**MACHINE LEARNING APPROACHES FOR ACCURATE BREAST CANCER CLASSIFICATION**

**Santaji Soujanya1, B. Sandhya Rani 2, J. Chaithanya3, Remalli Rohan4**

1M. Sc Data Science-PG Student

2,3Degree Lecturer, Department of Computer Science,

Telangana Social Welfare Residential Degree college for Women, Jagathgirigutta,

Hyderabad, India.

4Researcher, Computer Science Educator, Hyderabad, Telangana, India.

**ABSTRACT**

It is one of the most common types of cancer worldwide and can affect both women and men, though it is far more common in women. Breast cancer occurs when normal cells in the breast begin to grow uncontrollably, forming a tumor and can invade surrounding tissues or spread (metastasize) to other parts of the body. It is the leading cause of death for women globally, and in order to properly treat it and lower mortality rates, it must be identified early and diagnosed accurately. Using a variety of datasets and imaging modalities, including multipara metric Magnetic Resonance Imaging (mpMRI), ultrasound, and histopathology pictures, research have used artificial intelligence (AI) and deep learning approaches to help predict, diagnose, and classify breast cancer. A range of machine learning algorithms, including ensemble techniques like Ada Boost, Gradient Boosting, and Random Forest, as well as sophisticated models like Convolutional Neural Networks (CNN) and Vision Transformers (ViT), have been evaluated for their ability to distinguish between benign and malignant tumors. Techniques such as adaptive token sampling, semi-supervised learning, and optimized stacking ensemble learning (OSEL) have been used in recent research to maximize classification accuracy, with the findings showing notable gains in performance measures across several studies. These developments highlight the possibility of incorporating machine learning (ML) techniques to improve computer-aided diagnostic (CAD) systems; models have been shown to achieve classification accuracies of 91% to 99%, offering useful resources for medical practitioners managing breast cancer.

**Keywords**: Breast cancer diagnosis, Artificial Intelligence (AI), Machine learning algorithm, Convolutional Neural Networks (CNN) and Multiparametric MRI (mpMRI).

1. **INTRODUCTION**

According to the World Health Organization (WHO), breast cancer is the most prevalent disease diagnosed globally in 2020, accounting for 11.6% of all new cancer cases. The total number of new cases worldwide that year was close to 2.3 million. Among women, cancer is one of the leading causes of mortality, accounting for nearly 685,000 deaths in a single year. Breast cancer is one of the most prevalent cancers that can afflict both men and women, albeit it mostly affects women. Starting in the breast cells, it can spread to the surrounding tissues and other body parts.

* 1. **Types of Breast Cancer**

Breast cancer encompasses a wide range of types that are classified based on where the cancer begins and how it behaves. The two main categories are invasive and non-invasive (also called in situ) cancers. Breast cancer is classified according to its location, tumor kind, and cell type as shown in figure1.

**1. Ductal Carcinoma in Situ (DCIS) Unusual**: Cells in this non-invasive malignancy are limited to the milk ducts and have not migrated to neighboring tissue. A common method of detection is mammograph.

**2.s Invasive Ductal Carcinoma (IDC):**  It invades surrounding tissue after beginning in the milk ducts.

**3.** **Invasive Lobular Carcinoma (ILC): S**tarts in the glands (lobules) that produce milk. Less frequent than IDC and might be more difficult to find on mammograms.

**4. Inflammatory Breast Cancer (IBC):** This unusual and aggressive variant makes the breasts look red and puffy. Although it may not manifest as a noticeable lump, it damages the skin.

**5. Triple-Negative Breast Cancer:** Progesterone, estrogen, and HER2 protein receptors are absent. Harder to treat with traditional therapy and more aggressive.

**6. HER2-Positive Breast Cancer**: Characterized by the HER2 protein being produced in excess. It tends to spread more quickly and is treatable with targeted therapy.

