ISSN No:-2456-2165

How can Machine Learning be used to Classify Breast Cancer?

Krish Kapoor

Abstract - Breast cancer is a prevalent form of cancer that affects a significant number of individuals worldwide and can have severe consequences if not detected and treated early. The World Health Organization (WHO) estimates that breast cancer is the most common cancer among women globally, with an estimated 2.3 million new cases in 2020 alone. Early detection is crucial in improving survival rates and treatment outcomes. This paper explores the application of Machine Learning (ML) techniques for predicting breast cancer diagnosis in individuals. We utilize a publicly available dataset from the Kaggle machine learning repository, which contains data from breast cancer patients collected from various medical institutions. Several machine learning models, including Naive Bayes Algorithm, Decision Trees, Logistic Regression, Neural Networks, Random Forest, Stochastic Gradient, and Support Vector Machines, are employed to analyze the dataset. The performance of these models is assessed using 10-fold cross-validation. Furthermore, we propose the most suitable machine learning algorithm for breast cancer diagnosis based on specified input parameters and discuss the potential deployment of a breast cancer diagnostic tool.

***Keywords:-** Breast Cancer Detection, Supervised and Unsupervised Machine Learning, Artificial Intelligence.*

I. INTRODUCTION

Breast cancer is a prevalent and life-threatening disease that requires accurate classification for effective diagnosis and treatment planning. **[8]** The ability to classify breast tumors into malignant (cancerous) or benign (non-cancerous) categories is essential for determining the appropriate course of action. **[8]** Artificial intelligence (AI) and Machine Learning (ML) algorithms have shown great promise in assisting with the classification of breast cancer, providing accurate and efficient tools for healthcare professionals.

Traditionally, the classification of breast tumors relied on histological examination by pathologists, which is time-consuming and subject to inter-observer variability. **[9]** However, with the advancement of AI and machine learning, the development of automated systems for breast cancer classification has become possible. These systems leverage large datasets containing tumor features and corresponding diagnoses to learn patterns and make accurate predictions.

Machine learning algorithms, such as Naive Bayes, Decision Trees, Logistic Regression, Neural Networks, Random Forest, Stochastic Gradient, and Support Vector Machines, have been applied to breast cancer classification tasks. These algorithms analyze features extracted from breast tumor images, including size, shape, texture, and other characteristics, to differentiate between malignant and benign tumors. By learning from historical data, these algorithms can make predictions on new, unseen cases, aiding in the early detection and management of breast cancer.

In this paper, we focus on the classification of breast cancer using AI and machine learning algorithms. We utilize a publicly available dataset, such as the one provided by the Kaggle machine learning repository, which contains information on various tumor features and corresponding diagnoses. This dataset serves as the basis for training and evaluating the performance of different machine-learning models.

The dataset is pre-processed to handle missing values, normalize features, and split into training and testing sets. We then apply a range of machine learning algorithms to the training data, allowing them to learn from the patterns and relationships within the dataset. The performance of each algorithm is evaluated using appropriate metrics such as accuracy, precision, recall, and F1 score.

The goal of this research is to identify the most accurate machine-learning algorithm for breast cancer classification. The selected algorithm can then be used as a reliable tool for assisting healthcare professionals in diagnosing breast cancer cases. Early and accurate classification enables timely intervention, personalized treatment plans, and improved patient outcomes.

II. LITERATURE REVIEW

- Breast cancer type classification using machine learning - The study evaluated four machine learning algorithms for classifying breast cancer into triple negative and non-triple negative types. Among these algorithms, the Support Vector Machine (SVM) demonstrated higher accuracy and fewer misclassification errors compared to the other three algorithms. The findings suggest that machine learning algorithms, particularly SVM, are effective for accurately classifying breast cancer into triple negative and non-triple negative types.
- Breast cancer classification using machine learning - This paper discusses the significance of breast cancer classification due to its prevalence and high mortality rates. The study focuses on the application of machine learning techniques for breast cancer classification, specifically comparing two classifiers: Naive Bayes (NB) and k-nearest neighbor (KNN). The authors evaluate the accuracy of these classifiers using cross-validation and find that KNN achieves the highest accuracy (97.51%) with the lowest error rate, followed by the NB classifier (96.19%). This comparison underscores the effectiveness

of machine learning in accurately classifying breast cancer, supporting its potential for diagnostic applications.

