Journal of Physics: Conference Series

Breast cancer classification using digital biopsy histopathology images through transfer learning

Breast cancer classification using digital biopsy histopathology images through transfer learning

Ghulam Murtaza[1, 2*], Liyana Shuib[1*], Ainuddin Wahid Abdul Wahab[1], Ghulam Mujtaba[2], Ghulam Mujtaba[3], Ghulam Raza[4] , and Nor Aniza Azmi[5]

[1]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]Department of Computer Science, PAF-KIET Karachi, Pakistan.
[4]St. James's Hospital, James's Street, Ushers, Dublin 8, Ireland.
[5]Diagnostic Imaging & Radiotherapy Programme, Faculty of Health Sciences, Universiti Kebangsaan Malaysia

*gmurtaza@iba-suk.edu.pk, liyanashuib@um.edu.my

Abstract
Breast cancer (BC) infection, which is peculiar to women, brings about the high rate of deaths among women in every part of the world. The early investigation of BC has minimized the severe effects of cancer as compared to the last stage diagnosis. Doctors for diagnostic tests usually suggest the medical imaging modalities like mammograms or biopsy histopathology (Hp) images. However, Hp image analysis gives doctors more confidence to diagnose BC as compared to mammograms. Many studies used Hp images to develop BC classification models to assist doctors in early BC diagnosis. However, these models lack better and reliable results in terms of reporting multiple performance evaluation metrics. Therefore, the goal of this study is to create a reliable, more accurate model that consumes minimum resources by using transfer learning based convolution neural network model. The proposed model uses the trained model after fine tuning, hence requires less number of images and can show better results on minimum resources. BreakHis dataset, which is available publicly has been employed in overall experiments in this research. BreakHis dataset is separated into training, testing, and validation for the experimentation. In addition, the dataset for training was augmented followed by stain normalization. By using the concept of transfer learning (TL), AleNext was retained after fine-tuning the last layer for binary classification like benign and malignant. Afterward, preprocessed images are fed into the TL based model for training. The model training was performed many times by changing the hyper-parameters randomly until the minimum validation loss was achieved. Now the trained model was used for feature extraction. The extracted features were further evaluated by using six ML classifiers (i.e. softmax, Decision tree, Naïve Bayes, Linear discriminant analysis, Support vector machine, k-nearest neighbor) through five performance measures such as precision, F-measure, accuracy, specificity, and sensitivity for experimental evaluation. The softmax has outperformed among all classifiers. Furthermore, to reduce the wrong prediction, a misclassification reducing (MR)

algorithm was developed. After using the MR algorithm the proposed model produced better and reliable results. The observed accuracy, specificity, sensitivity, precision and F measure are 81.25%, 77.47%, 82.49%, 91.70%, and 86.80% respectively. These results show that the proposed TL based model along with misclassification reduction algorithm produced comparable results to the current baseline models. Hence, the expected model could serve as a second opinion for BC classification in any healthcare center.

1. Introduction

BC, the greatest disastrous kind of disease that affects women globally. In the United States of America, the 2016 statistical report on BC shows that about 0.04 million women had experienced death and about 0.25 million new cases have been diagnosed as BC patient (Miller et al., 2016). Moreover, one out of eight women will be affected during her life span in developed countries. Thus, it can increase drastically the burden of health in developed countries. Medical images like mammograms and biopsy HP images are used as a diagnostic test for BC. Mammograms can help in finding the location of cancerous regions but does not ensure the presence of cancer. However, HP images enable doctors to diagnose more confidently the presence of cancerous lesions. In a breast biopsy, the sample of breast cancer part was taken and preserved into microscopic slides for manual analysis. The microscopic analysis of BC tissue slides is performed by any expert pathologist. The final decision has been made after the consensus of more than two pathologists for better diagnosis. However, it may lead to increase diagnosis time and there may be a conflict of opinion among two pathologists' due to their experience, expertise and domain knowledge (Allison et al., 2014; Elmore, Longton, Carney, & et al., 2015). Therefore, the building of an automated method known as a computer-aided diagnostic system is required for Hp image classification. The CAD system is generally based on ML classification approaches. The ML-based BC diagnostic system will assist doctors as a second opinion for a better and early diagnosis of BC.

The ML-based BC classification can be categorized into three main approaches like traditional ML, deep neural network (DNN) and TL based approaches. Feature extraction is one of the major steps in any ML classification techniques in which the most useful features that will contribute to the classification accuracy are being extracted from the images. In traditional ML-based approaches, the handcrafted features descriptors like LBP (Ojala, Pietikainen, & Maenpaa, 2002), SIFT (Lowe, 1999), SURF (Bay, Tuytelaars, & Van Gool, 2006) were used to extract local and global features. To extract result oriented features is cumbersome tasks and needs a different preprocessing step for datasets collected from different sites (Rouhi, Jafari, Kasaei, & Keshavarzian, 2015). However, DNN based approaches, especially the convolution neural network (CNN) based model has resolved the issue of handcrafted feature extraction. CNN model extracts features during its training process automatically and proved better results for image data as compared to traditional ML-based approaches (Ciresan, Meier, Masci, Maria Gambardella, & Schmidhuber, 2011; Krizhevsky, Sutskever, & Hinton, 2012). However, when a CNN model is trained from scratch, it required a large number of annotated images and consumes very high resources like GPU (Litjens et al., 2017). However, the scenario is not possible for medical images because it is a trivial task in a real-life scenario to collect a huge number of images of all types of cancers with proper labels. Therefore, the researchers adopted a TL based approach in which, pretrained model like AlexNet can be used after fine tuning to classify the specific data. The TL based model required less number of image and the model can be retained on limited sources like a normal desktop computer. Thus, TL based models are suitable for the BC medical image classification.

