ADVANCEMENTS IN BREAST CANCER DETECTION: HARNESSING THE POWER OF MACHINE LEARNING ALGORITHMS

Hasna M[1], Gaya P Nair[2]

[1]*Assistant Professor Department of Computer Science and Engineering,AWH Engineering College Calicut,kerala*

[2]*Engineering Research Scholar,Department of Computer Science and Engineering,Karpagam Academy of Higher Education,Coimbatore,Tamilnadu*

***Abstract*—Breast Cancer (BC) stands as a pervasive global health concern and remains the foremost cause of mortality among women worldwide. Notably, a significant proportion of individuals afflicted by breast cancer lack any familial predisposition. The escalating incidence of breast cancer underscores the urgency of addressing it as a prevalent public health issue. With women facing approximately a 1/8 probability of a breast cancer diagnosis, factors such as aging, genetic predisposition, dense breast tissues, obesity, and exposure to radiation contribute to elevated risk. Distinguishing between malignant and benign tumours is paramount for accurate diagnosis, necessitating a reliable diagnostic procedure. Mammography currently serves as the primary method for breast cancer detection, supplemented by Fine Needle Aspiration Cytology (FNAC) in the diagnostic process. This paper explores the integration of machine learning models for early prediction of breast cancer, recognizing the pivotal role of early diagnosis in enhancing treatment success rates. The study compares the prediction accuracy and performance parameters of various machine learning algorithms, including Support Vector Machine (SVM), Logistic Regression, Decision Tree, and K Nearest Neighbours (KNN).). Through a comprehensive comparative analysis, the research aims to identify the most effective algorithm based on classifier performance, contributing valuable insights to the optimization of breast cancer detection methodologies.**

***Index Terms*—Fine Needle Aspiration Cytology, Convolutional Neural Network, Support Vector Machine, Mammography**

I. INTRODUCTION

Breast cancer, diagnosed in a staggering 7.8 million women, stands as a formidable global health challenge, as reported by the World Health Organization (WHO). This disease instigates the unbridled proliferation of cells within the breast, positioning it as the second most perilous cancer following lung cancer [2]. The inherent risk of breast cancer emanates from the aberrant growth of cells in the breast tissues, with a proclivity for invasion into neighbouring tissues.

The classification of breast cancer primarily hinges on the identification of anomalies, either through the palpation of lumps or via mammogram screening. The distinction between benign and malignant tumours, indicative of abnormal cellular growth, is pivotal for accurate diagnosis. In the past, Xrays were the cornerstone of breast cancer detection; however, the landscape has evolved. Contemporary diagnostic methodologies have embraced Fine Needle Aspiration Cytology (FNAC) alongside traditional techniques [7]. FNAC, a minimally invasive procedure, has gained widespread adoption, enriching the diagnostic arsenal for breast cancer detection.

Breast cancer poses a substantial challenge due to its potential for rapid progression and invasive behaviour. Recognizing these challenges, there is an ongoing imperative to enhance diagnostic accuracy and efficiency. Early detection significantly impacts prognosis, prompting continual refinement and augmentation of diagnostic tools and methodologies. This narrative sets the stage for a broader exploration into the evolving landscape of breast cancer diagnosis, emphasizing the transition from conventional Xray methods to the contemporary inclusion of FNAC, reflecting the continuous pursuit of precision in the face of this formidable health crisis [5]. It represents a scientific discipline dedicated to the creation and refinement of algorithms rooted in data analysis. Aiming to conduct a comparative analysis of various machine learning algorithms for both prediction and diagnosis, this study focuses on the application of ML in the context of breast cancer. To predict and diagnose breast cancer, the primary objective of this research is the implementation of diverse machine learning algorithms. The significance of this study lies in its potential to contribute to improved patient outcomes by enhancing the accuracy and efficiency of breast cancer prediction and diagnosis.[14] Employing a comparative study design, the research analyses the performance of different machine learning algorithms tailored for breast cancer prediction. Scrutinized for their predictive capabilities are algorithms such as Support Vector Machine (SVM), Logistic Regression, Decision Tree, and K Nearest Neighbours (KNN). Performance metrics, including accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUCROC), are considered for gauging the effectiveness of each algorithm.

