[1]C. Valarmathi

[2]S. John Justin Thangaraj

The Combination of Feature Extraction and Classification by Bag of Visual Words to Detect Breast Cancer for Improved Accuracy

Journal of Electrical Systems

Abstract: - The main objective is to improve diagnostic accuracy of the breast cancer. The goal is to improve the accuracy of breast cancer detection by leveraging both the discriminative power of feature extraction and the robustness of classification using BoVW. By combining these techniques, the system aims to effectively distinguish between cancerous and non-cancerous tissues in medical images, thereby aiding in early diagnosis and treatment planning for breast cancer patient. The breast cancer sample images were gathered from the city's hospitals Chennai 2023 as well as the mini-MIAS database and DDSM database. The BoVW classifier is used to identify and classify breast tumor images according to their size. This improves the efficacy of the classification method developed for the CAD system's mammography breast cancer picture categorization. Several Computer Aided Diagnosis (CAD) systems have been created to use mammography images to identify breast cancer in its early stages. ANN attained 93.6%,94.18%, 93.2% and SVM gained 92%,92.44%,93.02% for Precision, Recall and Accuracy Respectively.The proposed method display excellent accuracy of 98.83% over the other methods. The novelty could stem from the seamless integration of multiple modalities within the feature extraction and classification framework.

Keywords: BADF, K-means clustering, Bag of visual words, Computer Aided Diagnosis, Breast cancer detection.

I. INTRODUCTION

Breast cancer detection is crucial in medical imaging, as timely and accurate diagnosis greatly influences patient outcomes. In this field, there are several gaps in utilizing feature extraction and classification techniques for breast cancer detection [1]. One significant gap felt was to improve feature extraction methods specifically tailored for mammography images to capture subtle yet clinically relevant imaging characteristics more effectively [2]. CNNs have proven highly successful in various image classification tasks in improving breast cancer detection accuracy when combined with feature-based approaches [3]. This integration represents an active area of exploration within the field.

Milton et al [4] have introduced a pioneering approach utilizing machine learning techniques for the detection of breast cancer. Their method has shown an improved accuracy and effectiveness over the methodologies they compared with 94.7% . On a similar note, Anji et al. [5] delve into the influence of different machine learning algorithms on the automation of mammography image classification as 95.3%. [6] Discusses the impacts of these algorithms but also compiles a range of representative works that highlight the utilization of machine learning methodologies in advancing the field of medical diagnostics and analytical sciences, particularly in the realm of breast cancer detection and classification.

The integration of feature extraction and classification using the Bag of Visual Words (BoVW) model represents a significant advancement in the field of breast cancer detection. This innovative approach draws inspiration from both machine learning and computer vision techniques, offering a novel framework to enhance the accuracy and efficiency of diagnosis as 98.8%. Overall, the integration of feature extraction and classification using the Bag of Visual Words model represents a promising avenue for improving the accuracy of breast cancer detection. By harnessing the power of machine learning and computer vision, this approach has the potential to revolutionize breast cancer diagnosis, ultimately leading to better patient outcomes and improved survival rates. Addressing these research gaps will not only advance our understanding, the potential of the BoVW model for breast cancer detection but also contribute to the development of more accurate and reliable diagnostic tools for clinical practice.

II. RELATED WORKS

By using machine learning methods, Anji [5] provides a novel strategy for detecting breast cancer. When compared to the current methods, the proposed method has produced results that are both extremely accurate and effective. The impacts of various machine learning algorithms on the automation of mammography image classification are discussed by Meenalochini, G., and S. Ramkumar [6]. This research compiles representative works that demonstrate how machine learning methodology is used to the outcomes of various problems discovered through various analytical science examinations.

For the combined profiling of morphological, molecular, and clinical characteristics from breast cancer histology, Binder [7] provided an understandable machine-learning approach. First, our method enables accurate heatmap

[1]*Correspondingauthor: Research Scholar, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, India.Vinmathi20@gmail.com

[2]Professor , Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, India.johnjustinthangarajs.sse@saveetha.com

representations of the classifier decisions and the robust detection of cancer cells and tumour-infiltrating lymphocytes in histological images. Second, histology can be used to predict molecular characteristics such as DNA methylation, gene expression, copy number variations, somatic mutations, and proteins. Balanced accuracy of molecular predictions is up to 78%, whereas accuracy for patient subgroups can reach above 92%. Li [8] Cancer researchers face a significant challenge in accurately estimating the survival probability of individuals with breast cancer. The promise of precise outcomes has made machine learning (ML) a hot topic, although its modelling techniques and prediction ability are still debatable.

