Proceedings of the International Conference on Artificial Intelligence and Smart Systems (ICAIS-2021)
IEEE Xplore Part Number: CFP21OAB-ART; ISBN: 978-1-7281-9537-7

Transfer Learning Based Breast cancer Classification using One-Hot Encoding Technique

Karthiga R[1], Usha G[2],Raju N[3], and Narasimhan K[4]
[1, 3, 4]Department of ECE, SEEE, SASTRA Deemed University, Thirumalaisamudram, Thanjavur-613401, India
[2]Department of ECE, Srinivasa Ramanujan Centre, SASTRA Deemed University, Kumbakonam-612001, India
Corresponding author: Narasimhan K (knr@ece.sastra.edu)

2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) | 978-1-7281-9537-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICAIS50930.2021.9395930

Abstract— **Early diagnosis of breast cancer can be curable with precise techniques and improve patients prognosis with cancer. Most people failed to detect cancer early, leading to an increased death rate. In recent years, several studies have developed in various imaging modalities to predict breast cancer. The medical practitioner sometimes diagnosis diseases mistakenly due to misinterpretation. The automated assistance for practitioners using different computer-aided diagnosis (CAD) will give an accurate prediction for critical diseases. This paper presented a CAD system to perform automatic breast cancer diagnosis by employing the one-hot encoding technique. The system has experimented with BreakHis dataset partitioned into training and testing sets. The system performance can be measured with accuracy, sensitivity, specificity, precision, and recall. The system acquired 98.62% accuracy using the one-hot encoding technique, which is better than state-of-art methods. The proposed system outperformed the existing system by utilizing various performance metrics.**

Keywords—Breast cancer, One-hot encoding, Deep neural network, Transfer learning, VGG-16

I. INTRODUCTION

Breast cancer affects millions of women due to the increasing death rate, and it would vary in different countries. Breast cancer is the second deadly cancer after lung cancer among women in the age group of 20-59 years [1]. The annual survival rate of breast cancer is 1.7 million in the world, and it causes the highest death case to compare other types of cancer [2]. According to American cancer society data, around 2 68,600 new breast cancer cases were diagnosed in 2019 [3]. The early diagnosis is significant to increase the number of survivors. The primary stage diagnosis will increase the survival rate of up to 80%. The high morbidity and cost motivated researchers to investigate solutions to evolve more precise techniques with various breast cancer diagnosis modalities.

The modalities involved in breast diagnosis are mammography, Thermography, ultrasound, and tissue biopsy. Tissue biopsy produces the most promising results due to analyzing the suspected tissue samples. The pathologist performs Hematoxylin and Eosin (H&E) stains on tissue samples to establish the absolute diagnosis. Visual analysis is a complicated task, which makes the wrong prediction.

The process of multi-classification of histopathological images is achieved by the pathologists using the Computer-Aided Diagnosis (CAD). There are four important reasons why the task proves to be a difficult challenge [21, 22].

- Pathologists that have great experience and also have a reputational background are very rare. Most healthcare medical centers lack such experienced pathologists and have to adjust with low-skilled ones, which may increase human error during analysis.
- The whole process of diagnosis is often time-taking and is overpriced for some people.
- Since it takes a lot of time to confirm cancer cell existence, it may lead to other challenges such as pathologists weariness resulting in misdiagnosis.
- The various magnifications of each subclass may lead to high false positives due to their similarity in the structure shown in figure 1 and

The automated process in the CAD system will improve diagnosis efficiency and relieve pathologists workload. The traditional machine learning methods with classifiers predict the cancer labels from the histopathology images from the derived handcrafted features [4-7]. Due to the complexity of meaningful handcrafted feature extraction and the requirement of extensive domain knowledge, machine learning becomes unsatisfactory. In recent years, deep learning has evolved in the research field in which the features are extracted automatically and perform classification. The convolution neural networks are utilized in deep learning to extract the features from raw images. The challenging task in deep learning is it consumes more time during training [8].

The convolution layer comprises multiple maps of neurons or filters. The filter performs a reduced sample from the previous layers. However, the neuron in filters detects the same features from previous layers and mapping in different locations. The final pooling layer summarizes adjacent neurons, and the parameters are significantly reduced [9]. The fully connected layer after the pooling layer decides the class, which is binary or multi-class.

