Prediction of Cancer in Breast Using Machine Learning Techniques: A Comparative Study of Classification Algorithms

Vasu Bansal

Dept. of Computer Science and Engineering

Affiliation to AKTU

KIET Group of Institutions

Affiliation to AKTU

Ghaziabad,India

vasubansal3926@gmail.com

Shivangi Gaur

Dept. of Computer Science and Engineering

Affiliation to AKTU

KIET Group of Institutions

Affiliation to AKTU

Ghaziabad,India

shivangigaur1002@gmail.com

Dr. Seema Maitrey

Dept. of Computer Science and Engineering

Affiliation to AKTU

KIET Group of Institutions

Affiliation to AKTU

Ghaziabad,India

Seema.maitrey@kiet.edu

***Abstract*—Cancer in breast is a frequently occurring type of cancer that impacts individuals globally and can be fatal if not detected and treated in a timely manner. Depending on the tools, datasets, and conditions, researchers have investigated a variety of strategies, including ML techniques , DL methods, and data mining techniques, to predict breast cancer with variable degrees of accuracy. The main aim of this study is to analyze and distinguish the existing machine learning and data mining techniques to determine the most effective way for correctly predict breast cancer utilizing large datasets. The main objective is to give newcomers helpful knowledge on understanding machine learning algorithms and laying the groundwork for deep learning.**

This study looked at several studies that show the usefulness of DL and ML algorithms in detecting breast cancer across multiple datasets. SVM, KNN, RF, ANN, and CNN were discovered to have the highest accuracy rates among the algorithms. Future study will focus that allows users to calculate their risk of developing breast cancer.

Keywords-- Breast cancer prediction, Naïve Bayes

I. INTRODUCTION

A lot of people die each year from the dangerous and widespread disease known as breast cancer. According to the World Health Organization, 2.9 million women could pass away from breast cancer worldwide. After skin cancer, the second most dangerous cancer in wome in the US is breast cancer. In the U. S. A. in 2021, the Society of American Cancer predicts that there will be 48,530 new instances of non-invasive (in situ) breast cancer and approximately 284,200 new cases of invasive breast cancer. However, with an anticipated 685,000 deaths from breast cancer in 2020, it will continue to be a leading cause of death for women worldwide.[4] The World Health Organization (WHO) states that the leading cause of death for women is breast cancer (BC) [1].

A. The indications and manifestations of breast cancer include:-

Breast cancer symptoms can vary among individuals.

While some may not experience any symptoms, others may notice a new lump in the breast or underarm. Additional indications of breast cancer may involve enlargement or hardening of a specific breast area, puckering or inflammation of the breast skin, or the appearance of reddened or scaly skin surrounding the breast or nipple, nipple soreness or pulling inwards, discharge from the nipple that is not breast milk, any changes in the size or shape of the breast, and breast pain in any area.

Machine learning [2] and data mining methods are methods are utilized to predict breast cancer, and Identifying the appropriate algorithm for the job is the obstacle to overcome. Breast cancer emerges from malignant tumors that arise when cellular growth becomes unmanageable, resulting in a typical multiplication of adipose tissue. There are different types of breast cancer, such as DCIS, IDC, MTBC, LBC, MBC, and IBC.

These categories of breast cancer [4] happen when cancerous cells and tissues metastasize to other parts of the body.

DCIS is a form of breast cancer that is sometimes referred to as non-invasive cancer. It arises when abnormal cells grow beyond the breast. [5]

The second type of breast cancer is known as infiltrative ductal carcinoma (IDC) [7] or invasive ductal carcinoma, depending on the classification. IDC is more commonly found in men, and it happens when abnormal breast cells spread throughout the entire breast tissue.

Mixed Tumors Breast Cancer (MTBC) [9], also called invasive mammary breast cancer, is the third category of breast cancer. This type of cancer develops when abnormal cells in the ducts and lobules of the breast grow uncontrollably.

