CANCER TUMOR DETECTION USING MACHINE LEARNING

L.Pradeep
Information Technology
Institute Of Science
Technology
Hyderabad, Telangana, India.

Shaik Faisal
Information Technology
Sreenidhi Institute Of Science
and Technology
Hyderabad, Telangana, India.

Syed Obaid
Information Technology Sreenidhi
Sreenidhi Institute Of Science
and Technology
Hyderabad, Telangana, India.

Dr.B.Indira
Professor
Information Technology
Sreenidhi Institute Of Science and
Technology
Hyderabad, Telanganga, India.

Dr.M.Sreenivas
Associate Professor
Information Technology
Sreenidhi Institute Of Science and
Technology
Hyderabad, Telanganga, India.

***

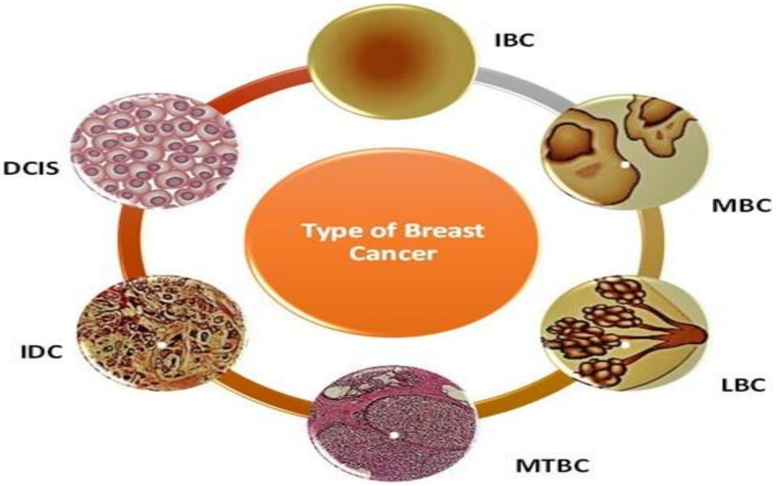
ABSTRACT :

Breast tissue can develop tumours of the breast cancer variety. It is the most prevalent type of cancer in females around the world and one of the main causes of death in females. This piece gives a comparison of the data mining, machine learning, and deep learning methods used to detect breast cancer.

Numerous researchers have worked to improve breast cancer diagnosis and prognosis; nevertheless, each technique has a distinct accuracy rate that changes depending on the circumstances, resources, and datasets employed. Our primary goal is to compare and contrast various Machine Learning and Data Mining approaches currently in use in order to identify the most effective approach that will support the enormous dataset with the best possible prediction accuracy. The major goal of this study is to highlightall the prior research on machine-learning algorithms that have been used to predict breast cancer, and this article gives all the knowledge a novice needs to understand machine learning algorithms and build a solid foundation for deep learning.

INTRODUCTION:

In the modern day, breast cancer is one of the most deadly and diverse diseases, killing a huge number of women all over the world. It is the second most common illness that kills women. Different machine learning and data mining methods are being applied for breast cancer prediction. One of the key tasks is to find the most acceptable and suitable algorithm for breast cancer prediction. Malignant tumours, which form when a cell's development spirals out of control, are the cause of breast cancer. Breast cancer is brought on by the abnormal proliferation of numerous fatty and fibrous breast tissues. Tumors that cause various stages of cancer have cancer cells that have spread throughout them. Breast cancer can take many distinct forms, and it develops when damaged cells and tissues are dispersed all across the body. DCIS, commonly referred to as non-invasive cancer, is a kind of breast cancer that develops when abnormal cells move outside the breast. The second kind is Infiltrative Ductal Carcinoma (IDC) , which is sometimes referred to as Invasive Ductal Carcinoma (IDC) . IDC cancer is typically observed in men, and it develops when breast aberrant cells expand throughout all breast tissues. The third subtype of breast cancer is known as Mixed Tumors Breast Cancer (MTBC), which is also referred to as invasive mammary breast cancer. Such cancers are brought on by abnormal duct and lobular cells.



Lobular Breast Cancer (LBC) [11] is the fourth form of cancer and develops inside the lobule. It raises the risk of developing more invasive malignancies. Colloid breast cancer, also known as mucinous breast cancer (MBC) [12], is the fifth kind of breast cancer that arises from invasive ductal cells. When aberrant tissues surround the duct, it happens [13]. IBC (Inflammatory Breast Cancer) is the most recent form that results in swelling and breast reddening When lymphatic channels become blocked in break cells, a breast cancer of this type begins to develop quickly [14].