Several approaches have been developed in recent years to identify an object to simulate human ability. One such method is the associative approach, which retrieves the input data complete patterns. CNN model is configured effectively with multi-dimensional input data. The designed model is utilized to solve problems of image recognition and computer vision [10]. In this work, one-hot method is utilized that employs CNN on histopathology datasets.

The BreakHis histopathology image dataset comprises two major classes (Benign and malignant) with eight subclasses. Each image is encoded with the binary vector of eight bits, with one of them is hot (1) While the others are 0. For instance sub class adenosis (SOB_B_A-14-22549AB-100-001) encoded as (1, 0, 0, 0, 0, 0, 0, 0). The values in vectors are considered as the probability of finding eight bases at the significant location. Once encoded, the image is considered a sequence and represented in an 8× 1 matrix. However, the 2D image is presented in one channel. In medical imaging, CNN has utilized adequately in the field of breast cancer diagnosis [11]. The input data is utilized in the transfer learning model to classify the benign and malignant.

In this presented work, fine-tuned VGG-16 transfer learning approach is used for classification. It is a sophisticated task to detect the exact type of breast cancer due to the existence of morphological parameters for identifying the stroma that is tumor affected. It is crucial to create an efficient and accurate algorithm that can identify and differentiate a tumor-bound stroma from normal stroma from the histopathological images. To achieve this objective, it may benefit from relying on deep learning algorithms vital to machine learning techniques. The CNN can recognize most of the segregate features from a broad range of diagnostic images without the requirement of pre-defined morphological parameters.

The paper is organized as follows. Section 2 discusses the related literature works in breast cancer prediction. Section 3 deals with materials and methods with fine-tuned VGG-16 model. Section 4 is about experimental results. The final section is the conclusions of the proposed approach.

II. RELATED WORKS

CNN is a useful tool utilized in the field of pattern recognition and applications [12][13], such as handwritten identification and object identification from a large dataset [14]. CNN has the advantages over traditional techniques in the field of biomedical imaging. Tuba *et al.* implemented a statistical neural network-based breast cancer diagnosis system. The radial basis function (RBF), statistical neural network, and regressions were utilized on the breast cancer dataset. The system achieves 98.8% classification results [15].

In [16], the authors proposed a back-propagation neural network with Levenberg–Marquardt algorithm. They accomplished superior classification accuracy of 99.28%. The authors performed pre-processing by using median filtering and min-max technique. They utilized 80% and 20% data for training and testing. In [17], the authors proposed a feed-forward neural network for breast cancer classification. Based on the splitting of training and testing data, the results were obtained.

In [18], the authors implemented an artificial neural network with multilayer perceptron and backpropagated neural network. Bewal et al. proposed a back propagated neural network with quasi-Newton and Levenberg–Marquardt algorithm with MLP to train the network [19]. The performance measure can be accomplished with the steepest descent back-propagation algorithm. The system obtained 94.11% classification accuracy. Spanhol et al. implemented CNN on extracted pixel patches from breakHis dataset using the sliding window method [20]. The patches achieved accuracies between 80.8% and 89.6%.