- Evaluating the Performance of Machine Learning Techniques in the Classification of Wisconsin Breast Cancer - This paper addresses the importance of accurate diagnosis in distinguishing between malignant and benign breast tumors due to the significant impact of breast cancer on women's health and mortality rates. The study focuses on the application of three machine learning algorithms (Support Vector Machine, K-nearest neighbors, and Decision tree) for breast cancer classification. Using the Wisconsin Breast Cancer (Diagnostic) dataset, the study compares the performance of these classifiers to determine the most effective one in terms of accuracy. The findings reveal that the quadratic support vector machine achieves the highest accuracy (98.1%) with the lowest false discovery rates. This research contributes to the literature by highlighting the superior performance of the quadratic support vector machine in breast cancer classification, demonstrating the potential of machine learning algorithms in this domain.
- Machine learning techniques to diagnose breast cancer - This paper explores the application of machine learning techniques in cancer diagnosis, specifically for distinguishing between benign and malignant breast tumors. The study combines support vector machines, K-nearest neighbors, and probabilistic neural network classifiers with various feature ranking, selection, and extraction methods. The research achieves high accuracy in breast cancer diagnosis, with support vector machine classifiers attaining an overall accuracy of 98.80% and 96.33% on two commonly used breast cancer benchmark datasets. This study demonstrates the effectiveness of machine learning algorithms in accurately differentiating between benign and malignant breast tumors, offering valuable insights into the field of cancer diagnosis.
- The above research primarily investigates the accuracy of machine learning models in classifying malign and benign breast cancer, comparing each model's performance to determine the most effective one. This study not only contributes to the existing body of knowledge on the application of machine learning in healthcare but also holds the potential to advance the medical industry. The observations and analyses from this research could significantly impact future medical investigations, facilitating quicker and more accurate diagnoses of breast cancer.

III. METHODOLOGY

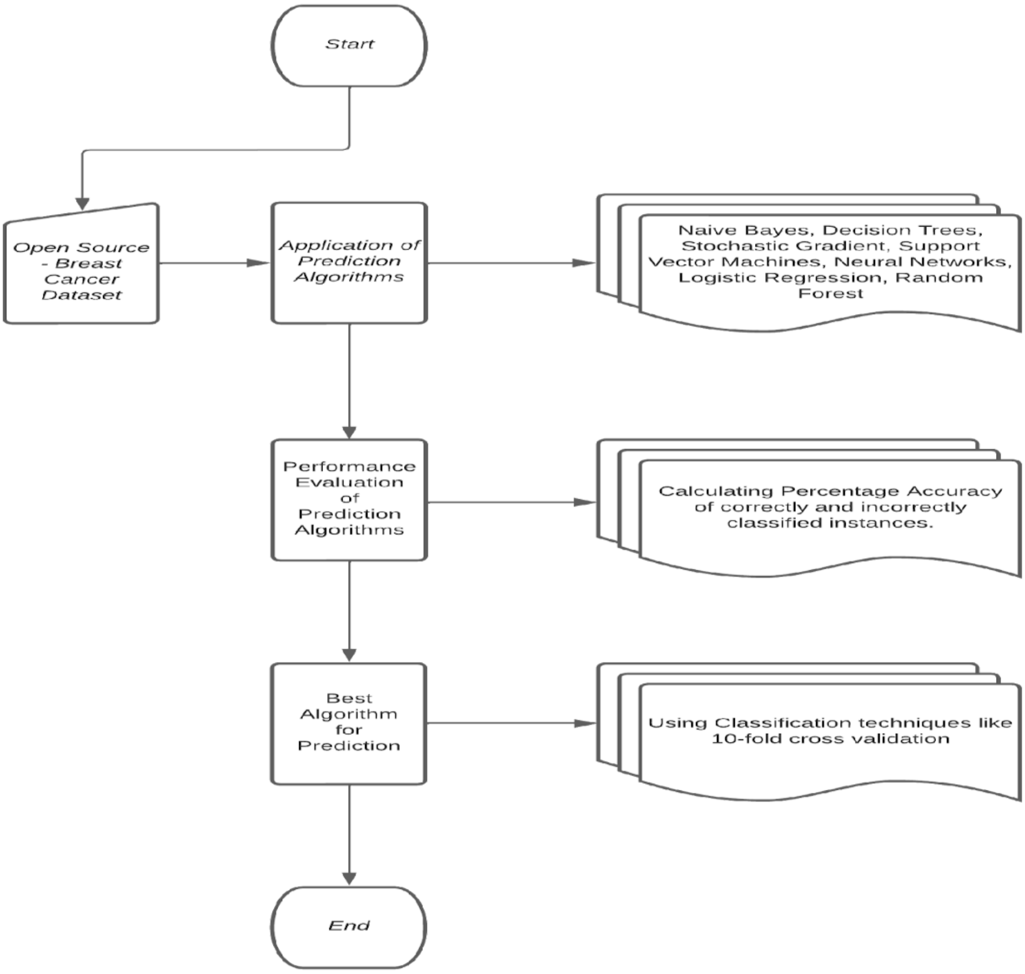
Start
Open Source - Breast Cancer Dataset
Application of Prediction Algorithms
Naive Bayes, Decision Trees, Stochastic Gradient, Support Vector Machines, Neural Networks, Logistic Regression, Random Forest
Performance Evaluation of Prediction Algorithms
Calculating Percentage Accuracy of correctly and incorrectly classified instances.
Best Algorithm for Prediction
Using Classification techniques like 10-fold cross validation
End

