A Survey on Methodologies Available for Predicting the Recurrence of Breast Cancer Post Treatment

Mamatha S
Assistant Professor, Dept. of ISE,
Cambridge Institute of Technology,
Bangalore, India

Dr Josephine Prem Kumar
Professor, Dept. of CSE,
Cambridge Institute of Technology,
Bangalore, India

ABSTRACT- Cancer is a growing global health concern and one of the leading causes of death. Recent studies highlight that breast cancer is among the most prevalent forms of cancer, particularly affecting women. Early detection plays a crucial role in lowering treatment costs and significantly enhancing survival rates for individuals diagnosed with breast cancer. Current early diagnosis methods in healthcare systems, while valuable, have certain limitations. These include long-term impacts, a heavy reliance on human resources, and challenges in ensuring widespread access to these services. This study aims to provide a comprehensive review of the latest models used for predicting the prognosis, diagnosing, and assessing the risk of breast cancer. It explores various machine learning techniques for breast cancer prediction, such as SVM, naïve Bayes, and random forests, along with different types of data, including clinical, genomic, and imaging data. The study also compares various algorithms and methodologies in this field.

Keywords- Breast cancer, recurrence, post treatment, machine learning techniques.

I. INTRODUCTION

Cancer, the leading cause of death globally, is a rapidly spreading disease characterized by the uncontrolled growth of abnormal cells. It is considered a non-lethal genetic condition that can lead to severe physical deterioration and, ultimately, death. Research shows that cancer is responsible for one in every six deaths worldwide, with breast cancer being the most prevalent among them. Cancer occurs when cells begin to multiply uncontrollably and spread throughout the body. Major side effects of breast cancer include fatigue, physical pain, and bone-related conditions such as osteoporosis. This study compares various machine learning models—CNN, Random Forest, Decision Trees, and Logistic Regression—while also recognizing that cancer progression involves changes in diseased tissues over time. Machine learning (ML) offers significant advantages over traditional pathology, including reducing human labor and aiding doctors in classifying images and text. It is regarded as a promising tool for future cancer screening. Additionally, there is evidence suggesting that a patient's emotional state can influence their health outcomes.

Breast cancer recurrence refers to the return of cancer after initial treatment, either in the same breast or elsewhere in the body. Recurrence can occur at any time after the primary treatment, but it is most common within the first 5 years.

II. RELATED WORK

A variety of machine learning techniques have been employed to predict the recurrence of breast cancer after treatment. Most studies have utilized datasets from the UCI Machine Learning Repository and the Wisconsin Breast Cancer Dataset, achieving accuracies above 95% by analyzing various parameters within these datasets. This section provides a comparative analysis of these methodologies, the datasets used, and the accuracy results obtained. One study, *Prediction of Recurrence and Non-recurrence Events of Breast Cancer using Bagging Algorithm* [1], employed the Bagging algorithm to predict recurrence and non-recurrence events. Using the UCI dataset, the study achieved an accuracy of 73.81%. Another paper, *An Analysis of Ensemble Machine Learning Algorithms for Breast Cancer Detection: Performance and Generalization* [2], compared Light Gradient Boosting (LightGBM) and Gradient Boosting, using the Wisconsin dataset and achieving a high accuracy of 99.41%. *Breast Cancer Classification Using Machine Learning* [3] analyzed various algorithms for predicting breast cancer recurrence, including logistic regression, decision trees, random forests, and convolutional neural networks (CNN). Based on the UCI dataset, the accuracy for each method was recorded as 95.6%, 92.9%, 93.8%, and 96.2%, respectively. In the study *Breast Cancer Prediction System Utilizing Machine Learning Algorithms* [5], multiple machine learning techniques were applied to forecast the return of breast cancer, further advancing prediction accuracy. *Unveiling Precision in Breast Cancer Prediction with Random Forest and Decision Trees* [15] focused on the use of Random Forest and Decision Trees to detect and predict breast cancer, showing strong performance. This research emphasized the effectiveness of these models when working with annotated data, demonstrating their practical potential. *The Role of Linear Discriminant Analysis for Accurate Prediction of Breast Cancer* [14] utilized Support Vector Machines (SVM) and Linear Discriminant Analysis (LDA) for classification. LDA was employed for dimensionality reduction, while SVM was used as the classifier, achieving accurate results through 5-fold cross-validation. Overall, these studies demonstrate the potential of machine learning to enhance the accuracy and efficiency of breast cancer predictions. The implementation of these techniques can significantly improve early detection, making the prediction process faster and more reliable, which is crucial for effective disease management.