Lobular breast Cancer (LBC) is the fourth type of breast cancer and it originates in the lobules of the breast. This cancer increases the risk of developing more invasive types of breast cancer.

Colloid breast cancer, also known as mucinousbreast.is the fifth kind of breast cancer. It develops from invasive ductal cells, and it takes place when anomalous tissue surrounds the duct.

Inflammatory Breast Cancer (IBC) is the last form which leads to breast inflammation and redness. IBC is a rapidly growing cancer that occurs when cancer cells block the lymph vessels.

Finding Data mining involves the retrieval of significant details from an extensive collection of data. Various techniques, such as machine learning, statistics, databases, fuzzy sets, data warehouses, and neural networks, can be applied to analyze and predict the prognosis of various types of cancer, including prostate cancer, lung cancer, and leukemia[15, 16]. Unlike the conventional method of cancer diagnosis, which involves three tests --, radiological imaging, clinical examination and pathology examination[18] -- modern machine learning approaches and algorithms focus on creating models to forecast expected outcomes using unknown data. Machine learning is based on three primary methodologies: preprocessing, classification, feature selection or extraction. Feature extraction is particularly important, as it can help to find between benign and malignant tumors. The primary goal of creating a model is to forecast and obtain accurate results during both the training and testing phases, using data that was previously unknown. [21].

II. LITERATURE REVIEW

This section discusses about few of the previous work on breast cancer treatment ways done by researchers using various ML methods.

Table I show some of the things previously done on breast cancer diagnosis by researchers using different methods, including ML, deep learning, which are discussed.

TABLE I. REVIEW OF PAPERS

Reference	**Study**	**Published Year**
[3]	Proposed BCAD Framework, Multilayer Perceptron Model, 99.12% Accuracy.	2022

[4]	Analysis of MRI imaging and gene sequencing methods applied to various data sources.	2022
[5]	When compared to deep learning models, the GB-based model offers more accuracy along with faster classification and requires less computer power.	2020
[6]	This study introduced an uncomplicated and efficient technique for categorizing HE stained histological images of breast cancer, even with limited training data. (328 samples).	2022
[7]	The SVM, NN, RF, DT and LR classification models are used. Random forest is most effective method.	2018
[8]	Random forest produced the best accuracy (82.7%), whereas decision trees produced the lowest accuracy (79.8%).	2019
[9]	Within this research, Random Forest classifier showed the best results in terms of precision and execution time.	2020
[10]	The paper comes to the conclusion that deep learning models perform more accurately than machine learning algorithms.	2020
[11]	Researchers can use several data augmentation strategies to address the problem of the small amount of available dataset. Researchers should take into account the issue of the disparity between positive and negative data since it can result in bias towards either a positive or negative prediction. For accurate breast cancer diagnosis and prognosis, an essential problem with an uneven number of breast cancer photos against affected patches needs to be resolved..	2020
[12]	SVM shows higher accuracy(97.13%) compared to C4.5, Naïve Bayes, and k-NN.	2020
[13]	SVM, KNN, and LR provide the most accurate results among the three methods implemented.	2019
[14]	KNN has the highest accuracy among Support Vector Machine, Naive Bayes, and Logistic Regression.	2020
[15]	SVM has the best performance with a 97.07% perfect prediction accuracy when compared to NB, KNN, RF, DT, and LR.	2020
[16]	The highest accuracy is achieved by Adam Gradient Learning, which combines the advantages of AdaGrad and RMSProp.	2020
[20]	KNN classifier performs best when used with the most predictive variables.	2018
[21]	In a study using Wisconsin Breast Cancer (original) datasets, SVM, NB, KNN, and C4.5 algorithms were employed, and SVM achieved the best performance in terms of precision and low error rate in 2016	2016

II. METHODOLOGY

We must first gather the data and execute pre-processing in order to create machine learning methods for breast cancer prediction. The integral components of the pre-processing phase are data cleaning, attribute selection, target role definition, and feature extraction.