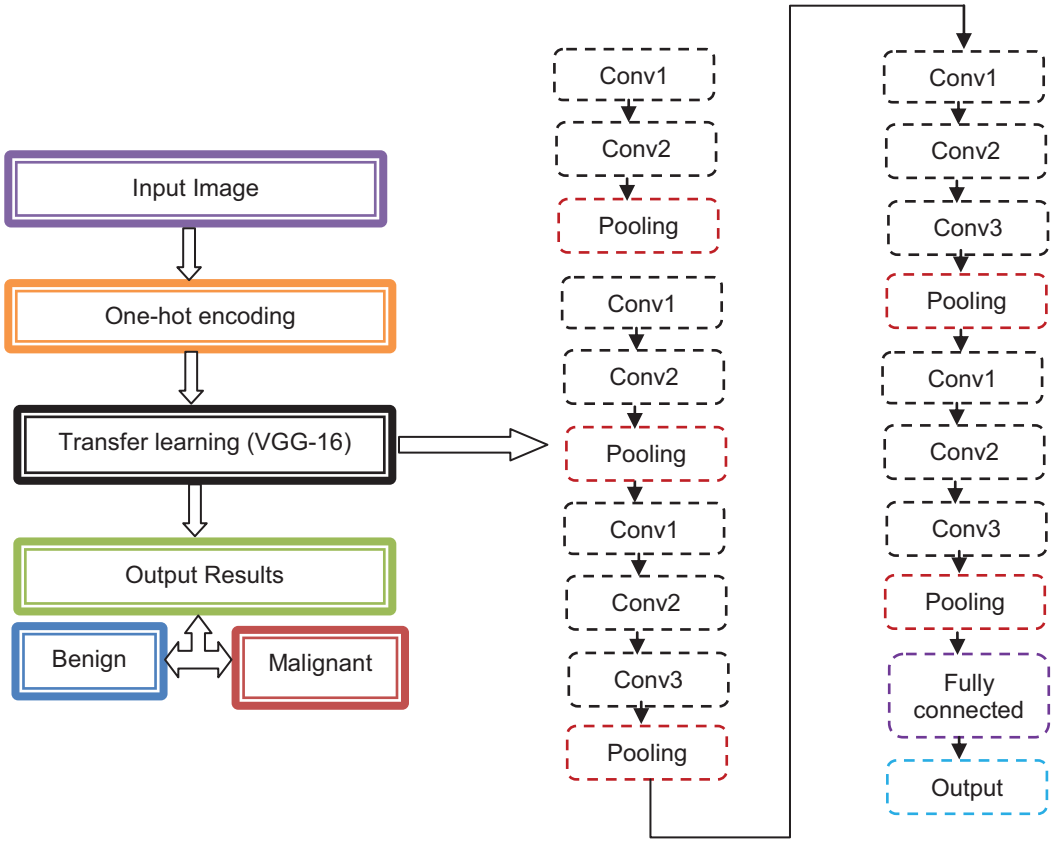


Fig. 1. Block diagram of the proposed scheme

III. MATERIALS AND METHODS

A. Dataset

The medical process of generating the dataset includes the following steps: For the diagnosis, the tissues and other materials are treated with formalin and then are stored in paraffin to stop the cultivation of microbes. Then, using a high precision instrument, the required tissues are separated. These tissues are then put on glass slides for inspection. These glass slides are also treated with hematoxylin and eosin (H&E) so that the internal constituents of the tissues, such as cell nuclei and cytoplasm, can be observed. These slides are then scrutinized under the microscope to generate histopathological slides by varying magnification. Lastly, these images are then manually diagnosed by pathologists by optical examination. BreakHis is a benchmark dataset created by P&D laboratory, Brazil.

The dataset consists of a total of 7,909 images collected from 82 patients with two major classes (benign and malignant). However, this data is far from being error-free. The primary issue is that it consists of the irregular ratio of two classes, with 31.36% of images belonging to the benign class and the rest 68.64% belonging to the malignant class. The detailed main class image distribution is depicted in the following table 1. Also, each of these classes contains four subclasses based upon their characteristics and imminence.

B. Proposed scheme

The input image converted into data matrices using one hot encoding strategy. In order to predict the breast cancer, we implement the VGG-16 pretrained network to learn higher order features by utilizing two basic blocks feature extraction and classification layers. The CNN layers perform spatial correlation between adjacent matrixes. As depicted in figure 1, the VGG-16 model comprises five blocks of convolution layers followed by five max pooling layers which accomplish feature extraction. In this proposed work, the classification layers execute flattening using fully connected layer with eight subclass output. The final classification layer delivers the classification result (Benign or malignant).

C. One-hot encoding

The encoding of categorical variables depends on a set of finite and collected categories with mutually exclusive elements [21]. Some of the literature considered encoding relies on high-cardinality categorical variables [22]. In CNN, the input images of all subclass are represented as a vector of numerical values. In this work, a one-hot encoding scheme is utilized in which the input length is denoted as l, and the total number of input values are given as 8×l.

The advantage of using this scheme is it requires a short time for a large dataset. However, it utilizes less memory with minimum computations. The one-dimension vector (1×l) can be reshaped into two-dimensional matrixes (m × n). The encoding scheme of the subclass is shown in table 2. The two major categories are benign and malignant, encoded as 1 and 0.

D. CNN Architecture

Transfer learning transfers knowledge learned from one domain to another for feature extraction and classification [23]. The transfer learning method is usually utilized to solve small dataset classification problems. The transfer learning performs precise classification task with fine-tuned parameters.

