www.irbas.academyirmbr.com April 2022
International Review of Basic and Applied Sciences Vol. 10 Issue.2

A Comparative Analysis of Classification and Prediction Methods for Breast Cancer

WAJID KHAN
Department of Computer Science, Abbottabad University of Science & Technology, Havelian, Pakistan.
Email: khanwajid48@gmail.com

AYESHA IRSHAD
Department of Computer Science, Abbottabad University of Science & Technology, Havelian, Pakistan.
Email: meayeshairshadkhan@gmail.com

KALSOOM AYYAZ
Department of Computer Science, COMSATS University Islamabad, Abbottabad Campus.
Email: kalsoomayaz@cuiatd.edu.pk

GOHAR ZAMAN
Faculty of Computer Science and Information Systems,
Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor Malaysia.
Email: gohar.zaman.isb@gmail.com

MUHAMMAD AYYAZ
Department of Computer Science, Abbottabad University of Science & Technology, Havelian, Pakistan.
Email: ayazsurkhdheri2@gmail.com

SYED ZOHAIB HASSAN
School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China.
Email: syedzohaibhasan@hotmail.com

Abstract

Cancer is a rouge disease causing billion of human lives every year. The disease occurs when the cells start dividing uncontrollably and spreading into different body tissues. Among carcinoma, breast tumor is the most common cancer in females and the vital reason for female death. Although the disease is curable and can be controlled, it requires an early diagnosis and prognosis for the patient survival. Artificial intelligence and machine learning have recently gained great success and have been successfully applied to solving real-life problems. Also, these technologies have been assisting the medical field in various areas, like disease classification and prediction, using multiple methods. For example, the models have been applied in radiology for breast cancer detection and prediction. In this paper, we will review the computer-aided diagnosis (CAD) models for breast cancer diagnosis using machine learning (ML) and deep learning (DL). Also, compare these methods, including their details, along with the pros and cons of each model. This review also highlights the open challenges, research issues, and future work to improve the breast cancer diagnosis system. ML and DL are the subfields of artificial intelligence (AI.) and are widely used in the healthcare industry for excellent results and cost-effective solutions. The primary purpose of this study is to identify the different machine learning and deep learning classifier for breast tumor diagnosis. We also compare the results of the classifiers for performance measures. The research gap will guide the researchers to make improvements for better results.

Keywords: *Breast Cancer Classification, Machine Learning Classification, Deep Learning, Convolution Neural network, Computer-aided Diagnosis, Breast Cancer.*

Introduction

The human body has billions of cells that perform different necessary functions. Cancer or tumor starts in the human body when a cell grows excessively, generating a bump or primary tumor (Goyal, 2019). According to a study in Cancer Statistics 2021, almost ten million people have cancer disease; among them, 2.27 million people have breast cancer. Nearly 0.7 million people die because of cancer. The danger of death from the tumor in males is almost 7.40% and 6.18% in women making it the most death rate disease around the world. In males, oral cavity and lung cancer have a ratio of 25%, whereas breast cancer in females has a ratio of more than 25% (Cancer, 2022). Figure 1 indicates the cancer ratio among male and female, and it shows that breast tumor is the second most prominent among both sexes and the most common cancer in females, with a ratio of 11.7% among all cancers. The most alarming situation is the increase in breast cancer in the 20s to 40s (Munir, 2019). The primary season behind the death rate due to breast cancer is its late detection. The early detection of cancer in the patient has high survival chances. Cancer prediction and analysis have recently gained great hype due to their increased occurrence rate.