Abbas [9] proposes a novel method known as BCD-WERT that effectively selects and categorizes features using the Extremely Randomized Tree and Whale Optimization Algorithm (WOA). WOA decreases the dataset's dimensionality and extracts the pertinent features for precise categorization. The efficiency of feature selection strategies in raising prediction accuracy is further demonstrated by experimental data. Ajay and Mishra [10] analysed recently proposed methods that have a few restrictions and potential workarounds. This study emphasises the value of identifying key traits that enhance the outcomes suggested by current approaches. The suggested model generates an improved performance label with a 90.41% accuracy score.

III. PROPOSED METHODOLOGY

Our method takes the segmented and preprocessed mammography image It gathers patient information about their illness, puts it in a database (mini MIAS), and then offers the information to the classifier. This collection includes benign and malignant images, both of which are in the DICOM file format. 86 pairs of images were used for the study and 86 pairs were used for the training collection and extract the hybrid feature set from it. This work categorization algorithm divides the image into benign and malignant conditions. The majority of the Focus should be placed on staging breast cancer when it is discovered the image data must be pre-processed for noise reduction from the images using Filterisation method.

The pre-processed satellite image is then segmented using the K-means clustering Segmentation approach to achieve inverse shape identification while utilizing the least amount of energy. Following segmentation, the breast cancer images are subjected to Bag of Visual Words combined with feature extraction and the Classification. By utilizing hybrid features, this approach avoids the requirement to explore a wide feature space. It makes use of a different collection of features while simultaneously increasing classification accuracy [6]. The system must employ methods and mechanisms that assure the highest level of accuracy in order to ensure the accurate classification. The main part of the system taken into consideration in this work is connected to feature extraction and classification, even though CAD contains many aspects, including preprocessing, segmentation, and classification. The primary benefit of the research is that it improves the radiologist's understanding of the type and degree of complexity of the tumor.

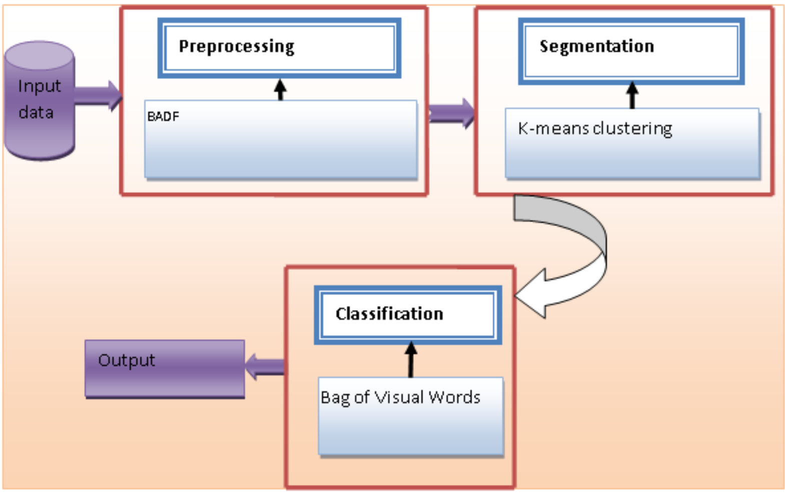


FIGURE 1. FLOW OF PROPOSED METHODOLOGY

A. *Pre-processing: BADF*

It mainly focuses on eliminating noise while keeping the image fine features. This BADF improves upon the previous anisotropic scattering filter by adding the Partial Differential Equation (PDE) after creating the scattered image. Diffusion that is lacking at the edges and boundaries can also be used to smooth surfaces [7].