**7. Paget's Disease of the Breast:** This uncommon type of cancer begins in the nipple and progresses to the areola. Changes in the nipple skin, like redness or flaking, are common symptoms.

**8. Phyllodes Tumors:** unusual growths that appear in the breast's connective tissue. Although usually benign, it can turn cancerous.



**Figure 1:** Breast Cancer

**1.2 Causes of Breast Cancer**

The exact causes of breast cancer are not fully understood, but several factors are known to increase the likelihood of developing the disease. These factors can be broadly classified into genetic, hormonal, environmental, and lifestyle factors. While some causes are beyond control, others can be managed to reduce the risk of developing breast cancer. When the genetic material of the breast cells is altered and destroyed, breast cancer develops. Although the precise causes are unknown, a number of risk factors are known to be involved as showed in the figure 2:

1. **Genetic Factors: The** chance of developing breast and ovarian cancer is greatly increased by Mutations in the BRCA1 and BRCA2 genes. Breast cancer in the family.
2. **Hormonal Influence:** Exposure to estrogen: Long-term exposure raises the risk. Breast cancer risk may rise with hormone replacement treatment (HRT).
3. **Age:** The risk of breast cancer increases with age, especially after 50
4. **Personal History:** Individuals who have experienced breast cancer in one breast are more likely to experience cancer in the other breast.
5. **Radiation Exposure:** The chance of developing breast cancer is increased by prior radiation therapy, particularly prior to the age of thirty.
6. **Lifestyle Factor:** Drinking alcohol: Drinking too much raises the danger. Obesity: Being overweight can raise the risk, especially after menopause. Physical inactivity: One of the contributing factors is not exercising.
7. **Reproductive History:** Individuals who have never conceived or who gave birth to their first kid later in life are at a higher risk. Higher risks are also associated with late menopause and early menstruation.
8. **Environmental Exposure:** There is evidence that exposure to substances that disrupt hormones, such as pesticides, increases the risk.



**Figure 2:** Causes of Breast Cancer

**1.3 Symptoms of Breast Cancer**

The symptoms of breast cancer can vary depending on the type, stage, and location of the cancer. Early-stage breast cancer may not cause noticeable symptoms, which is why regular screening (such as mammograms) is crucial for early detection. However, there are common signs and symptoms that may suggest breast cancer. It's important to note that these symptoms can also be caused by conditions other than cancer, so it is essential to consult a healthcare provider for proper evaluation. Common symptoms include as showed in the figure 3:

1. **Lump in the Breast:** A hard, painless lump in the breast is frequently the first sign that people detect.
2. Change in Breast Size or Shape: unexplained breast form distortion, expansion, or decrease.
3. Pain in the Breast: persistent discomfort or pain, albeit malignant tumors are less likely to cause this
4. Changes to the Skin of the Breast: dimpling or redness of the skin, which can occasionally resemble orange peel (peau d'orange).
5. Nipple Changes: Inversion of the nipple, unusual discharge (including blood), or itching and irritation.
6. Swelling or Lumps in the Armpit: Swelling or lumps in the armpit area may indicate lymph node involvement.
7. Unexplained Weight Loss: Unexpected weight loss could be a sign of advanced cancer
8. Fatigue and Malaise: Feeling unusually tired or unwell without a clear reason.



**Figure 3:** symptoms of breast cancer

**1.4 Complications of Breast Cancer:**

Breast cancer can lead to a range of complications, particularly if it is not diagnosed and treated early. These complications can arise both during the disease and as a result of the treatments used to manage it. Here are some of the key complications associated with breast cancer as showed in the figure 4:

****

**Figure 4:** Complications of Breast Cancer

1. **Metastasis:** The cancer may spread to the brain, liver, lungs, or bones, among other regions of the body.
2. **Lymphedema:** When lymph nodes are removed after surgery, the lymphatic outflow becomes blocked, which can result in arm or chest edema.
3. **Pain:** Prolonged discomfort in the breast or other locations may result from invasive cancer if it spreads.
4. **Skin Ulcers:** Breast ulcers or open sores may result from big tumors or inflammatory breast cancer.
5. **Bone Fractures:** Bones that have metastasized may become weaker and more prone to fracture.
6. **Psychological Impact:** Anxiety, depression, and other mental health problems can result from receiving a cancer diagnosis and receiving treatment.
7. **Secondary Infections:** Radiation, chemotherapy, and surgery can all impair immunity, increasing a patient's risk of infection.
8. **LITERATURE SURVEY**

Using machine learning to increase the accuracy of breast cancer diagnosis has been the subject of numerous studies as shown in the table-1. To categorize tumors as benign or malignant, for instance, Sannasi Chakravarthy and Harikumar Rajaguru [1] employed Decision Tree and K-Nearest Neighbor (KNN) algorithms on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. Likewise, in order to more accurately detect malignant tissues in mammography pictures, Hediyehzadeh and Safdarian [2] used automated detection and pattern recognition. A 3D computer-aided diagnostic (CAD) system was created by Megariani and Samuri [3] that tracks lesions over several mammograms using a deformable 3D model, enhancing early detection.

Additionally, feature selection methods have been applied to improve machine learning models. Boruta and LASSO were employed in conjunction with classifiers such as Support Vector Machine (SVM) and Logistic Regression (LR) by Mohd Ali et al. [4] to classify breast cancer; LASSO and LR produced the best results. EnSNR and Genetic Algorithm (GA) is an ensemble filter technique that Hengpraprohm and Jungjit [5] utilized to create a model with less features while maintaining high accuracy. Offering a viable technique for detecting cancer early.

Molecular data for classification has been the subject of certain investigations. Using immunohistochemistry, Mittal and Mani [6] discovered molecular indicators of breast cancer and connected them to conventional prognostic variables. Wu and Hicks [7] classified tumors as triple-negative or non-triple-negative using gene expression data and machine learning. Cai et al. [9] created a non-invasive extracellular vesicle-based plasma-derived mRNA classifier, offering a viable approach to early cancer detection. Boosting techniques such as LightGBM and XGBoost have demonstrated efficacy in diagnosing breast cancer. When applying these techniques to MRI scans, Vamvakas et al. [10] were successful in identifying lesions. When Kadhim and Kamil [11] used the WDBC dataset to test different machine learning classifiers, they discovered that Extra Trees outperformed the others in terms of F1-score. The stacking ensemble model presented by Kumar et al. [12] using the UCI Breast Cancer Wisconsin dataset had an accuracy of 99.45%.

Classifying breast cancer has shown success using deep learning models. The advantages of deep learning for improved diagnostics and clinical decision-making were illustrated by Wang et al. [13]. LightGBM was the model that Chatterjee et al. [14] utilized to attain good classification accuracy. In their comparison of SVM, KNN, Naive Bayes, Random Forest, and deep learning models, Sukmandhani et al. [15] discovered that deep learning with certain activation functions produced the greatest results, 93.14%. Barwal et al. [16] discovered that feature extraction and data preprocessing enhanced sensitivity and specificity using classifiers such as SVOF-KNN, KNN, and Naive Bayes, achieving 91% specificity. Another use of convolutional neural networks (CNNs) is the classification of breast cancer. Using CNNs, feature integration, and transfer learning, Hassan et al. [17] created a CAD system that achieved 99.24% accuracy. Similarly, Biçer and Yanik [18] demonstrated CNNs' capacity to manage complicated tissue properties by classifying breast tissue using CNNs in conjunction with KNN and fuzzy logic, obtaining 96% accuracy.

