SELF-SUPERVISED LEARNING FOR BREAST CANCER IMAGE ANALYSIS

Gitika Jha[1,a] , Manashree Jhawar[2,a], Vedant Manelkar[3,a], Radhika Kotecha[4,a], Ashish Phophalia[5,b]

[a] *K. J. Somaiya Institute of Engineering and Information Technology, India*

[b] *Indian Institute of Information Technology Vadodara, India*

[1]gitika.jha@somaiya.edu

[2]manashree.j@somaiya.edu

[3]vedant.manelkar@somaiya.edu

[4]radhika.kotecha@somaiya.edu

[5]ashish_p@iiitvadodara.ac.in

Abstract

Breast cancer, being one of the most widespread terminal diseases, has the potential to claim the lives of thousands of people each year. Machine Learning approaches have shown promise in the biomedical field because of recent advances. However, real-time medical data can be hard to obtain, and the number of samples might be small or unannotated for traditional machine learning approaches to make accurate predictions. The proposed methodology uses Self-Supervised learning to overcome this hurdle. Since this model works with unlabeled data and can be downstream from any standard dataset, the proposed approach does not need large labeled datasets. The open-source datasets, Digital Database for Screening Mammography (DDSM), The Mammographic Image Analysis Society (MIAS), and INbreast are taken, and samples are arbitrarily selected from them to make a smaller randomized dataset. An accuracy of 96.7% is achieved using BYOL as a pretext task on Resnet18 architecture. This paper demonstrates that the proposed algorithm surpasses the current state-of-the-art supervised learning methods without compromising on accuracy. This approach has the potential to avoid misdiagnosis and assist medical practitioners, in turn, saving innumerable precious lives.

Keywords: Self-supervised Learning, Breast Cancer Detection, Medical Image Analysis, Artificial Intelligence, Deep Learning

1. Introduction

One of the top reasons for unnatural deaths in every country of the world is cancer, as stated in a report by the World Health Organisation (WHO) [1]. Breast cancer is the second most commonly diagnosed cancer (11.7% of total cases) worldwide amongst women [2]. According to WHO, cancer can be cured in 30-50% of the patients if detected early [3]. The current strategy for combating this disease entails early detection and treatment. The 10-year survival rate for patients with stages 0 and I cancer is 98 percent, compared to a 65 percent 10-year survival rate for patients with stage III disease [4]. More patients must be identified at an early stage in order to improve the survival rate of this disease. Moreover, accurate categorization of benign lesions can save patients from having to undergo needless medications.

During their life, 8% of women are diagnosed with Breast cancer (BC). Following lung cancer, Breast Cancer is the second most common cause of death in developed and undeveloped worlds. It is characterized by the mutation of genes, constant pain, changes in the size, color (redness), and skin texture of breasts. Statistics demonstrate that 1 in every 8 women in the US will be diagnosed with breast cancer in their lifetime. It also accounts for 14% of cancers in Indian women. It is caused by the abnormal growth of cells in the breast [5]. The type of cell determines what type of cancer it will be. The cells can begin to grow in many different parts of the organ like the lobules, connective tissue, and duct. It metastasizes from one part to the other via blood and lymph vessels. There are various types of breast cancers like invasive ductal carcinoma, invasive lobular carcinoma, and Paget's disease.

Symptoms of breast cancer may or may not be present [6]. There are various unchangeable risk factors like - genetic mutations, aging, consumption of certain drugs like diethylstilbestrol (DES). However, risk factors can be changed for the better, like not taking external hormones, being physically active, not consuming alcohol, and others. Regular checkups are recommended by gynecologists, especially for people with genetic history. There is a much higher risk of breast and ovarian cancer for people with inherited changes in BRCA1 and BRCA2 genes. One of the ways of confirming the presence of breast cancer is via

screenings. Although screenings cannot prevent breast cancer, they can be helpful in its early detection. There are two types of screenings - Mammograms and MRIs (Magnetic Resonance Imaging). There are various treatments for breast cancers that are done with doctors' consultation like double mastectomy, surgery, chemotherapy, hormonal therapy, biological therapy, radiation therapy. Though breast cancer is commonly detected in females, 1% of the total patients are males.

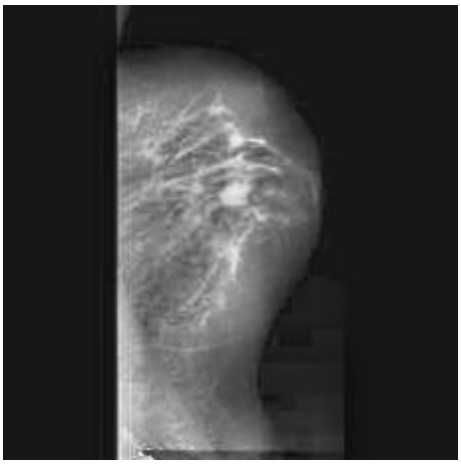

Fig.1. Benign mammogram from the dataset

Fig.2. Malignant Mammogram from the dataset

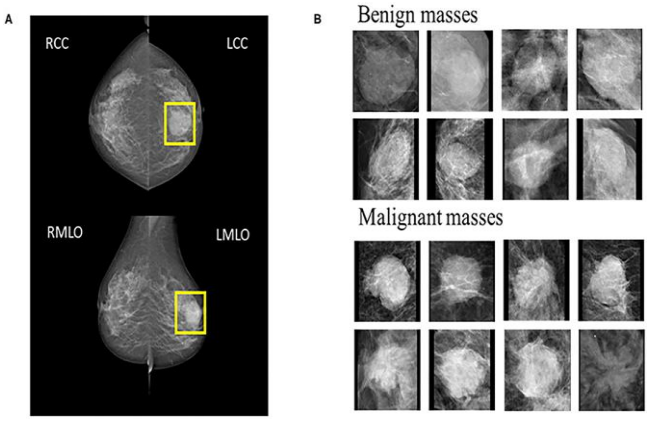

Fig.3. Benign masses vs Malignant masses in mammograms [7]