Data mining is the process of extracting useful information from large datasets. Data mining functions and techniques can be used to identify any type of disease. For example, machine learning, statistics, databases, fuzzy sets, data warehouses, and neural networks can be used to diagnose and predict the prognosis of various cancer diseases, including prostate cancer, lungs cancer, and leukaemia [15]. The "gold standard" approach, which entails three procedures (clinical examination, radiological imaging, and pathology test), forms the foundation of traditional cancer detection methodology [18]. While the latest machine learning approaches and algorithms are based on model creation, the conventional method uses regression to signal the existence of cancer. The model is created to predict unknown data and delivers the predicted results well throughout training and testing [19]. Preprocessing, features selection or extraction, and classification are the three primary methodologies on which machine learning is founded [20]. The main component of machine learning, feature extraction, aids in the diagnosis and prognosis of cancer and may distinguish between benign and malignant tumours [21]. We can diagnose and anticipate certain types of breast cancer, like the one depicted in Figure 1, thanks to data mining and machine learning algorithms. Classification, regression, and clustering are a few data mining techniques

[22] that assist us in obtaining useful data on breast cancer patients. These algorithms [23] include training datasets, and by using these datasets, we can determine the likelihood of predicting various types of breast cancer [24].

II. MACHINE LEARNING ALGORITHMS FOR BREAST CANCER PREDICTION:

We input a vast amount of data, the machine learning model analyses that data, and on the basis of that trained model, we can make a prediction about the future [24], [26], [27]. Machine learning is an automatic learning method [25]. The following are the main machine learning algorithms for predicting breast cancer:

A. ARTIFICIAL NEURAL NETWORK (ANN):

An efficient approach for data mining is the artificial neural network [28]. Input, hidden, and output layers make up a neural network. This method is employed to extract the too-complex patterns [29]. The algorithm is based on network architecture [32]–[34], distributed memory [31], collaborative solutions, and parallel processing [30].

B. LOGISTICS REGRESSION (LR):

The algorithm is supervised learning and has more dependent variables. This algorithm's output takes the form of a binary number. Regression in logistics [35] can offer a continuous result for a certain data. A statistical model with binary variables makes up this method [32].

C. K-NEAREST NEIGHBOR (KNN)

In order to recognise patterns, this method is utilised. It is an effective strategy for predicting breast cancer. Every class received the same amount of attention in order to spot the trend. From a sizable dataset, K Nearest Neighbor [36] extracts the related highlighted data. We classify a sizable dataset on the basis of feature similarity [32].

D. DECISION TREE (DT)

Classification and regression models are the foundation of decision tree [37]. The data set is broken up into fewer subsets. The best degree of precision in prediction may be achieved using these smaller sets of data. CART [38], C4.5 [39], C5.0 [40] and conditional tree [32, [41] are among the decision tree methods.

E. NAIVE BAYES ALGORITHM (NB)

With this approach, a sizable training dataset is assumed. The Bayesian approach is employed in the algorithm to calculate probability [42]. When determining the input probabilities of noisy data, it offers the maximum accuracy [43]. This classifier uses analogies to compare training datasets and training tuples [32].

F. SUPPORT VECTOR MACHINE (SVM)

Both classification and regression issues are addressed by this supervised learning system [44]. It uses mathematical and theoretical functions to address the regression issue. When making predictions using a huge dataset, it offers the highest accuracy rate. Based on 3D and 2D modelling, it is a powerful machine learning technique [32], [45].

G. RANDOM FOREST (RF)

The supervised learning-based Random Forest algorithm [46] is used to address classification and regression issues. It is a machine learning building component that is used to predict new data based on historical datasets [32].

H. K MEAN ALGORITHM

With the help of the clustering method K mean, data can be divided into small groups. To determine the degree of similarity between various data points, algorithms are used. The most appropriate cluster for evaluating a large dataset is present in every data point [48].

K. GAUSSIAN MIXTURE ALGORITHM

It is the unsupervised learning method that is most widely used. The method of computing the likelihood of various forms of clustered data is referred to as the soft clustering methodology. This algorithm's implementation is based on expectation maximisation [51].

III. ENSEMBLE TECHNIQUES FOR BREAST CANCER PREDICTION

Both homogeneous and heterogeneous ensemble techniques can be used; homogeneous ensemble techniques [52] combine one base method with two or more configuration methods, such as bagging and boosting technique, while heterogeneous ensemble techniques [53]-[55] combine two or more base methods. Ensemble techniques are based on supervised learning, which offers accurate predictions based on specific hypotheses.

A. BAGGING

The other name of the bagging technique is bootstrap aggregation which is used for the prediction of any disease. It is based on multiple models, [54] each model is trained separately and then combined together for prediction [52].

B. BOOSTING

Boosting is homogenous week learner that creates one strong classifier from some weak classifiers [52]. It is based on step by step strategies for building up the model from some training data [54], [55].

C. STACKING

For prediction on the same dataset, stacking is a heterogeneous [52] weak learner that integrates many machine learning techniques. It is made up of two or more basic models and combines their predictions [54, 55].

