Final Project: Breast Cancer Detection in Histopathological Images using various Convolutional Neural Networks

Jasmine Batra
Department of Computer Engineering
New York University
New York, NY 11201
jb7854@nyu.edu

Sachin Karthik Kumar
Department of Computer Engineering
New York University
New York, NY 11201
sk9622@nyu.edu

Rudrasinh Ravalji
Department of Computer Engineering
New York University
New York, NY 11201
rnr8237@nyu.edu

Abstract

Cancer is one of the world's most frequent and deadly diseases. One in every eight women and one in every eight hundred males is diagnosed with breast cancer. As a result, our first goal should be early cancer detection, as early detection can aid in the effective treatment of cancer. As a result, we present a saliency detection method based on sophisticated deep learning techniques, in which the machine is taught to mimic pathologists' movements for the localization of diagnostically relevant regions. We use a CNN to train to identify diagnostic types of breast cancer (ResNet architectures). To train our model, we used the BreakHis dataset. In histopathology imaging, we focus on both detecting and classifying malignant areas. The 'salient' areas are those that are diagnostically important. Pathologists and medical institutions will be able to use the detection technology as an open-source online application.

https://github.com/rudra717/Breast-Cancer-Detection-in-Histopathological-Images-using-various-Convolutional-Neural-Networks

1 Introduction

Cancer is one of the most frequent diseases that a person can contract, and it is also the second leading cause of mortality. Cancer cells divide and spread in the surrounding tissues in any part of the body and in all of its forms. Breast cancer is one of the most serious and prevalent cancers in women, and it can be caused by a variety of factors including lifestyle, screening, and family history [1]. Breast cancer affects one in every eight women and one in every 800 males, according to statistics. Breast exams, mammograms, ultrasounds, collecting a sample of breast tissue (biopsy), and breast magnetic resonance imaging are just a few of the procedures to test and diagnose breast cancer. When compared to high-income countries, the number of cases of breast cancer is increasing by the day in low and middle-income countries. One of the major reasons for this is the high cost of getting yourself diagnosed, as well as the substandard treatment provided in low and middle-income countries. Many of the resources needed for mammography diagnosis are not available in low- and middle-income countries [2]. Due to a lack of resources, the death rate from breast cancer has not decreased with diagnosis from clinical breast examination or self-examination. Even though we live in a time where digitization is prevalent in the field of diagnosis, pathology remains largely reliant on investigations conducted using microscopic examinations of tissues on glass slides. The advancement of machine vision and digitally scanning microscopes has enhanced the possibility of computer-assisted diagnosis in recent years, allowing slides of tissue histology to be recorded as

digital images. As a result, image analysis and machine learning approaches have benefited from digital tissue histology.

2 Background Study and Related Works

There hasn't been much research done in this area previously. Though there have been a lot of new approaches recently. This is a new field with a lot of room for advancement. As a result, our effort intends to make a substantial contribution to the discipline.

Breast Cancer Diagnosis Using Deep Learning Algorithm [1] is one of the few publications that uses the Wisconsin Breast Cancer Database. The database in question has 569 rows and 30 characteristics. For scaled datasets, they used pre-processing methods such as normalizers and label encoders. In 2018, their work was published. Breast Cancer Diagnosis from Histopathological Image based on Deep Learning is another work that leverages the BreakHis dataset. CNNs, or convolutional neural networks, were utilized to identify images as benign or cancerous. However, they employed the Inception V3 model, which has a poor initialization rate. A lot of computation time is wasted and adjusting is costly.

In the year 2019, such material was published. Nuclear Atypia Grading in Histopathological Images of Breast Cancer [2] is another paper. CNN-based approach was utilized with Convolutional Neural Networks. They graded mitotic cells, nuclear atypia, and tubule development using Nottingham's technique, which has three elements. Their model is made up of two parts: a feature extraction section with convolutional layers and an activation function section with ReLu and pooling layers.

In 2018, the same was published. Another approach that leverages the BreakHis dataset is Histopathological Image Analysis for Breast Cancer Detection Using Cubic SVM [3].

They examined six SVM variations, including support vector machine, random forest technique, and KNN (K Nearest Neighbors), for experimental purposes. They discovered that cubic SVM outperformed all of the other approaches. When the class size is big, however, SVM struggles to predict class labels. In the year 2020, this work was released.