In deep learning, transfer learning is performed by utilizing a convolution neural network that was already trained with a large database. The fine-tuning parameter is used to train further by utilizing a smaller dataset. Most of the research used with pre-trained models due to fast and trained easily. In general, initial layers identify curves, corners, edges, color blobs features, and classification performed in the final layer [24]. The new classification task was carried out by using three layers (classification layer, fully conned layer and softmax layers).

For our work, we utilized the VGG-16 convolution neural network; however, it achieves 92.7 % accuracy with 14 million images. Layer architecture is shown in table 3. The input image is resized with 224*224*3 for training. The eight sub-categories are classified under two major classes (benign and malignant). The fully connected layer is replaced with eight, and the final classification layer performs the output. The fine-tuning and classification process is shown in figure 2. The CNN model in transfer learning was integrated with 13 convolution layers in the VGG-16 pre-trained model with a fully connected layer shown in table 3.

As depicted in table 3, the CNN block weights were imported from the VGG-16 model without any modification. The fully connected layer was modified with random initialization and training.

IV. RESULTS AND DISCUSSION

The implementation of the one-hot encoding algorithm for multi-category classification that can predict a specific sub-category among Fibroadenoma (F), Adenosis (A), tubular adenoma (TA), and phyllodes tumor (PT), ductal carcinoma (DC), mucinous carcinoma (MC), lobular carcinoma (LC), and papillary carcinoma (PC) sub-classes of breast cancer. The custom model has been trained on a local GPU for over a total of 3 hours for different iterations of varying epochs. After the training is done, the resulting model will be saved as a weights file within a specified directory. This specific file can be further used to train the custom again, acting as a feedback structure and increasing the model accuracy.

The results of the designed custom model are as follows. Initially, we have set the training to classify only the major classes, and we got training accuracy of 0.8935 and validation accuracy of 0.7365. These low rates were mainly due to randomness that occurred during data splitting.

However, the main focus of the proposed work is multi-classification. So, we repeated the training process, including all eight sub-classes, to classify the data into ten classes (2 major categories and 8 sub-categories) and got the training and validation accuracy of 0.9714 and 0.8894 shown in figure 3(a). Figure 3(b) represents loss after 25 epochs.

Then the epoch has been increased to 50. At this stage, the superior results were accomplished for training and validation accuracy (0.9862 and 0.9511). The results were shown in figure 4(a). However, the loss is gradually reduced for training and validation, illustrated in figure 4(b).

The superior results are accomplished after 50 iterations, and the model was tested with random input images from various subclasses. The one-hot encoder performs error-free results shown in Figures 5. The values closer to 1 provide good results.

TABLE IV. DATASET DESCRIPTION.

Category	Benign				Total Benign	Malignant				Total Malignant	Total Both
Sub-category	A	F	PT	TA		DC	LC	MC	PC		
# images	444	1014	453	569	2480	3451	626	792	560	5429	7909
# patients	4	10	3	7	24	38	5	9	6	58	82

TABLE V. ENCODING SCHEME USED IN BREAKHIS DATASET

Images	Benign				Malignant			
	A	FA	TA	PT	DC	LC	MC	PC
SOB_B_A-14-22549AB-100-001	1	0	0	0	0	0	0	0
SOB_B_F-14-23060AB-400-008	0	1	0	0	0	0	0	0
SOB_B_TA-14-16184CD-100-027	0	0	1	0	0	0	0	0
SOB_B_PT-14-22704-100-025	0	0	0	1	0	0	0	0
SOB_M_DC-14-11031-400-012	0	0	0	0	1	0	0	0
SOB_M_LC-14-12204-400-032	0	0	0	0	0	1	0	0
SOB_M_MC-14-12773-40-020	0	0	0	0	0	0	1	0
SOB_M_PC-14-15704-100-003	0	0	0	0	0	0	0	1