III. INPUT DATASET

Acquiring appropriate datasets for developing machine learning (ML)-based systems can be particularly challenging, especially in the medical field, where concerns about privacy, confidentiality, ethics, and security are paramount. One widely used dataset in breast cancer (BC) detection and classification is the Wisconsin Breast Cancer Dataset (WBCD), available online at the UCI Machine Learning Repository, along with other related Wisconsin BC datasets. This dataset has become a benchmark for BC detection and classification tasks.

In addition to the raw data, the dataset includes features extracted from medical images after undergoing preprocessing steps such as image segmentation and feature extraction. These extracted features play a crucial role in improving the model's performance. To reduce computational complexity and enhance classification accuracy, feature selection or optimization can be applied to remove redundant or irrelevant features. This step helps reduce the amount of training data required, accelerating the learning process and potentially improving the model's prediction accuracy.

However, this optimization process can also impact the quality of the data, introducing additional time and computational costs, especially when only a small subset of features is selected. On the other hand, dimensionality reduction and feature transformation techniques are typically applied to high-dimensional datasets with many features, serving similar goals of reducing complexity and improving model performance.

The proposed Breast Cancer Prediction System leverages machine learning techniques, including Random Forest, Support Vector Machine (SVM), and Gradient Boosting Ensemble, to support early diagnosis and prognosis of breast cancer. To ensure data consistency and quality, multiple breast cancer datasets are first collected and then carefully preprocessed before being used for analysis.

IV. COMPARITIVE ANALYSIS OF VARIOUS MACHINE LEARNING TECHNIQUES

Ref no	Paper Title	Technique used	Data Set used	Accuracy Achieved
[1]	Prediction of Recurrence and Non-recurrence Events of Breast Cancer using Bagging Algorithm	Bagging Algorithm	UCI Machine Learning Repository	73.8182%.
[2]	An Analysis of Ensemble Machine Learning Algorithms for Breast Cancer Detection: Performance and Generalization	Comparative analysis of Light Gradient Boosting (LightGBM) and Gradient Boosting	Wisconsin breast cancer dataset	99.41%
[3]	Breast Cancer Classification Using Machine Learning	logistic regression, decision tress, random forest and CNN	UCI dataset	95.6%, 92.9%, 93.8%, 96.2%
[4]	Breast Cancer Modeling and Prediction Combining Machine Learning and Artificial Neural Network Approaches	SVM, Random Forest, KNN,	UCI dataset	97.2%
[5]	Breast Cancer Prediction System Utilizing Machine Learning Algorithms	Random Forest, Support Vector Machine (SVM), and Gradient Boosting Ensemble	UCI dataset	59%, 72%, 64%
[6]	A Comparative Study on Breast Cancer Prediction using Optimized Algorithms	Divide and Conquer Kernal Support Vector Machine (DCKSVM) and Hybrid Radial Basis Function Neural Network machine learning algorithms	Wisconsin breast cancer dataset	98%
[7]	Breast Cancer Risk Prediction using XGBoost and Random Forest Algorithm	Random Forest (RF) and Extreme Gradient Boosting (XGBoost).	UCI Machine Learning Repository and available in online.	90.01%
[8]	Machine Learning Approach for Breast Cancer Prediction: A Review	SVM, Random Forest, KNN,	WDBC data set	91.25%
[9]	Machine Learning Based Approach for Breast Cancer Detection	KNN, Logistic Regression, Naive Bayes, SVM, Decision Tress and Random Forest, Logistic Regression	Breast cancer Dataset.	97.13%, 98.83%, 97.37%
[10]	Optimized Ensemble Prediction Model for Breast Cancer	Stacked Generalization Ensemble Model	Breast cancer dataset from Wisconsin, maintained by the Universfornia.	99.41%

V. CONCLUSION

Breast cancer is the most common and hazardous cancer faced by large percentage of women all over the world. By this model, we can predict the symptoms of the cancer before it spreads in large number. In the advanced computational world, it is possible to predict using machine learning algorithms and can prevent and can be diagnosed properly at the starting stage of the cancer. This will allow us to slowly lower the mortality rate. Developing a prediction model for breast cancer is the major objective of this proposed model. The different classifiers are used in the research and it is observed that logistic regression outperformed comparing to other machine learning algorithm with the accuracy of 98.74%.