Fig. 1: Proposed System Architecture

A. *Proposed System Architecture*

The proposed system architecture is shown in the underlying figure. The dataset containing the information about the symptoms of the patients will be fed to the prediction algorithms like Naive Bayes, Decision Trees, Logistic Regression, Support Vector Machines, Neural Networks, Stochastic Gradient, and Random Forest algorithms. Then, the performance of the algorithms will be tested with an appropriate evaluation model, in particular, 10-fold Cross-validation. I will then choose the best algorithm to build the system for the end users using the dataset as Database. The tool will take the symptoms from the user as input and will display and classify whether the user has breast cancer or not.

B. *Dataset Details*

This dataset contains reports of breast-cancer symptoms of 570 persons. We have taken this dataset from the Kaggle machine learning repository.

Link: https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset?resource=download

Table 1: Description of Dataset

	Number of Attributes	**Number of Instances**
Breast Cancer Symptom Dataset	30	570

Table 2: Description of Attributes

Attributes	
Radius_Mean	Smoothness_Se
Texture_Mean	Compactness_Se
Perimeter_Mean	Concavity_Se
Area_Mean	Concave Points_Se
Smoothness_Mean	Symmetry_Se
Compactness_Mean	Fractal_Dimension_Se
Concavity_Mean	Radius_Worst
Concave Points_Mean	Texture_Worst
Symmetry_Mean	Perimeter_Worst
Fractal_Dimension_Mean	Area_Worst
Radius_Se	Smoothness_Worst
Texture_Se	Compactness_Worst
Perimeter_Se	Concavity_Worst
Area_Se	Concave Points_Worst
Fractal_Dimension_Worst	Symmetry_Worst

Table 3: Dataset Details

	Benign	**Malignant**
Diagnosis	0 (Negative)	1 (Positive)

IV. RESULTS

Performance of different Data Mining techniques on our dataset with detailed accuracy, information is represented in the following tables.

Table 4: Comparison of Evaluation Metrics using 10-Fold Cross Validation

Evaluation Metrics	**Cross Validation**						
Model	**NB**	**LR**	**DT**	**NN**	**RF**	**SG**	**SVM**
Total Number of Instances	570	570	570	570	570	570	570
Correctly Classified Instances	537	540	542	540	544	454	521
	94.2%	94.7%	95.0%	94.7%	95.4%	79.6%	91.4%
Incorrectly Classified Instances	33	30	28	30	26	116	49
	5.8%	5.3%	5.0%	5.3%	4.6%	20.4%	8.6%