Properly processed data is vital for creating a machine learning model that can accurately predict breast cancer. To evaluate the performance of the algorithms, a new set of labeled data is used. A common method is to divide the labeled data into two parts using the Train_test_split approach. The training data consists of three-fourths of the total data, while the remaining one-fourth is reserved for testing the model's accuracy. This is illustrated in the figure.

Based on the findings, the algorithm that predicts the presence of breast cancer with the highest degree of accuracy is chosen. The algorithms that offer the highest level of accuracy and dependability in predicting the existence of breast cancer are identified after models have been examined.

III. DATASET

Various datasets are used for breast cancer prediction, such as Coimbra Breast Cancer (CBC), Wisconsin (Prognostic) Breast Cancer (WPBC), Wisconsin (Diagnostic) Breast Cancer, Wisconsin Original Breast Cancer (WOBC), and Breast Tissue Dataset (BTD). The study in question utilized the WBCD, which contains information about both malignant and benign breast cancer. The dataset includes thirty features that are obtained from fine-needle aspiration (FNA) of the breast mass. Unlike typical cancer datasets that consist of images, the WBCD consists of feature vectors that describe the cell nuclei in the image. These vectors consist of ten real-valued features that characterize the cell nuclei

TABLE II. FEATURES TABLE

Features		
Radius_mean	Texture	Radius worst
Perimeter_mean	Smooth_mean	Texture worst
Concavity_mean	Concave_pts.	Perimeter worst
Fractal_dim.	SE_Compact	Area worst

IV. MACHINE LEARNING ALGORITHMS

Our research involves carrying out predictive analysis of data using machine learning algorithms. The algorithms that are used in our project are:

A. *SUPPORT VECTOR MACHINE*

It is a classification algorithm that partitions the dataset into different classes using the nearest points of data to identify the maximum margin hyperplane (MMH).

B. Random forest

Random forests or decision forests, on the other hand, are an ensemble method used for various tasks such as classification and regression. In this method, throughout the training phase, multiple decision trees are developed and outcome is mean class. Random forests are used to overcome the tendency of decision trees to overfit their training data.

C. k-Nearest Neighbors (K-NN)

To acquire the ability to label new data points, the algorithm requires a vast quantity of labeled data points. When presented with a new data point, the algorithm evaluates its nearest neighbors, or those in close proximity, and seeks their input in predicting its label.

•

D. Logistic regression

It is a widely used and effective modeling technique that extends linear regression. It is commonly used to evaluate the probability of a health condition or disease based on a risk factor. Logistic regression is employed to analyze the connection between independent variables, such as predictor or exposure variables (Xi), and a binary dependent variable, which is also known as an outcome or response variable (Y). Logistic regression's primary application is forecasting binary or multiclass dependent variables, and it is employed in both single and multiple logistic regression scenarios.

E. Decision Tree

It is a versatile predictive modeling technique that can be utilized in a variety of domains. By using an algorithmic approach, it can partition the dataset into multiple segments based on diverse standards.

V. RESULT AND DISCUSSION

On the Wisconsin dataset, several machine learning techniques were applied to predict breast cancer, including K-NN, SVM, Decision Tree, Nave Bayes Logistic Regression, and Random Forest. The precision attained by ANN, CNN, SVM, Random Forest, and KNN algorithms.

These algorithms reached more than 95%,and deep learning algorithms like Convolutional

Neural Network (CNN) and Artificial Neural Network (ANN) were used to improve prediction accuracy.

TABLE III. ALGORITHM ACCURACY TABLE

Algorithm	**Accuracy(%)**	**Precision(%)**	**Sensitivity(%)**
ANN	99	99	99
CNN	97	97	98
SVM	96	98	97
Random Forest	96	98	97
KNN	95	95	99
Decision tree	95	99	93
Logistic Regression	94	96	96
Naïve Bayes	92	93	94

VI. CONCLUSION AND FUTURE SCOPE

In this study, we looked at several studies that discovered that deep learning algorithms, machine learning approaches, and some proposed algorithms provide improved accuracy for detecting breast cancer on various datasets. However, machine learning algorithms like SVM, KNN, RF, ANN, and CNN provide the highest level of accuracy.