In a study conducted by (Abdullah-Al, Bin Ali, & Kong, 2017), the author proposed the Retinex algorithm to normalize the contrast and illumination inconsistencies of Hp images. Afterward, three types of CNN models were trained from scratch. Firstly, conventional CNN, secondly CNN devised with residual Blocks and thirdly by using MaxMin conventional CNN. Thus, the best results were reported on the first type of conventional model using BreakHis images for benign and malignant classes. The reported accuracy (Acc), specificity (Sp) and sensitivity (Sn) are 85.36%, 70.36%, and 91.36% respectively. However, the model showed better accuracy, but it is biased towards the majority class. This needs to be improved to make the results more reliable. In the research conducted by (Araujo

et al., 2017), the image was divided into many patches with 50% overlapping method. The model was trained from scratch to extract the deep convolution activation features (DeCAFs). Furthermore, SVM and softmax classifiers have been used to classify DeCAFs and the experimental analysis shows that SVM outperformed the softmax by showing 83.3% accuracy for BC classification on carcinoma and non-carcinoma using publicly available dataset like Bioimaging Challenge 2015 Breast Histology. However, the author reported better results, but only Acc and Sn have been reported. Moreover, Sn is very high than Acc, thus Sp can be assumed to be very low. Thus results need to be improved in terms of specificity. The author (Nejad, Affendey, Latip, & Ishak, 2017) proposed a model that consists of a completely connected convolution layer together with the softmax layer to perform BC binary classification. Images of BreakHis dataset were divided into two, (70% to train the model) and (30% to test the model). Training images are resized, normalized (subtracting mean from all images) and augmented before feeding to CNN model. Training was performed with and without augmented images. The model trained with augmented images has outperformed by showing Acc of 77.5% for binary classification of BC. However, the only accuracy has been reported and needs to be improved to enhance the performance of the classifier. In a study (Wan, Cao, Chen, & Qin, 2017) the nuclei segmentation was performed by enhanced hybrid active contour model by using exclusive dataset images for BC grading i.e. low, medium or high. CNN was used to extract an integrated set of features from the image pattern and item level features (construction). SVM is used in a cascade manner to combine two types of features to maximize the classification performance. The reported average accuracy for the two classes is 69%. However, only the accuracy has been reported and needs improvement for better reliable grading of BC. The aforementioned studies only showed accuracy while other performance metrics need to be measured to show the classification model reliability. Moreover, the reported poor performance in terms of accuracy and needed to be enhanced. Thus the aim of this study is to produce better and reliable results by using five performance measures like Acc, Sn, Sp, precision (Pr) and F-measure.

2. Proposed Method

This segment describes the overall research methodology (Figure 1) established to develop a BC classifier using histopathology images. The given result of the research methodology is based on five main steps like data collection, image preprocessing, transfer learning-based model development and feature extraction, breast cancer classification technique and performance evaluation metrics. In the data collection step, publicly available Break His dataset was downloaded and stored in local storage. Afterward, duplicate patient images were deleted. Thus, 81 patient images were left behind out of 82 patients. Noticeable in this study 40x magnification images were exploited,(see Table 1) because, it has shown better accuracy in baseline studies. Moreover, images of 81 patients with 40x magnifications were split into training (50%), validation (20%) and testing (30%) set through random sampling approach. In the second step, images were normalized by using Reinhard's (Reinhard, Adhikhmin, Gooch, & Shirley, 2001) method to remove the inherent Hp image inconsistencies. Reinhard's method preserves the structure of breast cancer lesions while image normalization process, thus it can show better performance to any other type of stain normalization method like Macenko (Macenko et al., 2009), Khan's (Khan, Rajpoot, Treanor, & Magee, 2014) and RGB histogram normalization methods. Afterward, only training set images were augmented by using 25 basic image processing techniques and a comprehensive number of augmented training images were created. The equal number of augmented images per class was exploited along with all original images of the training set for model training to mitigate the overfitting issue. In the preprocessing, all images were rescaled to fit in the input layer need while in step three, AlexNet has been fine-tuned by changing the last layer for binary classification of BC. However, the model training was performed many times by changing the hyper-parameters randomly until minimum validation loss has been achieved. Moreover, a separate model was stored on each epoch of the training process. Afterward, the best epoch model was chosen on the basis of testing set performance evaluation results. Finally, in step three, features to train the model and the features to test the model were extracted and utilized for further analysis. In step four, six ML classifiers such as softmax, Decision tree, Naïve Bayes, Linear discriminant analysis, Support vector machine, and a k-nearest neighbor were used to classify the extracted features of BC HP images. The best performing classifier was selected through five performance evaluation metrics like Acc, Sn, Sp, Pr, and F-measure.