Models like logistic regression have been used to classify tumors. By applying logistic regression to the WDBC dataset, Viswanatha et al. [19] were able to reliably classify benign and malignant tumors. To better understand tumor growth, metastasis, and therapy resistance for hormone receptor-positive (HR+) breast cancer, Nicotra et al. [20] developed rat models of the disease. A comparison study on the categorization of breast cancer using machine learning algorithms was carried out by Houfani et al. (2020) in [21]. A variety of classifiers, such as SVM, Random Forest, Decision Trees, and Multi-layer Perceptron, were examined using the Wisconsin Diagnostic Breast Cancer dataset. The study found that MLP and Logistic Regression were the most accurate (98%) models for classification tasks.

 **Table 1:** Summary of Methodologies

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Authors name** | **Dataset** | **Feature Extraction** | **Algorithm** | **Result** |
| Rajaguru H et al. [1] | Wisconsin Diagnostic Breast Cancer (WDBC) | PCA | Decision Tree, K-Nearest Neighbor (KNN) | KNN outperformed Decision Tree in accuracy. |
| Naser Safdarian et al [2] | Clinical mammography images | Pattern recognition methods | Automated classification system | 97% accuracy in cancer vs. non-cancer image classification |
| Samuri & Viananda Nova Megariani[3] | Clinical 3D breast models | Contrast enhancement, fuzzy logic | CAD-based classification | Effective at distinguishing cancerous vs. non-cancerous tissues |
| Supoj Hengpraprohm et al. [4] | Microarray dataset | Entropy and SNR function | Genetic Algorithm | 86.92% average prediction accuracy |
| Nursabillilah Mohd Ali et al. [5] | GEO dataset | Boruta, LASSO | SVM, Logistic Regression | LASSO with LR provided best accuracy |
| Ankit Mittal et al. [6] | 70 cases of invasive breast carcinoma | Immunohistochemistry with ER, PgR, Her2/neu, and Ki67 markers | Statistical analysis | The study found a substantial association between tumor grade and lymphovascular invasion (LVI) and the prevalence of Luminal B and Her2/neu subtypes. |
| Jiande Wu et al. [7] | 110 triple-negative and 992 non-triple-negative samples from The Cancer Genome Atlas (TCGA) | RNA-Sequence data analysis | Support Vector Machines, K-Nearest Neighbor, Naïve Bayes, Decision Tree | The most accurate method for differentiating between triple-negative and non-triple-negative breast cancer was Support Vector Machine (SVM). |
| Zuherman Rustam, et al. [8] | Publicly available breast cancer datasets | Pattern learning on tumor cell data | Support Vector Machine, Linear Discriminant Analysis | SVM's 98.77% accuracy rate beat LDA's, demonstrating its effectiveness in clinical categorization. |
| Geng-Xi Cai et al. [9] | 259 plasma samples (144 breast cancer, 72 benign, 43 healthy) | Next-generation sequencing, RT-qPCR | Support Vector Machine, Linear Discriminant Analysis, Logistic Regression | With an AUC of 0.737 in validation, the EXOBmRNA classifier demonstrated efficacy in non-invasive diagnosis with an accuracy of 71.9%. |
| Alexandros Vamvakas et al. [10] | mpMRI images of 140 patients (70 benign, 70 malignant) | 3D radiomic features via Pyradiomics | AdaBoost, Gradient Boosting, XGBoost, LightGBM, SVM | XGBoost and LightGBM had AUC = 0.95, 0.94 |
| Rania R. Kadhim, et al. [11] | Wisconsin Diagnostic Breast Cancer (WDBC) dataset | Data preprocessing | Decision Tree, AdaBoost, SVM, Extra Trees, Gaussian NB, MLP, etc. | Extra Trees had an F1-score of 96.77% |
| Mukesh Kumar et al. [12] | UCI Breast Cancer Wisconsin dataset | Data preprocessing | Stacking ensemble, AdaBoost, XGBoost, Gradient Boosting, CatBoost | OSEL model achieved maximum accuracy of 99.45% |
| Wei Wang et al. [13] | Ultrasound and histopathology images | Adaptive token sampling | Vision Transformer (ViT) | Outperformed CNN baselines in all tasks |
| Chatterjee et al. [14] | Breast cancer dataset | Swarm plots, Heat maps | Light GBM, Logistic Regression, Random Forest, XGBoost, Gradient Boosting | Accuracy: 96.98% (Light GBM) Sukmandhani et al. (2023) |
| Sukmandhani et al. [15] | Wisconsin Diagnostic Breast Cancer (WDBC) | Data pre-processing | SVM, KNN, Naive Bayes, Random Forest, Decision Tree, Deep Learning | Best Accuracy: 93.14% (Deep Learning) |
| Barwal et al. [16] | WDBC, Coimbra BC datasets | Feature extraction and selection | SVOF-KNN, KNN, Naive Bayes | Accuracy: 91% specificity, 90% sensitivity (SVOF-KNN) |
| Hassan et al. [17] | Histopathological images | Feature extraction, Transfer learning | DCNN, SVM, Chi-square feature selection | Accuracy: 99.24% |
| BİÇER and YANIK et al. [18] | Electrical Impedance Spectroscopy data (UCI) | Impedance spectroscopy features | KNN, SVM, Decision Tree, Fuzzy Logic, CNN | Accuracy: 96% (CNN) |
| Viswanath V., et al [19] | Wisconsin Diagnostic Breast Cancer dataset | Histograms, scatter plots, box plots, bar plots | Logistic Regression | Accurate classification of benign vs. malignant tumors. |
| Raquel Nicotra, et al. [20] | Rat mammary tumor models | Hormone receptor analysis | Logistic Regression | Established six distinct rat models. |
| Djihane Houfani,[21] | Wisconsin Diagnostic Breast Cancer Dataset | usage of features from the Wisconsin dataset directly | Kernel SVM (K-SVM), Linear SVM (L-SVM), Random Forest (RF), Decision Tree (DT), Multi-layer Perceptron (MLP), Logistic Regression (LR), k-Nearest Neighbors (k-NN) |