Another technique that employed BreakHis dataset was A Deep CNN Technique for Detection of Breast Cancer Using Histopathology Images[4].They also employ CNN, or convolutional neural networks, to differentiate between cancerous and benign images. Breast cancer is the most prevalent cancer diagnosed in women in the United States (excluding skin cancers), accounting for 30% of all new cancer diagnoses in women. Histopathology tries to differentiate between normal tissue, benign (benign) and malignant (carcinomas) abnormalities, as well as perform a prognosis evaluation.

In 2018, the same was published. Another approach that leverages the BreakHis dataset is Histopathological Image Analysis for Breast Cancer Detection Using Cubic SVM. They examined six SVM variations, including support vector machine, random forest technique, and KNN (K Nearest Neighbors), for experimental purposes. They discovered that cubic SVM outperformed all of the other approaches. When the class size is big, however, SVM struggles to predict class labels.

In another related work, Yusuf et al. suggested an automated detection of IDC method using deep transfer learner strategies with two CNN pre-trained models including ResNet-50 and DenseNet-161 on 277,524 image patches from the public histopathology dataset (Celik et al., 2020). Here DenseNet-161 and ResNet-50 achieved an accuracy level of 91.57% and 90.96%respectively, with only the last layers of the models utilized for training [6].

The pathologist used to undertake morphological assessment and tumor grading visually, but this method is time-consuming and subjective, resulting in inter-observer variances even among experienced pathologists. Because the use of morphological criteria in visual classification is subjective, computer-aided diagnosis (CAD) systems are used to improve diagnosis accuracy, reduce human error, boost inter-observer agreement, and improve reproducibility. For digital pathology picture analysis, a variety of methods have been developed, ranging from rule-based to machine learning applications [11].

Deep learning-based systems have recently been proven to outperform traditional machine learning methods in a variety of image analysis tasks, automating the entire process. Convolutional neural networks (CNN) have been successfully applied in the realm of medical imaging for diabetic retinopathy screening, bone disease prediction, age assessment, and other challenges. Previous deep learning-based applications in histological microscopic image analysis have shown their potential to aid in the diagnosis of breast cancer.

3 Methodology

This section illustrates how breast images classify by integrating different convolution neural networks. Figure 1. demonstrates proposed framework used in this classification. This framework starts with extracting images and loading labels from the dataset. After splitting the dataset, several Data augmentation techniques are performed. Finally, we train the model individual on Break-His dataset and evaluate the proposed Framework by doing validation on the test set. Now that the network is trained, we will save the model so we can load it later for making predictions. We will save the mapping of the classes to indices for which we can get from one of the image datasets: image_datasets['train'].class_to_idx. We will attach this to the model as an attribute which makes inference easier later. We even included the checkpoints that will include all the information which we can use it later. We have used predict function to use trained network for inference that take image and model then return the top K most likely classes along with the probabilities. After This method returns both the highest k probabilities and the indices of those probabilities corresponding to the classes. We need to convert from these indices to actual class labels using class_to_idx which we added to Image Folder. This method takes a path an image and a model checkpoint then return the probabilities and classes. Using matplotlib, we plot the probabilities for the top 5 classes as a bar graph along with input image . The details of blocks related to the proposed framework are given in the following subsections.