Afterward, 3 out of 25, best performing augmentation methods were selected to reduce the misclassification rate. In step four, the misclassification reduction (MR) algorithm was developed and exploited to reduce the incorrect predictions made by the selected ML classifier. Lastly, in step five, the comparisons of the existing models and the results of ML classifier were carried out by using the aforementioned five performance evaluation metrics.

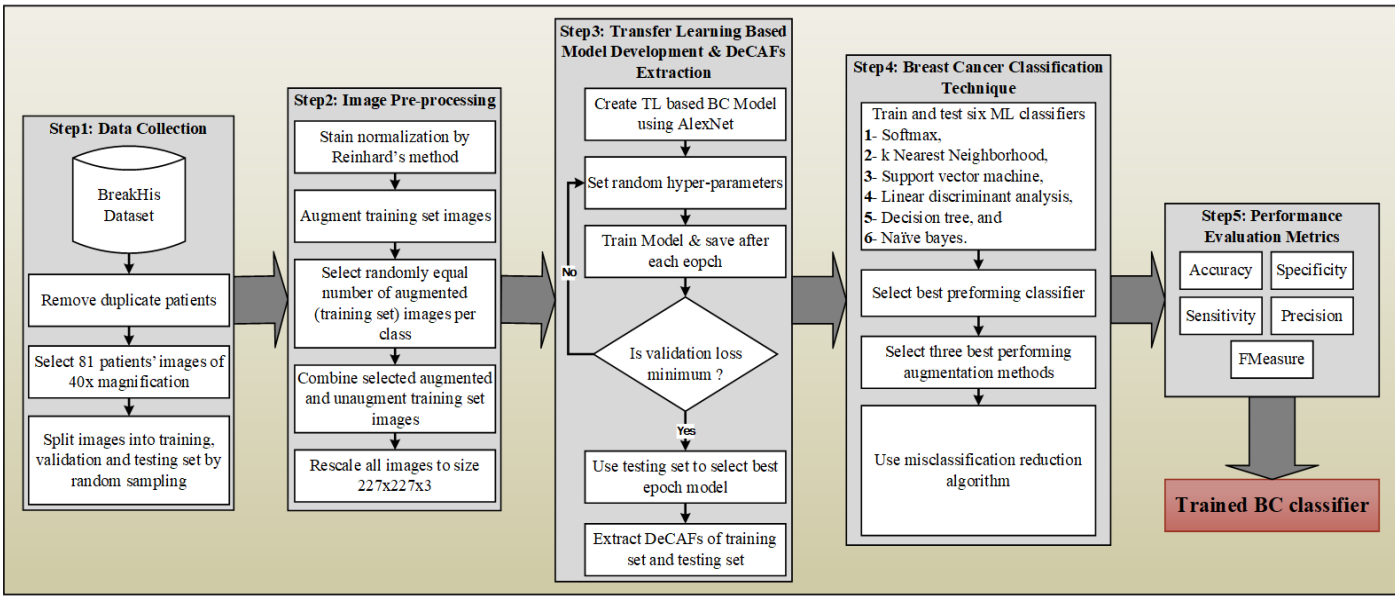

Figure 1. Illustrates research methodology

Table 1. BreakHis dataset images distribution

Magnifications / BC Types	40×	100×	200×	400×	Total images	Total patients
Benign	625	644	623	588	2480	**24**
Malignant	1370	1437	1390	1232	5429	**58**
Total images	**1995**	**2081**	**2013**	**1820**	**7909**	**82**

2.1 Development transfer learning based BC model

The TL is actually a process of reusing and pre-training model using data on a specific domain. For instance, the natural image was trained using AlexNet which uses neural network based classification similar to the study by (Krizhevsky et al., 2012) and can be retrained on medical images. AlexNet is a previously trained deep convolution neural network based classification model. It was trained on millions of natural images for months using GPU for thousand classes/objects. However, after fine-tuning AlexNet can be used to classify breast cancer images. The fine tuning is a process in which the last layer of AlexNet is replaced with a new layer that possesses a target number of classes. In this study, the last layer was fine-tuned for benign and malignant classes before performing and retraining of TL-based BC model. The basic structure of AlexNet consist of three types of layers such as the input, the output and the hidden layer as shown in Figure 2. The images (of size 227x227x3) can be made to be accepted directly by the input layer. The fully connected and the convolutional layer is the two categories of the hidden layers. The five convolution layer of AlexNet is developed to extract low level and middle level features present in the input vision such as color, edges, and blobs whereas the three fully connected layers are designed to learn high-level features (semantic feature) at the pixel level of the image. The normalization layer and the Rectified Linear Unit (ReLU) are used to equip each convolution layer of AlexNet whereas an additional layer like MaxPool layer is a component of the convolution layer 1, 2 and 5. The ReLU is an activation function that maps all negative value to positive on the results of

