FEATURE EXTRACTION METHODS BASED ON DEEP-LEARNING APPROACHES. APPLICATION TO AUTOMATIC DIAGNOSIS OF BREAST CANCER

Stefania BARBURICEANU, Romulus TEREBES
Technical University of Cluj-Napoca, Communications Department, Cluj-Napoca, Romania
Stefania.Barburiceanu@com.utcluj.ro, Romulus.Terebes@com.utcluj.ro

Abstract: This paper proposes the use of several pre-trained CNN architectures (Vgg16, AlexNet, and Resnet50) for texture feature synthesis. The information gained from training the models on a very large dataset (ImageNet) is transferred to a new classification problem (where the dataset is rather small) on the premise of transfer learning. For the classification part, the Support Vector Machine is considered. Since in the previous two decades, cancer has become more widespread, and by the year 2020, breast cancer was considered the most common disease on the planet, we are interested in applying the proposed strategy for detecting breast cancer. Various imaging modalities have been developed to diagnose this type of disease, with mammography being the most often employed. However, it is the expert's histological analysis that determines the ultimate diagnosis. As a result, pathologists may benefit from the development of autonomous diagnostic techniques based on computer vision technologies. Thus, a public database with histological images of malignant and benign breast tissues is employed in the conducted experimental analysis. For the same dataset, the present methodology outperforms other classical machine learning methods as well as CNN approaches.

Keywords: *medical image classification, deep-learning, breast cancer, disease classification.*

I. INTRODUCTION

Deep-learning approaches have recently demonstrated their ability to classify with a very high accuracy a variety of images from different domains including medical, agricultural, remote sensing, industrial, and more. The Convolutional Neural Network (CNN) is the most widely used deep-learning approach for image classification.

The CNN may be utilized as an end-to-end strategy that uses a multi-layered convolutional base network to automatically find the most relevant features of the considered image and a final network (based on fully-connected layers) that is capable of performing image classification. The primary downside of this strategy is that the classification performance is highly dependent on the amount of samples used to train the system. Usually, the classification results are not good enough due to the limited number of labeled medical images available online. As a consequence, the concept of transfer learning becomes quite valuable. This entails applying previously learned knowledge to a new task. The information is acquired from a task where a very large training database is used and thus, many key features are learned. The knowledge can be further employed in another classification problem. This implies using the CNN for generating the most significant features by using the weights already learned in the initial task. These features are then considered as input for a classification technique. Because extracting features based on the transfer-learning process involves a single passage of the image sample through the network, the computation time is drastically reduced. Figure 1 describes the transfer learning strategy.

According to the World Health Organization, the total number of cancer patients in 2020 is almost twice as much as in 2000 and the predictions suggest that the number of cancer cases will rise even more in the following years. Cancer mortality has also increased, with cancer accounting for more than one in every six deaths. The prevalence of cancer has grown as a result of poor nutrition, a sedentary lifestyle, and excessive tobacco and alcohol intake [1]. This requires the development of technologies to accelerate cancer detection in order to respond sooner and get the best possible results.

Breast cancer was the most common cancer in the world by the end of 2020, with 7.8 million women diagnosed in the previous five years [2]. This requires the development of automated technologies that are able to aid clinicians in making a diagnosis.

Various imaging modalities, such as the ones based on X-rays (e.g. mammography), ultrasound (e.g. sonography), magnetic fields (e.g. Magnetic Resonance Imaging - MRI), gamma radiation, or non-ionizing radiation, are utilized to make a diagnosis of breast cancer [3]. The most prevalent imaging procedure for breast cancer is mammography, with MRI and ultrasound as additional procedures that aid in the diagnosis process.

Imaging procedures can help in detecting abnormal tissue, but in most circumstances, a biopsy is the safest approach to make a diagnosis [4]. In most situations, if the considered imaging methods reveal a potential problem, a tissue sample is taken, which is subsequently inspected under a microscope by a specialist to determine the proper diagnosis and disease stage. In the case of breast cancer, the expert's histopathological investigation is the one that establishes the final diagnosis. One of the issues in this

Manuscript received March 14, 2022; revised June 13, 2022

scenario is the scarcity of professionals, particularly in smaller, less developed countries or towns. This lengthens the time it takes to generate the biopsy result, which has a negative impact in the event of a cancer diagnosis, where the time to act is critical. Secondly, because this procedure is time-consuming and very complex, exhaustion and tiredness may emerge as a result of the stress, which can lead to diagnostic mistakes. As a result, the development of autonomous diagnostic procedures based on computer vision approaches may help ease the pathologists' effort by screening out evident benign cases. As a result, pathologists can concentrate on cases where determining the diagnosis is more challenging.