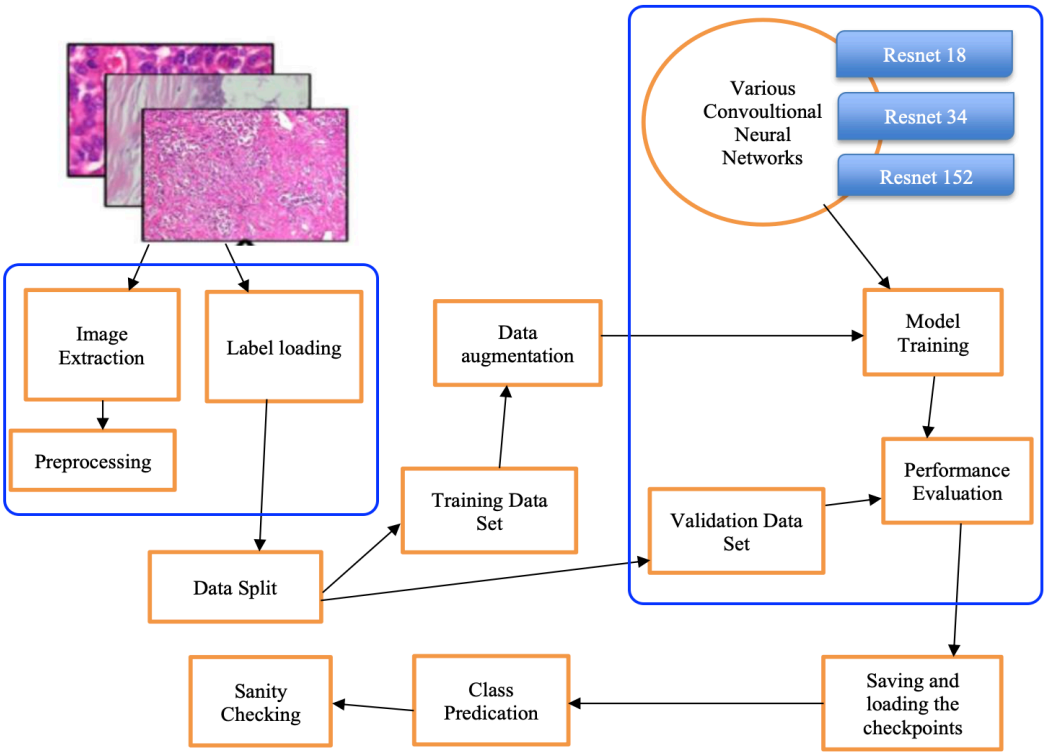

Figure 1: Proposed framework for the classification of breast images

3.1 Dataset

There are a variety of histopathological imaging data sets that may be used with deep learning approaches to detect tumor tissue. We used the Break His dataset, which consists of 4052 microscopic images of breast tumor tissue from 82 individuals using various magnification factors (40X,100X,200X,400X). It is divided into two types: benign and malignant. Tumor tissue is missing in benign tumors, but tumor tissue is present in malignant tumors. Figure 2 shows the structure of the BreaKHis Dataset. There are now 2,042 benign and 2010 malignant samples in the database (700X460 pixels, 3-channel RGB, 8-bit depth in each channel, PNG format). This database was created in partnership with Parana, Brazil's P&D Laboratory - Pathological Anatomy and Cytopathology.

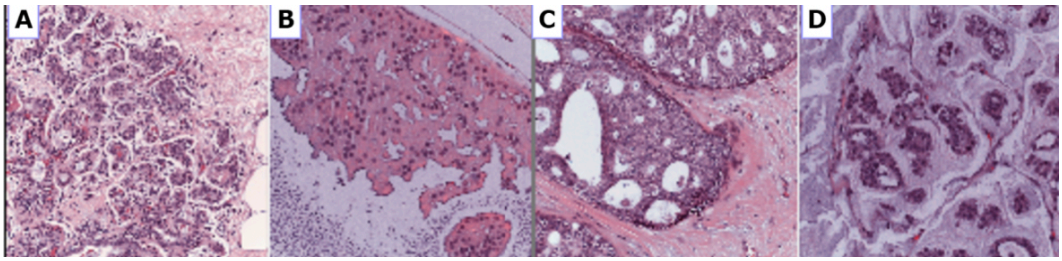

Figure 2: Example of images available in Break His Dataset.

3.2 Data Preprocessing stage

Several preprocessing steps have been employed before feeding the images into ResNet modules. All the microscopic images in the Break His dataset are in.png format with three channel RGB where each channel is in 8-bit depth. The images have 700 X 460 pixels, we convert all the images into a NumPy array so that the model can train faster and takes less space into the memory. After that, we have shuffled the images so that the model can train on some unordered data. This dataset is divided into two steps, including training, and validation, respectively. In this work, 80% of the data used for training, and the further 20% of the data used for validation and testing purposes. Several data augmentation techniques perform for different reasons, solving the overfitting issue, and making the model more robust.