The following PDE guiding method is the foundation for the standard linear smoothing technique known as anisotropic filtering.

$$\frac{\partial I_m}{\partial t} = div(T\nabla I_m) \qquad (1)$$

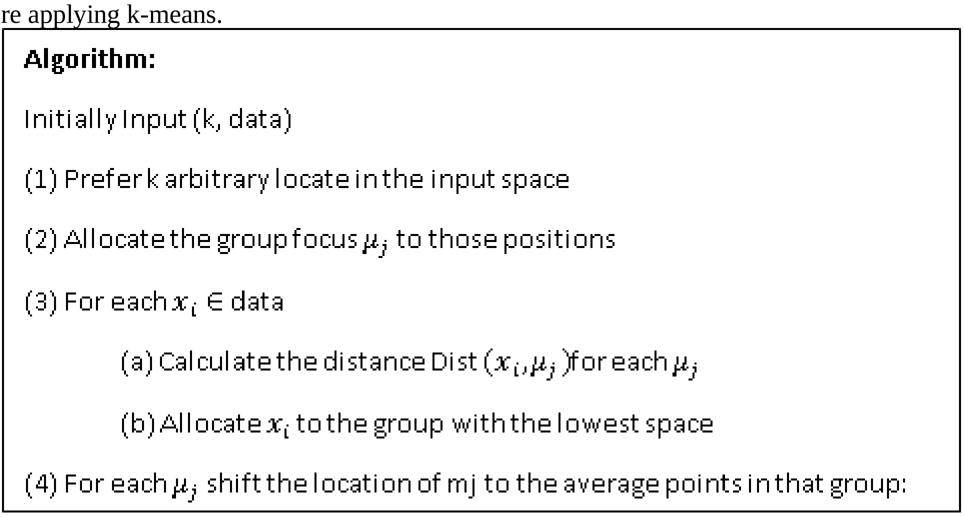
In the context of the weighting direction denoted as m, where Im signifies image strength, ∇ represents the gradient operator, div is the divergence operator, t corresponds to time, and T is a tensor indicating the smoothing directionality, it is formed from a gradient tensor G. This G is derived by convolving the sum of outer products of ∇Im across all weighting directions with a Gaussian kernel Kρ..

$$G = k_p * \sum_m (\nabla I_m \oplus \nabla I_m) \qquad (2)$$

Here, ⊗ signifies the outer product operation, and ρ stands for the standard deviation of the Gaussian kernel, determining the spatial scale of the gradient tensor. The key benefit of employing an anisotropic diffusion filter lies in its ability to smooth homogeneous regions of the image while simultaneously enhancing edges.

B. Segmentation: K-means clustering

A popular unsupervised technique called K-means divides an image into k sections depending on the mean of each sector. The mean for each cluster will be calculated when the data have been separated into k clusters. Each datum is placed in the cluster that is closest to the Euclidean distance of the cluster mean [8]. A vector of input data and a vector of k vectors are the results. The pixels in a two-dimensional MRI picture should be placed in a vector before applying k-means.



```
Algorithm:
Initially Input (k, data)
(1) Prefer k arbitrary locate in the input space
(2) Allocate the group focus $\mu_j$ to those positions
(3) For each $x_i$ ∈ data
        (a) Calculate the distance Dist ($x_i$,$\mu_j$ )for each $\mu_j$
        (b) Allocate $x_i$ to the group with the lowest space
(4) For each $\mu_j$ shift the location of mj to the average points in that group:
```

C. Combination of Feature Extraction and Classification: Bag of visual words

In bag of visual words is produced using the computer vision system toolbox. An image histogram of visual word occurrences is produced using an image category classifier. Following are the steps for using the bag of visual words classifier to categorize the image:

Step 1: Image category setup

The images are separated into training and test individuals in this step and kept for training purposes. Large data sets can be handled with ease.

Step 2 : Bag of feature creation

By removing features from each image in each category, the vocabulary or "bags of features" are formed.

Step 3: Train the image with Bag of Visual Words

This method uses a binary support vector machine to train a multiclass classifier using error-correcting output code. In the bag of visual words approach, which is used to extract the features from images, the image is encoded from a training set. The nearest neighbor approach is used to produce the feature histogram in the image. The image's histogram serves as the feature vector.

Step 4: Classify an Image Set or image

In this step, the CT image is classified. Two steps make up this classifying method.