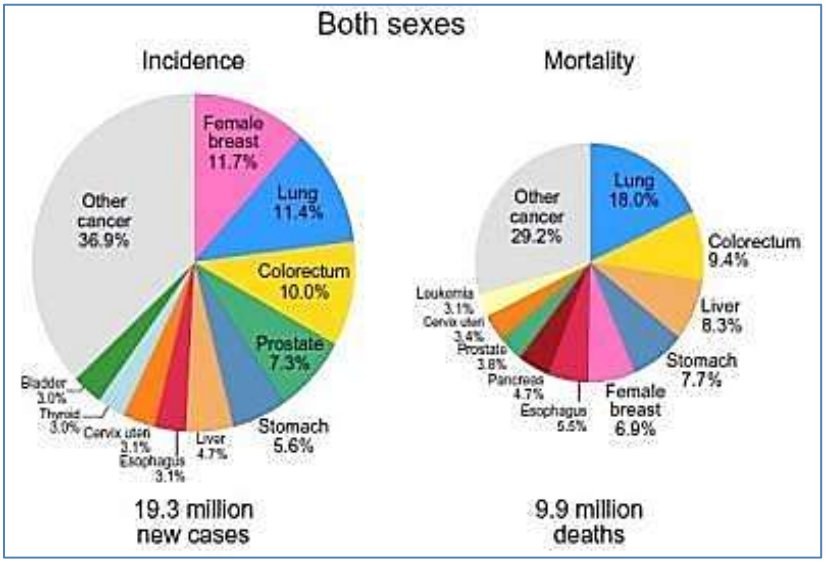

Figure 1: The ratio of different cancer for both males and females (Cancer, 2022).

The researchers have applied different methods for cancer prediction before its symptoms appear. But most of the processes are not accurate or require time and resources. Therefore, the precise diagnosis and prediction of breast cancer is a very challenging task for the healthcare industry. The manual diagnosis is a time-consuming and lengthy process with fewer accuracy chances (Selvathi D. &., 2019). Therefore, it required an automated diagnosis system. The current systems are good at diagnosis but have errors, and the process is very lengthy. To solve this problem, medical image classification using computer-aided diagnosis (CAD) has gained great success in assisting doctors with image classification for premature diagnosis and treatment (Henriksen, 2019). In this scenario, machine learning (ML) models can make more accurate diagnosis as early as possible, increasing the patient's survival chances. These models can also improve the treatment process and reduce the cost due to misdiagnosis (Khan, 2020) .The purpose of this work is to review the most recent developments in early breast cancer diagnosis by applying machine learning and deep learning methods. The paper has the following structure, including the breast cancer details, CAD using machine learning and deep learning, comparison of different ML and DL models, Research gaps, and future directions concluded the paper.

Breast Cancer

Breast cancer imitates beginning the lobules to the mammilla of the female chest. But the cancer cells spread from breast tissues to the arms lymph glands from where they can transmit to another part of the body. The physical signs are breast tenderness, change in breast size and color, irritation, and soreness. (Simsek, 2020) calculated the impacts of different factors that contributed the breast cancer. They applied logistic regression and artificial neural networks and highlighted the importance of various parameters significance for the survival of the tumor. They concluded that the age, grade, and positive lymph nodes are the factors for survival. Also, breast cancer can be classified as malignant or benign. Both required different diagnoses and treatment procedures. Table 1 highlighted the breast cancer stages with their description for better understanding.

Table:1: The breast cancer stages along their description

S. No	Stage	Description
01	Stage 0	The abnormal cell existed in the body but with no movement.
02	Stage 1 Situ	The abnormal cells transmit into Lobules and the duct of the breast.
03	Stage 2 Localized	Cell moves to lymph nodes, and tumor size are below 50mm.
04	Stage 3 Regional	Cancer is transmitted into the skin, chest, and lymph more than 50mm in size.
05	Stage 4 Distant	Cancer can be spread to different parts of the body at any length.

Computer-Aided Design (CAD)

Manual image processing has three major issues. The first one requires more pathology knowledge, the second manual image investigation was tough and labor work, third, the expert knowledge and domain expertise were very important. Therefore, it may lead to misdiagnosis in the early stage (Debelee, 2020). So, early detection and classification are very vital. The CAD is a solution for such a problem. The CAD systems also need processing for image classification. Very extensive computational models are required for this image processing. But the CAD is still more reasonable while detecting breast cancer in early stages. The CAD systems can be applied for the detection and localization of anomalies in images. But these systems can also be used for classification as benign or malignant. In the future, these systems can also provide extra information for future conclusion (Fourcade, 2019).

CAD using Machine Learning

Machine learning models are used in CAD systems for image preprocessing, segmentation, feature extraction and classification. We are going to discuss these processes in detail:

Image Preprocessing

Image preprocessing is the process of removing the noise and outliers from the image to clean the image for further processing. These processing included the different methods for images cleaning (Yassin N. I., 2019).

Image Segmentation

Image segmentation is the process of separating the image into different parts. It allows the system to work on a specific segment region known as the region of interest (ROI) (Yassin N. I., 2019).

Feature Selection and Extraction

Feature selection and extraction are the processes of extracting required features from segmented images. This process has many advantages including cost reduction and improved system accuracy (Turgut, 2018).