each convolution. The normalization layer helps to normalize each input channel throughout a mini-batch. It speeds up the training process of CNN's and minimizes the dependency of network initialization. The MaxPool layer reduces the spatial size of receptive fields, thus it reduces the number of parameters and computations of CNN. In addition, the first two fully connected layers contain ReLU and dropout layers. The dropout layer is required to reduce the overfitting issue by selecting the neurons to ignore their processing to speed up the overall training process. The last fully connected layer contains the softmax classifier to predict the probabilities for intended class labels. Softmax is a non-linear variant of multinomial logistic regression whereas the output layer holds the results of classification.

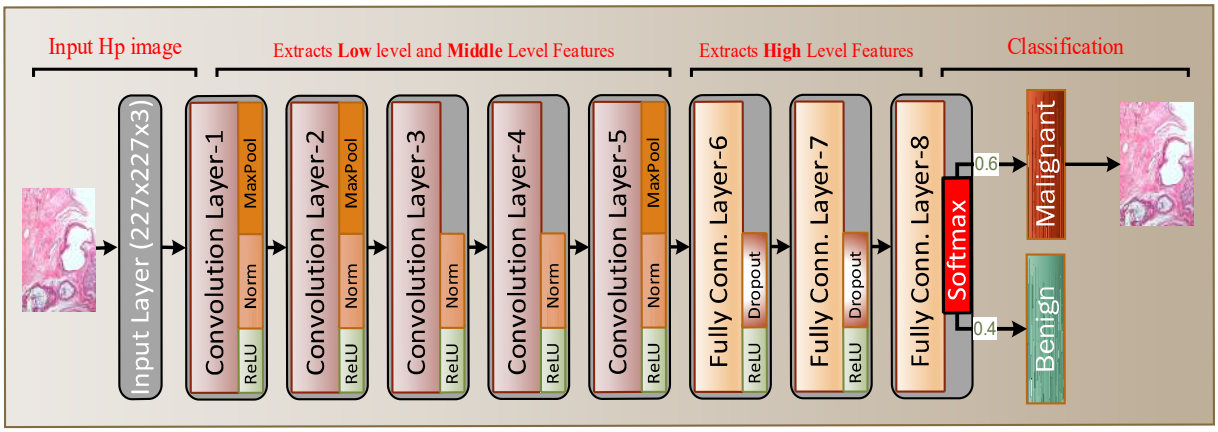

Figure 2. Proposed transfer learning based (fine-tuned AlexNet) model for BC Classification.

2.2 Training and feature extraction

The TL-based model was exploited to extract the distinct and result oriented features using BC Hp images. Numerous experiments were carried out to get the optimum results by tuning a few hyper-parameters for model training. However, it has been kept in considerations while setting up the hyper-parameters, that the model must be trained on a normal desktop computer. The proposed model was trained through gradient descent solver at momentum 0.9. The rest of modified parameters are like a number of epochs to be trained up to 10, 64 was the mini batch size, while learning rate start from 0.001, whereas the validation patients were set to 5, and drop of learning rate by 0.1 after two epochs. Finally, the model was compelled to stop if the validation loss does not reduce up to five successive iterations while training process. Thus, the proposed model was trained in such a manner, to get the best-generalized features for BC classification.

2.3 Misclassification Reduction

A higher misclassification that observed test images are supplied to TL-based proposed model. It is because the classifier yielded many false-positives and false-negatives while prediction. Thus, it is required to reduce the misclassification rate by using some technique to produce the most reliable results. Therefore, an algorithm was developed to mitigate the false prediction for BC classification using Hp images.

MR Algorithm: Before using an algorithm out of 25 augmentation methods, three top-performing augmentation methods were selected for trial and error basis. Because more than augmentation methods were unable to get the best results. Each input image is augmented three times by using selected augmentation methods before feeding into the classifier. Now, the label of the actual image was decided on the basis of labels of three augmented image labels. If the majority number of the augmented images were classified as benign, then the original image will be considered as benign otherwise malignant, see algorithm 1.

Algorithm 1. Misclassification Reduction (MR)

```
Input: TrainedClassifier, TestingImages
Output: OrignalImageLabel
Function MR
 1  NoImages  = count(TestingImages)
 2        For NoI = 1:NoImages
 3              Img = ReadImage(TestingImages(NoI))
 4              AugImages =
    AugmentImage3Times(Img)
 5              PredictLbls =
    Predict(Classifier,AugImages)
 6              CntB=CountBenign(PredictLbls)
 7              CntM=CountMalignant(PredictedLbls)
 8              If CntB > CntM
 9                    OrignalImageLabel = 'Benign'
10              Else
11                    OrignalImageLabel = 'Malignant'
12              End
13        End
14  Return OrignalImageLabel
End
```

2.4 Setup of experiments

In this study, the evaluation of the performance measures was carried out using three experimental setups with the overall 17 (9+6+2) analyses of the proposed model.