Classification of breast cancer leads pathologists to find a systematic and objective prognostic; generally, the most frequent classification is binary (benign tumor/malign tumor). Today, Machine Learning (ML) techniques are being broadly used in the breast cancer classification problem. They provide high classification accuracy and effective diagnostic capabilities. The most efficient way for diagnosing breast cancer is medical imaging examination. Digital Mammography, Magnetic Resonance Imaging (MRI), Microscopic (histological) pictures, Ultrasound, and Infrared Thermography (IRT) are some of the medical imaging modalities employed for diagnosis. Mammograms are X-Rays of the breast, and they are best for early detection, unlike MRIs and other screening techniques, which are recommended for high-risk patients. These modalities create images that have reduced mortality rates by 30–70% as a means of assisting radiologists and clinicians in recognizing abnormalities [8]. However, because image interpretation is operator-dependent and requires skill, IT is needed to speed up and improve the accuracy of diagnosis while also providing a second opinion to the experts. Further, as diagnoses are based on a doctor's subjective judgment, the efficiency of diagnosis varies, giving rise to an increased chance of misdiagnosis. To counter this, Machine Learning and Deep Learning techniques have rapidly risen in biomedical applications to assist medical personnel on tasks like such. The use of classification algorithms and automated feature extraction, called Computer-Aided Diagnosis (CAD), can be a handy tool for doctors and experts in spotting anomalies.

Mammography-based Computer-Aided Diagnosis (CAD) has proven to be successful in terms of reliability. Convolution neural networks (CNNs) have demonstrated their potency in the field of mammography-based CAD as a traditional Deep Learning method. However, even so, collecting and annotating a large number of mammograms can indeed be costly and time-consuming. As a result, the limited training samples are insufficient to train a reliable and robust model. As a result, improving the diagnostic accuracy of a mammography-based CAD with a small dataset remains a difficult task.

In order to create a good model for detecting Breast Tumours accurately, a sufficient number of samples are required for training. High-quality annotated medical data are not readily available at times. Moreover, the datasets might be smaller, or unlabeled data might outnumber the labeled data. Supervised learning requires manual data labeling, which is a tedious task and also error-prone; this challenge is solved by self-supervision. Self-Supervised learning is a process in which the model uses one part of the data to make accurate predictions about the other part and make the labels accordingly. It is instrumental when the datasets are large and thus becomes a faster approach.

The recent advances in Machine Learning algorithms like Deep Learning, Random Forest, and others have put light on its applications in medical imaging. Such methods make it promising to apply advanced techniques in medical applications and surpass the current state-of-the-art schemes.

Supervised learning requires manual data labeling, which is a tedious task and also error-prone; this challenge is solved by self-supervision. Self-Supervised learning is a process in which the model uses one part of the data to make accurate predictions about the other part and make the labels accordingly. It is instrumental when the datasets are large and thus becomes a faster approach. The recent advances in Machine Learning algorithms like Deep Learning, Random Forest, and others have put light on its applications in medical imaging. Such methods make it promising to apply advanced techniques in medical applications and surpass the current state-of-the-art schemes. Moreover, cancer-related mortality is increasing significantly all around the world. The proposed method is incorporated into the process of diagnosing patients and acts as an assistant to the medical personnel. It has the potential to save a significant number of lives.

This paper aims to create a novel approach towards Image Analysis used in Breast Cancer Detection from Tumor Segmentation and to apply self-supervised learning as a method of analysis in which smaller medical datasets can be leveraged to find better accuracy by manipulating data instances to populate the dataset further. Several attempts have been made to correctly evaluate the accuracy of data classification by various algorithms in terms of effectiveness and efficiency to analyze the results, creating a tool that will help radiologists, doctors, and other researchers study Breast Cancer further and save lives.

1.1 Contributions of this paper

The major contributions of this paper are -

1. Proposition of a novel approach in analyzing scans that can help detect tumors and determine if they are malignant or benign.
2. Comprehensive analysis and comparison of the application of algorithms on different open source mammogram and ultrasound datasets like DDSM, MIAS, and INbreast.
3. Assisting radiologists and doctors in interpreting breast tumors in screening exams using Machine Learning.

The following is an overview of the structure of the paper: Section II gives a summary of the related works in the field in the form of a literature survey and the concept of Self-Supervised Learning. Section III explains the proposed approach and datasets used. Section IV explores the results and analyzes them. Finally, Section VI deals with the discussions, limitations, future scope, benefits to the society, and conclusions.

2. Related Works

2.1 Semi-Supervised Learning

A learning task with a small number of labelled instances and a large number of unlabeled examples is known as semi-supervised learning. This type of learning problem is difficult to solve since neither supervised nor unsupervised learning algorithms can effectively use a mixture of labelled and untellable data. As a result, semi-supervised learning methods that are specialised and effective are necessary. It refers to a learning challenge in which a model must learn and make predictions on fresh examples from a small number of labelled examples and a large number of unlabeled examples.

The ability of a semi-supervised learning algorithm to outperform a supervised learning algorithm based solely on labelled training instances is an indicator of its effectiveness. When working with data in which labelling instances is difficult or expensive, we need semi-supervised learning techniques. Furthermore, semi-supervised learning can be used to contrast inductive learning and transductive learning methods. Inductive learning is a learning algorithm that generalizes from labeled training data to new data, as a test dataset. Learning from labelled training data and generalizing to unlabeled (training) data is known as transductive learning. A semi-supervised learning method can handle both types of jobs. Integrating clustering and classification algorithms is a popular semi-supervised learning strategy. Clustering algorithms are a type of unsupervised learning tool that groups data based on similarities. These methods aid in the identification of the most important samples within a data set. The samples can then be labelled and utilised to train a classification problem using a supervised learning model.