Classification

The classification process is the process of assigning labels or classes into different groups based on their properties. The support vector machine (SVM), random forest (R.F.), K-nearest neighbor (KNN), and Logistic regression (L.R.) are the model applied for classification (Hu, 2019).

CAD using Deep learning

Deep learning or deep neural networks are models with a hierarchical structure and used for data characterization. Therefore, deep learning is also called as an illustration learning method. The deep learning model has three layers: input, output and hidden. These layers work together to find hidden features. The Convolution Neural Network (CNN) has achieved great performance in cancer detection and diagnosis. The input layers calculate the low-level features and pass them to the hidden layer, where they find the more hidden and detailed features and passes them to the output. At the output layer, these values were estimated to find the class or value for the input. Figure 2 explains the ML and Dl-based CAD systems (Geras, 2019).

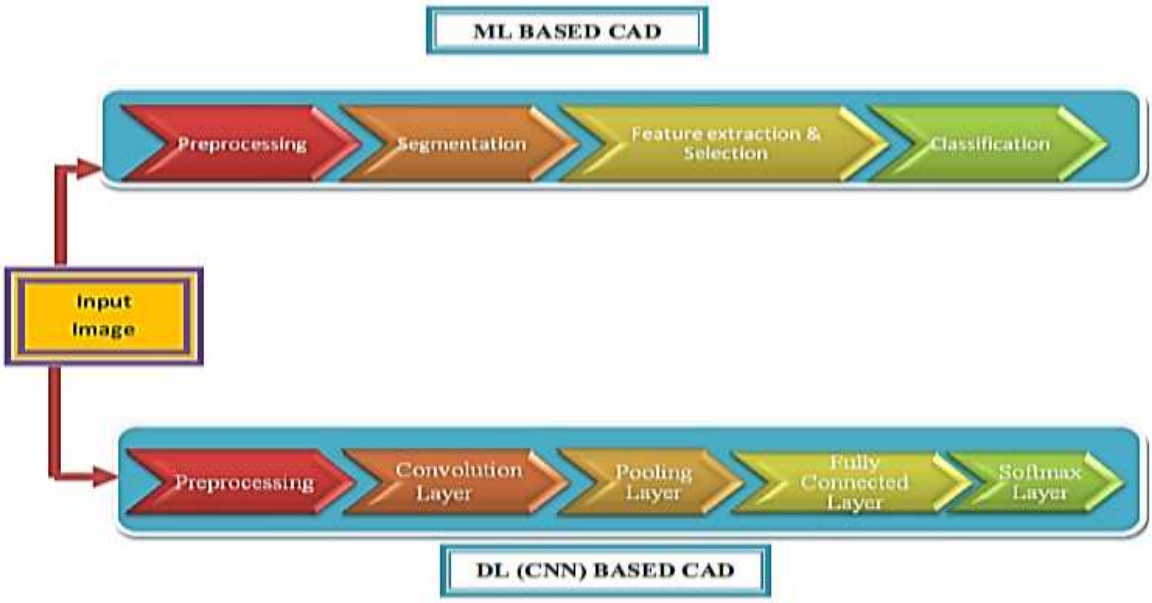

Figure 2: The application of ML and DL in CAD systems.

Dataset Description

In this section, we will discuss the dataset and image repository used for ML and DL models experiment. The details of public datasets used for researchers in current years are given in table 2.

Table 2: The details of the datasets used available for breast cancer research.

Depository Details	Dataset Name	Description
IRMA (Oliveira, 2008)	Image Retrieval in Medical Applications	The dataset has 10500 breast images.
InBreast (Moreira, 2012)	InBreast	The dataset has 115 different breast cases, along with 400 images for each cases.
SEER (NIC, n.d.)	Surveillance Epidemiology and End Result	The dataset has the stage-level analysis of breast cancer data.
Mini-MIAS (Suckling, 2012)	Mini-MIAS	The dataset has MIAS images of 200 microns.
BCDR (BCDR, n.d.)	Breast Cancer Digital Repository.	The dataset has the mammography and ultrasound images of 1700 patients.
BreakHis (Spanhol, 2019)	Breast Cancer Histopathological Database	The dataset has 7000 micro images of breast tissues having 2500 benign and 5500 malignant cases.