1. In the setting I, a total number of 9 analysis were made, see Figure 3. The performance of the proposed model was evaluated by using validation loss and validation data accuracy. Moreover, the validation loss (see Figure 3 (a)) and validation data accuracy (see Figure 3 (b)) was computed on each iteration of four epochs in the course training process. To acquire the best model, a trained model is created at the end of each epoch of the training process. Apart from this, testing data was also used to select the best model out of four models, which were created on each epoch while training, see Figure 3 (c). In such a way, these 9 different analyses enabled us to select the best epoch model for further analysis.
2. In setting II, 6 different more analysis was carried out, see Figure 4. The performance of extracted features of the proposed model was examined by using six ML classifiers namely softmax, KNN, LDA, DT, SVM, and NB. Here, six analysis was made through five performance metrics like Ac, Sn, Sp, Pr, and F-measure. These metrics assisted in selecting the top performing classifier among all classifiers.
3. In setting III, last 2 different analysis was performed as shown in Figure 5. Here, the performance of the best classifier selected in setting II will be examined with MR algorithm. The results of softmax with and without using MR algorithm were compared.

2.4.1 Performance Metrics

Here, the performance of the overall classifiers has been evaluated by using five metrics namely, Acc, Sp, Sn, Pr and F measure. These five metrics are derived from the confusion matrix of each classifier by using actual and predicted labels. The reason for using five metrics is to ensure that the classifier is not biased and can produce reliable results. However, in medical science, Sn is more important as compared to any other performance metrics because Sn determines the correctly classified images of a cancer patient. In other words, misdiagnosis of a malignant BC will be more hazardous as compared to a benign misdiagnosis. Moreover, the harmonic mean of Sp and Sn is referred to as F-measure and it showed a value that represents a biased or unbiased classifier.

In this study, MATLAB R2017b version was utilized to perform all tasks like image preprocessing, TL-based model creation, training and testing, and development of MR algorithm. Mostly default parameters are used in all experiments except the few options which are specifically reported in this study.

3. Experimental Results

This section reports and analyzes the results of the previously mentioned experimental settings. The results were measured in terms of Ac, Sp, Sn, Pr and F measure.

3.1 Experimental results of setting I

This section elaborates the results of 09 analyses in order to select the best model from four models stored at each epoch while training. The validation loss and accuracy of validation data was computed and analyzed at the end of each epoch, throughout the overall training process of AlexNet model. Nonetheless, the training process is terminated when validation loss is not reduced for five consecutive epochs using training data.

As it can be seen in Figure **3** (a), the validation loss of the AlexNet was initially very low (i.e., 0.8614, 0.8939, 0.887) and suddenly fluctuated from 1.1138 to 0.8391. Afterwards it has been stabilized remain almost constant (i.e., 0.9419, 0.9691, 0.9771, 0.9406, and 0.9489). Thus, the validation loss results showed that the AlexNet model performed best (i.e., 0.8391) at epoch two in the overall training process. Similarly, it can be observed from Figure 3 (b), that the validation set accuracy was very low (i.e., %%, while and suddenly increased up to 54.64% while the trend shows that the validation accuracy will remain increasing gradually throughout the training process. However, the highest validation accuracy was observed during epoch two training process i.e. 55.7%). On the other hand, the performance of AlexNet using a testing set is shown in Figure 3 (c). It has been observed that the model initially performed very poor (i.e., 66.32%) on epoch one trained model while accuracy has been drastically increased (up to 72.92%) when epoch two trained models were used on the testing set. However, it has been observed from the overall analysis that the epoch two models have shown reliable and best results among all epochs. Thus, the AlexNet trained model at epoch two was selected for DeCAFs extraction and further analysis.

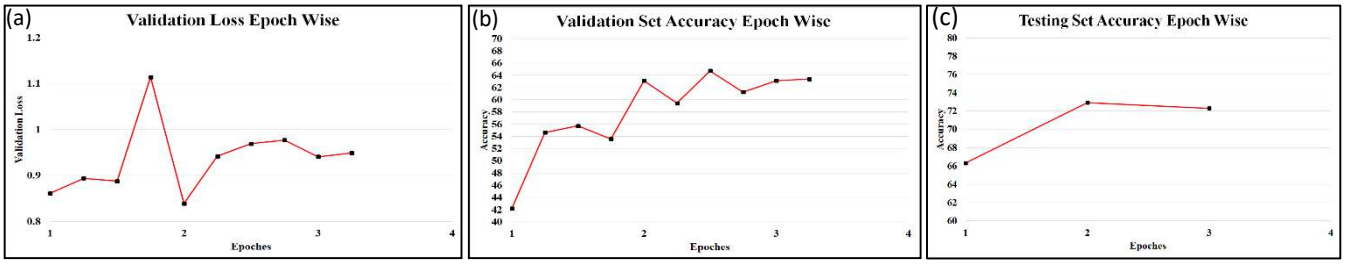

Figure 3. Epoch-wise training and testing performance of TL based model.

