapillary carcinoma
- Mixed carcinoma (has features of both invasive ductal and lobular)

L*ess Common* Breast Cancer

Inflammatory breast cancer is an uncommon type of invasive breast cancer. It accounts for about 1% to 5% of all breast cancers.

Paget disease of the nipple starts in the breast ducts and spreads to the skin of the nipple and then to the areola (the dark circle around the nipple). It is rare, accounting for only about 1-3% of all cases of breast cancer.

Phyllodes tumors [XXVII**]** are rare breast tumors. They develop in the connective tissue (stroma) of the breast, in contrast to carcinomas, which develop in the ducts or lobules. Most are benign but there are others that is malignant cancer.
Angiosarcoma-Sarcomas[V] of the breast are rare making up less than 1% of all

breast cancers. Angiosarcoma starts in cells that line blood vessels or lymph vessels. It can involve the breast tissue or the skin of the breast. Some may be related to prior radiation therapy in that area.

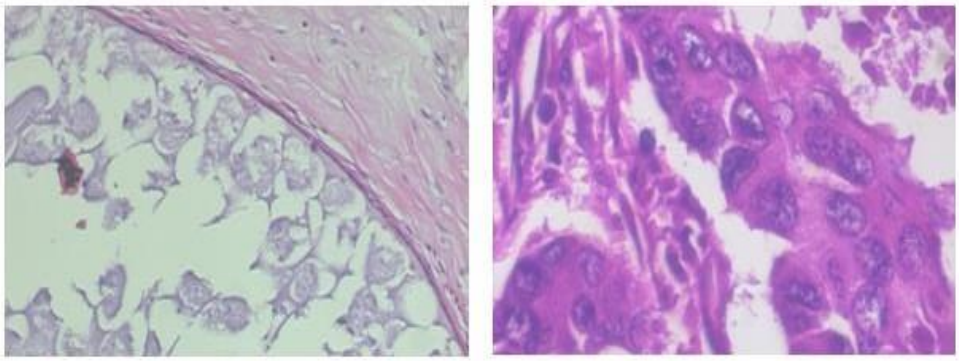

Fig. 1. Benign and Malignant Breast cancer cell histo-pathological image

II. Residual Learning [XIV]

The analysis of the cause and effects of vanishing gradients demonstrates an effective solution to enable the training of ultra-deep networks. [XIX]

Assuming you have a seven-layer network. In a residual setup, you would not only pass the output of layer 1 to layer 2 and on, but you would also add up the outputs of layer 1 to the outputs of layer 2.

Denoting each layer by ***f(x)***, in a standard network ***y = f(x)***
However, in a residual network,

$$y = f(x) + x$$

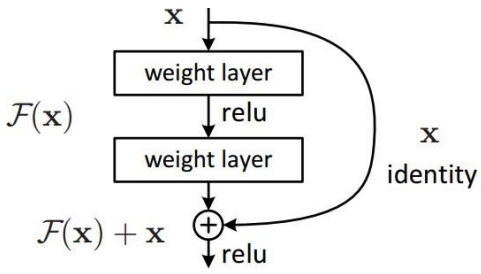

Fig. 2. Residual block

The idea has since been expanded into all other domains of deep learning including speech and natural language processing.

Support Vector Machine (SVM)

Linear support vector machines (SVM) is originally formulated for binary classification. Given training data and its corresponding labels $(x_n; y_n)$, n =1…..N, x_n

2 RD, t_n 2 f(1); +1g, SVMs learning consists of the following constrained optimization: [XXVIII]

$$min_{w,\xi_n} \frac{1}{2} w^T + C \sum_{n=1}^{N} \xi_n$$

$$\text{s.t.} \quad w^T x_n t_n \geq 1 - \xi_n \quad \forall_n$$

$$\xi_n \geq 0 \quad \forall n$$

$$min_w \frac{1}{2} w^T w + C \sum_{n=1}^{N} \max(1 - w^T x_n t_n, 0)$$

The objective of Eq. 5 is known as the primal form problem of L1-SVM, with the standard hinge loss. Since L1-SVM is not differentiable, a popular variation is known as the L2-SVM which minimizes the squared hinge loss:

$$min_w \frac{1}{2} w^T w + C \sum_{n=1}^{N} \max(1 - w^T x_n t_n, 0)^2$$

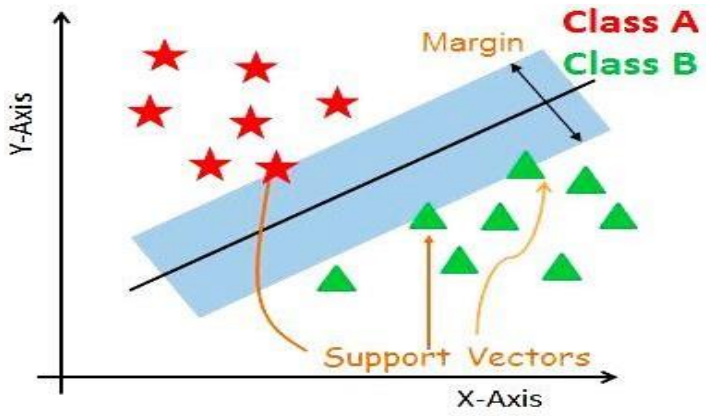

Fig.3. Support vector machine explained

Random Forest

The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set *X* = *x1* ... *xn* with responses *Y* = *y1*... *yn*, bagging
Repeatedly (*B* times) selects a random sample with replacement of the training set and fits trees to these samples: [XXI]