|  |
| --- |
| Best accuracy: 98% achieved by Multi-layer Perceptron (MLP) and Logistic Regression (LR). |

 |

1. **CONCLUSION AND FUTURE SCOPE**

Breast cancer is a complex and multifaceted disease that affects millions of people worldwide. While its exact causes remain incompletely understood, a combination of genetic, hormonal, lifestyle, and environmental factors can influence its development. Early detection through regular screenings, such as mammograms and self-examinations, is crucial for improving outcomes, as breast cancer is more treatable when diagnosed in its early stages. It can save lives if it is detected early since prompt diagnosis greatly enhances treatment results. Using data from nine distinct studies, this study emphasizes the significance of machine learning (ML) algorithms in identifying and classifying breast cancer as either benign or aggressive. The paper presents a literature survey on machine learning models used to improve diagnosis accuracy. Numerous research has successfully used algorithms like Random Forest, Support Vector Machines, and deep learning models, highlighting the promise of AI-driven systems to transform the diagnosis of breast cancer. More complex methods, like deep learning, may be developed as machine learning models improve, perhaps increasing classification accuracy even more. Also, even with minimal labeled data, diagnostic capacities could be improved by combining big, diverse datasets and using semi-supervised and unsupervised learning techniques. When incorporated into clinical practice, real-time AI-assisted diagnostic technologies may result in quicker and more accurate tests. Additionally, future studies might concentrate on tailoring therapy recommendations according to patient-specific information, supporting precision medicine in the treatment of breast cancer.