3.3 Feature Learning Model

Feature learning (FL) is a set of techniques that allows a system to learn the representations needed for feature detection, prediction, or classification from a preprocessed dataset automatically. This enables a computer to learn the features and apply them to tasks like classification and prediction. In deep learning, feature learning may be performed by either building a complete convolutional neural network (CNN) to train and test a set of images or adapting a pretrained CNN in the classification/prediction for a new image set [7].

This module oversees feature extraction and learning in order to complete a specific picture classification task. However, with the transfer learning technique, there is no need to repeat the lengthy training process for a new classification task; instead, any well-known deep neural network that has been pre-trained on a specific dataset can be used to perform new classification tasks, with the learning parameters fine-tuned.

Transfer learning is the latter strategy. Both approaches' concepts are demonstrated in. With transfer learning, you use the convolutional basis and only retrain the classifier to your dataset, as shown in the Figure 3.

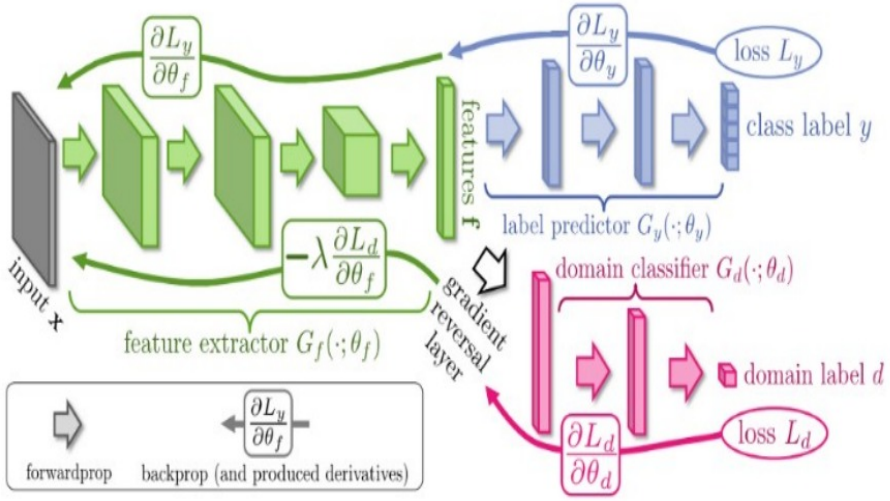

Figure 3: Illustration of no transfer learning vs transfer learning CNN.

In this module, we used the transfer learning technique to finetune the network parameters and hyperparameters of the ResNet-18, ResNet-34, and ResNet-152 CNNs. This is performed by preprocessing the obtained dataset (histopathologic pictures) and generating the models with

pretrained parameters (weights) from ImageNet Dataset [8].

3.4 Data Classification Modeling

Data categorization is an important feature for separating big datasets into classes for decision-making, pattern identification, and other purposes [9]. For multi-class classification problems with mutually exclusive classes, a classification layer uses a fully linked layer to compute the crossentropy loss. LogSoftMax is probability distribution function that assigns numerical probabilities to each class in a multi-class problem.

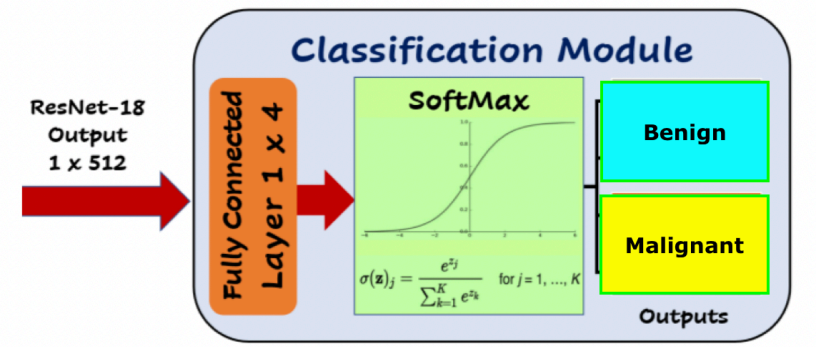

Figure 4 Classification module of ResNet 18.

3.5 Training and parameter optimization

For the Break His Dataset, Figure reflects the simulation result during the training the proposed Framework. One of the key hyperparameters to train a model is the optimizer function and the gradient loss function. We implemented the Resnet 18, 34 and 152 with SGD and ADAM combines the key properties of Adagrad and RMSProp optimizer that can handle sparse gradient on a large dataset. The hyperparameters that we used for the training the model is presented in the Table 1.