3.2 Experimental results of setting II

The DeCAFs were extracted using the best AlexNet model selected through experimental results of setting I. Here, the performance of six ML classifiers like softmax, kNN (k = 1, 5, 9), LDA, DT, SVM, NB was evaluated. Moreover, the performance is analyzed by reporting results of five measures like Ac, Sp, Sn, Pr and F measure which were computed from the confusion matrix of each classifier. The adopted performance measures enabled us to select the best performing classifier among six ML classifiers to perform further analysis for BC classification. Figure 4 shows that the softmax performed better (with an accuracy of 72.92%) than the rest of the classifiers. Therefore, Sp (i.e., 81.69%), Sn (i.e., 70.05%), Pr (i.e., 92.12%) and F measure (i.e., 79.58%) of softmax is also higher than all other classifiers. In contrast, the lowest performance (i.e., Ac=38.02%, Sp=43.66%, Sn=66.25%, F measure=46.8%) was reported by NB classifier. Furthermore, the second highest performance is shown by kNN (k = 9) and LDA classifier. Both kNN and LDA classifiers had shown almost the same performance as kNN had achieved Ac 55.56%, Sp 42.25%, Sn 59.91%, Pr 76.02%, and F measure was measured 67.01%. Likewise, LDA got Ac 55.04%, Sp 46.48%, Sn 57.83%, Pr 76.76%, and F measure

was measured 65.97%. Thus, it is clearly observed that the performance of softmax is very high compared to the kNN and LDA classifiers. Moreover, the DT and SVM also showed very lower accuracy as compared to softmax i.e. 50.17% and 47.22%. Due to very low accuracy of both DT and SVM classifiers, it is obvious that the rest of the performance measures will be lower than softmax, like DT which showed Sp = 47.89%, Sn = 50.92%, Pr = 74.92%, and F measure was computed 60.63%. Similarly, SVM got Sp = 40.85%, Sn = 49.31%, Pr = 71.81% and F measure was reported as 58.47%. Thus, it can be concluded from the analysis and performance of six ML classifier that softmax had outperformed the rest of the classifiers. Therefore, softmax is selected for further analysis of BC classification using Hp images.

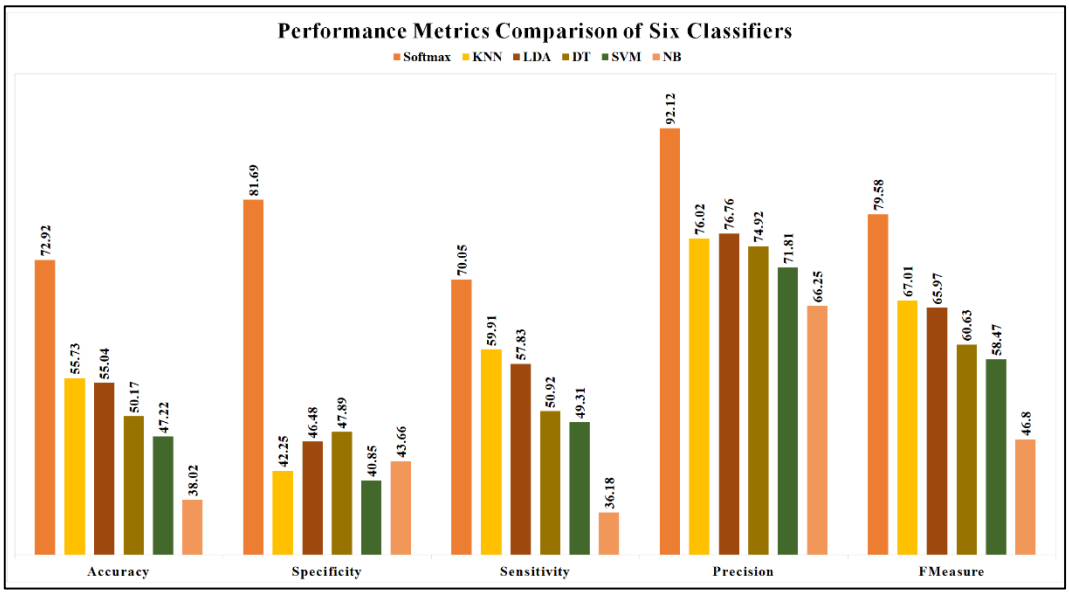

Figure 4. Illustrates the results of six ML classifiers, like softmax, kNN, LDA, DT, SVM, and NB

3.3 Experimental results of setting III

The comparison in performance of softmax before and after the application of MR algorithm is discussed in this section. However, it can be clearly observed from Figure 5 that the performance of softmax before using the MR algorithm was very low and biased toward majority class i.e. Malignant. The accuracy of softmax classifier was drastically improved from 72.92% to 81.25% when the wrong prediction is reduced. In contrast, Sp is reduced from 81.69% to 77.46% whereas Sn has been increased from 70.055% to 82.49%. However, it can be seen that Sn and Sp after using MR algorithm are more closes (i.e. Unbiased) to each other as compared to the Sn and Sp observed before applying the MR algorithm. Therefore, Pr has been slightly reduced from 92.12% to 91.97% due to the balance in Sn and Sp. However, F measure can show a clear picture of an unbiased/reliable classifier. Hence, the F measure is increased drastically from 79.58% to 86.89% when the MR algorithm was utilized.