TABLE VI. DESCRIPTION OF VGG-16 LAYERS

	Blocks	Layers	Filter size	Output Shape
VGG-16	Input layer (224, 224, 3)			
	Conv block 1	Convolution 2D (Conv1)+ReLU	3×3	(224, 224, 3)
		Convolution 2D (Conv2)+ReLU	3×3	(224, 224, 3)
		Maxpooling 2D	2×2	(112, 112, 64)
	Conv block 2	Convolution 2D (Conv1)+ReLU	3×3	(112, 112, 128)
		Convolution 2D (Conv2)+ReLU	3×3	(112, 112, 128)
		Maxpooling 2D	2×2	(56, 56, 128)
	Conv block 3	Convolution 2D (Conv1)+ReLU	3×3	(56, 56, 256)
		Convolution 2D (Conv2)+ReLU	3×3	(56, 56, 256)
		Convolution 2D (Conv3)+ReLU	3×3	(56, 56, 256)
		Maxpooling 2D	2×2	(28, 28, 256)
	Conv block 4	Convolution 2D (Conv1)+ReLU	3×3	(28, 28, 512)
		Convolution 2D (Conv2)+ReLU	3×3	(28, 28, 512)
		Convolution 2D (Conv3)+ReLU	3×3	(28, 28, 512)
		Maxpooling 2D	2×2	(14, 14, 512)
	Conv block 5	Convolution 2D (Conv1)+ReLU	3×3	(14, 14, 512)
		Convolution 2D (Conv2)+ReLU	3×3	(14, 14, 512)
		Convolution 2D (Conv3)+ReLU	3×3	(14, 14, 512)
		Maxpooling 2D	2×2	(7, 7, 512)
Fine-tuning layers	Fully connected layer (8)+ReLU			
	Output layer			

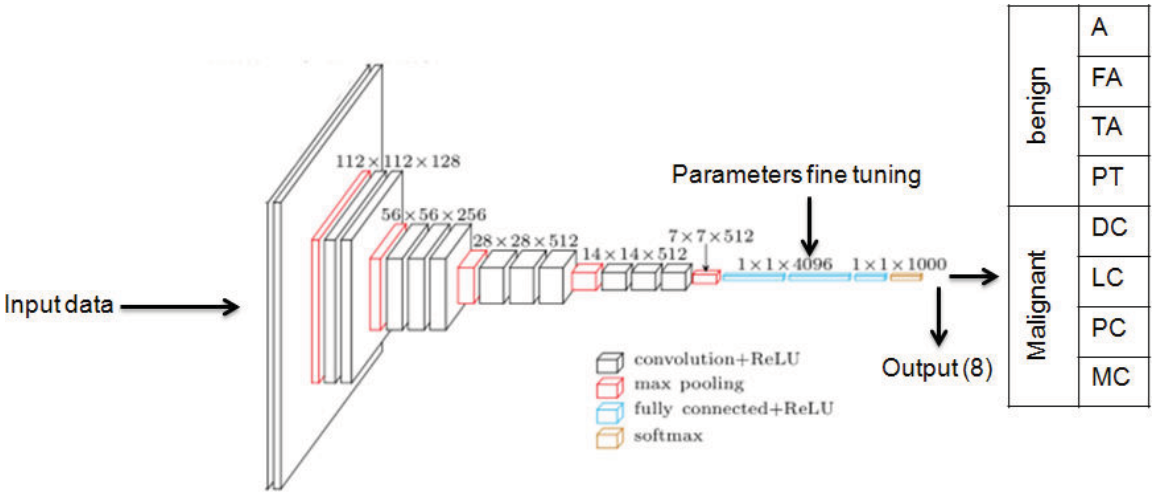


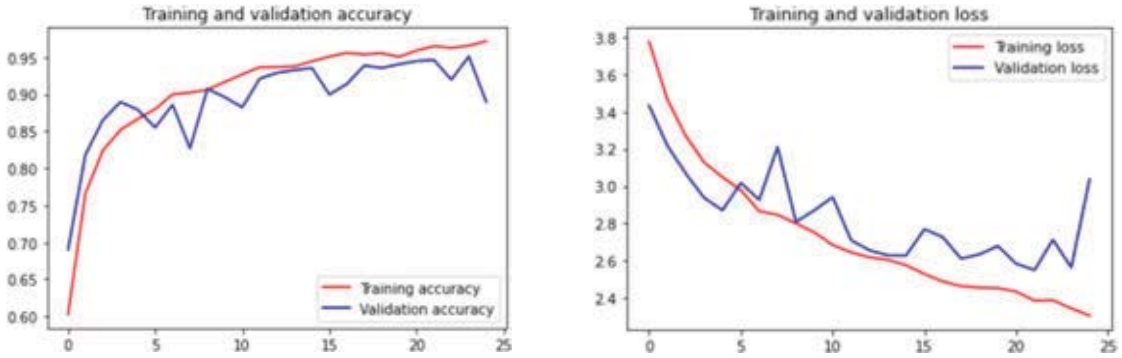
Fig. 2. Classification based on transfer learning using VGG-16