For *b* = 1... *B*:

Sample, with replacement, *n* training examples from *X*, *Y*; call these *Xb*, *Yb*.

Train a classification or regression tree *fb* on *Xb*, *Yb*.

After training, predictions for unseen samples *x′* can be made by averaging the predictions from all the individual regression trees on *x′*:

$$\hat{f}=\frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

An estimate of the uncertainty of theprediction can be made as the standard deviation of the predictions from all the individual regression trees on *x'*:

$$\sigma=\sqrt{\frac{\sum_{b-1}^{B}(f_b(x')-\hat{f})^2}{B-1}}$$

The number of samples/trees, B, is a free parameter

To compute the Gini impurity for a set of items with classes, suppose, and let be the fraction of items labeled with class in the set.

$$I_g(p) = \sum_{i=1}^{j} P_i \sum_{k\neq i} P_k = \sum_{i=1}^{j} P_i(1\text{-}P_i)$$

K-Nearest Neighbor (KNN)

Parameter selection

The best choice of *k* depends upon the data; generally, larger values of *k* reduce the effect of the noise on the classification but make boundaries between classes less distinct. A good *k* can be selected by various heuristic techniques. The special case where the class is predicted to be the class of the closest training sample (i.e. when *k* = 1) is called the nearest neighbor algorithm.[I]

The accuracy of the *k*-NN algorithm can be severely degraded by the presence of noisy or irrelevant features, or if the feature scales are not consistent with their importance. Much research effort has been put into selecting or scaling features to improve the classification. A particularly popular approach is the use of evolutionary algorithms to optimize feature scaling. Another popular approach is to scale features by the mutual information of the training data with the training classes.

In binary (two-class) classification problems, it is helpful to choose *k* to be an odd number as this avoids tied votes. One popular way of choosing the empirically optimal *k* in this setting is via the bootstrap method.

There are many results on the error rate of the *k* nearest neighbor classifiers. The *k*-nearest neighbor classifier is strongly consistent provided k := kn diverges and kn / n converges to zero as n ∞

Let C^{knn} denote the *k* nearest neighbor classifier based on a training set of size *n*. Under certain regularity conditions, the excess risk yields the following asymptotic expansion

$$\mathcal{R}_{\mathcal{R}}(C_{nknn})\text{-}\,\mathcal{R}_{\mathcal{R}}(C_{Bayes})=\left\{B_1\frac{1}{K}+B_2\left(\frac{K}{n}\right)^{\frac{4}{d}}\right\}\{1+O(1)\}$$

for some constants B1 and B2.

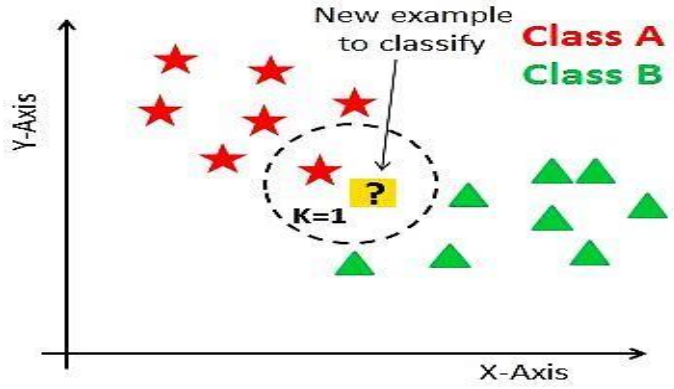

Fig.4. KNN explained

III. Overview

In our proposed method, firstly we have pre-processed the Break HIS dataset using a sharpening filter and data augmentation techniques. We fed the pre-processed data to the pre-trained ResNet-50 model. We used the ResNet-50 model as a feature generator. The CNN model itself consists of a feature extractor as well as a classifier, with convolution and pooling layers working as feature extractors and a softmax [III] layer working as a classifier. In this program, we altered the softmax layer to support the vector machine classifier, random forest classifier, decision tree classifier, and k-nearest neighbor classifier, to evaluate the classifier performance of each algorithm instead of softmax.