**4. REFERENCES**

1. Rajaguru, Harikumar, and Sannasi Chakravarthy SR. "Analysis of decision tree and k-nearest neighbor algorithm in the classification of breast cancer." *Asian Pacific journal of cancer prevention: APJCP* 20.12 (2019): 3777
2. Safdarian, Naser, and Mohammadreza Hedyezadeh. "Detection and classification of breast cancer in mammography images using pattern recognition methods." *Multidisciplinary Cancer Investigation* 3.4 (2019): 13-24.
3. Samuri, Megariani, Try Viananda Nova. "Intelligent 3D analysis for detection and classification of breast cancer." *JITCE (Journal of Information Technology and Computer Engineering)* 3.02 (2019): 96-103.
4. Hengpraprohm, Supoj, and Suwimol Jungjit. "Ensemble feature selection for breast cancer classification using microarray data." *Intelligence Artificial* 23.65 (2020): 100-114.
5. Nursabillilah Mohd Ali et al**.** "Comparison of Microarray Breast Cancer Classification Using Support Vector Machine and Logistic Regression with LASSO and Boruta Feature Selection." Medical Technologies Journal, vol. 4(2) pg.no 535-544. (2020)
6. Mittal, A et.al. “Molecular classification of breast cancer.” Bharati Vidyapeeth Deemed University, (2021).
7. Wu, Jiande, and Chindo Hicks. "Breast cancer type classification using machine learning." *Journal of personalized medicine* 11.2 (2021): 61.
8. Rustam, Z et al. “Linear discriminant analysis and support vector machines for classifying breast cancer”. IAES International Journal of Artificial Intelligence, vol 10(1), pg. no 253-256. (2021)
9. Cai, G. X et al. “A plasma-derived extracellular vesicle mRNA classifier for the detection of breast cancer”. Gland Surgery, vol 10(11), pg. no 275-284. (2021)
10. Alexandros Vamvakas et al. “Boosting Ensemble Methods for Breast Cancer Classification on MRI”. *International Journal of Reconfigurable and Embedded Systems*, vol 11(2), 166-174 .pg.no166-174. (2022)
11. Kadhim, R. R., et.al. “Comparing Breast Cancer Classification Models Using the Wisconsin Dataset”. *International Journal of Reconfigurable and Embedded Systems*, vol11(2), pg.no166-174. (2022)
12. Mukesh Kumar et al. “ Optimized Stacking Model for Breast Cancer Detection and Classification”. *University of California, Irvine Repository*./ijres.vol,11. i2.pg.no 166-174. (2022)
13. Wei Wang et al. “Vision Transformer for Semi-Supervised Breast Cancer Classification”. Frontiers in Pharmacology, (2022)
14. Chatterjee, V et al . “An efficient approach for breast cancer classification using machine learning”. Journal of Decision Analytics and Intelligent Computing, vol 4(1), pg.no32-46. (2023)
15. Sukmandhani, A. et al. “Classification algorithm analysis for breast cancer”. International Journal on Recent and Innovation Trends in Computing and Communication, vol. 11(3s). (2023)
16. Barwal, R. K et al. “Analysis and classification of breast cancer disease via different datasets and classifier models.” IJRITCC, vol. 11(3s). (2023)
17. Hassan, Abdallah M., Ahmed Yahya, and Ashraf Aboshosha. "A framework for classifying breast cancer based on deep features integration and selection." *Neural Computing and Applications* 35.16 (2023): 12089-12097.
18. Biçer, M. B. et.al. “A convolutional neural networks model for breast tissue classification”. Bitlis Eren University Journal of Science, vol.11(3), pg. no. 798-811(2022)
19. Viswanatha, V et.al. “Breast Cancer Classification Using Logistic Regression”. Journal of Soft Computing Exploration, pg. no: 2746-7686, (2024)
20. Nicotra, Raquel, et al. "Rat models of hormone receptor-positive breast cancer." *Journal of Mammary Gland Biology and Neoplasia* 29.1 (2024): 12.
21. D. Houfani, et.al, "Breast cancer classification using machine learning techniques: a comparative study," Medical Technologies Journal, vol. 4, no. 2, pg. no. 535–544, (2020)