REFERENCES

[1] S. Kabiraj, L. Akter, M. Raihan, N. J. Diba, E. Podder and M. M. Hassan, "Prediction of Recurrence and Non-recurrence Events of Breast Cancer using Bagging Algorithm," *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, 2020, pp. 1-5, doi: 10.1109/ICCCNT49239.2020.9225440.

[2] R. Kumar, M. Chaudhry, H. K. Patel, N. Prakash, A. Dogra and S. Kumar, "An Analysis of Ensemble Machine Learning Algorithms for Breast Cancer Detection: Performance and Generalization," *2024 11th International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, India, 2024, pp. 366-370, doi: 10.23919/INDIACom61295.2024.10498618

[3] Hasan, Srwa & Sagheer, Ali , Veisi, Hadi. (2021). Breast Cancer Classification Using Machine Learning Techniques: A Review. Turkish Journal of Computer and Mathematics Education (TURCOMAT). 12. 1970-1979.

[4] Dubey, Chhaya & Shukla, Nidhi & Kumar, Dharmendra Singh, Ashutosh & Dwivedi, Vijay. (2022). Breast Cancer Modeling and Prediction Combining Machine Learning and Artificial Neural Network Approaches. 119-124. 10.1109/ICCCIS56430.2022.10037709.

[5] Bista, Chirayou M, Asreetha ,Slimanzay, Salahuddin Sheikh, Md & Rao, P. (2024). Breast Cancer Prediction System Utilizing Machine Learning Algorithms. 80-84. 10.1109/IEEECONF61558.2024.10585589.

[6] Nathiya, S. & Sumitha, J.. (2021). A Comparative Study on Breast Cancer Prediction using Optimized Algorithms. 1401-1405. 10.1109/ICOSEC51865.2021.9591787.

[7] Raihan, M.. (2020). Breast Cancer Risk Prediction using XGBoost and Random Forest Algorithm. 10.1109/ICCCNT49239.2020.9225451.

[8] Choudhari, Gauri , Desai, Rutuja , Dagale, Pratik , Dashetwar, Isha. (2023). Breast Cancer Prediction Using Various Machine Learning Algorithms and Their Comparative Analysis. 1-6. 10.1109/PuneCon58714.2023.10450082.

[9] Hegde, Harshita Kodipalli, Ashwini. (2022). Machine Learning Based Approach for Breast Cancer Detection. 782-786. 10.1109/ICCCIS56430.2022.10037645.

[10] Aditya, Jatin. (2021). Optimized Ensemble Prediction Model for Breast Cancer. 1-4. 10.1109/ITSS-IoE53029.2021.9615269.

[11] Rovshenov, Atajan, Peker, Serhat. (2022). Performance Comparison of Different Machine Learning Techniques for Early Prediction of Breast Cancer using Wisconsin Breast Cancer Dataset. 1-6. 10.1109/IISEC56263.2022.9998248.

[12] Yifeng, Dou , Jinsong, Lv , Wentao, Meng. (2024). Study on Breast Cancer Classification Prediction based on XGBoost. 411-416. 10.1109/IMCEC59810.2024.10575645.

[13] Padmapriya, B ,Thambusamy, Velmurugan. (2016). Classification Algorithm Based Analysis of Breast Cancer Data. International Journal of Data Mining Techniques and Applications. 5. 10.20894/IJDMTA.102.005.001.010.

[14] Egwom, Onyinyechi , Hamada, Mohamed , Yusuf, Saratu , Hassan, Mohammed. (2021). The Role of Linear Discriminant Analysis for Accurate Prediction of Breast Cancer. 340-344. 10.1109/MCSoC51149.2021.00057.

[15] Kaur, Arpanpreet ,Gupta, Sheifali. (2024). Unveiling Precision in Breast Cancer Prediction with Random Forest and Decision Trees. 1232-1236. 10.1109/ICOSEC61587.2024.10722493.

[16] B. S., M. A. M. J., P. B. N., S. F. P. and K. A, "Breast Cancer Classification and Recurrence Prediction Using Artificial Neural Networks and Machine Learning Techniques," 2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), Trichirappalli, India, 2023, pp. 1-4, doi: 10.1109/ICEEICT56924.2023.10157890.

[17] Prem Kumar, Josephine. (2020). Identification of Factors Causing Breast Cancer using Factor Analysis. International Journal of Engineering Research and. V9. 10.17577/IJERTV9IS020031.