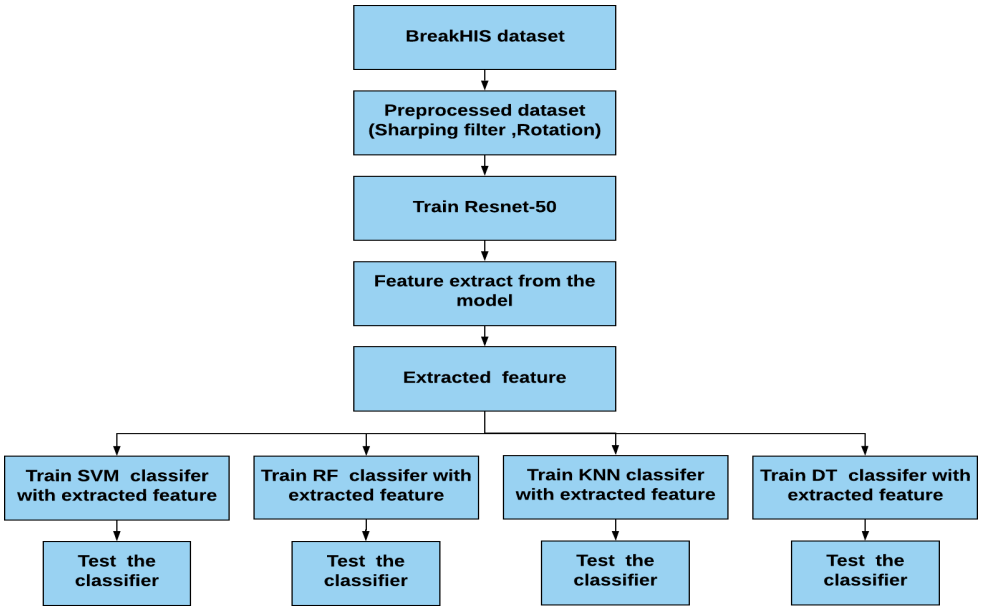

Fig. 5. Block diagram of proposed methodology.

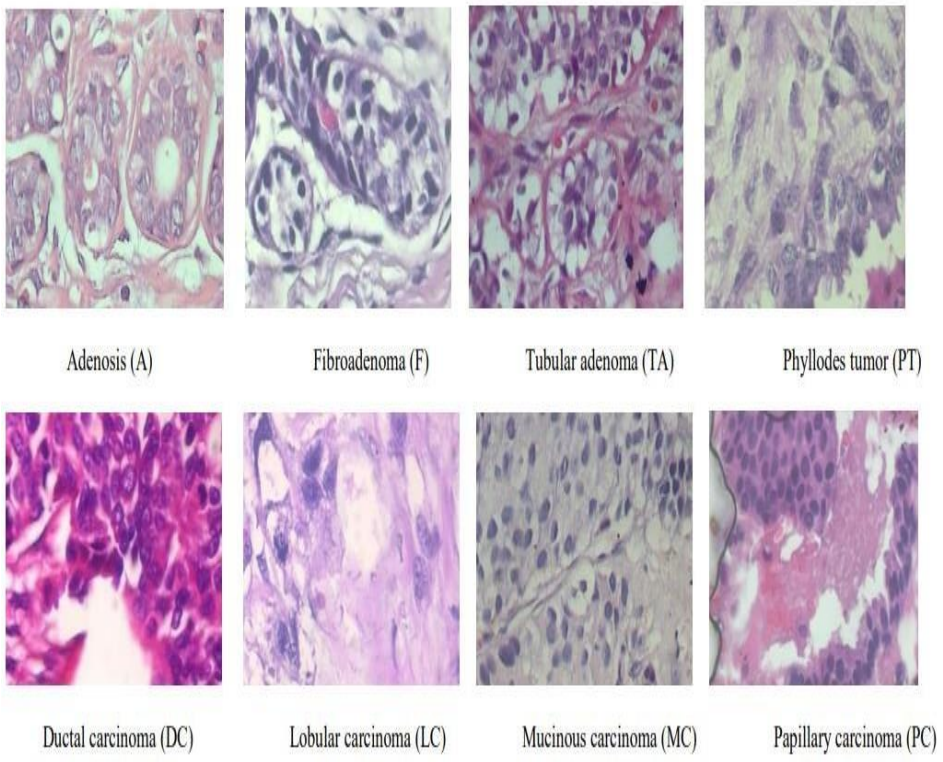

Fig. 6. Different types of Breast cancer cell

Generally, in the medical domain, the dataset consists of limited samples due to a complicated and expensive procedure of data collection. So it is very challenging to access a relevant and large-scale well-annotated dataset. A well-annotated dataset provides a common platform for the comparison and validation of purposed models. Thus, to develop a robust model, the availability of a large-scale well-annotated dataset is the foremost requirement. In this work, a publicly available BreakHis dataset is used built in collaboration with the Prognostics and Diagnostics (P&D) Laboratory, Parana, Brazil.