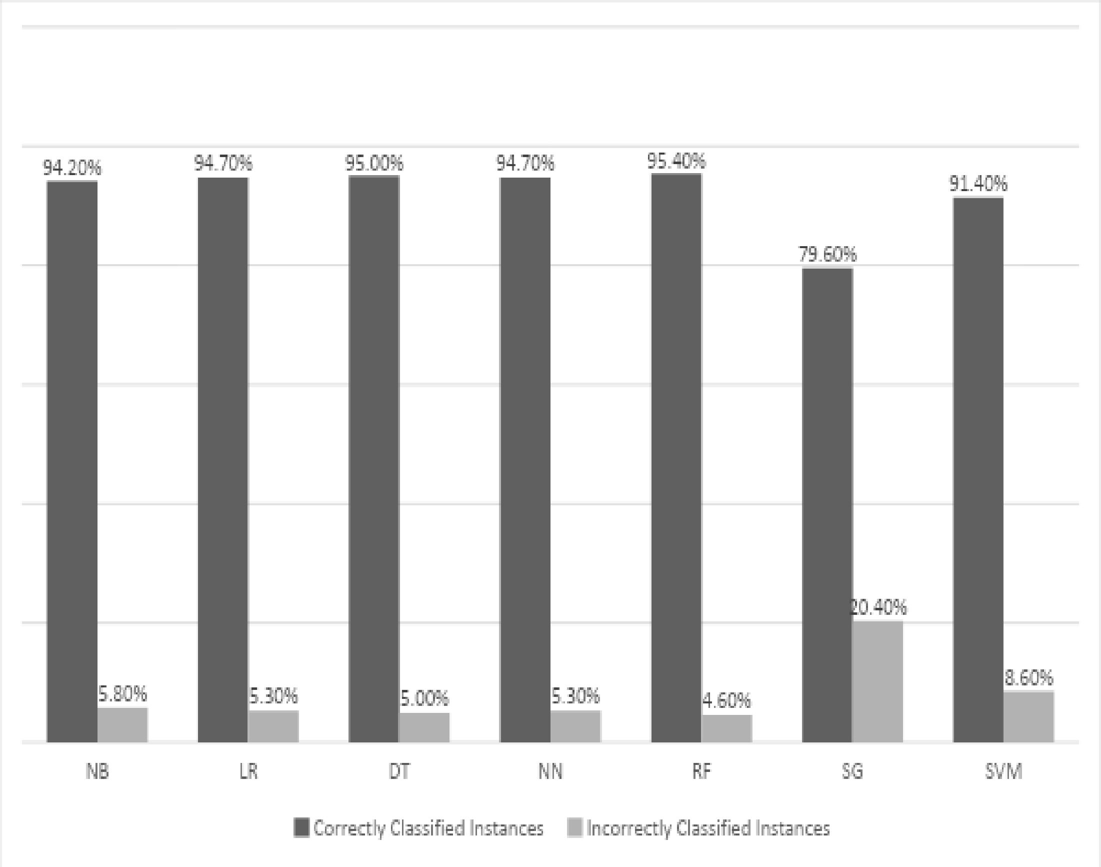

Fig. 2: Performance of Classification Algorithms Using Cross-Validation Technique

Table 5: Comparison of Performance Parameters using 10-Fold Cross Validation

Performance Parameters	**Class**	**Weighted Average**						
		NB	**LR**	**DT**	**NN**	**RF**	**SG**	**SVM**
Precision	**Positive (Malignant)**	0.92	0.90	0.92	0.93	0.93	1.00	0.99
	Negative (Benign)	0.95	0.90	0.91	0.94	0.91	0.66	0.85
	Weighted Average	0.94	0.90	0.91	0.94	0.92	0.79	0.90
Recall	**Positive (Malignant)**	0.92	0.83	0.87	0.90	0.85	0.13	0.72
	Negative (Benign)	0.95	0.95	0.94	0.96	0.96	1.00	0.99
	Weighted Average	0.94	0.90	0.91	0.94	0.92	0.67	0.89
F-measure	**Positive (Malignant)**	0.92	0.86	0.90	0.91	0.89	0.24	0.83
	Negative (Benign)	0.95	0.92	0.93	0.95	0.94	0.79	0.92
	Weighted Average	0.94	0.90	0.91	0.94	0.92	0.58	0.89