ML for Breast Cancer Diagnosis

This section represents the comparative analysis of different machine learning models used for breast tumor diagnosis and prediction. In this regard, (Gupta, 2018) proposed a joint framework for breast tumor classification by applying KNN, SVM, D.T. and L.R. The model achieved the accuracy of 97.88% accuracy for breast cancer classification. Similarly, (Agustian, 2020) emphasise the improvement of accuracy applying particle swarm optimization method for feature selection. The KNN and Naïve Bayes were used for classification. They spread their ways to Saudi Arabian Females for breast tumor classification. Their results suggested that women over 40 are the main target of the disease. The author also implemented FVE phase-based data examination on the WBCD dataset and compared an analysis of classification with features and classification without features selection. They achieved 71% and 76.4% accuracy for both ML models. (Yue, 2018) reviewed the SVM, KNN and decision tree models for breast cancer detection on the WBCD dataset. The SVM model used a two-step clustering model with SVM achieved 99% accuracy for prediction . Whereas KNN and decision tree were unable to achieve an accuracy of over 95%. (Thirumalaikolundu, 2020) applied Naïve Bayes methods for breast cancer prediction and performed a comparison analysis of Tree Augmentednaïvee Bayes, Bayes Belief Network. The WBCD dataset was applied for the experiment. According to their results the models achieved 91.6%, 91.7% and 94% of accuracy. For Naïve bayes, TAN and BBN network. (Chaurasia, 2018) used the Naïve Bayes, RBF network and J48 models on WBCD for breast cancer classification. They applied the WEKA tool for analysis. The model achieved great accuracy on simulation models. The Naïve Bayes achieved 97.30%, while the RBF network reached 96% and J48 achieved 93% accuracy, respectively.

Deep Learning for Cancer Diagnosis

Deep learning methods are also widely applied for breast tumor prediction and classification. For this purpose, the (Al-Antari, 2020) made a CAD design named YOLO for breast cancer detection. They applied the deep convolution neural network for the detection for segmentation and classification. The model attained an accuracy of 94.80% with the adequate time of 12 sec per image. (Al-Masni, 2019) also used the YOLO model for the simulation of breast masses for the DDSM dataset. The model achieved great performance with mass detection for solid tissues. Another mass detection model was used by (Jung H, 2020) based on Retina Net. The model has achieved a higher sensitivity value for false-positive images. The model was applied to the GURO dataset for performance evaluation. (Jiang Y, 2019) proposed a Breast Cancer Histopathology Image Classification Network (BHC Net). The author also used a new model, Squeeze and Excitation Residual Network to prevention of overfitting. The model has achieved an accuracy of 98.80% for binary classification and 93% for multi-level classification. This method can be

applied for cell duplication and irregular studies. (Nahid A Al, 2019) used the LSTM model for histopathology image classification. The LSTM and CNN model were combined to find the unseen structure and statistical information from the data. The model achieved an accuracy of 91% with the SoftMax layer. We have seen that both the ML and DL models are imposing when classifying the breast cancer disease using the image data. In table 3, We are going to make a comparison of these models while applied for cancer prediction.

Table 3: The comparison of ML and DL-based models for breast tumor classification.

Ref No	Model Applied	Dataset Details	Performance	Advantages	Disadvantages
(Huang, 2019)	Genetic Algorithm for feature selection SVM for classification	UCI ML repository .	G.A. + SVM 96%	G.A. reduces the computation Time SVM achieved good accuracy	They are applied on a small scale.
(Mihaylov, 2020)	SVR and Decision Tree regression	2000 patients' clinical data records	Not mention	Both models are suitable for survival time prediction	Small Scale application
(Kate, 2019)	Naïve bayes, Linear Regression, Decision Tree	SEER	Survival prediction at all stages.	Stage level model performs better than the joint model	Not good in terms of accuracy and performance
(Ram, 2020)	UCI ML repository	SVM, NB, D.T., LR	NB and LR= 97% SVM and PCA=96%	Highly accurate More features can improve the model's accuracy.	Not applied to clinical data
(Chen HL, 2017)	WBCD	RS and SVM	99.40%	Very high accuracy Implemented on a large scale	
(Shaikh TA, 2020)	WBCD	NB, KNN, SVM	SVM=97.4%	The hybrid model was applied for feature selection and classification	Except for SVM, no other model achieved good results.
(Selvathi D. &., 2019)	Mammogram Mini-MIAS	CNN and auto-encoder	CNN=97.2% Autoencoder= 98%	Very good accuracy. Auto-encoder help in measuring small masses.	
(Sun, 2019)	Mammogram data	CNN with semi-supervised learning	Accuracy=0.86 AUC= 0.88	Very effective for mixed data. Useful when training data is in a small amount.	
(Yurttakal, 2020)	MRI images	DCNN	Acc= 98.1%	It takes 36 seconds to classify 60 images. Time-saving	Small Scale Application.