For classification of Hp images of BC, the DeCAFs of training and testing data are extracted by using transfer learning-based model. Afterward, the MR algorithm was applied to reduce the incorrect prediction made by the proposed model.

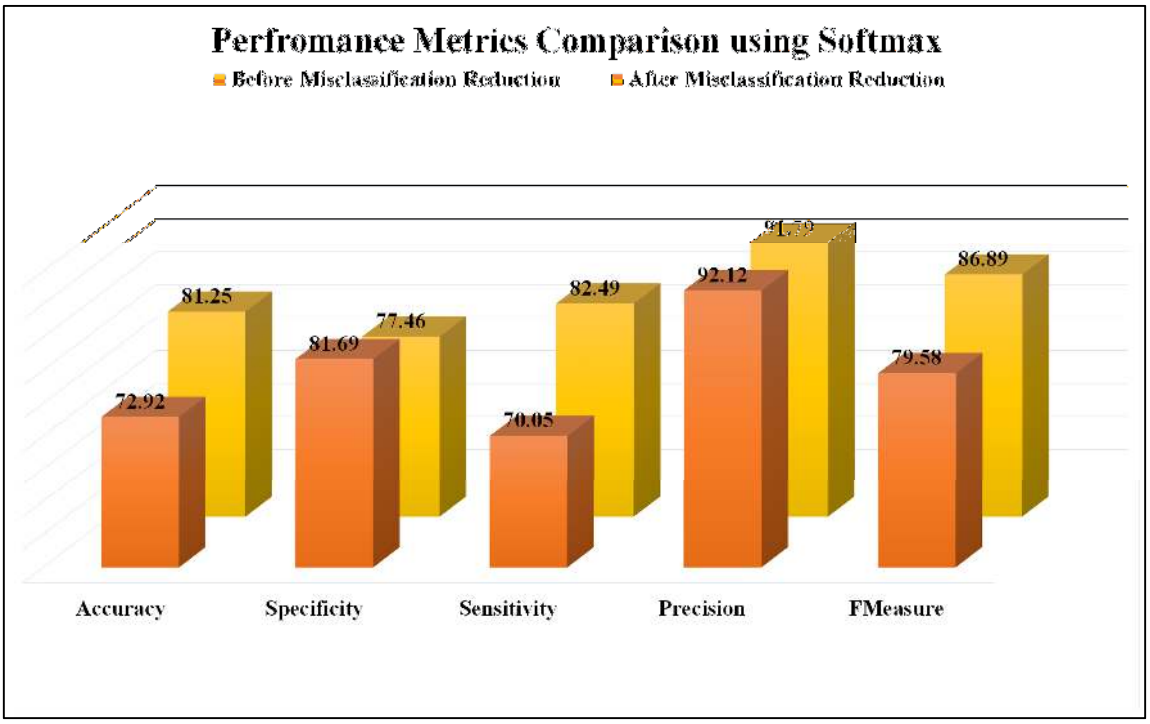

Figure 5. Demonstrates the performance comparison of softmax using misclassification reduction algorithm.

4. Discussion

This segment elaborated the theoretical analysis of the proposed TL-based model with misclassification algorithm by using BreakHis Hp images for binary classification. The proposed model produces a better result by using minimum resources. Due to transfer learning, the computational time of the model can be reduced drastically. Moreover, it requires less number of images to show better results. It also consumes less computational power, storage and can be trained in lesser time. Many studies reported the only accuracy, but other evaluation metrics are also important to show the unbiased results of the classification model. Furthermore, the reported accuracy is low, thus need to be improved for BC classification. The baseline studies results of the expected study can be seen clearly from Table-2. The recent studies showed comparable accuracy, but unable to report other performance measures like Sn, Sp, Pr and F measure. Thus higher accuracy is more prone to be biased toward majority class which can lead to a higher misclassification. Thus, our proposed model had shown comparable and reliable results in term of Acc, Sp, Sn, Pr and F measure as compared other baseline studies.

Table 2. Existing state-of-the-art baseline studies.

Study Reference	Results	Limitation
(Abdullah-Al et al., 2017)	Acc = 85.36%, Sp = 70.36, Sn = 91.36	• Better accuracy but it is highly biased towards majority class.
(Araujo et al., 2017)	Acc=83.3% Sn = 95.6%	• Better results but only Ac and Sn have been reported. • However, Sn is very high than Sp. Thus, Sp becomes very low.
(Nejad et al., 2017)	Acc= 77.5%.	• Only accuracy has been reported. Thus it can be biased. • Requires a better accuracy.
(Wan et al., 2017)	Average Acc=69%	• Only accuracy has been reported. Thus it can be biased. • Requires a better accuracy.