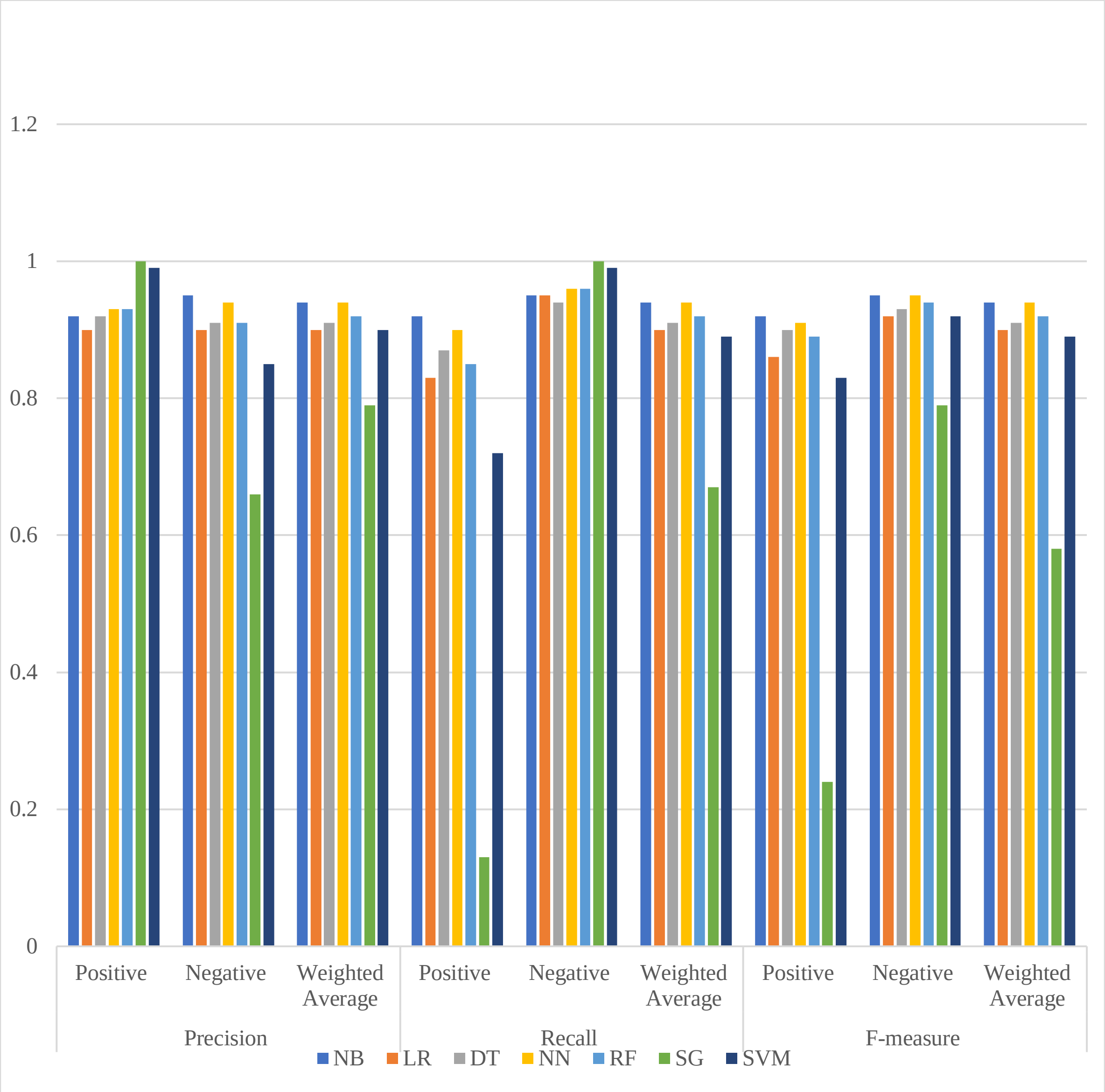

Fig. 3: Performance of Classification Algorithms Using Cross-Validation Technique

Table 4 shows us the pure accuracy of each model using 10-fold cross-validation. We can clearly see that the Random Forest model classified the greatest number of instances correctly with 544 correct instances out of 570 (95.4% accuracy). This is followed closely by Decision Trees, classifying 542 instances correctly with a 95.0% accuracy. The least accurate models were Support Vector Machines and Stochastic Gradient, classifying 521 and 454 instances correctly respectively.

Table 5 shows us the precision, recall, and f-scores of each model. The models with the highest average precision scores (proportion of positively predicted labels that are actually correct) are Naive Bayes, Neural Networks, and Random Forest. These three models also have the highest average recall scores (the ability to correctly predict the positives out of actual positives). Moving on to f-scores (mean of a system's precision and recall values), the same 3 models appear again.

In statistics, precision, recall, and F-measure are common metrics used to evaluate the performance of a classification model. Precision measures the proportion of true positives (TP) among the instances that are predicted as positive (TP + false positives, FP), and thus reflects the accuracy of the positive predictions. Recall, on the other hand, measures the proportion of true positives among the instances that are actually positive (TP + false negatives, FN), and thus reflects the completeness of the positive predictions.