				model	
(Samala, 2020)	SFM, D.M.	Multi-level transformer learning	ROI and AUC for different stages. Single stage: 0.84 Multi-Stage: 0.90	Simple transformer improves learning.	Hyper-parameters are not optimized due to small-scale datasets.

The table 3 shows that ML and DL models achieved great success in image classification for breast cancer prediction. But there are several issues regarding the clinical implementation of ML models for breast cancer classification.

Research Gaps and Future Directions

Breast tumor cases are growing at a huge rate around the globe. This ratio is very high in underdeveloped countries (Yassin N. I., 2018). Premature detection and diagnosis are very crucial for survival. CAD has been very effective for early diagnosis and treatment. Experiment results identified that ML and DL models are proven to be best while diagnosis. Also, CNN and DCNN can handle complex, unstructured data (Miotto, 2019). Despite having advantages, these models suffer from several challenges that need to be addressed while applying these models.

The unavailability of large-scale expert datasets

We know that deep learning models required massive data for training. But there is no independent dataset with expert doctors' images. Therefore, the public dataset can't be considered an authentic source from exports (Murtaza G. S.-g., 2020).

Imbalanced Dataset

An additional problem for the deep neural networks is the imbalanced dataset. This allows the model to become biased to another common class while trained on the imbalanced dataset. These issues need to be addressed to effectively apply deep learning models (Sharma, 2020).

Unavailability of pre-trained medical images models

As CNN can be applied in a CAD system for different purposes. Many of these models required pre-training on medical images for better performance. The parameters fine-tuning is also performed during the training process. Therefore, this process is time-consuming and costly and requires a large number of resources (Murtaza G. S., 2020).

These problems are the main reason behind the industrial application of ML and DL models for breast cancer classification and prediction. Addressing them will open the new gates for these technologies to perform better in the current area.

Conclusion

Breast tumor is the primary factor of decease in female considering carcinoma diseases. The application of CAD has allowed the healthcare system to make early diagnoses and prognoses. The motivation of this review is to highlight the current application and research in breast tumors using ML and DL. While studying, we have identified that these models perform very well while working on a public dataset. The

main problem is the unavailability of the expert datasets for the training and testing these models. We have also seen the deep learning model, including LSTM and CNN, have achieved the highest level of accuracy. The research gap also indicated these issues that need research for the improvement of healthcare and CAD systems for chest tumor diagnosis and classification.

References

Agustian, F. &. (2020). Particle swarm optimization feature selection for breast cancer prediction. *8th International Conference on Cyber and IT Service Management*.

Al-Antari, M. A.-M. (2020). A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification. *International journal of medical informatics*.

Al-Masni, M. A.-A. (2019). Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system. *Computer methods and programs in biomedicine*.

BCDR. (n.d.). *Breast Cancer Digital Repository*. (BCDR) Retrieved 11 12, 2022, from https://bcdr.eu/

Cancer. (2022, February 3). Retrieved from https://www.who.int/news-room/fact-sheets/detail/cancer

Chaurasia, V. P. (2018). Prediction of benign and malignant breast cancer using data mining techniques. *Journal of Algorithms & Computational Technology*.

Chen HL, Y. B. (2017). A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. *Expert Syst Appl*.

Debelee, T. G. (2020). Survey of deep learning in breast cancer image analysis. *Evolving Systems*.

Fourcade, A. &. (2019). Deep learning in medical image analysis: A third eye for doctors. *Journal of stomatology, oral and maxillofacial surgery*.

Geras, K. J. (2019). Artificial intelligence for mammography and digital breast tomosynthesis: current concepts and future perspectives. *Radiology*.

Goyal, K. S. (2019). Comparative analysis of machine learning algorithms for breast cancer prognosis. *In Proceedings of 2nd International Conference on Communication, Computing and Networking*.

Gupta, M. &. (2018). An ensemble model for breast cancer prediction using sequential least squares programming method (slsqp). *In 2018 Eleventh International Conference on Contemporary Computing*.