2.1.1 Semi-Supervised Learning vs. Self-Supervised Learning:

The most striking similarity between semi-supervised and self-supervised strategies is that neither relies solely on data that has been manually labelled. The resemblance, however, ends here. In the self-supervised learning technique , the model relies on the underlying structure of data to anticipate outcomes . It does not include any data that has been labelled. However, we still offer a tiny quantity of labelled data in semi-supervised learning [9].

2.2 Self-Supervised Learning

Recently, machine learning has had several breakthroughs and continues to have an impact across a variety of disciplines. Data availability is a common thread that runs through all of these fields. Machine learning algorithms have reached or even outperformed human performance [10]. Due to the recent growth in the availability of clinically relevant datasets, researchers have applied machine learning techniques to a wide range of clinical activities, from identification/diagnostic tasks to prediction tasks. Mainly, there are three major categories of machine learning approaches: Supervised Learning, Unsupervised Learning, and Reinforcement Learning.

This paper explores a specific Machine Learning approach called Self-Supervised Learning. In the past, a significant portion of clinical data was overlooked (or not collected at all). This constraint stemmed from the data's bulk and complexity, as well as the lack of means for gathering and storing it. Medical sample collection is always a time-consuming procedure that necessitates a lot of overhead and time from skilled personnel for labeling purposes. Self-supervision allows to utilization of a fraction of the labeled samples required for deep learning classification tasks. Due to the expense of gathering medical data and labeling them, datasets for identifying cancer with mammography images are often tiny and potentially non-representative, unlike in conventional object identification, where big diversified datasets can be leveraged. This makes training and evaluating such models difficult, raising concerns about their reliability and generalizability.

Self-supervised learning improves the performance of machine learning models by utilizing both annotated and unannotated data. This approach does not use information already known or labeled beforehand, but, instead, learns from scratch. This enables the model to uncover different aspects of the medical scans that were not apparent to the classifier beforehand.

2.2.1 SimCLR:

To make the best from a smaller set of unlabeled data, unsupervised pretraining is followed by supervised fine-tuning. In contrast to popular techniques, this paradigm employs unlabeled data in a task-agnostic manner. By employing this method using task-specific method, the huge network can be refined and distilled into a much smaller network without any loss in classification accuracy after fine-tuning [11].

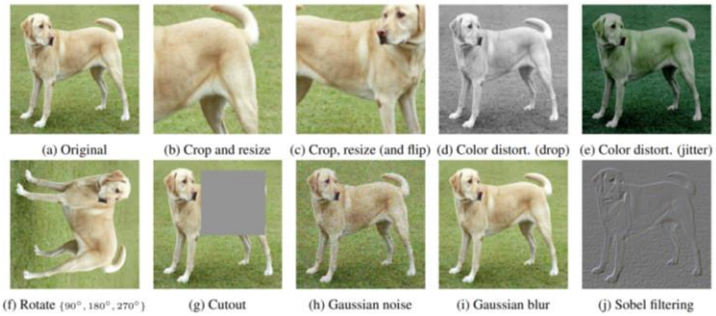

Fig.4. Some types of Image augmentations. Taken from [11]

2.2.2 Drawbacks of SimCLR:

The choice of image augmentations has an impact on contrastive approaches. SimCLR struggles to remove color distortion from its image augmentations. It indicates why the color histograms of crops from the same image are almost always the same. Color histograms fluctuate among images. As a result, when a contrastive task uses random cropping as image augmentations, it can usually be done by concentrating solely on color histograms. As a result, there is no reward for the representation to maintain information. Instead, BYOL stores any data gathered by the target representation in its online network in order to boost predictions. retains more characteristics in its representation even if augmented views of the same image share the same color histogram. As a result, BYOL is more resistant to image augmentation selections than contrastive approaches. This model is also less susceptible to systematic biases in the training data. This usually suggests that it is more generalizable to unobserved examples. In most benchmarks, BYOL outperforms SimCLR [12].

2.3 Existing Literature

The existing state-of-the-art in both traditional methods and self-supervised learning is described in this paper. Moreover, the paper presents the most recent developments in these fields in this section along with certain limitations faced by the current literature.

2.3.1 Traditional Computer-Aided Diagnosis (CAD) Systems:

Wang et al. [13] use a CNN-based Computer-Aided Diagnosis (CAD) for breast cancer classification. The paper uses Inception-v3 for effective feature extraction, with satisfactory results. The preferred approach gets an AUC value of 0.9468 after assessing 316 breast lesions, basing their verification process on five human reviewers' diagnoses. Compared to the standard feature extraction approaches like PCA (principal component analysis) or HOG (histogram of oriented gradients), the model achieves greater than 10 percent improved AUC value. In the future, the paper plans to compare their current work with specialized detection algorithms for breast lesion identification and classification to see whether an end-to-end CAD solution can be created.

Shen et al. [14] used the CBIS-DDSM dataset to Improve Breast Cancer Detection on Screening Mammography. In this method, Lesion annotations are only required during the initial training step, and later stages just require image-level labels, removing the dependency on scarce lesion annotations On an independent test set of digitised film mammograms from the Digital Database for Screening Mammography, the best single model had a per-image AUC of 0.88. (CBIS-DDSM), while four-model averaging increased the AUC to 0.91. However the drawback for this paper was that to develop this algorithm, the mammograms were downsized to work on the available GPU.

Ribli et al. [15] have used deep learning techniques to detect and classify lesions in mammograms. The approach method in this paper is Faster R-CNN. On top of the last convolutional layer of the original network, the faster R-CNN contains a branch of convolutional layers called Region Proposal Network (RPN), which is trained to detect and localize objects on the picture, regardless of their class. On the public INbreast database, the suggested technique achieves classification performance with an AUC of 0.95. With an AUC of 0.85, the method presented here

came in second place in the Digital Mammography DREAM Challenge. On the INbreast dataset, the system achieves great sensitivity with low false-positive marks per image when employed as a detector. This algorithm's detection performance was only tested on the small INbreast dataset, and this is a limitation of the proposed method.