Many works [5]–[8] have addressed the transfer learning strategy by employing features retrieved from pre-trained CNNs. To generate features for the current classification problem, a pre-trained network with the already learned weights is used. For this network, only the classification layers are removed. Various CNN models have been developed by researchers and are successfully used for feature extraction such as VggNet [9], AlexNet [10], and ResNet [11].

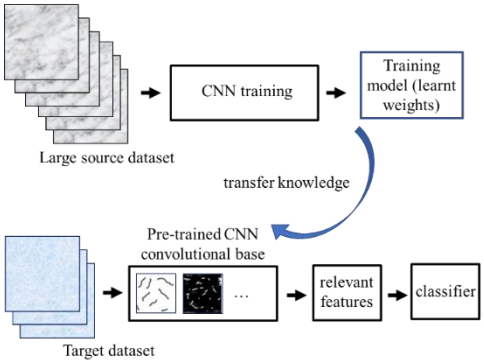

Figure 1. The transfer learning strategy

Many researchers focused on developing automatic systems for the classification of breast cancer. Spanhol et al. [12] describe their proposed database, BreaKHis, which comprises images of malignant and benign breast tumors. The images are captured at four different magnification factors. This database is described in detail in Section III. The authors use different techniques such as the Local Binary Patterns (LBP), the Local Phase Quantization, the Gray-Level Co-occurrence Matrix (GLCM), and others for feature generation. For the classification stage, there are employed four algorithms: the 1-Nearest Neighbor, the Quadratic Linear Analysis, the Support Vector Machine (SVM), and the Random Forest technique. The highest accuracy of 85.1% is obtained for the 200X magnification case. In [13], the same images are classified using an end-to-end CNN technique which is trained on image patches. They also explore the fusion of results obtained by using four CNNs trained on different image patches.

In [14], the authors use another breast cancer database and propose a novel CNN architecture for the classification of images in four classes. The results are also compared to the ones achieved by feeding the features generated by using the proposed CNN into an SVM classifier. In [15], Belsare et al. used 40X magnification breast histopathology images to perform binary classification. After the segmentation of the breast duct, texture characteristics are employed as input for the SVM, the k Nearest Neighbours, and the Linear Discriminant Analysis classifiers, with the latter achieving the best results. In [16], the authors employed a CNN to classify invasive ductal carcinoma in breast cancer images. When compared to classical handcrafted approaches, the network trained on 100×100 image patches produced better results.

II. PROPOSED APPROACH

In this paper, pre-trained CNN models are employed for feature vector generation. This procedure involves the use of a CNN model that has been pre-trained on a database containing a very large number of samples. Because the CNN is not retrained, the weights remain the same. The usage of pre-trained models provides a number of benefits. First of all, the processing time for the feature generation process is reduced since the images only traverse the network once. Secondly, promising results may be achieved even for a limited number of training images.

In the proposed approach, a pre-trained CNN model is used as a feature extractor, and for the classification stage, the Support Vector Machine is considered. Three pre-trained models are explored in our work: Vgg16 [9], AlexNet [10], and Resnet50 [11]. They are described in detail in Figure 3-Figure 5.

Convolution is the basic process of applying a filter to an input to produce an activation. This entails multiplying the input image with a 2D matrix containing the weights (referred to as a filter). The position and strength of an observed feature are shown by a feature map created by repeatedly applying the same filter to the input data. During the training phase, the CNN learns which are the relevant features (the filter weights). In the considered approach, several convolutional layers are used to produce multiple feature maps. Each feature map corresponds to a different filter and each filter is able to detect different features in the input data. The input image passes through a series of convolutional layers and pooling layers (used for downsampling) and, in the proposed strategy, the feature vector is obtained by computing the mean value over each obtained feature map. The considered approach is described in detail in Figure 2.