Breast Cancer Histopathological Dataset (BreakHIS)

The Breast Cancer Histopathological Image Classification (BreakHIS)[IX] is composed of 7,909 microscopic images of breast tumor tissue collected from 82 patients using different magnifying factors (40X, 100X, 200X, and 400X). To date, it contains 2,480 benign and 5,429 malignant samples (700X460 pixels, 3-channel RGB, 8-bit depth in each channel, PNG format).

The dataset BreakHIS is divided into two main groups: benign tumors and malignant tumors. Histologically benign is a term referring to a lesion that does not match any criteria of malignancy – e.g., marked cellular atypia, mitosis, disruption of basement membranes, metastasize, etc. Normally, benign tumors are relatively "innocents", present slow growing and remains localized. Malignant tumor is a

synonym for cancer:[XII] lesion can invade and destroy adjacent structures (locally invasive) and spread to distant sites (metastasize) to cause death.

The dataset currently contains four histological distinct types of benign breast tumors: adenosis (A), fibroadenoma (F), phyllodes tumor (PT), and tubular adenona (TA); and four malignant tumors (breast cancer): carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC) and papillary carcinoma (PC).

In the current version, samples present in the dataset were collected by the SOB method [XVIII], also named partial mastectomyor excisional biopsy. This type of procedure, compared to any method of needle biopsy, removes the larger size of the tissue sample and is done in a hospital with a general anesthetic.

Table 2: Samples and number of patients of BreakHIS dataset.

Class	Sub classes	Number of patients	Magnification factors				Total
			40x	100x	200x	400x	
Benign	A	4	114	113	111	106	344
	F	10	253	260	264	237	1014
	TA	3	109	121	108	115	453
	PT	7	149	150	140	130	569
Malignant	DC	38	864	903	896	788	3451
	LC	5	156	170	163	137	626
	MC	9	205	222	196	169	792
	PC	6	145	142	135	138	560
Total		82	1995	2081	2013	1820	7909

The statistics for this dataset are given in Table 1. In this experiment, we used 70% of the samples for training and 30% of the samples for testing. To generalize the classification task to perform successfully when testing new patients, we ensure that the patients selected for training are not used during testing.

Sharpening filter

Human perception is highly sensitive to edges and fine details of an image and since they are composed primarily of high-frequency components, the visual quality of an image can be enormously degraded if the high frequencies are attenuated or completely removed. In contrast, enhancing the high-frequency components of an image leads to an improvement in the visual quality. Image sharpening refers to any enhancement technique that highlights edges and fine details in an image.

In image processing filters are mainly used to suppress either the high frequencies in the image, *i.e.* smoothing the image, or the low frequencies, *i.e.* enhancing or detecting edges in the image. An image can be filtered either in frequency or in the spatial domain.

The first involves transforming the image into the frequency domain, multiplying it with the frequency filter function, and re-transforming the result into the spatial domain. The filter function is shaped to attenuate some frequencies and enhance others. For example, a simple low pass function is *1* for frequencies smaller than the cut-off frequency and *0* for all others.

The corresponding process in the spatial domain is to convolve the input image *f*(*i*,*j*) with the filter function *h*(*i*,*j*). This can be written as $g(i, j) = h(i, j)f(i, j)$

The mathematical operation is identical to the multiplication in the frequency space, but the results of the digital implementations vary since we have to approximate the filter function with a discrete and finite kernel. For a square kernel with size M× M, We can calculate the output image with the following formula:

$$g(i,j)=\sum_{m=-\frac{M}{2}}^{\frac{M}{2}}\sum_{n=-\frac{M}{2}}^{\frac{M}{2}} h(m,n)f(i-m,j-n)$$

The discrete convolution can be defined as a *`shift and multiply'* operation, where we shift the kernel over the image and multiply its value with the corresponding pixel values of the image.

$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$
$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$
$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$

Mean

0	-1	0
-1	4	-1
0	-1	0

Median

Fig. 7. Convolution kernel for a mean filter and discrete Laplacian.

In contrast to the frequency domain, it is possible to implement non-linear filters in the spatial domain. In this case, the summations in the convolution function are replaced with some kind of non-linear operator:

$$g(i, j) = O_{m,n}[h(m, n)f(i - m, j - n)]$$

For most non-linear filters the elements of *h* (*i*, *j*) are all 1. A commonly used non-linear operator is the median, which returns the `middle' of the input values.

The operators included in this section are those whose purpose is to identify meaningful image features based on distributions of pixel gray levels. The two categories of operators included here are:

Edge Pixel Detectors that [XXVI] assign a value to a pixel in proportion to the likelihood that the pixel is part of an image edge (*i.e.* a pixel that is on the boundary between two regions of different intensity values).

Line Pixel Detectors that assign a value to a pixel in proportion to the likelihood that the pixel is part of an image line (*i.e.* a dark narrow region bounded on both sides by lighter regions, or vice-versa).