V. SURVEY ON BREAST CANCER

The world's most populous nation is China. Males have breast cancer at a rate of 8.6%, whilst females experience it at a rate of 19.2%, according to a recent organisation report (GLOBOCAN-2018) [65]. Everyyear, 1.2 million people pass away from this illness. The American Cancer Society identified 48,100 incidences of DCID cancer in female patients. According to a US 2019 study, 41,760 women and 500 men are anticipated to pass away from breast cancer [66]. According to a US survey, there are 3.8 million women who are still living but are battling breast cancer. 2019 saw 59,838 incidences of Ductal Carcinoma in Situ (DCIS) breast cancer in US women [67]. 458,000 people have died from breast cancer worldwide. Chinese women died from breast cancer at a rate of 48% in 2012, compared to a global death rate of 52% [68]. Data from 1,517 women were examined in 2015 to determine the breast cancer survival and recurrence rates; the breast cancer recurrence rate was 100 and the mortality rate was 132[69].

VI. REVIEW OF MACHINE LEARNING ALGORITHMS FOR BREAST CANCER PREDICTION

The major goal of this study is to evaluate several machine learning and data mining methods that have aided in breast cancer prediction. Finding the most precise and appropriate algorithm for breast cancer prediction is our main goal. In order to do this, we've gone over and examined previous research on breast cancer prediction algorithms. additionally examined research publications based on linear, nonlinear, naive bayes, K-nearest neighbour, support vector machine, and certain ensemble algorithms (Linear Regression, Logistic Regression, Linear Discriminant Analysis) (Decision Tree, Random Forest, Boosting and AdaBoost). The vast majority of researchers combined linear and nonlinear or nonlinear and ensemble techniques. As a result, we have divided our review article into sections that will compare and contrast each algorithm based on its accuracy level. Following that comparison, we will highlight the best machine learning method for predicting breast cancer.

CONCLUSION

In this paper, we have examined various data mining, machine learning, and deep learning methods for the prediction of breast cancer. Finding the best algorithm to more accurately forecast the onset of breast cancer is our key goal. This article's main goal is to showcase all of the prior research on machine learning algorithms that have been used to predict breast cancer. It also gives newcomers all the information they need to understand machine learning algorithms and provide the groundwork for deep learning. The review of this article begins with a discussion of the many forms of breast cancer. To learn more about the main forms, symptoms, and causes of breast cancer, fourteen research publications were examined. Following that, a review of the most important machine learning, ensemble, and deep learning approaches was given. These techniques greatly elaborate the algorithms that are used to forecast breast cancer. There are still certain problems that will need to be resolved in future development. 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 afflicted patches needs to be resolved.

REFERENCES:

Wang, D. Zhang and Y. H. Huang "Breast Cancer Prediction Using Machine Learning" (2018), Vol. 66, NO. 7.

B. Akbugday, "Classification of Breast Cancer Data Using Machine Learning Algorithms," 2019 Medical Technologies Congress (TIPTEKNO), Izmir, Turkey, 2019, pp. 1-4.

Keles, M. Kaya, "Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: A Comparative Study." Tehnicki Vjesnik - Technical Gazette, vol. 26, no. 1, 2019, p. 149+. [4] V. Chaurasia and S. Pal, "Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability", IJCSMC, Vol. 3, Issue. 1, January 2014, pg.10 – 22.

Delen, D.; Walker, G.; Kadam, A. Predicting breast cancer survivability: A comparison of three data mining methods. Artif. Intell. Med. 2005, 34, 113–127.

R. K. Kavitha1, D. D. Rangasamy, "Breast Cancer Survivability Using Adaptive Voting Ensemble Machine Learning Algorithm Adaboost and CART Algorithm" Volume 3, Special Issue 1, February 2014 [7] P. Sinthia, R. Devi, S. Gayathri and R. Sivasankari, "Breast Cancer detection using PCPCET and ADEWNN",CIEEE' 17, p.63-65

Vikas Chaurasia and S.Pal, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis" (FAMS 2016) 83 (2016) 1064 – 1069

N. Khuriwal, N. Mishra. "A Review on Breast Cancer Diagnosis in Mammography Images Using Deep Learning Techniques", (2018), Vol. 1, No. 1.

Y. Khourdifi and M. Bahaj, "Feature Selection with Fast Correlation-Based Filter for Breast Cancer Prediction and Classification Using Machine Learning Algorithms," 2018 International Symposium on Advanced Electrical and Communication Technologies (ISAECT), Rabat, Morocco, 2018, pp. 1-6.

R. M. Mohana, R. Delshi Howsalya Devi, Anita Bai, "Lung Cancer Detection using Nearest Neighbour Classifier", International Journal of Recent Technology and Engineering (IJRTE), Volume-8, Issue-2S11, September 2019

Ch. Shravya, K. Pravalika, Shaik Subhani, "Prediction of Breast Cancer Using Supervised Machine Learning Techniques", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8 Issue-6, April 2019.

Haifeng Wang and Sang Won Yoon, "Breast Cancer Prediction Using Data Mining Method", Proceedings of the 2015 Industrial and Systems Engineering Research Conference, [14] Abdelghani Bellaachia, Erhan Guven, "Predicting Breast Cancer Survivability Using Data Mining Techniques"