Table 1: Accuracies on different optimizers and different learning rate for Resnet 18.

Optimizer	Learning Rate	Accuracy
Adam	0.01	89.66
SGD	0.01	90.465
Adam	0.001	96.474
SGD	0.001	96.955

After training only 10th epoch, the model managed to achieve more than 95% and 96.9% validation accuracy by using Learning Rate as 0.001 and SGD as the optimizer.

Table 2: Accuracies on different optimizers and different learning rate for Resnet 34

Optimizer	Learning Rate	Accuracy
Adam	0.01	96.38
SGD	0.01	97
Adam	0.001	94.792
SGD	0.001	97.67

After training only for 10th epoch, the model managed to achieve more than 96% and 96.67 % validation accuracy by using Learning Rate as 0.001 and SGD as the optimizer.

Table 3: Accuracies on different optimizers and different learning rate for Resnet 152

Optimizer	Learning Rate	Accuracy
Adam	0.01	76.23
SGD	0.01	86.45
Adam	0.001	79.52
SGD	0.001	87.33

After training only for 50th epochs, the model managed to achieve more than 89% training accuracy and 87% validation accuracy by choosing Learning Rate 0.01 and SGD as Optimizer.

We chose negative log likelihood as a loss function since our work is based on binary classification

in the Break His Dataset. To decrease the loss function we need optimal learning rate value. However, choosing the learning rate is the challenging since it produces undesirable behavior is quite large. If the learning rate is small, the model takes more time for training and makes tiny updates to the weights. In this work our models work best on 0.001 learning rate which limit such issues. Furthermore, a small batch size of 32 has been used which shows a suitable generalization of the model.

Table 4: Parameter value used during Training.

Resnet Model	Parameter	Value
18	Optimizer	SGD
	Learning rate	0.001
	Loss Function	Negative log likelihood
	Epochs	10
34	Optimizer	SGD
	Learning Rate	0.001
	Loss Function	Negative log likelihood
	Epochs	10
152	Optimizer	Adam
	Learning Rate	0.01
	Loss Function	Cross Entropy Loss
	Epochs	50

4 Results

The plot for loss function comparing the behaviour of training loss and validation loss the plot for loss function comparing the behavior of training loss and testing loss obtained during the training process is presented in Figure 5b, Figure 6b Figure 7b. It can be clearly seen, both losses are systematically decreasing. The plot for accuracy metric comparing the performance of training accuracy and validation accuracy obtained during the training process is given in Figure 5a, Figure 6a Figure 7a.

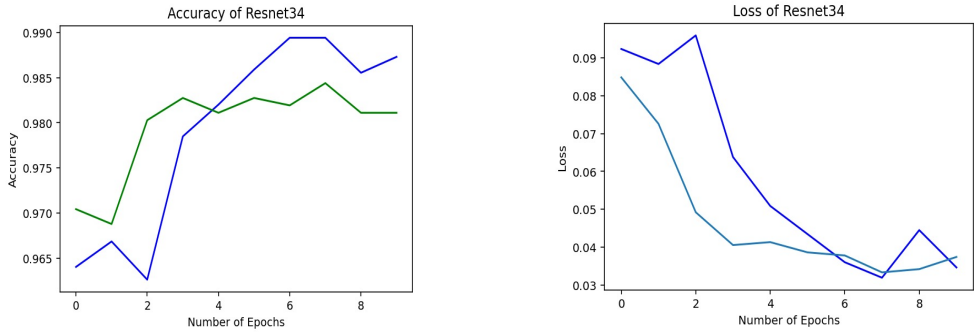

Figure 5: a) Training and validation accuracy and b) training and validation loss