Harvey et al. [16] discuss the role of deep learning in breast cancer screening. The approach shows how from the beginning when practitioners used CAD (computer-aided detection) models to what all challenges were faced and how better systems were eventually developed to train these datasets. It mainly discusses all the relevant material that is available on breast cancer screening and how it can be used for further research purposes. The paper also talks about a deep learning system made by the authors called Mia (Mammography Intelligent Assessment), which understands the models and acts as a second reader that gives out case-wise call-back decisions to patients.The paper also states that while the research on using 3-D Digital Breast Tomosynthesis (DBT) is going in full traction, it will require more time to interpret, which is a major reason why it is not as commonly used.

Geras et al. [17] in their paper discuss why the deep learning models that give high accuracy to natural images cannot be used for mammograms. In medical photos, precise details are required for detection, whereas in natural images, coarse structures are most important. Because of this discrepancy, existing network designs designed for natural images are insufficient because it functions on significantly downscaled images to save memory needs which conceals information needed to make good forecasts. Thus, the paper uses BI-RADS dataset to achieve a probabilistic accuracy of 0.688 whereas the radiologists achieved 0.704. One of the constraints was the fact that because of limited computing resources, any systematic survey for optimal hyperparameters could not be performed.

Muduli et al. [18] propose a deep CNN (Convolutional Neural Network) model for both mammograms and ultrasound images. The proposed approach tests their unique model with only five learnable layers on INbreast, DDSM, and MIAS mammogram datasets as well as BUS-1 and BUS-2 ultrasound datasets. Using fewer layers, the suggested approach succeeded in achieving lower computational cost , maximized learning speed, and improved accuracy compared to the pre-defined CNN models. The proposed model attains an accuracy of 96.55% on the MIAS, 90.68% on the DDSM , 91.28% on the INbreast, 100% on the BUS-1, and 89.73% on the BUS-2 dataset. The model also eliminates the need for manual feature extraction and eliminates feature reduction activities. As a result, the proposed approach is more time-efficient. The method is made up of five learnable layers: four convolutional layers and a fully linked layer.

Wu et al. [19] present a Deep CNN model based on over 1 million scans and achieve an average AUC value of 0.895 in predicting whether a breast tumor is malignant or not. The proposed approach verified the results from the novel two-stage network design by presenting radiologists with a portion of the screening mammogram exams. The hybrid approach also demonstrates that the average likelihood of malignancy by the neural network and radiologists, provides better accuracy than either of them independently.

Yassin et al. [20] surveyed various Machine Learning Techniques used in image modalities to detect Breast Cancer. The approach discovered that Digital Mammograms (DMs) had been used in the majority of the papers. Support Vector Machine (SVM) technique achieved an accuracy between the range 90% - 99.5% for DMs, with two achieving 100% accuracy. Artificial Neural Networks (ANNs) used for DMs gave 90 - 98.14% accuracy. K-nearest Neighbour (KNNs) registered the highest accuracy of 98.69% for DMs. According to the data that was gathered, it was difficult to compare approaches extensively due to a number of issues. The datasets used for evaluation, the image samples selected for evaluation, the number of samples used, and the assessment process employed are some of these considerations. The main limitation of the paper is that it is difficult to compare approaches fully due to a number of issues. Furthermore, the tuning of parameters used in different approaches differs from one method to the next, creating still another barrier to a meaningful comparison of diverse methods.

While these methods have been extremely beneficial in furthering the research in the field of breast cancer, there came a need to address the limitations of medical data availability as well as employing the approach of self-supervised learning in this use case.

2.3.2 The promise of Self-Supervised Learning:

Yin et al. [21]used the University of California, Irvine (UCI) machine learning database for finding Pattern Classification Techniques for Breast Cancer Detection using MRI scans. The review paper also discussed self-supervised and semi-supervised deep learning strategies as well as GANs for differentiating the tumor types. The approach proposed in the paper can help in the early identification of heterogeneous tumors. The suggested approach employs innovative loss functions that serve as the foundation for a produced confrontation learning methodology for tensorial DCE-MRI. Because some of the methods mentioned are based on time-lapse imaging, inferences about the disease's rate of progression are conceivable. This paper concludes DCE-MRI as one of the most effective modalities for breast cancer imaging considering all the various approaches used to detect it but also states that the specificity of detection is low. Thus, it proposes that deep convoluted neural networks should be used in order to make full use of these datasets.

Gong et al. [22] proposed a method of Task-Driven Self-Supervised Bi-Channel Networks framework to overcome the challenge of pretext task and fine-tuning mechanism. The pretext task is a new gray-scale image mapping task that incorporates the mammogram class label information into the restoration task in

order to enhance the discriminative feature representation. The proposed system then merges various network architectures, including the image retrieval network and the classifier, into a single SSL framework in a unique manner. It trains the bi-channel network models to transmit information from the pretext task network to the downstream task network with better precision. The results on the open-source INbreast dataset show that the proposed framework outperforms traditional fine-tuning-based SSL techniques. The pretext and downstream bi-channel networks are trained, and the learned feature representations in both networks are communicated to each other collaboratively. This SSL technique not only increases downstream job performance, but also solves the problem of limited samples. The suggested TSBN outperforms traditional fine-tuning-based SSL algorithms, according to experimental findings on the public INbreast dataset. The paper further plans to look into other ways to design different types of pretext tasks based on label information and medical data characteristics in the TSBN framework. The researchers also strive to increase the transfer performance across two separate networks in order to improve the downstream CAD model's classification accuracy.

Self-supervised learning for Breast Cancer Image analysis and detection has been attempted however several open issues remained.