1) Training phase
2) Testing phase

The classifier is prepared on the extracted features to create a classification model during the training stage, where they choose features are segregated since all of the training images. The test image is then organized using this in the testing stage into the predefined classes. One classification model created after the training phase is coupled to the test image during the testing stage in order to characterize into the preset categorization. Any number of likely classes can be produced by this classification [9]. Based on this output and the knowledge that classes were used to create each and every classification model, it is possible to determine the actual classification model. This section examines the phases of breast cancer by using mammography pictures from the mini-database of benign and malignant stages. The sample image is preprocessed, then the ROI is measured. Implementing segmentation and removing characteristics allows for straightforward classification. Stages of mammography images are categorized using the BOVW system. Analyses and comparisons of classifier performance are conducted [10]. Determine performance indicators like sensitivity, specificity, accuracy, and precision.

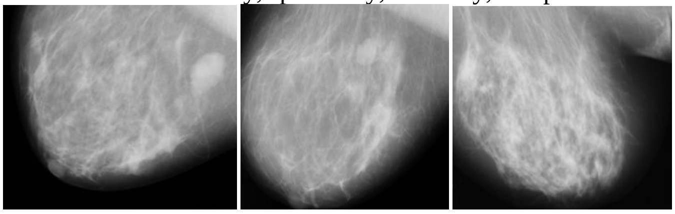


Figure 2: Benign Samples

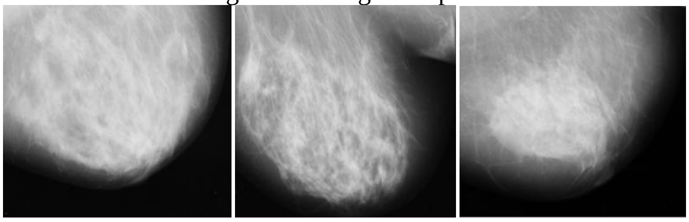


Figure 3: Malignant Samples

The dataset consists of images collected from city hospitals, the mini-MIAS, and DDSM databases, featuring both benign and malignant images in DICOM format. For the study, 86 pairs of images were employed, with an additional 86 pairs reserved for training. The BoVW classifier demonstrated remarkable accuracy of 95.6%, enabling the identification and categorization of breast tumor images according to their size. This advancement notably enhances the effectiveness of the classification technique in the Computer-Aided Detection (CAD) system for sorting mammography breast cancer images.

IV. Results and Discussion

The results acquired by providing the original breast mammography image from the database are demonstrated in the following steps. Preprocessing techniques are used to improve the image quality and minimize noise on the sample images.

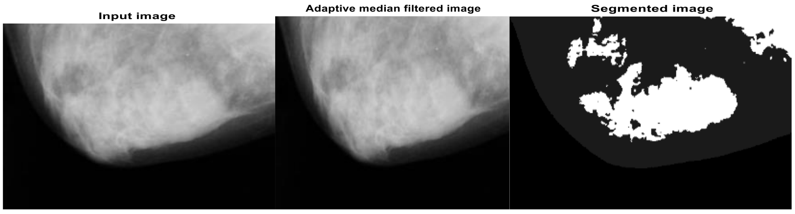


Figure 4: Benign Samples outputs

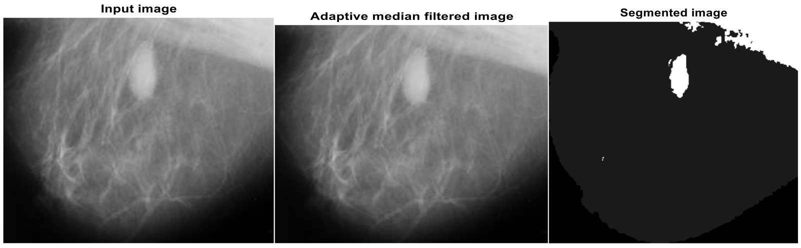


Figure 5: Malignant Samples outputs

Image enhancement techniques are added to the original image to remove the unwanted elements and increase the feature of images. Traditional diagnostic methods involve skilled physicians examining medical images for characteristic symptoms, but this process is time-consuming and not all doctors possess expertise in symptom identification. Consequently, there's a pressing need for a reliable and automated diagnostic system to precisely predict tumors. However, the available data for manual diagnosis is often noisy and raw, requiring preprocessing before applying feature selection methods to reduce management costs and error rates. Leveraging state-of-the-art machine learning techniques can offer a solution, empowering life scientists to extract essential information from tumor image databases efficiently.