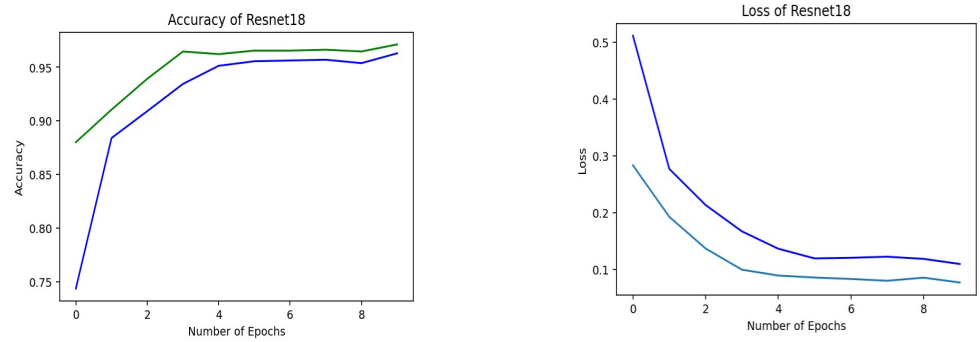

Figure 6: a) Training and validation accuracy and b) training and validation loss

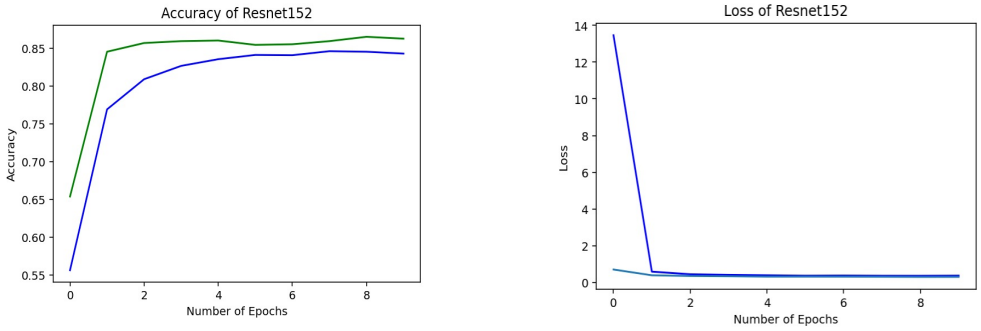

Figure 7: a) Training and validation accuracy and b) training and validation loss