The networks are trained on a very large dataset that has 1.2 million samples and 1000 different image categories, ImageNet [17]. The images are resized in order to match the size required by each considered CNN model.

We test the extraction of features from various layers in each network under consideration, and only the best results are presented. The layers supporting the best classification accuracy are detailed in Table ***I***.

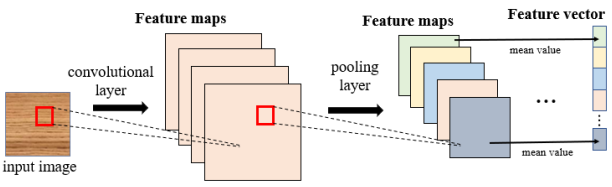

Figure 2. The feature extraction process

Table I. Considered feature extraction layers

CNN model	Considered layer for feature extraction
Vgg16	relu4_1
AlexNet	relu4
Resnet50	activation_22_relu

We also consider the concatenation of feature vectors extracted from Vgg16 and Ressnet50.

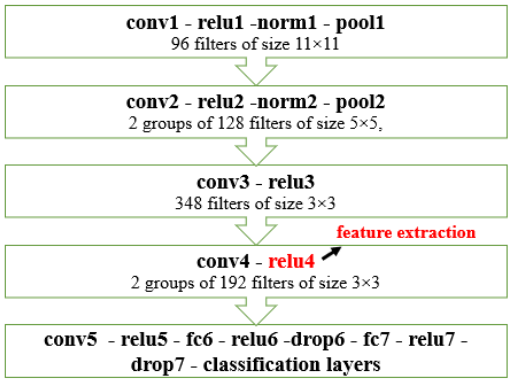

Figure 3. AlexNet structure

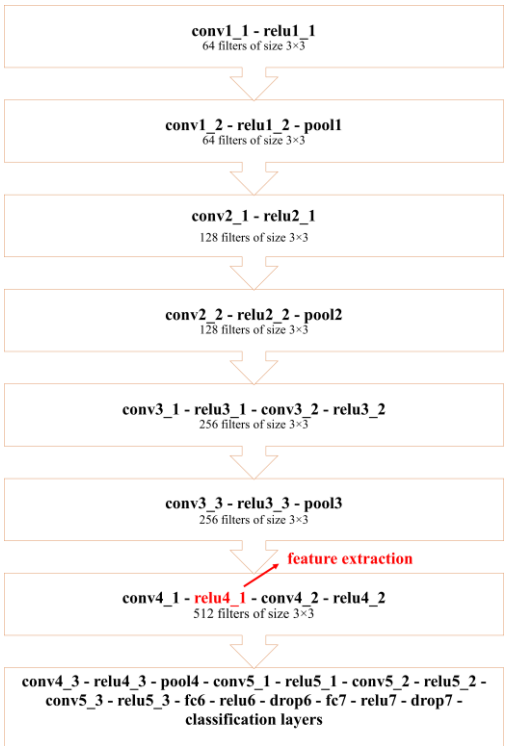

Figure 4. Vgg16 structure

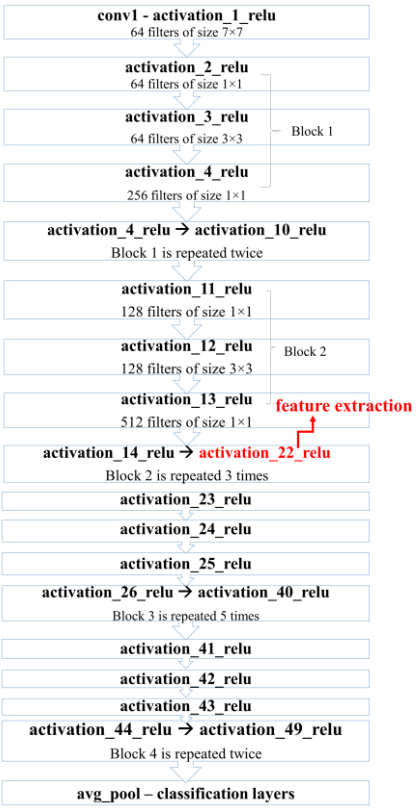

Figure 5. Resnet50 structure

III. EXPERIMENTAL RESULTS

1. Dataset

The BreaKHis dataset [12] is used for assessing the performance of the approaches described in the previous section. It comprises histopathological images of benign and malignant breast cancer obtained from 82 individuals. The dataset's 7909 samples are split into two categories: benign and malignant tumors. The following four magnification factors are considered when capturing the images from the BreaKHis dataset: 40X, 100X, 200X, and 400X. All samples are RGB images of size 700×460 pixels. The structure of this dataset is shown in Table II.