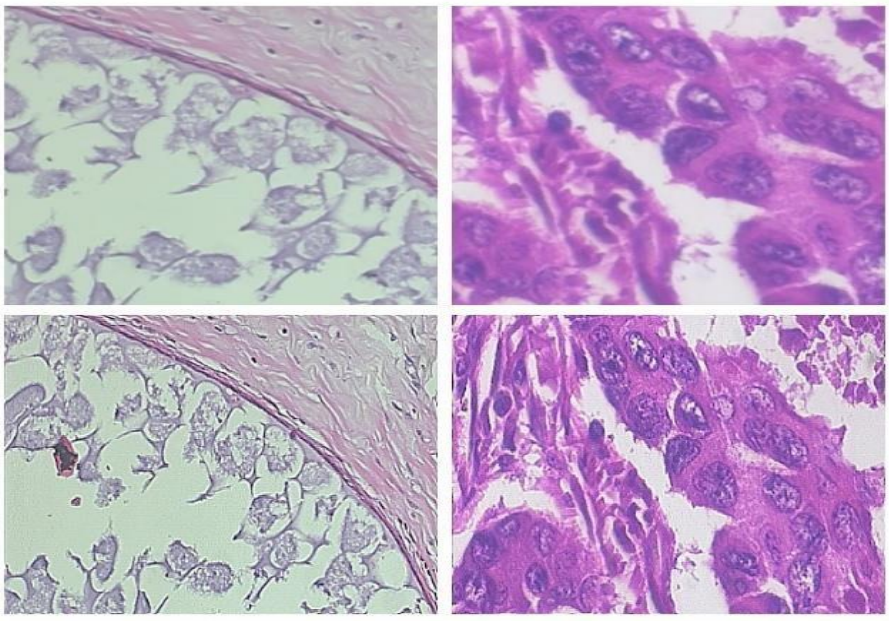

Fig. 8 1st row indicates the original image and 2nd row preprocessed image.

IV. Data Augmentation

Unbalanced and limited data size is the major challenge in the development of robust computer-aided diagnosis (CAD) systems. Data augmentation is an approach used in deep models to enlarge the dataset to alleviate the problem of limited data size. Some popular data augmentation techniques like flipping, cropping, scaling, rotation, interpolation, translation, and noise insertion have already been applied in many previous studies. But all augmentation approaches used for natural images would not work over medical images, as a lot of medical images are a top-down solved problem

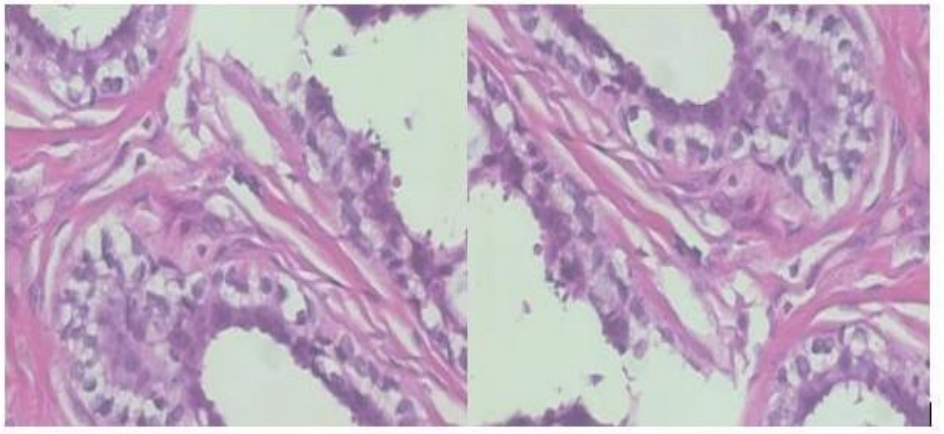

Figure 9: Rotation of image about 180°

Since the histopathological images possess rotation and reflection symmetry, so there is always a possibility to eliminate some inherent properties from the image if other augmentation techniques are used which may cause a loss of discriminating features or information. Consequently, for both the full training and transfer learning, only rotation is employed as a data augmentation technique. Rotations of the images are carried out about their center with the three angles: 90◦, 180◦, and 270◦. Another advantage of using the augmentation process lies in the minimization of overfitting, which is a major issue in the successful implementation of a machine-learning model. After preprocessing, we fed the dataset to our pre-trained ResNet-50 model.

