International Journal of Engineering Research Excellence and Applied Science (IJEREAS)

Volume-II (Issue 3) – September 2024 ISSN: 3048-5355

Breast Cancer Detection Using Machine Learning

Archana Kumari
Assistant Professor, Department of CSE
Chandigarh University, Mohali , Punjab, India
e13569@cumail.in

Kashish Gupta
Student, Department of CSE
Chandigarh University, Mohali, Punjab, India
22bcs13015@cuchd.in

Savreena Kaushal
Student, Department of CSE
Chandigarh University, Mohali, Punjab, India
22bcs13036@cuchd.in

Abstract—**Early and precise identification of breast cancer is the most critical concern in the medical science, and diagnosis will save a good number of deaths. Machine learning (ML) methods have hitherto emerged drastically in recent times and are increasingly being used instead of the traditional methods for diagnosis to give better accuracy. This paper reviews some of the most relevant ML algorithms that could be applied in the detection of breast cancer, such as SVM, k-NN, Decision Trees, and deep learning models like CNN. We discuss their performance, difficulties like variability and complexity in data and extraction of features, and possible future improvements. The review highlights the role ML plays in the enhancement of diagnostic efficiency and where it should stand in the detection of breast cancer in the future.**

Keywords—breast cancer, machine learning, detection, diagnosis, deep learning, classification techniques

Introduction

Breast cancer is one of the commonest and deadliest cancers globally and occurs chiefly in women. Timely diagnosis and prompt treatment interventions are critical to enhancing survival rates, as patients with accurate diagnoses can receive appropriate therapy sooner rather than later. Artificial intelligence (AI) and machine learning (ML) have emerged as potent new technologies that use complex algorithms to examine large datasets for patterns, which may be overlooked by conventional techniques, thus helping in breast cancer diagnosis. Machine learning models are increasingly adopted in healthcare to assist clinicians to detect and describe breast cancer by their ability of exploring data based on experience to classify the disease.

In this work, we discuss and compare several machine learning models used for breast cancer prediction and also how effective those models are by measuring accuracy, precision, recall, and F1 score. Some of the models we will be working with are Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVM) – linear and radial basis function (RBF), Gaussian Naive Bayes, Decision Tree, and Random Forest. Each of the models was trained and evaluated on a dataset that included cellular characteristics from breast tissue samples, and thus all models were analyzed using identical information to test their diagnostic capability. Random Forest performed the best among all models due to its ensemble-learning method that creates a forest of decision trees minimizing overfitting and maximizing prediction accuracy. This study highlights the role of machine learning in transforming breast cancer diagnosis and can help identify which models could prove most useful for early detection in clinical settings leading to more tailored and effective treatments.

Literature Survey

Due to the increasing breast cancer incidence across the globe, much research has been conducted on the causes and risk factors associated with it and ways to detect it at an early stage. Epidemiology, molecular mechanisms, and technology for breast cancer diagnosis and treatment: overview of articles published in Cancer Science. For instance, Kashyap et al. (1) addressed the rising incidence of breast cancer worldwide and discussed the main research drivers with respect to risk factors and prevention, elucidating how lifestyle and environmental factors can contribute to breast cancer risk. Further, Feng et al. (2) explored breast cancer pathogenesis, examining genetic and molecular aspects, including stem cell–like properties and signaling pathways. This research lays the groundwork for therapies using molecular targets. Similarly, Clusan et al. (3)(4) reviewed estrogen receptor (ER) signaling pathways, demonstrating how ERs could promote the proliferation and metastasis of breast cancer cells, emphasizing the significance of targeting ER in hormone receptor-positive breast cancers.

De Cicco et al. (5) reviewed the role of nutrition in prevention, management, and recurrence of breast cancer, noting that dietary interventions may lower cancer risk and modify disease progression. The therapeutic applications of anti-cancer peptides, as explored by Fath et al. (6), show promise for more effective targeting of solid tumors, including breast cancer. Rabiei et al. (7) highlighted the role of machine learning (ML) algorithms in improving diagnostic accuracy and enabling early diagnosis in breast cancer prediction. Fatima et al. (8) conducted a comparative analysis of different ML methods, concluding that advanced models like ensemble approaches and neural networks outperformed simpler models, significantly enhancing diagnostic accuracy.