Proposed TL-based model with MR algorithm	Acc = 81.25% Sp = 77.46% Sn = 82.49% Pr = 92.70% F measure = 86. 89%	• Requires more images to learn better-generalized features because AlexNet requires a huge number of images for the training of the model. Thus, more images are required for learning of particular medical image features.

5. Conclusion

This research has carried out a study on the problem of BC disease that commonly affects women globally. However, the study has utilized benign or malignant as a BC classification to tackle the problem by using Hp images of BreakHis dataset. In this regard, a transfer learning model was created and trained on stain normalized and augmented training set. The proposed model was trained multiple times after changing a few training parameters in order to get the best model with the lowest possible validation loss. Subsequently, features were extracted by using the trained proposed model. The extracted feature was evaluated by using six machine learning classifiers to choose the best classifier for BC classification. However, it has been observed by comparing the results that softmax performed better than the rest of the classifiers. Furthermore, to improve the results of BC HP image classification, a misclassification reduction algorithm has been developed and implemented along with softmax. The good result observed based on the accuracy, the specificity, the sensitivity, precision and F measure was 81.25%, 77.46%, 82.49%, 91.79%, and 86.89% respectively. The developed TL-based model along with the MR algorithm has produced more reliable and comparable results with baseline models. Moreover, it consumed less training time and limited resources like a normal desktop computer. As a future work classification model, the study will be extended in developing a classification for BC by using multimodalities of medical images.

REFERENCES

1. Abdullah-Al, N., Bin Ali, F., & Kong, Y. N. (2017). *Histopathological Breast-Image Classification With Image Enhancement by Convolutional Neural Network.* Paper presented at the 2017 20th International Conference of Computer and Information Technology, New York.
2. Allison, K. H., Reisch, L. M., Carney, P. A., Weaver, D. L., Schnitt, S. J., O'Malley, F. P., . . . Elmore, J. G. (2014). Understanding diagnostic variability in breast pathology: lessons learned from an expert consensus review panel. *Histopathology, 65*(2), 240-251. doi:doi:10.1111/his.12387
3. Araujo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., . . . Campilho, A. (2017). Classification of breast cancer histology images using Convolutional Neural Networks. *Plos One, 12*(6), 14. doi:10.1371/journal.pone.0177544
4. Bay, H., Tuytelaars, T., & Van Gool, L. (2006). *Surf: Speeded up robust features.* Paper presented at the European conference on computer vision.
5. Ciresan, D. C., Meier, U., Masci, J., Maria Gambardella, L., & Schmidhuber, J. (2011). *Flexible, high performance convolutional neural networks for image classification.* Paper presented at the IJCAI Proceedings-International Joint Conference on Artificial Intelligence.
6. Elmore, J. G., Longton, G. M., Carney, P. A., & et al. (2015). Diagnostic concordance among pathologists interpreting breast biopsy specimens. *JAMA, 313*(11), 1122-1132. doi:10.1001/jama.2015.1405
7. Khan, A. M., Rajpoot, N., Treanor, D., & Magee, D. (2014). A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. *IEEE Transactions on Biomedical Engineering, 61*(6), 1729-1738.
8. 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.

9. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., . . . Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis, 42*, 60-88. doi:https://doi.org/10.1016/j.media.2017.07.005
10. Lowe, D. G. (1999). *Object recognition from local scale-invariant features.* Paper presented at the Computer vision, 1999. The proceedings of the seventh IEEE international conference on.
11. Macenko, M., Niethammer, M., Marron, J. S., Borland, D., Woosley, J. T., Guan, X., . . . Thomas, N. E. (2009). *A method for normalizing histology slides for quantitative analysis.* Paper presented at the Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on.
12. Miller, K. D., Siegel, R. L., Lin, C. C., Mariotto, A. B., Kramer, J. L., Rowland, J. H., . . . Jemal, A. (2016). Cancer treatment and survivorship statistics, 2016. *CA: a cancer journal for clinicians, 66*(4), 271-289.
13. Nejad, E. M., Affendey, L. S., Latip, R. B., & Ishak, I. B. (2017). *Classification of histopathology images of breast into benign and malignant using a single-layer convolutional neural network.* Paper presented at the ACM International Conference Proceeding Series.
14. Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence, 24*(7), 971-987.
15. Reinhard, E., Adhikhmin, M., Gooch, B., & Shirley, P. (2001). Color transfer between images. *IEEE Computer graphics and applications, 21*(5), 34-41.
16. Rouhi, R., Jafari, M., Kasaei, S., & Keshavarzian, P. (2015). Benign and malignant breast tumors classification based on region growing and CNN segmentation. *Expert Systems with Applications, 42*(3), 990-1002. doi:10.1016/j.eswa.2014.09.020
17. 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