The study anticipates revealing insights into breast cancer prediction and diagnosis by comparing the efficacy of various machine learning algorithms [15]. By elucidating the strengths and limitations of each classifier, the research aims to provide a foundation for future advancements in the integration of ML in healthcare, specifically in the domain of breast cancer management. As technology continues to evolve, the intersection of machine learning and healthcare holds immense promise. Through the judicious application of machine learning algorithms, this research endeavours to contribute valuable knowledge to the ongoing efforts aimed at refining and optimizing breast cancer prediction and diagnosis

II. METHODS

Breast cancer indeed poses a significant health challenge globally, including in countries like India. Early detection is crucial for improving survival rates, and integrating data science solutions can play a vital role in addressing this issue. Here are several ways in which data science can contribute to the early recognition and management of breast cancer:

Machine Learning for Diagnosis: Machine learning algorithms can analyze mammography images and other diagnostic data to assist in the early detection of breast cancer. These algorithms can learn patterns indicative of cancerous growths, improving the accuracy of diagnosis. Absolutely, machine learning (ML) algorithms have shown promising results in the analysis of medical imaging data, such as mammography images, for the early detection of breast cancer. Here are some key ways in which machine learning is being applied in this context:

Image Classification: ML algorithms can be trained to classify mammography images into categories such as benign, suspicious, or malignant. A type of deep learning algorithm, Convolutional Neural Networks (CNNs) have demonstrated effectiveness in learning and recognizing patterns indicative of cancerous growths.

Feature Extraction: ML algorithms, will automatically extract the relevant features from mammography images that might be challenging for human observers to discern. These features could include irregularities in shape, texture, or the presence of microcalcifications.

Risk Stratification: Machine learning models can assist in stratifying patients based on their risk of developing breast cancer. By analyzing a combination of clinical data and imaging results, these models can identify individuals who may benefit from more frequent screenings or additional diagnostic tests.

Computer Aided Detection (CAD): To assist radiologists in interpreting mammography images, computer aided detection systems integrate ML algorithms. CAD systems can automatically highlight areas of interest that may require closer examination, aiding in the early identification of potential abnormalities.

Continual Learning: ML models can be designed for continual learning, adapting to new data and evolving in their ability to detect patterns associated with breast cancer. This adaptability is valuable in the dynamic field of medical diagnostics where new insights and data are continually emerging.

Reducing False Positives and Negatives: ML algorithms aim to reduce false positives (incorrectly identifying a noncancerous condition as cancer) and false negatives (missing the presence of cancer). By learning from large datasets, these algorithms can improve the overall accuracy of breast cancer detection.

Integration with Clinical Data: Combining imaging data with other clinical data, such as patient history and genetic information, can enhance the accuracy of machine learning models. This holistic approach allows for a more comprehensive understanding of an individual's risk profile.

Interpretable AI: Efforts are being made to develop interpretable AI models that can provide insights into the decision-making process of the algorithm. This transparency is crucial for gaining trust among healthcare professionals and ensuring responsible deployment in clinical settings. While machine learning holds great potential in improving the early detection of breast cancer, it's important to note that these algorithms are tools to assist healthcare professionals rather than replace them. Collaboration between AI systems and human expertise can lead to more accurate and timely diagnoses, ultimately benefiting patient outcomes. This analysis helps to observe which features are most helpful in predicting malignant or benign cancer.

Using Google Collab as the integrated development environment provides the advantage of working in a cloud-based environment, allowing to run Python code and machine learning models without the need for high-end local hardware. It also provides access to GPU resources for faster training of machine learning models. Additionally, the minimum hardware requirements mentioned here are suitable for general Python programming and basic machine learning tasks. Adjustments may be needed based on the scale and complexity of the specific projects.