Table II. Dataset structure

Magnification factor	Classes		Total number of samples
	Benign	Malignant	
40×	625	1370	1995
100×	644	1437	2081
200×	623	1390	2013
400×	588	1232	1820
Number of patients	**24**	**58**	**82**

The different types of benign and malignant tumors for this database are presented in

Table III. Types of cancer

Benign tumors	Malignant tumors
adenosis (A)	ductal carcinoma (DC)
fibroadenoma (F)	lobular carcinoma (LC)
phyllodes tumour (PT)	mucinous carcinoma (MC)
tubular adenoma (TA)	papillary carcinoma (PC)

Figure 6 depicts some instances of benign and malignant breast cancer images obtained at various magnifications.

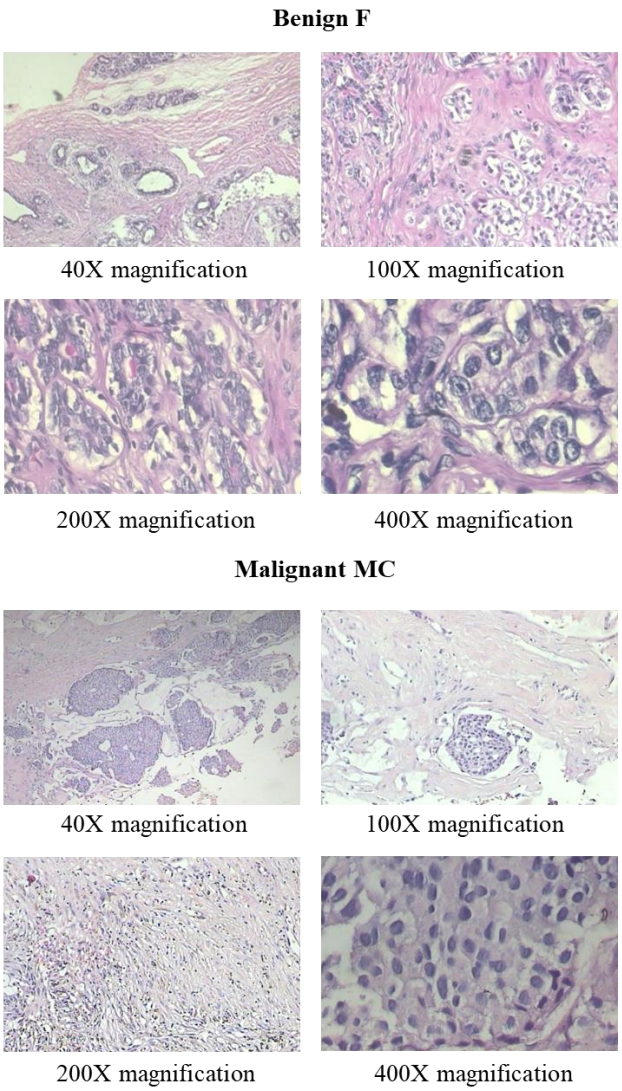

Figure 6. Examples of image samples [12]

2. Results and discussion

We used the Support Vector Machine (SVM) technique for the classification stage of our approach. The SVM parameters were determined using a grid search strategy for achieving the best results.

In [12] and [13], the authors used the same dataset to evaluate different feature extraction and classification techniques. In the aforementioned papers, the patients used for training are not used in the testing phase. In order to observe if the classification system is able to generalize well for other patients, we consider in our approach the following rule: all images belonging to a patient are either used for training or testing.