3. Methodology

3.1 Proposed Approach

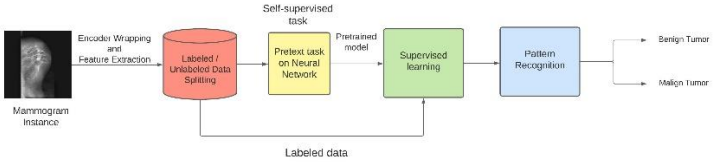

Fig. 5. Proposed Approach for Breast Cancer Image Analysis and Prediction

The proposed approach entails the following:

The necessary image augmentations are made first. A random patch of the images is selected and resized into a target size of 224 x 224 with a horizontal flip. Then, in a random order, the other image augmentation techniques are applied. Colour dropping is performed where the image is converted into grayscale. Gaussian blurring is done where for a 224 x 224 image, a 23 x 23 square Gaussian kernel is employed using a standard deviation uniformly sampled over [0.1, 2.0]. Colour jittering is also performed which includes shifting the saturation, contrast, brightness, and hue of the image by a random offset on all pixels.

To implement these transformations, Kornia in PyTorch is used, which is a differentiable CV library. The Encoder is responsible for feature extraction from the base model and projecting them into a latent space with a lower dimension. Feature extraction will collect outputs from one of the last model layers. This is implemented using hooks which are functions that execute automatically after a particular event.

The projector is another module that is in charge of projecting the outputs down the lower dimensions. For BYOL's training code, PyTorch Lightning is used which is a library for deep learning projects, which includes conveniences like multi-GPU training, experiment logging, model checkpointing, and mixed-precision training. BYOL class will define the target, updated_target, optimizers, training_step, epochs, validation steps, learning rate, and other parameters. Code for processing data samples can get messy and hard to maintain; ideally, the dataset code should be decoupled from the model training code for better readability and modularity. PyTorch provides two data primitives: torch.utils.data.DataLoader and torch.utils.data.The dataset that allow you to use preloaded datasets as well as your own data. Dataset stores the samples and their corresponding labels, and DataLoader wraps an iterable around the Dataset to enable easy access to the samples.

This is used to create a custom dataset and dataloader that can be easily iterable and can be used with ease with Pytorch lightning neural network modules. The dataset is divided into Train, test (for supervised learning), and unlabeled data (for BYOL). The model is passed after pretext class to do supervised learning and use resnet18's 50-200 epochs with a 1e-4 and 1e-6 learning rate, weight decay is used.

3.2 Bootstrap Your Own Latent (BYOL)

Bootstrap Your Own Latent (BYOL) is a self-supervised learning approach that is similar to contrastive learning with the exception that it does not worry about whether dissimilar samples have differing representations (which is the contrastive part of contrastive learning). The method merely ensures that similar samples have comparable representations. This difference may appear insignificant, but it has significant implications for training efficiency and generalization. Because BYOL does not require negative sampling, training is more effective. The negative equivalents can be ignored entirely because each training example is sampled just once in each epoch. Furthermore, the suggested method is less vulnerable to systematic biases in the training dataset, implying that it generalizes better to unobserved occurrences.

BYOL additionally closes the gap between each sample's representations and transformations. Translation, rotation, blurring, color inversion, color jitter, Gaussian noise, and other transformations are examples. Though this paper has used images as an example, BYOL can also be utilized with other forms of data. Models are often trained using a combination of transformations that can be used together or separately. If the model is expected to be invariant under a specific transformation, it should be incorporated in the training.

BYOL has two Encoder networks that are identical. The first one is trained as normal, with every training batch updating its weights. A continuous mean of the first Encoder's weights is used to update the second, also referred to as the "target" network. During training, a raw training batch is supplied to the target network, and a modified version of the very same batch is delivered to the other encoder. For its respective data, each network develops a low-dimensional, latent representation. Then, using a multi-layer perceptron, we try to anticipate the outcome of the target network. The correlation between this forecast and the output of the target network is maximized by BYOL.

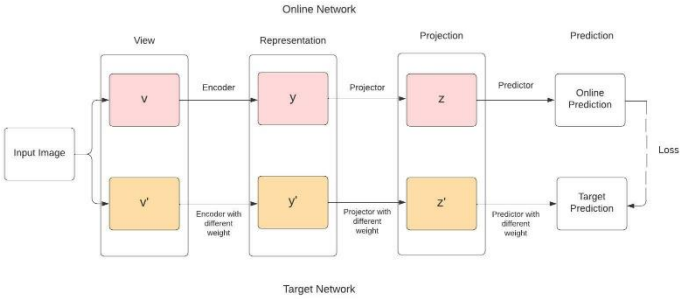

Fig.6. BYOL's Architecture

The multilayer perceptron layer, a feedforward artificial neural network that learns to identify input modifications and estimates the target latent vector, is a type of feedforward artificial neural network. As a result, weights no longer collapse to zero, allowing the model to learn self-consistent representations. This way the overall performance of the model can be improved using this self-supervised learning method.

3.3 Resnet

A series of advancements in the field of computer vision have occurred in recent years. Researchers are receiving state-of-the-art results on tasks like picture classification and image recognition, especially since the advent of deep Convolutional neural networks. As a result, in order to perform such complex tasks and enhance classification accuracy, researchers have tended to build deeper neural networks (adding more layers) over time. We usually stack some additional layers in Deep Neural Networks to address a complex problem, which improves accuracy and performance. The idea behind adding more layers is that these layers would learn increasingly complicated features as time goes on. For example, in the case of image recognition, the first layer might learn to detect edges, the second one might learn to identify textures, and the third layer might learn to detect objects, and so on. However, it has been discovered that the classic CNN model has a maximum depth threshold. This implies that a network's performance diminishes as additional layers are added on top of it. When more layers are added to the neural network, it gets more difficult to train them, and its accuracy begins to saturate and ultimately decline. Here, ResNet or Residual Network comes to the rescue and assists in the resolution of this issue [23]. The emergence of ResNet or residual networks, which are made up of Residual Blocks, has relieved the challenge of training very deep networks. The first thing that happens is that there is a direct connection that bypasses several levels (which may change depending on the model) in between. This is known as the 'skip connection,' and it is at the heart of residual blocks. ResNet's skip connections help the issue of vanishing gradient in deep neural networks by enabling the gradient to flow through an additional shortcut channel.