Henriksen, E. L. (2019). The efficacy of using computer-aided detection (CAD) for detection of breast cancer in mammography screening: a systematic review. *Acta Radiologica,*.

Hu, Z. T. (2019). Deep learning for image-based cancer detection and diagnosis− A survey. *Pattern Recognition*.

Huang, M. W. (2019). SVM and SVM ensembles in breast cancer prediction. *PloS one*.

Jiang Y, C. L. (2019). Breast cancer histopathological. *PLoS One*.

Jung H, K. B. (2020). Detection of masses in mammograms using a one-stage object detector based on a deep convolutional neural network. *PLoS One*.

Kate, R. J. (2019). Stage-specific predictive models for breast cancer survivability. *International journal of medical informatics*.

Khan, W. A. (2020). AdaDiffGrad: An Adaptive Batch Size Implementation Technique for DiffGrad Optimization Method. *2020 14th International Conference on Innovations in Information Technology (IIT)*.

Mihaylov, I. N. (2020). Machine learning techniques for survival time prediction in breast cancer. *In International Conference on Artificial Intelligence: Methodology, Systems, and Applications*.

Miotto, R. W. (2019). Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*.

Moreira, I. C. (2012). Inbreast: toward a full-field digital mammographic database. *Academic radiology*.

Munir, K. E. (2019). Cancer diagnosis using deep learning: a bibliographic review. *Cancers*.

Murtaza, G. S. (2020). Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges. *Artificial Intelligence Review*.

Murtaza, G. S.-g. (2020). Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges. *Artificial Intelligence Review*.

Nahid A Al, M. M. (2019). Histopathological breast cancer image classification by deep neural network techniques guided by local clustering. *Biomed Res Int*.

NIC, N. C. (n.d.). *SEER Data*. (U.S. Department of Health and Human Services) Retrieved 11 16, 2022, from https://seer.cancer.gov/data/access.html

Oliveira, J. E. (2008). Toward a standard reference database for computer-aided mammography. *In Medical imaging 2008: Computer-aided diagnosis*.

Ram, S. &. (2020). Building Machine Learning Based Diseases Diagnosis System Considering Various Features of Datasets. *In Emerging Trends in Expert Applications and Security*.

Samala, R. K. (2020). Breast cancer diagnosis in digital breast tomosynthesis: effects of training sample size on multi-stage transfer learning using deep neural nets. *IEEE transactions on medical imaging*.

Selvathi, D. &. (2019). Deep learning techniques for breast cancer detection using medical image analysis. *Biologically rationalized computing techniques for image processing applications*.

Selvathi, D. &. (2019). Deep learning techniques for breast cancer detection using medical image analysis. *In Biologically rationalized computing techniques for image processing applications*.

Shaikh TA, R. A. (2020). Applying Machine Learning Algorithms for Early Diagnosis and Prediction of Breast Cancer Risk. *2nd International Conference on Communication, Computing*.

Sharma, S. &. (2020). Conventional machine learning and deep learning approach for multi-classification of breast cancer histopathology images—a comparative insight. *Journal of digital imaging*.

Simsek, S. K. (2020). A hybrid data mining approach for identifying the temporal effects of variables associated with breast cancer survival. *Expert Systems with Applications*.

Spanhol, F. O. (2019). *Breast Cancer Histopathological Database (BreakHis)*. (Laboratório Visão Robótica e Imagem) Retrieved from https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/

Suckling, J. (2012, 11 12). *The mini-MIAS database of mammograms*. Retrieved from http://peipa.essex.ac.uk/info/mias.html

Sun, W. T. (2019). Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. *Computerized Medical Imaging and Graphics*.

Thirumalaikolundu, S. (2020). Comparison of Bayes classifiers for breast cancer classification. *Asian Pacific journal of cancer prevention*.

Turgut, S. D. (2018). Microarray breast cancer data classification using machine learning methods. *In 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting*.

Yassin, N. I. (2018). Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. *Computer methods and programs in biomedicine*.

Yassin, N. I. (2019). Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. *Computer methods and programs in biomedicine*.

Yue, W. W. (2018). Machine learning with applications in breast cancer diagnosis and prognosis. *Designs,*.

Yurttakal, A. H. (2020). Detection of breast cancer via deep convolution neural networks using MRI images. *Multimedia Tools and Applications*.