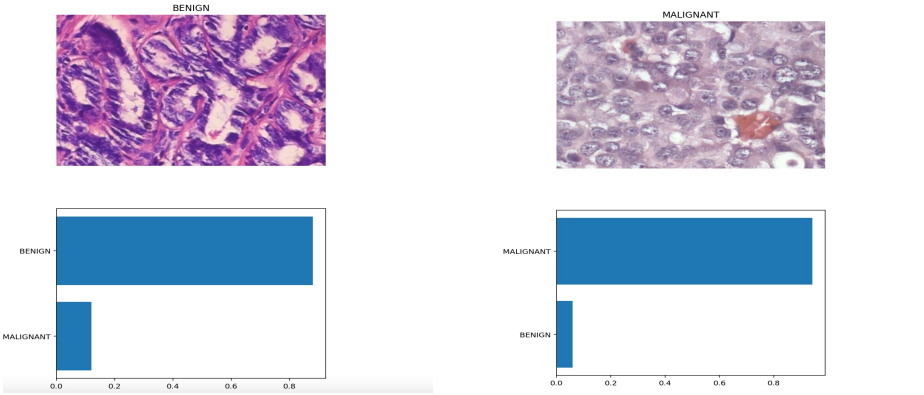

Figure 8: Prediction images of Resnet 152

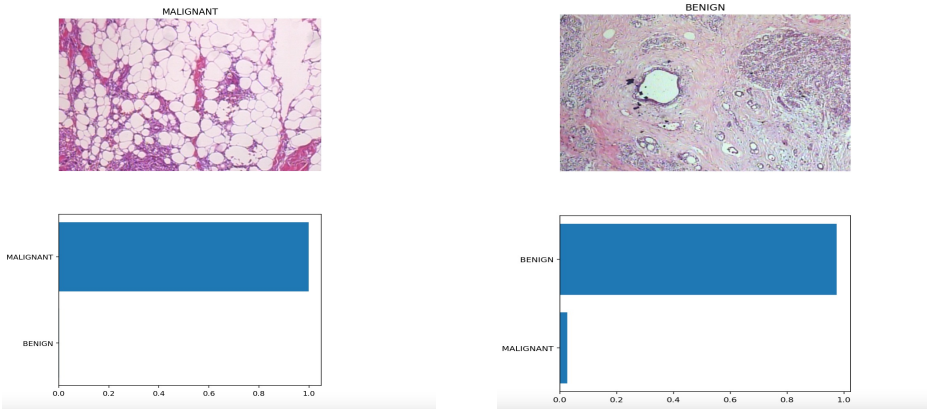

Figure 9: Prediction of images of Resnet 34

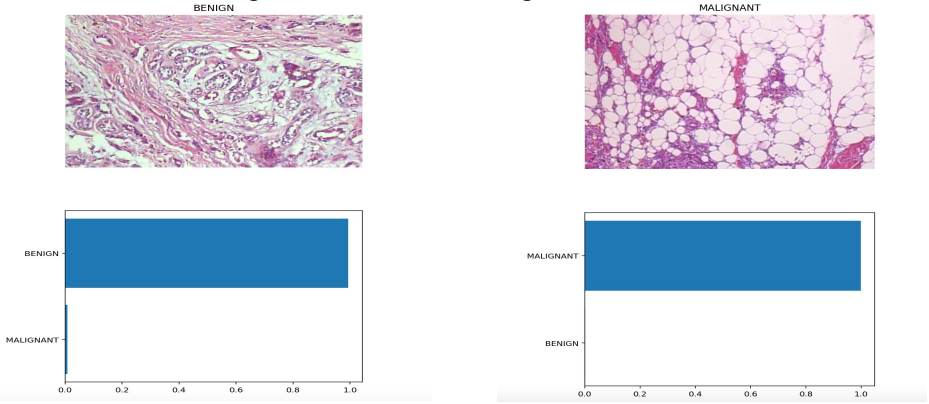

Figure 10: Prediction of images of Resnet 18

5 Conclusion

In this paper, we present a Resnet 18, Resnet34, Resnet 152 framework for more accurately and reliably detecting breast cancer patients from microscope pictures. The suggested framework is based on a feature extraction and transfer learning technique that allows several pre-trained CNN models to extract features independently and then combine them for the classification job. The model has been trained and tested on a variety of image datasets, both small and large. It is built in such a way that it can operate efficiently on a wide range of datasets. The experimental results show that our framework for different Resnet models achieves percent and 98 percent classification accuracy in the BreakHis and respectively, beating both individual CNN pretrained architectures and all other state-of-the-art models discovered in the literature. Furthermore, it performs well in recognizing cancerous pictures, boosting the likelihood of survival. Based on these positive results, we feel that our model would be an excellent option for assisting doctors in quickly diagnosing and detecting breast cancer.

References

[1] N. Khuriwal and N. Mishra, "Breast Cancer Diagnosis Using Deep Learning Algorithm," 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2018, pp. 98-103.

[2] S. K. Jafarbiglo, H. Danyali and M. S. Helfroush, "Nuclear Atypia Grading in Histopathological Images of Breast Cancer Using Convolutional Neural Networks," 2018 4th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS), 2018, pp. 89-93.

[3] S. Singh and R. Kumar, "Histopathological Image Analysis for Breast Cancer Detection Using Cubic SVM," 2020 7th International Conference on Signal Processin

[4] G. Wadhwa and A. Kaur, "A Deep CNN Technique for Detection of Breast Cancer Using Histopathology Images," 2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA), 2020, pp. 179-185[7]

[5] 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.

[6] Yusuf Çelik, Muhammed Talo, Özal Yildirim, Murat Karabatak, U. Rajendra Acharya. Automated invasive ductal carcinoma detection based using deep transfer learning with whole-slide images. Pattern Recognition Letters, 133:232-239, 2020.

[7] Y. Bengio; A. Courville; P. Vincent. Representation Learning: A Review and New Perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence. 35 (8): 1798–1828, 2013. arXiv:1206.5538. doi:10.1109/tpami.2013.50. PMID 23787338.

[8] K. He, X. Zhang, S. Ren and J. Sun . Deep Residual Learning for ImageRecognition. arXiv:1512.03385v1, 2015.

[9] O. Knocklein. Classification Using Neural Networks. Medium: Toward Data Science. Jun. 2019

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in Advances in Neural Information Processing Systems 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[11] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. J. Mach. Learn. Res., 15(1):1929–1958, Jan. 2014.