In [12] and [13], 70% of the images are used for training and 30% for testing and the presented results represent the average over five randomly chosen sets. In order to achieve good statistical confidence, we consider 50 random partitionings of the training and test sets. In order to guarantee that in the testing step there are used only images from unseen patients in the training phase, instead of considering 70% of the images from each class for training, we use 70% of the patients for the training step. The number of acquired images is different from patient to patient. This means that the number of images considered for training in each partition can be different. We also report the results obtained using the same five folds used in [12] and [13].

As in [12] and [13], we will consider the sets corresponding to the magnification factors independently. We explore two types of classification metrics, similarly to the work in [13]: average accuracy at image level (denoted by *AI*) and average accuracy at patient-level (denoted by *AP*), given in Eq. (1) and (2):

$$\mathrm{AI} = \frac{\text{Number of correctly classified images}}{\text{Total number of images in the test set}} \quad (1);$$

$$\mathrm{AP} = \frac{\mathrm{PS}}{\text{Total number of patients}} \quad (2),$$

where PS is the patient score given in Eq. (3):

$$\mathrm{PS} = \frac{\text{number of correctly classified images of a patient}}{\text{total number of images of a patient}} \quad (3).$$

Table IV, VI, VIII, and X show the obtained classification results for the images obtained using the four different magnification factors by considering the same five folds as in [12] and [13]. We also show in

Table ***V***, VII, IX, and XI the scores obtained by considering 50 random partitionings.

Table IV. Classification scores obtained considering five folds on 40× images [%]

Metric/ Operator	Vgg16 (relu4_1)	AlexNet (relu4)	Resnet50 (activation_22_relu)	Concatenation Vgg16 Resnet50
AI	89.6±1.6	84.57±1.08	89.41±3.2	**90.17±1.41**
AP	90.5±3.2	86.7±2.26	89.82±4.04	**91.07±2.48**

Table V. Classification scores obtained considering 50 random partitionings on 40× images [%]

Metric/ Operator	Vgg16 (relu4_1)	AlexNet (relu4)	Resnet50 (activation_22_relu)	Concatenation Vgg16 Resnet50
AI	87.79±5.06	83.57±4.81	88.8±4.07	**89.16±4.16**
AP	88.03±5.65	85.05±4.74	89.47±4.08	**89.79±4.56**

Table VI. Classification scores obtained considering five folds on 100× images [%]

Metric/ Operator	Vgg16 (relu4_1)	AlexNet (relu4)	Resnet50 (activation_22_relu)	Concatenation Vgg16 Resnet50
AI	89.6±2.73	85.86±2.02	88.89±3.39	**89.79±2.73**
AP	89.86±2.95	86.63±2.46	**90.09±4.02**	89.97±3.38

Table VII. Classification scores obtained considering 50 random partitionings on 100× images [%]

Metric/ Operator	Vgg16 (relu4_1)	AlexNet (relu4)	Resnet50 (activation_22_relu)	Concatenation Vgg16 Resnet50
AI	88.14±3.26	86.19±3.8	**88.66±3.87**	87.55±4.19
AP	88.4±3.47	87.09±3.68	**89.44±3.73**	87.65±4.13

Table VIII. Classification scores obtained considering five folds on 200× images [%]

Metric/ Operator	Vgg16 (relu4_1)	AlexNet (relu4)	Resnet50 (activation_22_relu)	Concatenation Vgg16 Resnet50
AI	89.86±2.69	87.24±3.04	**90.47±1.8**	90.05±1.44
AP	89.72±3.38	88.21±4.03	90.03±2.41	**90.89±1.34**

Table IX. Classification scores obtained considering 50 random partitionings on 200× images [%]

Metric/ Operator	Vgg16 (relu4_1)	AlexNet (relu4)	Resnet50 (activation_22_relu)	Concatenation Vgg16 Resnet50
AI	88.54±4.39	85.54±4.29	88.52±4.2	**89.8±3.99**
AP	88.78±4.29	86.05±4.11	88.42±4.49	**90.13±3.68**

Table X. Classification scores obtained considering five folds on 400× images [%]