TABLE IV. METRICS EVALUATION AFTER 25 EPOCHS

	Training	Validation
Accuracy	0.89350	0.736458
AUC	0.95809	0.816399
F-Score	0.89350	0.726458
MCC	0.78700	0.472917
Precision	0.89350	0.736458
Specificity	0.89350	0.736458
Recall	0.89350	0.736458

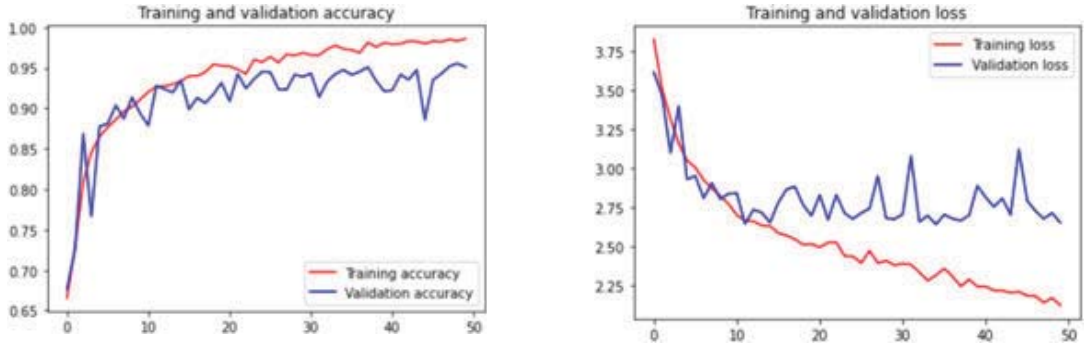
TABLE V. METRICS EVALUATION AFTER 50 EPOCHS

	Training	Validation
Accuracy	0.986249	0.951149
AUC	0.981978	0.961103
F-Score	0.847473	0.821496
MCC	0.812639	0.781980
Precision	0.886656	0.871475
Specificity	0.974063	0.971354
Recall	0.811607	0.776940



(a) (b)

Fig. 3. Accuracy and loss function for 25 epochs.



(a) (b)

Fig. 4. Accuracy and loss function for 50 epochs.

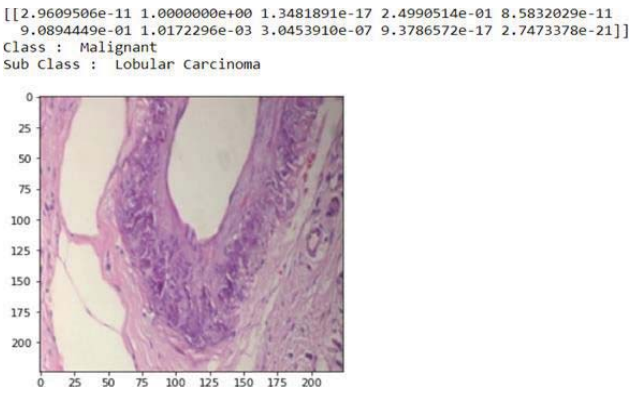


Fig 5. Predicted output Lobular Carcinoma

V. CONCLUSION

Breast cancer has become the most diagnosed cancer, beating lung cancer among women around the world. The primary cause is mainly because the mutations of this cancer are significantly increasing. The notified reason from literature is mainly due to delayed diagnosis. This late diagnosis can prove to be deadly as the consequences of it remains for a lifetime. However, the deep learning model has been developed based on VGG-16 to help pathologists diagnose breast cancer's specific sub-category and avoid misdiagnosis. The valuable lives are under threat because of lack of awareness, improper diagnosis, and lack of assessment. The observation rate of breast cancer is very low in India because of its high charge and lack of portable self-check-up devices. In such cases, the proposed model can recognize the stage of cancer and eight different subclasses with 98% (training) and 95% (validation) accuracy. To get maximum validation accuracy even with low-resolution reports so that the model can be automated in portable devices so that Check-up becomes handier at low expenditures. In the future, the work was modified with multiple algorithms to increase accuracy with precise predictions.

REFERENCES

[1]. F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J. Clin. 68 (November (6)) (2018) 394–424, https://doi.org/10.3322/caac.21492.

[2]. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics. CA A Cancer J. Clin. **2016**, 66, 7–30.