A. Dataset

The project is based on the Wisconsin Diagnosis Breast Cancer dataset, comprising 569 instances and 32 attributes which extracted from the digitized images of fine needle aspirates (FNAs) of the breast masses. These attributes describe characteristics of the cell nuclei present in the images. There are no missing values present in the dataset. The output variable is binary, indicating either benign or malignant diagnosis. Class distribution: 357 benign, 212 malignant

Table 1: Wisconsin Diagnosis Breast Cancer (WDBC)

Dataset	No. of Attributes	No. of Instances	No. of Classes
WDBC	32	569	2

To create the dataset, Dr. Wolberg used fluid samples taken from patients with solid breast masses and a graphical computer program called Xcyt. The influential variables include diagnosis, radius, texture, perimeter, area, symmetry, concave points, etc. In this dataset, the positive class is used for benign cases, while the negative class is used for malignant cases. The attribute information for the Wisconsin Diagnosis Breast Cancer dataset is: ID number: Unique identification number for each instance. Diagnosis (Target Variable):M = malignant B = benign. Radius (mean): Mean of distances from the center to points on the perimeter. Perimeter: Perimeter of the cell nucleus. Area: Area of the cell nucleus. Smoothness: Local variation in radius lengths. These features capture various aspects of the characteristics of cell nuclei present in the digitized images of fine needle aspirates (FNAs) of breast masses. They provide a quantitative description of cell morphology, which can be valuable for predicting whether a breast mass is malignant or benign. When working with this dataset, you can use the features to train machine learning models to classify instances as either malignant (M) or benign (B) based on the observed characteristics.

B. Methodology

Here, we have chosen to apply Support Vector Machine (SVM), Decision Tree, Logistic Regression, and K Nearest Neighbors (KNN) on the Breast Cancer Wisconsin Diagnostic dataset to identify the most efficient algorithm for the breast cancer detection using machine learning. In order to build and assess machine learning algorithms for breast cancer prediction, this methodology offers a methodical and comprehensible approach. Let us go through each step:

Data Acquisition: The initial step involves obtaining the Breast Cancer Wisconsin Diagnostic dataset, which serves as the raw data for the analysis.

Preprocessing: Preparing the data for analysis and model training. In the context of breast cancer prediction using the Breast Cancer Wisconsin Diagnostic dataset, here are common preprocessing steps you might consider

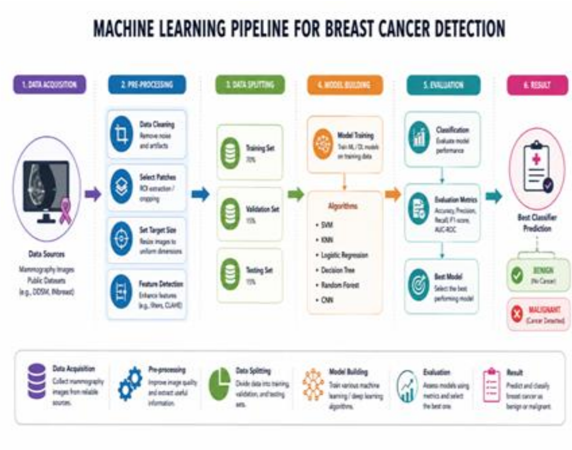

Fig 1: Process Flow Diagram

a. Data Cleaning: This step involves cleaning the dataset to ensure its quality and integrity. This may include handling missing values, dealing with outliers, and addressing any inconsistencies in the data.
b. Select Attributes: The selection of relevant attributes is crucial for model performance. This step likely involves assessing the importance of each feature and deciding which ones to include in the model.
c. Set Target Role: Identifying and designating the target variable, which in this case is likely the "Diagnosis" (malignant or benign). This variable is what your machine learning models will aim to predict.
d. Features Extraction: Extracting relevant features from the dataset, which could involve additional processing or transformation of existing features to enhance their predictive power.

Machine Learning Algorithm Construction: Using the prepared data to build machine learning algorithms for breast cancer prediction. It would seem reasonable to investigate a variety of algorithms, such as Support Vector Machine (SVM), Decision Tree, Logistic Regression, and K Nearest Neighbors (KNN), in order to determine which, one works best for this particular task.

Model Evaluation: Evaluating the performance of the constructed machine learning models. This is typically done by using a separate set of data (not seen during training) for which you have labels. The models predict the labels, and their predictions are compared to the true labels to assess accuracy and other relevant metrics.