Metric/ Operator	Vgg16 (relu4_1)	AlexNet (relu4)	Resnet50 (activation_22_relu)	Concatenation Vgg16 Resnet50
AI	86.31±0.78	84.93±3.07	87.76±1.65	**88.61±2.42**
AP	85±1.66	84.67±4.42	86.73±2.46	**87.58±2.95**

Table XI. Classification scores obtained considering 50 random partitionings on 400× images [%]

Metric/ Operator	Vgg16 (relu4_1)	AlexNet (relu4)	Resnet50 (activation_22_relu)	Concatenation Vgg16 Resnet50
AI	85.59±4.92	84.37±3.61	85.73±4.16	**86.88±4.34**
AP	85.02±5.16	84.51±3.99	85.6±4.15	**85.78±4.29**

In [12], the authors used different handcrafted feature extraction techniques for the classification of images in the BreaKHis dataset, while in [13], the same images are classified using an end-to-end CNN technique which is trained on image patches. In [12], only the average accuracy at patient level is reported while in [13], both the average accuracy at patient and image level are presented. For comparison purposes, we show in Table XII the best results reported in [12] and [13]. The authors report only the classification accuracy and other figures such as the false positive or false negative rate are not considered in [12] and [13].

Table XII. Comparison to other works

Work in literature	Classification metric	40X	100X	200X	400X
[12]	Patient level	83.8%	82.1%	85.1%	82.3%
[13]	Patient level	88.6% for CNN and 90% for CNN fusion	84.5% for CNN and 88.4% for CNN fusion	85.3% for CNN and 84.6% for CNN fusion	81.7% for CNN and 86.1% for CNN fusion
[13]	Image level	89.6% for CNN and 85.6% for CNN fusion	85% for CNN and 83.5% for CNN fusion	84% for CNN and 83.1% for CNN fusion	80.8% for CNN and 80.8% for CNN fusion
Proposed approach	Patient level	91.07%	90.09%	90.89%	87.58%
Proposed approach	Image level	90.17%	89.79%	90.47%	88.61%

From the obtained results, it can be observed that by using pre-trained deep-learning models, the obtained classification scores are improved compared to [12] and [13]. Furthermore, the processing time is also decreased compared to [13] because the end-to-end CNN training is very time-consuming, especially when considering the fusion strategy. Because the input images only run through the network once in our scenario, the feature extraction and the classification procedures are more efficient.

We show in Figure 7 the confusion matrix obtained on one run on the dataset containing 200× images when using the concatenation of Vgg16 and Resnet50 models. The confusion matrix presents an overview of the classification outcomes and assists us in creating a solid data visualisation because it reveals not only the amount but also the type of errors produced by our method. In Figure 7, the true class is located on rows and the predicted class is situated along the columns. Class 1 represents a benign tumor (NO) and class 2 denotes a malignant cancer (YES). By observing Figure 7, we can summarize the values of the following important indicators given in Eq. (4):

$$\begin{aligned} \text{true negatives} &= TN = 164 \\ \text{false positives} &= FP = 28 \\ \text{false negatives} &= FN = 31 \\ \text{true positives} &= TP = 409 \end{aligned} \quad (4).$$

Figure 7. Confusion matrix obtained on one run on 200× images

Two important rates can be computed from the values in Eq. (4): False negative rate $=\frac{FN}{FN+TP}=7\%$ and False positive rate $=\frac{FP}{FP+TN}=14.6\%$. The false-positive rate is the percentage of people who are mistakenly diagnosed as sick but are healthy, while the false-negative rate is the percentage of people who are mistakenly diagnosed as healthy but are sick. Of course, both rates should be as low as possible, but in a medical classification problem, a low false-negative rate is critical for preventing the incorrect identification of a patient as healthy when he has in fact cancer. Additional medical checks may be required if a healthy individual is diagnosed as being sick. However, it is possible that if a sick individual is labeled as healthy, extra testing may not be performed. In this case, we can see that the false-negative rate is relatively small (7%), which is desirable in a computer-aided diagnostic system.

Other figures derived from the sensitivity and specificity analysis are 93% and 85.4% respectively. Sensitivity refers to the capacity to accurately identify individuals who have the disease, whereas specificity refers to the ability to correctly identify those who do not have the condition.