Support Vector Machine

To compute Gini impurity for a set of items with classes, suppose, and let be the fraction of items labeled with class in the set.

$$I_G(p)=\sum_{i=1}^{J} P_i \sum_{k\neq i} P_k=\sum_{i=1}^{J} P_i\,(1-P_i)=1-\sum_{i=1}^{j} {P_i}^2$$

Information gain

Information gain is based on the concept of entropy and information content from information theory. Entropy is defined as below:

$$\mathrm{H(T)}= I_E(P_1, P_2 \ldots\ldots, P_j)=\sum_{i=1}^{j} P_i log_2 P_i$$

Where are fractions that add up to 1 and represent the percentage of each class present in the child node that results from a split in the tree.

$$\mathrm{IG(T,a)} = \mathrm{H(T)} - \mathrm{H(T|a)}$$

$$-\sum_{i=1}^{j} P_i log_2 P_i - \sum_a p(a) \sum_{i=1}^{J} -P_r(i|\mathrm{a})\, log_2 P_r(i|\mathrm{a})$$

Variance reduction

Introduced in CART variance reduction is often employed in cases where the target variable is continuous (regression tree), meaning that the use of many other metrics would first require discretization before being applied. The variance reduction of a node

N is defined as the total reduction of the variance of the target variable *x* due to the split at this node:

$$I_v(N) = \frac{1}{|s|^2}\sum_{i\in S}\sum_{j\in S}\frac{1}{2}(x_i - x_j)^2 = \frac{1}{|S_t|^2}\sum_{i\in S}\sum_{j\in S}\frac{1}{2}(x_i - x_j)^2 \left| \frac{1}{|S_f|^2}\sum_{i\in S}\sum_{j\in S}\frac{1}{2}(x_i - x_j)^2 \right.$$

V. Performance Evaluation

Once a classification model is obtained by using the aforementioned process[33], it is important to assess the classifier's performance. In this process, 70% of randomly selected samples are utilized as the training set and the remaining 30% is utilized as the testing set. . Here performance is measured with respect to accuracy (Acc), precision (Pre), recall (Rec), F-score (F1). Acc is a measurement corresponding to the total number of correct predictions. Pre means the ratio of predicted positives that are true positives, Rec denotes the ratio of true positives that can be correctly predicted, and F1 is the harmonic mean of Pre and Rec and can reflect the overall performance. These definitions of those statistical terms, Pre, Rec, F1, and Acc are shown in the following equation:

$$P_{re} = \frac{TP}{TP+FP}\%$$

$$R_{ec} = \frac{TP}{TP+FN}\%$$

$$A_{cc} = \frac{TP+TN}{TP+TN+FP+FN}\%$$

$$F_1 = \frac{2*P_{re}*R_{ec}}{P_{re}+R_{ec}}\%$$

$$\text{Specificity} = \frac{TN}{FP+TN}\%$$

Where,

TP: true positive, the classification result is positive in presence of malignancy.
TN: true negative, the classification result is negative in being benign.
FP: false positive, the classification result is positive in being benign.
FN: false negative, the classification result is negative in presence of malignancy.
According to the above definitions the equations related to specificity (accuracy of negative class), sensitivity (accuracy of positive class)

Accuracy: Accuracy is a scoring measurement of a machine learning model. It is the proportion of correctly defined patients.

Sensitivity: Sensitivity is the proportion of actually the presence of malignancy of histological image.

Specificity: Specificity is the proportion of those who are free from malignant in the histological image.

Table 3

Actual	Predicted	
	Negative	Positive
Negative	True Negative (TN)	False Positive (FP)
Positive	False Negative (FN)	True Positive (TP)

VI. Result

The performance of the different classifiers is measured from the confusion matrix which is obtained after testing the model with BreakHIS [IX, XII] dataset using SVM, Random forest, Decision tree, and KNN classifiers. Table 2 shows the accuracy and execution time of the four different classifiers we used. Table 3 shows various performance parameters of four different classifiers.

The confusion matrix of SVM, Random Forest, Decision Tree, and KNN is given below:

Table 4: **Accuracy and execution time**

Confusion Matrix(SVM)	
101	23
25	149

Confusion Matrix(RF)	
88	11
13	186

Confusion Matrix(DT)	
102	33
33	130

Confusion Matrix(KNN)	
103	22
23	150

Model	Classifier	Accuracy (%)	Execution time (sec)
Resnet_50	SVM	84	167
	Random forest	92	41
	Decision tree	78	82
	K nearest neighbor	85	108

Table 5: **Performance evaluation of different classifiers**

Classifier	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)
SVM	84	86.9	85.4	81.45
Random forest	92	94.2	93.3	87.13
Decision tree	78	80	79.6	75.56
KNN	85	87.3	86.6	80