Table: 1 Performance Metrics Comparison of the Database Images

Images	**MSE**	**PSNR**	**SNR**	**SSIM**
C1	59.1	35.56	30.31	0.9756
C2	38.47	35.27	32.16	0.9735
C3	69.66	35.55	29.79	0.9733
C4	25.17	35.86	34.03	0.9766
C5	43.27	35.13	31.67	0.9740
C6	40.81	35.12	32.11	0.9742

Partitioning of the tumor region is done following preprocessing of the model. For segmentation in this case, the k-means clustering algorithm is applied. The segmented results are shown in Figures 4 and 5. The combination of robust feature extraction techniques with advanced classification algorithms plays a crucial role in improving the accuracy and reliability of breast cancer detection systems. Additionally, continuous research and development in this field contribute to the refinement of existing method and the explorationof methodologies for more accurate diagnosis and improved patient outcomes. In Table 1, when comparing the performance of different combinations of feature extraction and classification techniques for breast cancer detection, it is essential to consider these metrics comprehensively, taking into account the specific characteristics of the dataset, the intended use case, and the relative importance of minimizing false positives and false negatives. In Figure 6, additionally, cross-validation and validation on independent datasets are crucial to ensure the reliability and generalizability of the results.

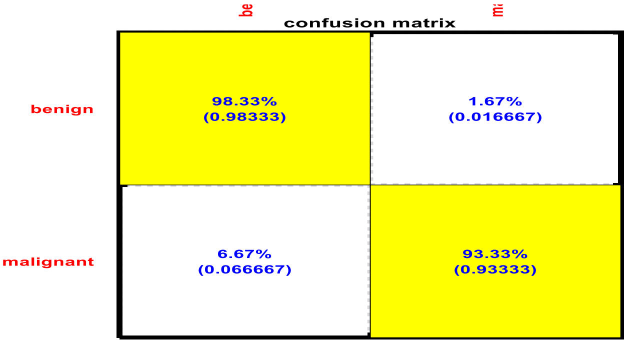


Figure 6: Confusion Matrix

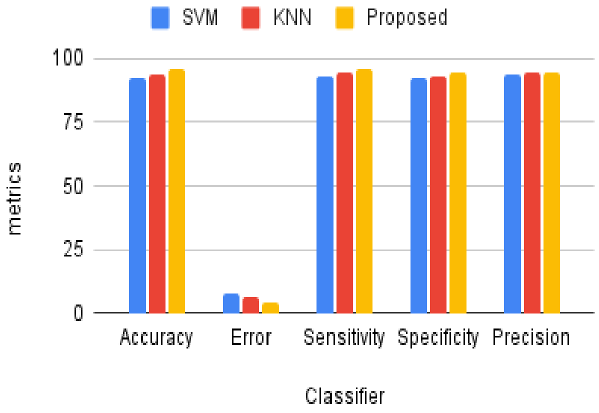


Figure 7: Performance plot

In figure 7, Our experiments indicate that the integration of the mentioned feature extraction techniques and the utilization of a classifier yielded the highest prediction accuracy at 95.8%. This outcome suggests that the accurate detection of IDC type breast cancer from histopathological images is achievable through the application of appropriate methods.

Table 2: Performance Comparison

Number of Samples	Classification Accuracy		
	SVM	**ANN**	**Proposed**
300	92.44%	93.6%	95.8%
400	93.02%	94.18%	95.17
500	92%	93.2%	95.3%
600	94.04%	94.11%	96.27%
700	92.44%	93.6%	98.8%