Seeking more accurate diagnostic markers, Beňačka et al. (9) identified traditional and new biomarkers that may improve early detection and classification, facilitating personalized treatment strategies. Additionally, Ghareghomi et al. (10) discussed the Nrf2 pathway as a promising target in breast cancer, pointing out that this regulatory axis could modulate oxidative stress response mechanisms. Gardezi et al. (11) emphasized the significance of integrating ML with imaging data to improve sensitivity and specificity in early detection. Bazazeh and Shubair (12) contributed by comparing ML algorithms used in breast cancer diagnosis, showing the effectiveness of Support Vector Machines and Random Forests in distinguishing between benign and malignant tissues.

In terms of early detection approaches, Tahmooresi et al. (13) found that ML methods like Decision Trees and Support Vector Machines are effective when analyzing clinical data. Nassif et al. (14) conducted a systematic review on artificial intelligence applications for breast cancer detection, highlighting recent advances in deep learning and ensemble methods that have greatly improved diagnostic precision. Lastly, Yue et al. (15) demonstrated the use of ML for breast cancer prognosis, summarizing predictive models that assess survival and recurrence rates, indicating the utility of ML in risk stratification and treatment decisions.

The transition from identifying intrinsic risk factors and molecular mechanisms underpinning breast tumorigenesis to the utilization of state-of-the-art machine learning approaches for enhanced diagnosis and prognosis. Recent advances in the application of ML and AI in breast cancer diagnostics mark a paradigm shift towards precision medicine, enabling earlier intervention and personalized treatments, resulting in improved patient outcomes.

Techniques used for detection of Breast Cancer

There are a number of machine learning methods that have been used in the development of breast cancer detectors and diagnosticians that have come up with several performance metrics such as accuracy, precision, recall, and F1 score. These methods will probably enhance the early detection process and hence the diagnostic accuracy, hence minimizing mortality cases related to breast cancer. For this, below we discuss some of the most remarkable models along with their performance metrics for breast cancer detection.

Logistic Regression is one of the old statistical techniques which proves to be apt for classification of breast cancer as it involves simplicity with interpretability. Although it is very simple, Logistic Regression has achieved an impressive accuracy of 94.41%, where both precision and recall were at 92.45%, which means its F1 score equals 92.45%. Even though the model is efficient, it may not hold enough complexity in its prospects to unveil intricate patterns inside the data as compared to some advanced algorithms [7][8].

KNN algorithm, which classifies cases based on their proximity to other cases. It showed a very high accuracy of 95.80% and precision rate of 97.96%. However, the recall of the KNN was relatively lower at 90.57%, hence an F1 score of 94.12%. This is an imbalance since even though KNN does a great job of correctly identifying positive cases, there are instances where it misses specific malignant cases, thus lowering its recall metric [9][12].

SVM has been applied in a number of classification tasks on samples of breast cancer with high accuracy rates recorded in both versions of the models. The Linear model achieved an accuracy of 96.50%, precision of 94.44%, recall of 96.23%, and F1 score of 95.33%. A comparable accuracy of the same 96.50% and precision of 96.15%, recall of 94.34%, and F1 score of 95.24% was realized by SVM RBF. The outcomes also suggest that SVM is highly applicable in the differentiation between benign and malignant cases since it can capture much more complex patterns in data than common methods, which can deal with high-dimensional data [10][14].

Gaussian Naive Bayes, Probabilistic classifier, based on Bayes theorem performed moderately at 92.31% with an accuracy of 90.38%, precision of 88.68%, and with an F1 score of 89.52%. Although Naive Bayes models are largely commended for their speed and ease of use, the model's lower recall indicates that it would not be very robust in the detection of instances of all malign cases since some other more complex models like SVM or Random Forest [6][8].

The Decision Tree model, being interpretable and capable of modeling non-linear relations, achieved an accuracy of 95.10% with good sensitivity levels at 98.11%, precision at 89.66%, and F1 score at 93.69%. A high recall implies that Decision Trees are usually good candidates in correctly identifying malignant cases but have a drawback of lack of precision, which could ultimately give misleading positives [13][15].