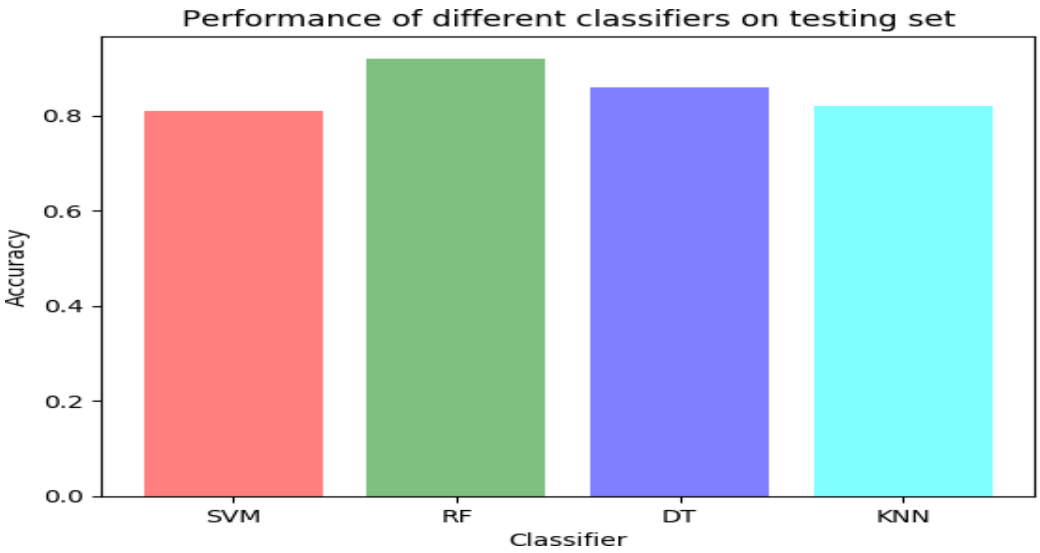

Figure 10: Test accuracy of different classifiers

VII. Discussion

To evaluate the effectiveness of classifiers, we made experiments on the BreakHIS dataset mentioned above. We compare our results with different classifiers on the BreakHIS data set. Table 2 gives the classification accuracies of our used classifiers for breast cancer classification. As we can see from the results, we used the ResNet-50 model to extract features from the dataset. The performance of the classifier can be increased significantly by applying fine-tuning for each layer. In this study, there were two diagnosis classes: Benign and Malignant which are classified using the histopathology images. Classification results of the system were displayed by using a confusion matrix. From the above results, we conclude that the classifiers show obtain very promising results in classifying breast cancer. We believe that the proposed system can be very helpful to physicians in their final decision on their patients.

VIII. Conclusion

This work determines the possibility of knowledge transfer from natural to histopathological images, by employing a pre-trained network ResNet-50. For transfer learning, features are extracted from the pre-trained model. We have switched from softmax to other classifiers namely support vector machine, random forest, decision tree, and k-nearest neighbor for the classification of breast cancer. Our main goal was to compare the performance of the classifiers on such a large-scale dataset. Due to the complex nature of the histopathological image dataset, the performance of the classifiers is significantly reduced. In conclusion, we have shown that the random forest classifier outperforms the SVM, decision tree, and KNN classifier on the BreakHIS dataset. Random forest gains an accuracy of 92% where SVM has 84%, KNN has 85% and decision tree has 78% accuracy.

Conflict of Interest:

There was no relevant conflict of interest regarding this paper.

References

I. Altman, N. S. (1992). "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression" (Pdf). The American Statistician.

II. Araújo, Teresa, Guilherme Aresta, Eduardo Castro, José Rouco, Paulo Aguiar, Catarina Eloy, António Polónia, and Aurélio Campilho. "Classification of breast cancer histology images using Convolutional Neural Networks." PloS one 12, no. 6 (2017): e0177544.

III. "Automatic white blood cell classification using pre-trained deep learning models: ResNet and Inception," Proc. SPIE 10696, Tenth International Conference on Machine Vision (ICMV 2017), 1069612 (13 April 2018);

IV. Balestriero, R. Neural Decision Trees. Arxiv E-Prints, 2017.

V. Bayramoglu, Neslihan, Juho Kannala, and Janne Heikkilä. "Deep learning for magnification independent breast cancer histopathology image classification." In Pattern Recognition (ICPR), 2016 23rd International Conference on, pp. 2440- 2445. IEEE, 2016.

VI. B. E. Bejnordi, G. Zuidhof, M. Balkenhol et al., "Contextaware stacked convolutional neural networks for classifcation of breast carcinomas in whole-slide histopathology images," Journal of Medical Imaging, vol. 4, no. 04, p. 1, 2017.

VII. Diaz-Uriarte R, Alvarez De Andres S: Gene Selection And Classification Of Microarray Data Using Random Forest. Bmc Bioinformatics 2006, 7:3.

VIII. George, Yasmeen Mourice, Hala Helmy Zayed, Mohamed Ismail Roushdy, and Bassant Mohamed Elbagoury. "Remote computer-aided breast cancer detection and diagnosis system based on cytological images." IEEE Systems Journal 8, no. 3 (2014): 949-964.

IX. Gupta, Vibha, and Arnav Bhavsar. "Breast Cancer Histopathological Image Classification: Is Magnification Important?." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 17-24. 2017.

X. Hall P, Park Bu, Samworth Rj (2008). "Choice Of Neighbor Order In Nearest-Neighbor Classification". Annals Of Statistics. 36 (5): 2135–2152.