3.4 Data Sources and Sample Size

To get the best results, the 3 different open-source datasets will be combined: DDSM (Digital Database for Screening Mammography), MIAS (Mammographic Image Analysis Society), and INbreast. Randomly selected scans from these datasets will be used to form a smaller, combined dataset to perform self-supervised learning efficiently.

3.4.1. DDSM:

Digital Database for Screening Mammography is a resource used by the mammographic image analysis research community. The database consists of around 2,500 cases. Each case consists of an image of each breast along with some information about the patient such as the age of the subject, ACR breast density rating, abnormalities and image information, etc. Images containing suspicious areas have associated pixel-level "ground truth" information about the suspicious regions. These mammograms and truth images can be used for calculating performance figures for automated image analysis algorithms.

3.4.2 MIAS:

The Mammographic Image Analysis Society (MIAS) [24] is a society of researchers who have helped in collecting mammograms and created a digital database of the same. These images have been digitized to 50-micron pixel edges with a Joyce - Loebl microdensitometer. This database is available in 2.3GB 8mm tape and has 322 digitized films. The images that contain abnormalities have been marked by the radiologists. The images in the database are reduced to 200-micron pixel edges and clipped so that the images are all of size 1024 x 1024. Mammographic images are available via the Pilot European Image Processing Archive (PEIPA) at the University of Essex.

3.4.3 InBreast:

The INBreast database [25] consists of 115 cases, which includes a total of 410 images. Out of this, women affected with both breasts make up around 90 cases, i.e 4 images in every case and about 21% of the cases are of mastectomy patients. This database was made for works related to breast cancer imaging and was made publicly available with a variety of cases. The INbreast database is built with Full-field digital mammography (FFDM), in that it uses electrical signals to form images of the breast which can then be viewed on screens. These mammograms are much better as opposed to digitized mammograms as they can be

maneuvered to improve their contrast and resolution. The clarified image then eliminates the need for advanced imaging and improves interpretations for medical practitioners.

3.5 Tools Used

For the implementation of proposed approach, the tools used are PyTorch, TensorFlow, Keras. PyTorch libraries used were Kornia, Lightning module, and dataloaders.

4. Results & Analysis

The basic architecture that has been used for the model is Resnet. Since resnet is designed for 3 input image channels, for the proposed approach , it had to be configured for 1 image channel size. We have trained this model for a period of 25 epochs with a batch size of 128. And The time taken for the model to train was roughly 300 mins. The learning rate is set to 1e-4 and the weight decay has been set to 1e-6. The highest accuracy achieved was on Resnet18, which was 96.7%.

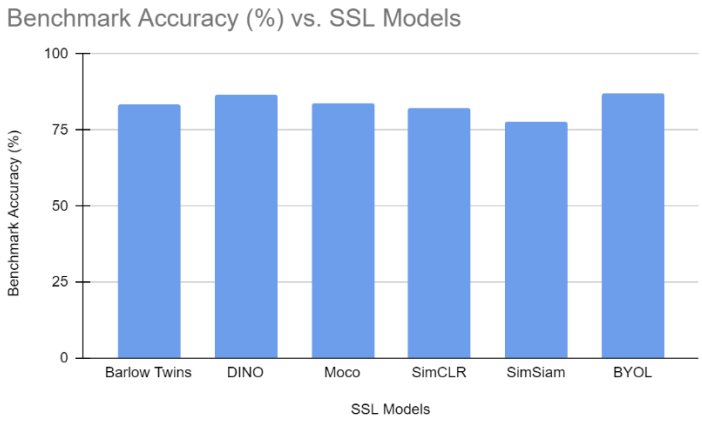

Fig.7. Bar plot of accuracies of various SSL models on CIFAR-10 dataset [26]

Cifar10 is a compilation of 50000 training and 10000 testing images. On the training data, the self-supervised models are trained from scratch. On the test set, all training images are embeded at the conclusion of each epoch and use the features for a kNN classifier. The reported accuracy is the model's maximum accuracy over all epochs. Due to the modest image sizes, all tests employ the very same ResNet-18 backbone and the gaussian blur augmentation is disabled. In this model, the ResNet-18 backbone is slightly different, it is a common variation for cifar10 SSL method benchmarks. The bar graph shows that the BYOL has the best performance on CIFAR10 with an accuracy of 87.2%, other models like BarlowTwins , DINO , Moco and SimCLR give the respective accuracies of 83.5% , 86.8% , 83.8% and 82.2%. And the worst performing model is SimSiam with an accuracy of 77.9%.