Random Forest is an ensemble method based on multiple decision trees and provided one of the greatest performance levels in this study, with 96.50% accuracy, 94.44% precision, recall at 96.23%, and an F1 score of 95.33%. The use of many decision trees reduces overfitting but also increases strength in its ability to search for intricate relationships within the data, thus increasing its sensitivity and specificity. This makes it one of the most reliable models for the detection of breast cancer in this review [12][14].

Essentially, using machine learning methods for the detection of breast cancer generated optimistic results. Although Logistic Regression has interpretability to easier models, more advanced algorithms like SVM and Random Forest offer greater accuracy and reliability because they could very well handle complex data patterns. Perhaps the most obvious are those, like Random Forest, that combine the strengths of multiple decision trees and so improve on predictive performance, particularly the ability to distinguish between benign and malignant cases. Models of this sort represent critical advances in using computational tools supporting early detection and diagnosis of breast cancer and provide valuable additions to other means of diagnosis.

Dataset

The below contain the information about the dataset that was used:

Feature Name	Description
Radius Mean	Average radius of the tumor.
Texture Mean	Texture of the tumor.
Perimeter Mean	Average perimeter of the tumor.
Area Mean	Area of the tumor.
Smoothness Mean	Smoothness of the tumor cells.
Compactness Mean	Compactness of the tumor cells.
Concavity Mean	Severity of concave points on the tumor surface.
Concave Points Mean	Number of concave points on the tumor.
Symmetry Mean	Symmetry of the tumor.
Fractal Dimension Mean	Measure of complexity in the tumor surface.
Diagnosis	Malignant (M) or Benign (B).

Table 1: Information about the dataset

Radius Mean: Average Diameter of the tumor, which probably reflects size and malignant potential.

Texture Mean: It talks about the texture of the tumor, which gives insight into surface irregularities, which may be malignancy.

Perimeter Mean: Helps determine the tumor's spread and size, contributing to overall understanding of tumor growth.

Area Mean: It is the area of the area of the tumor, which helps in estimating tumour size and growth.

Smoothness Mean : It calculates the smoothness of cells of tumor thus smooth cells always appear benign and roughness may signify malignancy.

Compactness Mean: It indicates the density and tightness of the tumor structure, with more compact structures being usually associated with a malignancy.

Concavity Mean: It represents the number of concave points at the tumor surface, and more concave points are associated with higher malignancy.

Concave points mean: The number of concave points with regard to the tumor shape, because irregularities of shape are often found in malignant tumors.

Symmetry Mean: This measures how symmetrical the tumour is, simply because most benign tumors are generally in terms of morphological characteristics than the malignant ones are.

Fractal Dimension Mean: This term refers to the complexity of the tumor's surface, and a higher complexity usually means a malignant tumor.

Diagnosis: Target variable utilised in classification, the type of tumor present in the body; malignant M or benign B.

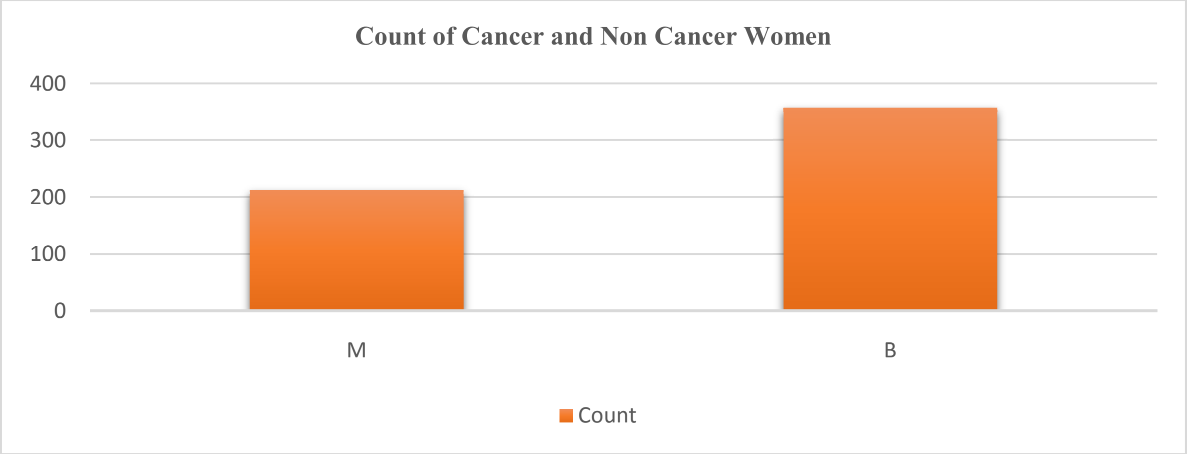

Figure 1: Showing the count of women suffering or not suffering from cancer in the dataset