[3]. U.S. Breast Cancer Statistics. Available online: https://www.breastcancer.org/symptoms/understandfgbc/statistics

[4]. P. Filipczuk, T. Fevens, A. Krzyzak, R. Monczak, Computer-aided breast Cancer diagnosis based on the analysis of cytological images of fine needle biopsies, IEEE Trans. Med. Imaging 32 (December (12)) (2013) 2169–2178, https://doi.org/ 10.1109/TMI.2013.2275151.

[5]. A.D. Belsare, M.M. Mushrif, M.A. Pangarkar, N. Meshram, Classification of breast cancer histopathology images using texture feature analysis, in: TENCON 2015 - 2015 IEEE Region 10 Conference, Macao, 2015, pp. 1–5, https://doi.org/10.1109/ TENCON.2015.7372809. Nov

[6]. S. Wan, X. Huang, H.-C. Lee, J.G. Fujimoto, C. Zhou, Spoke-LBP and ring-LBP: New texture features for tissue classification, in: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), Brooklyn, NY, USA, 2015, pp. 195–199, https://doi.org/10.1109/ISBI.2015.7163848. Apr.

[7]. Y.M. George, H.H. Zayed, M.I. Roushdy, B.M. Elbagoury, Remote computer-aided breast Cancer detection and diagnosis system based on cytological images, IEEE Syst. J. 8 (September (3)) (2014) 949–964, https://doi.org/10.1109/ JSYST.2013.2279415.

[8]. N. Tajbakhsh, et al., Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans. Med. Imaging 35 (May (5)) (2016) 1299–1312,

[9]. C. Angermueller, T. P¨arnamaa, L. Parts, O. Stegle, F. Albert, S. Treusch, A. Shockley *et al.*, "Deep learning for computational biology." *Molecular* systems biology, vol. 12, no. 7, p. 878, 2016.

[10]. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015.

[11]. Byra M. Discriminant analysis of neural style representations for breast lesion classification in ultrasound. Biocybernetics Biomed Eng 2018;38:684–690.

[12]. X.Ma, P. K. Wong, and J. Zhao, "Practicalmulti-objective control for automotive semi-active suspension systemwith nonlinear hydraulic adjustable damper,"Mechanical Systems and Signal Processing, vol. 117, pp. 667–688, 2019.

[13]. F. Liu, G. Lin, and C. Shen, "CRF learning with CNN features for image segmentation," Pattern Recognition, vol. 48, no. 10,pp. 2983–2992, 2015.

[14]. D. C. Cires¸an,U.Meier, L. M.Gambardella, and J. Schmidhuber, "Convolutional neural network committees for handwritten character classification," in Proceedings of the __th International Conference on Document Analysis and Recognition, pp. 1135–1139, September 2011.

[15]. Kiyan, T., & Yildirim, T. (2004). Breast cancer diagnosis using statistical neural networks. Journal of Electrical and Electronics Engineering, 4(2), 1149–1153.

[16]. Paulin, F., & Santhakumaran, A. (2011). Classification of breast cancer by comparing back propagation training algorithms. International Journal on Computer Science and Engineering, 3(1), 327–332.

[17]. Jhajharia, S., Varshney, H. K., Verma, S., & Kumar, R. (2016). A neural network based breast cancer prognosis model with pca processed features. In: 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1896–1901.

[18]. Jouni, H., Issa, M., Harb, A., Jacquemod, G., & Leduc, Y. (2016). Neural network architecture for breast cancer detection and classification. In: IEEE International Multidisciplinary Conference on Engineering Technology (IMCET), pp. 37–41.

[19]. Bewal, R., Ghosh, A., & Chaudhary, A. (2015). Journal of Clinical and Biomedical Sciences, 5(4), 143–148.

[20]. F.A. Spanhol, L.S. Oliveira, C. Petitjean, L. Heutte, Breast cancer histopathological image classification using convolutional neural networks, in: 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 2016, pp. 2560–2567.

[21]. Manoharan, Samuel. "Improved Version of Graph-Cut Algorithm for CT Images of Lung Cancer With Clinical Property Condition." Journal of Artificial Intelligence 2, no. 04 (2020): 201-206.

[22]. Pandian, A. Pasumpon. "Identification and classification of cancer cells using capsule network with pathological images." Journal of Artificial Intelligence 1, no. 01 (2019): 37- 44.