In classifying malignant and benign breast cancer using machine learning, false positives, and false negatives have different consequences. A false positive occurs when the model predicts malignancy when the tumor is benign, leading to unnecessary procedures and anxiety. A false negative occurs when the model predicts benignity when the tumor is malignant, resulting in delayed diagnosis and treatment. Balancing precision and recall is crucial, prioritizing recall if the cost of false negatives is higher and precision if the cost of false positives is higher. High recall is often favored in medical settings to minimize missed malignant cases, but the specific application and potential consequences should be considered to determine the optimal trade-off between precision and recall.

V. DISCUSSION

A. Result Analysis

The best result was achieved using Random Forest Algorithm where using 10-fold cross-validation, 95.4% of instances were classified correctly. It also had the highest average precision, recall, and f-measure percentages. In the figure above, the performance of the algorithms using Cross-validation evaluation is depicted.

B. Proposed Tool

Based on the study's findings, a user-friendly tool that utilizes machine learning algorithms is proposed, specifically the Random Forest, to classify malignant and benign breast cancer. This tool would allow individuals to input relevant medical information and receive a prediction regarding the nature of their breast tumor. By harnessing the power of machine learning, this tool aims to provide an accurate and convenient solution for predicting breast cancer risk. Given the increasing prevalence of breast cancer globally, this tool would empower individuals to monitor their health proactively. The intuitive design and user-friendly interface would ensure that users can easily interpret the results and take appropriate actions.

By leveraging this technology, individuals can self-assess their breast tumor's nature and subsequently seek medical advice from a healthcare professional. This approach saves time and resources, enabling healthcare providers to focus on cases requiring immediate attention. Moreover, in regions where breast cancer poses a significant health challenge, this tool can alleviate the strain on healthcare systems by enabling individuals to self-diagnose and manage their condition proactively. This not only benefits the individual but also helps ensure timely and adequate medical care while relieving the burden on healthcare systems.

VI. CONCLUSION

In this paper, we utilized open-source machine learning algorithms on a public dataset to classify breast tumors as malignant or benign. Through evaluation using metrics such as accuracy, precision, recall, and F1 score, we found that machine learning models, specifically Random Forest, demonstrated high accuracy in classifying breast cancer. This has significant implications for early detection, treatment planning, and improved patient outcomes. We proposed a classification tool that utilizes these models to aid healthcare professionals in making informed decisions. While limitations exist, our study highlights the potential of machine learning in accurately classifying breast cancer, paving the way for enhanced diagnostic accuracy and more effective treatment strategies.

REFERENCES

[1.] Wu, J., & Hicks, C. (2021). Breast cancer type classification using machine learning. Journal of Personalized Medicine, 11(2), 61.

[2.] Amrane, M., et al. (2018). Breast cancer classification using machine learning. In 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT). IEEE.

[3.] Obaid, O. I., et al. (2018). Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer. International Journal of Engineering & Technology, 7(4.36), 160-166.

[4.] Osareh, A., & Shadgar, B. (2010). Machine learning techniques to diagnose breast cancer. In 2010 5th International Symposium on Health Informatics and Bioinformatics. IEEE.

[5.] World. (2023, July 12). Breast Cancer. Who.int. World Health Organization: WHO. URL: www.who.int/news-room/fact-sheets/detail/breast-cancer.

[6.] Early Detection of Breast Cancer: Importance of Regular Self-Exams and Mammograms. (2023). Medanta.org. URL: www.medanta.org/patient-education-blog/early-detection-of-breast-cancer-importance-of-regular-self-exams-and-mammograms/#:~:text=Breast%20cancer%20affects%20millions%20of,when%20it%20is%20most%20treatable.

[7.] Yasser, M. (2015). Breast Cancer Dataset. Kaggle.com. URL: www.kaggle.com/datasets/yasserh/breast-cancer-dataset?resource=download.

[8.] Łukasiewicz, S., et al. (2021, August 25). Breast Cancer-Epidemiology, Risk Factors, Classification, Prognostic Markers, and Current Treatment Strategies-an Updated Review. Cancers. URL: www.ncbi.nlm.nih.gov/pmc/articles/PMC8428369/.

[9.] Ginter, P. S., et al. (2021, April). Histologic Grading of Breast Carcinoma: A Multi-Institution Study of Interobserver Variation Using Virtual Microscopy. Modern Pathology : An Official Journal of the United States and Canadian Academy of Pathology, Inc. URL: www.ncbi.nlm.nih.gov/pmc/articles/PMC7987728/.