XI. Hammer B, Gersmann K: A Note On The Universal Approximation Capability Of Support Vector Machines. Neural Processing Letters 2003, 17:43-53.

XII. Han, Zhongyi, Benzheng Wei, Yuanjie Zheng, Yilong Yin, Kejian Li, and Shuo Li. "Breast cancer multi-classification from histopathological images with structured deep learning model." Scientific reports 7, no. 1 (2017): 4172

XIII. Kahya, Mohammed Abdulrazaq, Waleed Al-Hayani, and Zakariya Yahya Algamal. "Classification of breast cancer histopathology images based on adaptive sparse support vector machine." Journal of Applied Mathematics and Bioinformatics 7.1 (2017): 49

XIV. Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun**,"**Deep Residual Learning for Image Recognition**,"** arXiv:1512.03385 [cs.CV]**,** 10 Dec 2015**;**

XV. Kowal, Marek, Paweł Filipczuk, Andrzej Obuchowicz, Józef Korbicz, and Roman Monczak. "Computer-aided diagnosis of breast cancer based on fine needle biopsy microscopic images." computers in biology and medicine 43, no. 10 (2013): 1563-1572.

XVI. Mehdi Habibzadeh, Mahboobeh Jannesari, Zahra Rezaei, Hossein Baharvand, Mehdi Totonchi,

XVII. Pan, S.J. And Yang, Q., 2010. A Survey On Transfer Learning. Ieee Transactions On Knowledge And Data Engineering, 22(10), Pp.1345–1359.

XVIII. Prinzie, A., Van Den Poel, D. (2008). "Random Forests For Multiclass Classification: Random Multinomial Logit". Expert Systems With Applications

XIX. Qicheng Lao, Thomas Fevens,"Cell Phenotype Classification using Deep Residual Network and its Variants," International Journal of Pattern Recognition and Artificial Intelligence © World Scientific Publishing Company, 01/17/19;

XX. Rawat, W. And Wang, Z., 2017. Deep Convolutional Neural Networks For Image Classification: A Comprehensive Review. Neural Computation, 29(9), Pp.2352–2449.

XXI. Rokach, Lior; Maimon, O. (2008). Data Mining With Decision Trees: Theory And Applications. World Scientific Pub Co Inc.

XXII. Shalev-Shwartz, Shai; Ben-David, Shai (2014). "18. Decision Trees". Understanding Machine Learning. Cambridge University Press.

XXIII. Scornet, Erwan (2015). "Random Forests And Kernel Methods".

XXIV. Simonyan, K. And Zisserman, A., 2014. Very Deep Convolutional Networks For Large-Scale Image Recognition. Arxiv Preprint Arxiv:1409.1556.

XXV. Spanhol, Fabio A., Luiz S. Oliveira, Caroline Petitjean, and Laurent Heutte. "A dataset for breast cancer histopathological image classification." IEEE Transactions on Biomedical Engineering 63, no. 7 (2016): 1455-1462.

XXVI. Spanhol, Fabio Alexandre, Luiz S. Oliveira, Caroline Petitjean, and Laurent Heutte. "Breast cancer histopathological image classification using convolutional neural networks." In Neural Networks (IJCNN), 2016 International Joint Conference on, pp. 2560-2567. IEEE, 2016.

XXVII. Spanhol, Fabio A., Luiz S. Oliveira, Paulo R. Cavalin, Caroline Petitjean, and Laurent Heutte. "Deep features for breast cancer histopathological image classification." In Systems, Man, and Cybernetics (SMC), 2017 IEEE International Conference on, pp. 1868-1873. IEEE, 2017.

XXVIII. Tang, Y., 2013. Deep Learning Using Linear Support Vector Machines. Arxiv Preprint Arxiv: 1306.0239.

XXIX. T. Araujo, G. Aresta, E. Castro et al., "Classifcation of breast cancer histology images using convolutional neural networks," PLoS ONE, vol. 12, no. 6, Article ID e0177544, 2017**.**

XXX. Veta, Mitko, Josien PW Pluim, Paul J. Van Diest, and Max A. Viergever. "Breast cancer histopathology image analysis: A review." IEEE Transactions on Biomedical Engineering 61, no. 5 (2014): 1400-1411

XXXI. Yosinski, J., Clune, J., Bengio, Y. And Lipson, H., 2014. How Transferable Are Features In Deep Neural Networks?. In Advances In Neural Information Processing Systems (Pp. 3320–3328).

XXXII. Zeiler, M.D. And Fergus, R., 2014, September. Visualizing And Understanding Convolutional Networks. In European Conference On Computer Vision (Pp. 818–833). Springer, Cham.

XXXIII. Zhang, Yungang, Bailing Zhang, Frans Coenen, and Wenjin Lu. "Breast cancer diagnosis from biopsy images with highly reliable random subspace classifier ensembles." Machine vision and applications 24, no. 7 (2013): 1405-1420.