In the future, models based on these methodologies will be created in an effort to provide software for the breast cancer prediction so that the user can access themselves.

REFERENCES

[1]‘WHO|Breastcancer’,WHO, http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/

[2] “UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set.”

[3]S. Aamir, A. Rahim, Z. Aamir, S. F. Abbasi, M. S. Khan, M.Alhaisoni, M. A. Khan, K. Khan, and J. Ahmad “Predicting Breast Cancer Leveraging Supervised Machine Learning Techniques,” August 16,2022.

[4]A.B. Nassif, M.A. Talib, Q. Nasir, Y.Afadar,and O.Elgendy“Breast cancer detection using artificial intelligence techniques: A systematic literature review,” March 5,2022.

[5]H.E. Agouri, M. Azizi, H. E. Attar, M. E. Khannoussi, A. Ibrahimi, R. Kabbaj, H. Kadiri,

[6]S. BekarSabein, S. EchCharif, C. Mounjid, B. E. Khannoussi, “Assessment of deep learning algorithms to predict histopathological diagnosis of breast cancer: first Moroccan prospective study on a private dataset,” Feb 19,2022.

[7]M.Mangukiya, A.Vaghani, M.Savani, “Breast Cancer Detection with Machine Learning,” Feb 2022.

[8]A. Rasool, C.Bunterngchit, L.Tiejian, Md. R. Islam, Q. Qu, and Q. Jiang, “Improved Machine Learning-Based Predictive Models for Breast Cancer Diagnosis” March 9,2022.

[9]H. Zhang, H. Liu, L. Ma, J. Liu, D. Hu “Ultrasound Image Features under Deep Learning in Breast Conservation Surgery for Breast Cancer” Sept 17,2021

[10]S. Bhise, S. Bepari, S. Gadekar, D. Kale, A. S. Gaur, Dr. S. Aswale, “Breast Cancer Detection using Machine Learning Techniques,” July 7,2021.

[11]S. Aryal, B. Paudel, “Supervised Classification using Gradient Boosting Machine: Wisconsin Breast Cancer Dataset,” June 2020.

[12] M. Srivenkatesh, “Prediction of Breast Cancer 0Disease using Machine Learning Algorithms,” Feb 4,2020.

[13]N. Fatima, L. Liu, S. Hong, H. Ahmed, “Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis,” Aug 14,2020.

[14]R. Rawal, “BREAST CANCER PREDICTION USING MACHINE LEARNING,” May 2020.

[15]Gaurav Singh, “Breast Cancer Prediction Using Machine Learning,” July 30,2020.

[16] M.Javed Mehedi Shamrat, Md. Abu Raihan, A.K.M.Sazzadur Rahman, Imran Mahmud, R. Akter, “An Analysis on Breast Disease Prediction Using Machine Learning Approaches,” Feb 2020.

[17] P. Gupta, S. Garg, “Breast Cancer Prediction using varying Parameters of Machine Learning Models,” June 4,2022.

[18]M. D. Ganggayah, N.A Talib, Y. C. Har, P. Lio, S. K. Dhillon, “Predicting factors for survival of breast cancer patients using machine learning techniques,” March 22, 2019.

[19]Ch. Shravya, K. Pravalika, Shaik Subhani, “Prediction of Breast Cancer Using Supervised Machine Learning Techniques,” April 2019.

[20]Yixuan Li, Zixuan Chen, “Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction,” Oct 18,2018.

[21]Madhu Kumari, Vijendra Singh, “Breast Cancer Prediction System,” June 8,2018.

[22] H. Asri, H. Mossanif, H. Al Moatassime, T. Noel, “Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis,” May 12,2016.

[23]K. Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, Dimitrios I. Fotiadis, “Machine learning applications in cancer prognosis and prediction,” Nov 15,2014.