When evaluating the performance of a combination of feature extraction and classification techniques for breast cancer detection, several metrics are commonly used to assess the effectiveness of the model. Here a comparison of some key performance metrics.This study calculates the accuracy of each classifier listed in Table 2. The classification accuracy of the proposed method was associated with further classification algorithms SVM, ANN.The BoVW produces the best results when compared to other classifiers. The classifiers employed in this investigation explicitly produce better results than other classifiers, as shown by similarity.D.J Kumari [11] demonstrated that the Artificial Bee Colony Optimization technique exhibits superior Accuracy, Sensitivity, and Specificity, with an accuracy rate of 94.9%. Bartsch R [12] introduced a comprehensible machine-learning method. Our approach facilitates precise heatmap visualizations of classifier decisions and robust detection of cancer cells and tumor-infiltrating lymphocytes in histological images. Moreover, histology can predict molecular characteristics such as DNA methylation, gene expression, copy number variations, somatic mutations, and proteins. Molecular predictions achieve a balanced accuracy of up to 78%, while subgroup accuracy for patients can surpass 92%. Abelman RO et al [13] highlight the challenge of accurately estimating survival probability in breast cancer patients. Despite ongoing debates about machine learning (ML) modeling techniques and prediction accuracy, ML holds promise for precise outcomes. Frenel JS et al [14] proposed a model which achieves an improved performance label with a 90.41% accuracy score. In summary, the analysis of parameters in image classification and the proposed Bag of Visual Words (BoVW) demonstrate superior results.

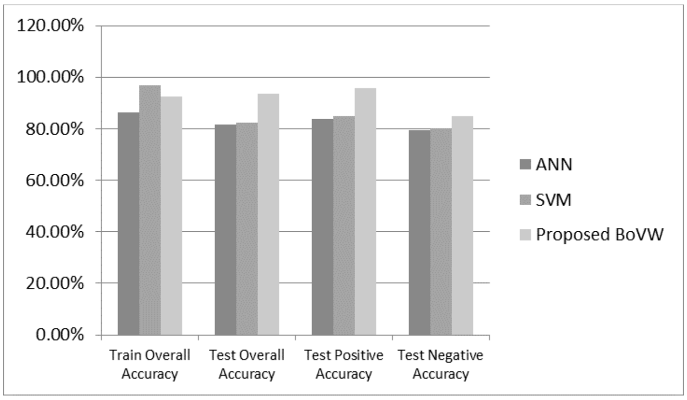


Figure 8: Performance Metrics of Different Classifiers

Figure 8, presents the matrices, and the accuracy of various classification models has been compared to determine the most appropriate one for breast cancer prediction. As illustrated in this figure, the BovW classifier exhibits the highest accuracy among the models compared.

V. CONCLUSION

This study enables the accurate detection and classification of breast cancer with enhanced precision 95.8% and specificity 95.17%. Various data processing techniques, including data cleaning, feature selection, classification, cross-validation, and fine-tuning, have been employed to ensure maximum accuracy. The study highlights that the BovW classifier achieves the highest accuracy 98% with a reduced subset of features as 93%. Additionally, Support Vector Machine is 92.4%, and Artificial Neural Network classifiers is 93.6% demonstrate reasonable performance in diagnosing breast cancer. It's noted that parameter selection plays a crucial role in correct classification, as multi-collinearity among attributes can compromise the effectiveness of the model. Cross-validation and fine-tuning are essential to prevent overfitting of the data. The analysis of the BoVW classifier's has been employed and various evaluation metrics have been scrutinized, including False Positive (FP), False Negative (FN), True Positive (TP), True Negative (TN), Negative Predictive Value (NPV), False Positive Rate (FPR), False Negative Rate (FNR), Positive Predictive Value (PPV), False Discovery Rate (FDR), Accuracy, Sensitivity, and Specificity. Within this evaluation, the Accuracy metric stands at 98.83% .Future research should focus on conducting prospective clinical trials and health economic evaluations to demonstrate the clinical utility and cost-effectiveness of automated breast cancer detection systems.

REFERENCES

[1] Swaminathan, Savitha, et al. "Breast cancer nutritional chemistry cachexia oncology–A clinical trials perspective." Indian Journal of Chemistry-Section A (IJCA) 59.9 (2022): 1369-1371.2021: 504-513, https://doi.org/10.56042/ijca.v59i9.41289.