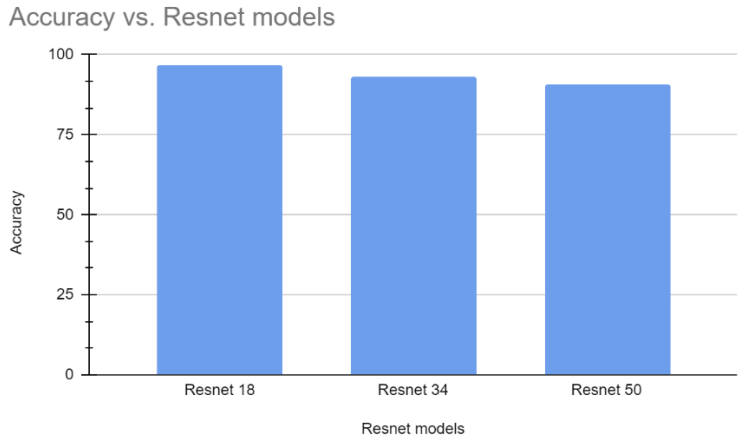

Fig.8. Comparison of Resnet Architetures with respect to their accuracies

Different resnet models have shown varying accuracies on the same dataset. Resnet 18 gave the best accuracy of 96.7% and the worst accuracy of 90.7% was given by Resnet 50. Resnet 34 gave an accuracy of 92.9%

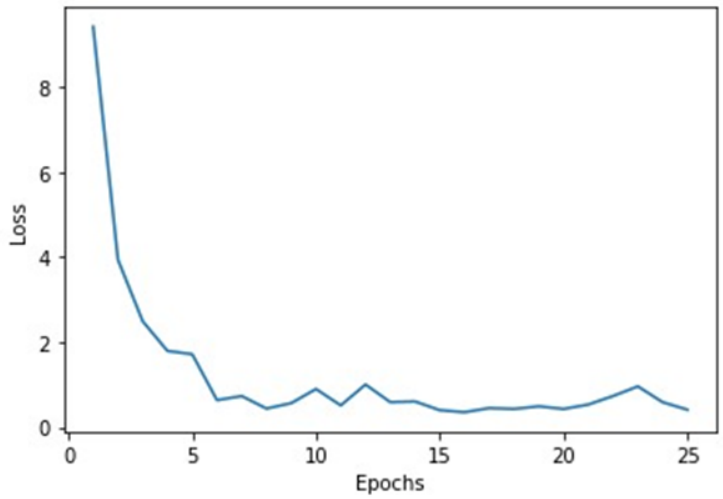

Fig.9. Line chart of Loss vs. Epochs

During training, a Loss curve is one of the most commonly used charts to debug a neural network. It provides an overview of the training process as well as the network's learning trajectory. The loss can be logged in two periods:

- After each epoch
- After each iteration

The loss vs. epoch curve tells us whether the model is underfitted or overfitted. If the model has trained really fast , then it might be overfitted and if trains extremely slow, then underfitiing might be the issue. The curve for model proposed in this paper shows that after the 10th epoch , the curve has flattened, which means that the model has trained and all the other iterations are redundant.

5. Conclusion and Future Scope

The results of implementation carried out demonstrate the efficacy of self-supervised learning for breast cancer prediction and analysis. The pretext task, powered by self-supervised learning, helps in populating the dataset to an extent that will help the model to learn sufficiently from the dataset. This novel approach will provide better accuracy on the dataset as compared to the original dataset which is smaller and contains many unannotated data instances. This approach has performed better than the state-of-the-art methods on these datasets without compromising on accuracy.

Moreover, this model will be of extreme help to radiologists, other medical professionals as well as researchers, hopefully giving rise to more research in the field and aiding in saving many lives by predicting whether a tumor in a mammogram is malign or benign. As the predictions will become more accurate, better diagnosis of the patients can be done and it will then help in decreasing the mortality rate. Also in India, the doctor-to-patient ratio is imbalanced, therefore this strategy will assist in reducing the workload of the medical personnel while also improving performance and transparency in the process.

The main goal of this research is to provide a tool that medical practitioners and other researchers may use and work on finding ways to cure Breast Cancer. Usually what happens after medical screenings is that different doctors have different opinions on the type of tumor in the initial phase, as the tumor is only visible to the naked eye. Traditionally, a biopsy is recommended only after a tumor is visible in the mammogram. Therefore, getting a clear idea from the mammogram plays a huge role in the medical process recommended further. The proposed Machine Learning algorithm uses data and on the basis of that, it gives approximate average confidence on which type of tumor it is (malign or benign) using various algorithms. This will, thus, help doctors come to a similar conclusion, further avoiding misdiagnosis if any.

Through the usage of these algorithms and machine learning models on a limited working dataset, the model can better help doctors or radiologists to analyze anomalies as compared to traditional machine learning approaches like Supervised Learning or Unsupervised Learning. This will help reduce medical misdiagnosis and will act as a tool to aid doctors in giving accurate predictions. The proposal is a novel method to analyze scans that can help detect tumors and determine if they are malignant or benign. The comprehensive analysis and comparison of the application of algorithms on different mammogram datasets like DDSM, MIAS, INbreast make it more reliable and accurate.

In the future, validation of the model's utility should be done in real-time by interpreting screening mammograms as they are produced. Moreover, among the various tasks of interest, the one investigated in this study, that is, predicting if the patient had a visible malignancy at the time of the screening mammography exam, is one of the simplest. A clear next step would be to anticipate the development of breast cancer before it is even detectable to a trained human eye.

This model can also be modified to make it suitable to be tested on other organs like the brain, lungs, etc. as well. Further, various options would be analyzed to design different types of pretext tasks such as rotating images, zooming in or out, etc. based on label information and medical data characteristics. The final task would be to look for ways to maintain or improve accuracy on real-time, unseen data because if the data encountered by the model has never been seen before, that could affect the learned anatomical features of the model.

The current weakness of this model is that it cannot be relied on completely, the model will need a human operator to validate the results. This is because the model does not take into account the extreme cases which are not already fed into it, so for this, specialized help from a medical practitioner is required. Also, medical projects can be very uncertain because of the implications involved. Many life and death decisions can depend on these models and so the results and the final validations always lie with the subject matter expert.

The main challenge while working on this research paper was the procurement of real-time medical data. Due to the difficulty in getting the data because of doctor-patient confidentialities and data privacy issues, it becomes arduous for the data processed to follow all the compliances strictly. Also, mammograms have to be downsized in order to be processed by the available GPU. Hence, it is difficult to retain the original image resolution, which might affect the results of the model.

REFERENCES

[1] *World Health Organization*, "The top 10 causes of death," Available at: www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death, Accessed on Dec 12, 2021.