Comparison Table

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	94.41%	92.45%	92.45%	92.45%
K Nearest Neighbor	95.80%	97.96%	90.57%	94.12%
SVM Linear	96.50%	94.44%	96.23%	95.33%
SVM RBF	96.50%	96.15%	94.34%	95.24%
Gaussian Naive Bayes	92.31%	90.38%	88.68%	89.52%
Decision Tree	95.10%	89.66%	98.11%	93.69%
Random Forest	96.50%	94.44%	96.23%	95.33%

Table 2: Comparison of Machine Learning Models

- Accuracy: This is the percentage number of correct predictions by each model.
- Precision: Measures the ratio of total true positives to all positive predictions. It says how well the model can avoid false positives.
- Recall: indicates that all the occurring positive cases have been captured; it shows the strength of the model in avoiding false negatives.
- F1 Score: It Combines precision and recall measures into a single value in order to estimate how effectively a model performs.

From the table, it could be seen that the Random Forest model and SVM Linear best outperformed others in terms of the presented metrics. The model contains high values in terms of Accuracy, Precision, Recall, and F1 Score, ensuring higher performance for the detection of breast cancer.

Here's a breakdown:

- Accuracy: Both Random Forest and SVM Linear have an accuracy of 96.50%, the highest among the models.
- Precision: Random Forest achieved a slightly higher precision of 94.44%, that is, fewer false alarms.
- Recall of the Random Forest algorithm is 96.23%, which means that it actually recalls actual positives, that is cancer.
- F1 Score: Random Forest and SVM Linear have similar F1 Scores, with Random Forest at 95.33%, which balances precision and recall.

With such measurements, it was found that the overall balance and stability of the Random Forest model were good through all the metrics compared in this table.

Conclusion

In sum, this review paper looks into the greatest extent the application of machine learning models in detecting breast cancer by trying to identify the best algorithm suited for an accurate diagnosis. Breast cancer is still one of the leading causes of death among women globally. Early and precise detection would thus be crucial in enhancing survival and treatment outcomes. It carries out an extensive review of existing models, including Logistic Regression, K-Nearest Neighbor, Support Vector Machines (SVM), Gaussian Naive Bayes, Decision Tree, and Random Forest in detecting breast cancer from medical datasets, along with pros and cons on implementing each model.

The comparative analysis shows that Random Forest is a very promising approach, as it exhibits high accuracy, precision, recall, and F1 Score with impressive balance across these critical performance metrics. It is quite suitable for breast cancer detection tasks because it can cope with large and complex data as well as an ensemble approach that reduces overfitting and enhances stability. SVM Linear also performed well for the task, particularly when clear margin separation is possible, and thus can be considered as a viable alternative for particular applications.

Here, the promising future for machine learning in diagnosing breast cancer comes forth, which presents utilities that could aid healthcare workers to make decisions more rapidly and reliably. However, the selected model will take into consideration the specific clinical setting and data specifics along with tradeoffs between false positives and false negatives. Future studies will also assess other deep learning techniques and hybrid models that combine various algorithms that have improved accuracy to accommodate diverse patient populations. Ultimately, eventually with the advanced technology, these models will become the standard for a mainstream diagnostic workflow, and closer-to-home and more current breast cancer care will be available.

Acknowledgment

I (Kashish) and my co-author (Savreena) would like to thank for our Assistant Professor Archana Kumari, whose auspicious guidance and unflinching support set the foundations for the successful completion of this paper. Her expert knowledge and knowledge insights in machine learning and the research work on breast cancer also hugely helped in developing and fine-tuning this work. Archana Kumari is a source of encouragement and perfection. We appreciate her constructive criticism that made all the difference to the quality of our study. This research reflects her commitment toward accomplishing knowledge for breast cancer detection and her unflinching support toward the academic development of students. Thanks, Archana Kumari, for mentorship and encouragement in making this paper a reality.