[2] Gibb, Adam, Sarah J. Pirrie, Kim Linton, Victoria Warbey, Kathryn Paterson, Andrew J. Davies, Graham P. Collins et al. "Results of a UK National Cancer Research Institute Phase II study of brentuximabvedotin using a response-adapted design in the first-line treatment of patients with classical Hodgkin lymphoma unsuitable for chemotherapy due to age, frailty or comorbidity (BREVITY)." British Journal of Haematology 193, no. 1 (2023): 63-71, https://doi.org/10.1111/bjh.17073.

[3] Dorling, Leila, Sara Carvalho, Jamie Allen, Anna González-Neira, Craig Luccarini, Cecilia Wahlström, Karen A. Pooley et al. "Breast Cancer Risk Genes-Association Analysis in More than 113,000 Women." The New England journal of medicine 384, no. 5 (2021): 428-439, https://doi.org/10.1056/nejmoa1913948.

[4] Milton, Lauren, Tara Behroozian, Natalie Coburn, Maureen Trudeau, Yasmeen Razvi, Erin McKenzie, Irene Karam, Henry Lam, and Edward Chow. "Prediction of breast cancer–related outcomes with the Edmonton Symptom Assessment Scale: A literature review." Supportive Care in Cancer 29, no. 2 (2024): 595-603, https://doi.org/10.1007/s00520-020-05755-9.

[5] Vaka, Anji Reddy, Badal Soni, and Sudheer Reddy. "Breast cancer detection by leveraging Machine Learning." ICT Express 6, no. 4 (2020): 320-324, https://doi.org/10.1016/j.icte.2020.04.009.

[6] Meenalochini, G., and S. Ramkumar. "Survey of machine learning algorithms for breast cancer detection using mammogram images." Materials Today: Proceedings 37 (2021): 2738-2743, https://doi.org/10.1016/j.matpr.2020.08.543.

[7] Binder, Alexander, Michael Bockmayr, Miriam Hägele, Stephan Wienert, Daniel Heim, Katharina Hellweg, Masaru Ishii , "Morphological and molecular breast cancer profiling through explainable machine learning." Nature Machine Intelligence 3, no. 4 (2024): 355-366, https://doi.org/10.1038/s42256-021-00303-4.

[8] Li, Jiaxin, Zijun Zhou, Jianyu Dong, Ying Fu, Yuan Li, Ze Luan, and Xin Peng. "Predicting breast cancer 5-year survival using machine learning: a systematic review." PloS one 16, no. 4 (2021): e0250370, https://doi.org/10.1371/journal.pone.0250370.

[9] Abbas, Shafaq, ZuneraJalil, Abdul RehmanJaved, IqraBatool, Mohammad Zubair Khan, AbdulfattahNoorwali, Thippa Reddy Gadekallu, and Aqsa Akbar. "BCD-WERT: a novel approach for breast cancer detection using whale optimization based efficient features and extremely randomized tree algorithm." PeerJComputer Science 7 (2021): e390, https://doi.org/10.7717/peerj-cs.390.

[10] Sharma, Ajay, and Pramod Kumar Mishra. "Performance analysis of machine learning based optimized feature selection approaches for breast cancer diagnosis." International Journal of Information Technology (2021): 1-12, https://doi.org/10.1007/s41870-021-00671-5.

[11] Kumari, D. Jaya. "Structural redesign of artificial neural network for predicting breast cancer with the aid of artificial bee colony." Indian Journal of Science and Technology (2017), Vol 10(15), DOI: 10.17485/ijst/2017/v10i15/108270.

[12] Bartsch, R, “ASCO 2023: highlights in breast cancer”. memo (2023). https://doi.org/10.1007/s12254-023-00936-8.

[13] Abelman RO, Spring L, Fell GG, “Sequential use of antibody-drug conjugate after antibody-drug conjugate for patients with metastatic breast cancer: ADC after ADC (A3) study” Journal of Clinical study, 2023;41(16):1022, https://doi.org/10.1200/JCO.2023.41.16_suppl.1022.

[14]. Frenel JS, Zeghondy J, Guerin C, “Efficacy of tucatinib+trastuzumab+capecitabine (TTC) after trastuzumab-deruxtecan (T-DXd) exposure in her2-positive metastatic breast cancer: a French multicentre retrospective study”, J ClinOncol. 2023;41(16):1014, https://doi.org/10.1200/JCO.2023.41.16_suppl.1014.