[2] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.

[3] *World Health Organization*, "Cancer," Available at: www.who.int/news-room/fact-sheets/detail/cancer Accessed on Feb 05, 2022.

[4] *Cancer.Net*, "Understanding Statistics Used to Guide Prognosis and Evaluate Treatment," Available at: www.cancer.net/navigating-cancer-care/cancer-basics/understanding-statistics-used-guide-prognosis-and-evaluate-treatment, Accessed on Dec 15, 2021.

[5] *Centers for Disease Control and Prevention*, "Basic Information About Breast Cancer," Available at: www.cdc.gov/cancer/breast/basic_info/index.htm Accessed on Dec 20, 2021.

[6] *Cancer Treatment Centers of America*, "Bilateral or Double Mastectomy: What to Expect and Recovery," Available at: www.cancercenter.com/cancer-types/breast-cancer/treatments/surgery/double-mastectomy , Accessed on Feb. 04, 2022.

[7] Y. Cui, Y. Li, D. Xing, T. Bai, J. Dong, and J. Zhu, "Improving the Prediction of Benign or Malignant Breast Masses Using a Combination of Image Biomarkers and Clinical Parameters," *Frontiers in Oncology*, vol. 11, Mar. 2021, doi: 10.3389/fonc.2021.629321.

[8] Z. Zhang and E. Sejdić, "Radiological images and machine learning: Trends, perspectives, and prospects.," *Computers in biology and medicine*, vol. 108, pp. 354–370, 2019, doi: 10.1016/j.compbiomed.2019.02.017.

[9] S. Goled, "Self-Supervised Learning Vs Semi-Supervised Learning: How They Differ," *AIM*, Available at: https://analyticsindiamag.com/self-supervised-learning-vs-semi-supervised-learning-how-they-differ/ Accessed on Dec. 07, 2021.

[10] J. Wiens and E. S. Shenoy, "Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology," *Clinical Infectious Diseases*, vol. 66, no. 1, pp. 149–153, Jan. 2018, doi: 10.1093/cid/cix731.

[11] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations." http://proceedings.mlr.press/v119/chen20j.html, Accessed on Mar. 09, 2022.

[12] J.-B. Grill *et al.*, "Bootstrap your own latent: A new approach to self-supervised Learning," Available at http://arxiv.org/abs/2006.07733, Accessed on Dec. 10, 2021.

[13] Y. Wang, E. J. Choi, Y. Choi, H. Zhang, G. Y. Jin, and S. B. Ko, "Breast Cancer Classification in Automated Breast Ultrasound Using Multiview Convolutional Neural Network with Transfer Learning," *Ultrasound in Medicine and Biology*, vol. 46, no. 5, pp. 1119–1132, May 2020, doi: 10.1016/j.ultrasmedbio.2020.01.001.

[14] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, "Deep Learning to Improve Breast Cancer Detection on Screening Mammography," *Scientific Reports*, vol. 9, no. 1, Dec. 2019, doi: 10.1038/s41598-019-48995-4.

[15] D. Ribli, A. Horváth, Z. Unger, P. Pollner, and I. Csabai, "Detecting and classifying lesions in mammograms with Deep Learning," *Scientific Reports*, vol. 8, no. 1, Dec. 2018, doi: 10.1038/s41598-018-22437-z.

[16] H. Harvey *et al.*, "The Role of Deep Learning in Breast Screening," *Current Breast Cancer Reports*, vol. 11, no. 1. Current Medicine Group LLC 1, pp. 17–22, Mar. 01, 2019. doi: 10.1007/s12609-019-0301-7.

[17] K. J. Geras *et al.*, "High-Resolution Breast Cancer Screening with Multi-View Deep Convolutional Neural Networks,". Available at: http://arxiv.org/abs/1703.07047, Accessed on Dec. 15 2021.

[18] D. Muduli, R. Dash, and B. Majhi, "Automated diagnosis of breast cancer using multi-modal datasets: A deep convolution neural network based approach," *Biomedical Signal Processing and Control*, vol. 71, Jan. 2022, doi: 10.1016/j.bspc.2021.102825.

[19] N. Wu *et al.*, "Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening," *IEEE Transactions on Medical Imaging*, vol. 39, no. 4, pp. 1184–1194, Apr. 2020, doi: 10.1109/TMI.2019.2945514.

[20] N. I. R. Yassin, S. Omran, E. M. F. el Houby, and H. Allam, "Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review," *Computer Methods and Programs in Biomedicine*, vol. 156. Elsevier Ireland Ltd, pp. 25–45, Mar. 01, 2018. doi: 10.1016/j.cmpb.2017.12.012.

[21] X. Yin, L. Yin, and S. Hadjiloucas, "Pattern Classification Approaches for Breast Cancer Identification via MRI: State-Of-The-Art and Vision for the Future," *Applied Sciences*, vol. 10, no. 20, p. 7201, Oct. 2020, doi: 10.3390/app10207201.

[22] R. Gong, J. Wang, and J. Shi, "Task-driven Self-supervised Bi-channel Networks for Diagnosis of Breast Cancers with Mammography," Jan. 2021.

[23] GreatLearning Blog, "Introduction to Resnet or Residual Network," Available at: www.mygreatlearning.com/blog/resnet/, Accessed on Mar. 09, 2022.

[24] *Mammoimage.org*, "Mammographic Image Analysis," Available at: www.mammoimage.org/ Accessed on September 12, 2021.

[25] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "INbreast: Toward a Full-field Digital Mammographic Database," *Academic Radiology*, vol. 19, no. 2, pp. 236–248, Feb. 2012, doi: 10.1016/J.ACRA.2011.09.014.

[26] *Lightly.ai*, “Benchmarks — lightly 1.2.8 documentation,” Available at: https://docs.lightly.ai/getting_started/benchmarks.html, Accessed on February 10, 2022