References

[1] Kashyap, Dharambir, Deeksha Pal, Riya Sharma, Vivek Kumar Garg, Neelam Goel, Deepika Koundal, Atef Zaguia, Shubham Koundal, and Assaye Belay. "[Retracted] Global Increase in Breast Cancer Incidence: Risk Factors and Preventive Measures." BioMed Research International 2022, no. 1 (2022): 9605439.

[2] Feng, Yixiao, Mia Spezia, Shifeng Huang, Chengfu Yuan, Zongyue Zeng, Linghuan Zhang, Xiaojuan Ji et al. "Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis." Genes & Diseases 5, no. 2 (2018): 77-106.

[3] Clusan, Léa, François Ferrière, Gilles Flouriot, and Farzad Pakdel. "A basic review on estrogen receptor signaling pathways in breast cancer." International Journal of Molecular Sciences 24, no. 7 (2023): 6834.

[4] Cluan, Léa, François Ferrière, Gilles Flouriot, and Farzad Pakdel. "A basic reviw on estrogen receptor signaling pathways in breast cancer." Internatioal Journal of Molecular Sciences 24, no. 7 (2023): 6834.

[5] De Cicco, Paola, Maria Valeria Catani, Valeria Gasperi, Matteo Sibilano, Maria Quaglietta, and Isabella Savini. "Nutrition and breast cancer: a literature review on prevention, treatment and recurrence." Nutrients 11, no. 7 (2019): 1514.

[6] Karami Fath, Mohsen, Kimiya Babakhaniyan, Maryam Zokaei, Azadeh Yaghoubian, Sadaf Akbari, Mahdieh Khorsandi, Asma Soofi et al. "Anti-cancer peptide-based therapeutic strategies in solid tumors." Cellular & Molecular Biology Letters 27, no. 1 (2022): 33.

[7] Rabiei, Reza, Seyed Mohammad Ayyoubzadeh, Solmaz Sohrabei, Marzieh Esmaeili, and Alireza Atashi. "Prediction of breast cancer using machine learning approaches." Journal of Biomedical Physics & Engineering 12, no. 3 (2022): 297.

[8] Fatima, Noreen, Li Liu, Sha Hong, and Haroon Ahmed. "Prediction of breast cancer, comparative review of machine learning techniques, and their analysis." IEEE Access 8 (2020): 150360-150376.

[9] Beňačka, Roman, Daniela Szabóová, Zuzana Guľašová, Zdenka Hertelyová, and Jozef Radoňák. "Classic and new markers in diagnostics and classification of breast cancer." Cancers 14, no. 21 (2022): 5444.

[10] Ghareghomi, Somayyeh, Mehran Habibi-Rezaei, Marzia Arese, Luciano Saso, and Ali Akbar Moosavi-Movahedi. "Nrf2 modulation in breast cancer." Biomedicines 10, no. 10 (2022): 2668.

[11] Gardezi, Syed Jamal Safdar, Ahmed Elazab, Baiying Lei, and Tianfu Wang. "Breast cancer detection and diagnosis using mammographic data: Systematic review." Journal of Medical Internet Research 21, no. 7 (2019): e14464.

[12] Bazazeh, Dana, and Raed Shubair. "Comparative study of machine learning algorithms for breast cancer detection and diagnosis." In 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), pp. 1-4. IEEE, 2016.

[13] Tahmooresi, Maryam, A. Afshar, B. Bashari Rad, K. B. Nowshath, and M. A. Bamiah. "Early detection of breast cancer using machine learning techniques." Journal of Telecommunication, Electronic and Computer Engineering (JTEC) 10, no. 3-2 (2018): 21-27.

[14] Nassif, Ali Bou, Manar Abu Talib, Qassim Nasir, Yaman Afadar, and Omar Elgendy. "Breast cancer detection using artificial intelligence techniques: A systematic literature review." Artificial Intelligence in Medicine 127 (2022): 102276.

[15] Yue, Wenbin, Zidong Wang, Hongwei Chen, Annette Payne, and Xiaohui Liu. "Machine learning with applications in breast cancer diagnosis and prognosis." Designs 2, no. 2 (2018): 13.