IV. CONCLUSIONS

This paper is focused on the classification of medical images using different pre-trained CNN models. We evaluate them on histopathological images of benign and malignant breast tumors. We considered images captured at four magnification factors and in all situations, we achieved a better performance than other handcrafted machine-learning methods and CNN approaches for the same dataset. The considered strategy also provides time efficiency.

As future work, we plan to also evaluate the presented techniques in the classification of each type of cancer (adenosis, fibroadenoma, phyllodes tumor, tubular adenoma, ductal carcinoma, lobular carcinoma, mucinous carcinoma, and papillary carcinoma) and we also intend to perform a clinical validation to assess the performance of our approaches.

REFERENCES

[1] “Breast cancer now most common form of cancer: WHO taking action.” https://www.who.int/news/item/03-02-2021-breast-cancer-now-most-common-form-of-cancer-who-taking-action (accessed Feb. 28, 2022).
[2] “Breast cancer.” https://www.who.int/news-room/fact-sheets/detail/breast-cancer (accessed Feb. 28, 2022).
[3] S. Iranmakani et al., “A review of various modalities in breast imaging: technical aspects and clinical outcomes,” Egypt. J. Radiol. Nucl. Med., vol. 51, no. 1, p. 57, Dec. 2020, doi: 10.1186/s43055-020-00175-5.
[4] Y.-J. Zhang, L. Wei, J. Li, Y.-Q. Zheng, and X.-R. Li, “Status quo and development trend of breast biopsy technology,” Gland Surg., vol. 2, no. 1, pp. 15–24, Feb. 2013, doi: 10.3978/j.issn.2227-684X.2013.02.01.
[5] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, “Return of the Devil in the Details: Delving Deep into Convolutional Nets,” ArXiv14053531 Cs, Nov. 2014, Accessed: Mar. 01, 2022. [Online]. Available: http://arxiv.org/abs/1405.3531
[6] M. Cimpoi, S. Maji, and A. Vedaldi, “Deep filter banks for texture recognition and segmentation,” in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2015, pp. 3828–3836. doi: 10.1109/CVPR.2015.7299007.
[7] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, “Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks,” in 2014 IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2014, pp. 1717–1724. doi: 10.1109/CVPR.2014.222.
[8] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, “CNN Features off-the-shelf: an Astounding Baseline for Recognition,” ArXiv14036382 Cs, May 2014, Accessed: Mar. 01, 2022. [Online]. Available: http://arxiv.org/abs/1403.6382
[9] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” ArXiv14091556 Cs, Apr. 2015, Accessed: Mar. 01, 2022. [Online]. Available: http://arxiv.org/abs/1409.1556
[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, Red Hook, NY, USA, Dec. 2012, pp. 1097–1105.
[11] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.
[12] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, “A Dataset for Breast Cancer Histopathological Image Classification,” IEEE Trans. Biomed. Eng., vol. 63, no. 7, pp. 1455–1462, Jul. 2016, doi: 10.1109/TBME.2015.2496264.
[13] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, “Breast cancer histopathological image classification using Convolutional Neural Networks,” in 2016 International Joint Conference on Neural Networks (IJCNN), Jul. 2016, pp. 2560–2567. doi: 10.1109/IJCNN.2016.7727519.
[14] T. Araújo et al., “Classification of breast cancer histology images using Convolutional Neural Networks,” PLOS ONE, vol. 12, no. 6, p. e0177544, Jun. 2017, doi: 10.1371/journal.pone.0177544.
[15] A. D. Belsare, M. M. Mushrif, M. A. Pangarkar, and N. Meshram, “Classification of breast cancer histopathology images using texture feature analysis,” in TENCON 2015 - 2015 IEEE Region 10 Conference, Nov. 2015, pp. 1–5. doi: 10.1109/TENCON.2015.7372809.
[16] A. Cruz-Roa et al., “Automatic detection of invasive ductal carcinoma in whole slide images with Convolutional Neural Networks,” Prog. Biomed. Opt. Imaging - Proc. SPIE, vol. 9041, Feb. 2014, doi: 10.1117/12.2043872.
[17] O. Russakovsky et al., “ImageNet Large Scale Visual Recognition Challenge,” Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/s11263-015-0816-y.