We typically achieve this by using the Traintest split method to divide the labeled data into two halves. The machine learning model is then constructed using 75 percentage of the data, referred to as the training data. The remaining quarter of the data, known as the testing set, is used to evaluate the performance of the model. By testing the models and evaluating the results, we can determine which algorithm has the highest accuracy and is the most predictive for detecting breast cancer

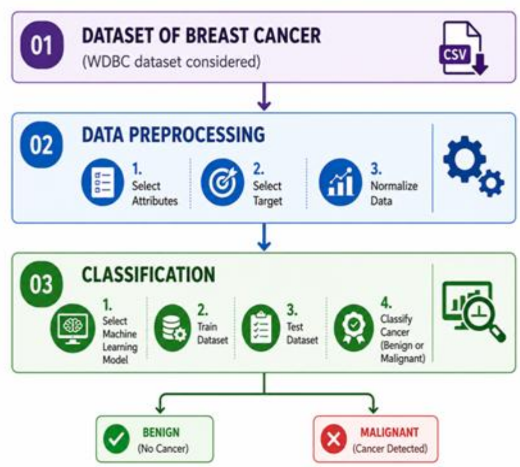

Fig 2: Methodology

III. RESULTS AND DISCUSSION

This part aids the three classifiers under investigation in this study by outlining the parameters and presenting the findings.

Accuracy: It can be characterized as the ratio of the right forecast to the incorrect one.

Precision: The model's accuracy in classifying a sample as positive is evaluated.

F1 Score: Like all other error metrics, the F1 score is a classification error statistic that helps assess algorithm performance. In terms of binary classification, it enables us to evaluate the performance of the machine learning model.

Table 2: RESULT ANALYSIS

Sl No	Algorithm	Accuracy Score	AUC Score
1	KNearest Neighbours (KNN)	0.953	0.941
2	Support Vector Machine (SVM)	0.968	0.957
3	Decision Tree	0.954	0.939
4	Logistic Regression	0.947	0.943

IV CONCLUSION

The research paper has successfully demonstrated the significance of early detection in combating breast cancer, given its high prevalence with a 12 percentage of chance for a randomly chosen woman to be diagnosed with the disease. Using the Wisconsin Breast Cancer Diagnostic dataset (WBCD), the study conducted a comparative analysis of major machine learning algorithms Support Vector Machine (SVM), Logistic Regression, Decision Tree, and K Nearest Neighbors (KNN). The evaluation criteria included essential metrics such as accuracy, precision, Area Under the Curve (AUC), and confusion matrix. Notably, all four algorithms achieved an accuracy of more than 90 percentage, indicating their effectiveness in breast cancer detection. The study concludes that Support Vector Machine (SVM) outperformed the other algorithms, demonstrating superior efficiency in breast cancer prediction and diagnosis. The choice of SVM as the most effective algorithm is based on its ability to provide both high accuracy and precision. The potential of machine learning in enhancing breast cancer diagnostic capabilities is underscored by these findings. This study nonetheless contributes valuable insights into the application of machine learning for breast cancer detection, paving the way for improved early intervention and, consequently, saving lives.

Table 3: Evaluation Metrics

Algorithm	Precision		Recall		F1 Score		Support	
	Malignant	Benign	Malignant	Benign	Malignant	Benign	Malignant	Benign
K-NN	0.93	0.95	0.97	0.87	0.95	0.91	108	63
SVM	0.97	0.95	0.91	0.99	0.94	0.97	43	71
Logistic Regression	0.93	0.96	0.93	0.96	0.93	0.96	43	71
Decision Tree	0.96	0.97	0.98	0.94	0.97	0.95	108	63

REFERENCES

[1] K. S. Hughes, J. Zhou, and Y. Bao, "Natural language processing to facilitate breast cancer research management," International Journal of Cancer.

[2] J. M. C. Ortiz-Rodriguez, M. Guerrero-Mendez, M. D. R. Martinez-Blanco, M. Moreno-Lucio, M. Jaramillo-Martinez, and L. Garza-Veloz, "Breast cancer detection by means of artificial neural networks," 2018. DOI: 10.5772/intechopen.71256

[3] M. S. Yarabarla, L. K. Ravi, and A. Sivasangari, "Breast cancer prediction via machine learning," in Proc. Third Int. Conf. on Trends in Electronics and Informatics (ICOEI), 2019.

[4] S. Nayak and D. Gope, "Comparison of supervised learning algorithms for RF-based breast cancer detection," in Proc. Computing and Electromagnetics International Workshop (CEM), Barcelona, Spain, 2017.

[5] S. Vishnu, O. Chantarakasemchit, and P. Meesad, "A breakup machine learning approach for breast cancer prediction," in Proc. 11th Int. Conf. on Information Technology and Electrical Engineering, 2019.

[6] V. Chaurasia, S. Pal, and B. B. Tiwari, "Prediction of benign and malignant breast cancer using data mining techniques," Journal of Algorithms & Computational Technology, vol. 12, no. 2, pp. 119–126, 2018.

[7] S. A. Mohammed, S. Darrab, S. A. Noaman, et al., "Analysis of breast cancer detection using different machine learning techniques," in Proc. Int. Conf. on Data Mining and Big Data, Singapore: Springer, 2020, pp. 108–117.

[8] M. R. Alhadidi, A. Alarabeyyat, and M. Alhanahnah, "Breast cancer detection using k-nearest neighbor machine learning algorithm," in

Proc. 9th Int. Conf. on Developments in eSystems Engineering, 2016, pp. 35–39.
[9] L. Liu, "Research on logistic regression algorithm of breast cancer diagnose data by machine learning," in Proc. Int. Conf. on Robots and Intelligent System, 2018, pp. 157–160.
[10] M. Rana, P. Chandorkar, A. Dsouza, and N. Kazi, "Breast cancer diagnosis and recurrence prediction using machine learning techniques," International Journal of Research in Engineering and Technology, vol. 4, no. 4, 2015.
[11] B. Dai, R. C. Che, S. Zhu, and W. Zhang, "Using random forest algorithm for breast cancer diagnosis," in Proc. Int. Symp. on Computer, Consumer and Control, 2018, pp. 449–452.
[12] E. M. F. El Houby, "A survey on applying machine learning techniques for management of diseases," Journal of Applied Biomedicine, pp. 165–174, 2018.
[13] B. M. Gayathri and C. P. Sumathi, "Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer," 2016.
[14] H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," Procedia Computer Science, vol. 83, pp. 1064–1069, 2016. DOI: 10.1016/j.procs.2016.04.224
[15] A. H. Osman, "An enhanced breast cancer diagnosis scheme based on two-step SVM technique," International Journal of Advanced Computer Science and Applications, vol. 8, no. 4, pp. 158–165, 2017.
[16] Y. Khoudfi and M. Bahaj, "Applying best machine learning algorithms for breast cancer prediction and classification," in Proc. IEEE, 2018.
[17] A. Sharma and R. Rani, "Classification of cancerous profiles using machine learning," in Proc. Int. Conf. on Machine Learning and Data Science, 2017, pp. 31–36.
[18] B. M. Gayathri, C. P. Sumathi, and T. Santhanam, "Breast cancer diagnosis using machine learning algorithms – A survey," International Journal of Distributed and Parallel Systems (IJDPS), vol. 4, no. 3, pp. 105–112, May 2013.
[19] H. L. Chen, B. Yang, J. Liu, and D. Y. Liu, "A support vector machine classifier with rough set based feature selection for breast cancer diagnosis," Expert Systems with Applications, vol. 38, pp. 9014–9022, 2011.
[20] P. Shah, R. Deshpande, and N. Rao, "Breast cancer detection system," International Research Journal of Engineering and Technology (IRJET), vol. 7, no. 5, May 2020.
[21] M. Shilpa and C. Nandini, "Breast cancer diagnosis and prediction using machine learning algorithm," International Journal of Science and Research (IJSR), vol. 9, no. 4, Apr. 2020.
[22] V. M. Botcha and B. P. Kolla, "Predicting breast cancer using modern data science methodology," vol. 8, no. 10, Aug. 2019.
[23] G. Chugh, S. Kumar, and N. Singh, "Survey on machine learning and deep learning applications in breast cancer diagnosis," Cognitive Computation, vol. 13, no. 6, pp. 1451–1470, 2021.
[24] N. Rane, J. Sunny, R. Kanade, et al., "Breast cancer classification and prediction using machine learning," International Journal of Engineering Research and Technology, vol. 9, no. 2, pp. 576–580, 2020.
[25] N. Al-Azzam and I. Shatnawi, "Comparing supervised and semi-supervised machine learning models on diagnosing breast cancer," Annals of Medicine and Surgery, vol. 62, pp. 53–64, 2021.