CANCER TUMOR DETECTION USING MACHINE LEARNING

Mandava Vasu
Compter Science Engineeering
Sir Padampat Singhania University
Udaipur, India
mandava.vasu@spsu.ac.in

***Abstract*—Breast cancer-related tumours can form in breast tissue. It is one of the leading causes of death for women worldwide and the most common type of cancer in females. This article contrasts the detection of breast cancer using data mining, machine learning, and deep learning. However, each technique has a unique accuracy rate that varies based on the circumstances, resources, and datasets used. Many researchers have sought to enhance breast cancer detection and prognosis. Our main objective is to evaluate and contrast the many Machine Learning and Data Mining approaches that are currently in use in order to determine which strategy is the most efficient and will support the massive dataset with the highest prediction accuracy. This article provides all the information a novice needs to comprehend machine learning algorithms and lay a strong foundation for deep learning. The main objective of this study is to highlight all the earlier research on machine-learning algorithms that have been used to detect breast cancer.**

I. INTRODUCTION

Today, one of the most terrible and diversified diseases is breast cancer, which claims the lives of a great deal of people worldwide. It ranks as the second most frequent cause of death for women [1]. For the purpose of predicting breast cancer, various data mining and machine learning techniques [2] are being used. Finding the best and most appropriate algorithm for breast cancer prediction is one of the most important jobs. Breast cancer is brought on by malignant tumours, which arise when a cell's growth spirals out of control [3]. The aberrant development of multiple fatty and fibrous breast tissues is what causes breast cancer. Cancer cells have spread all over tumours that create different stages of cancer. Breast cancer occurs when damaged cells and tissues are disseminated throughout the body, and it can appear in a variety of unique ways [4]. Breast cancer that develops when abnormal cells spread outside the breast is known as DCIS, sometimes known as non-invasive cancer [5]. Infiltrative Ductal Carcinoma (IDC) [7], also known as Invasive Ductal Carcinoma (IDC) [6], is the second kind. IDC cancer develops when breast aberrant cells spread throughout all breast tissues and is frequently seen in men [8]. Mixed Tumours Breast Cancer (MTBC), commonly known as invasive mammary breast cancer, is the third subtype of breast cancer The fourth type of cancer, called lobular breast cancer (LBC) [11], starts inside the lobule. It increases the chance of getting more aggressive forms of cancer. The fifth type of breast cancer to develop from invasive ductal cells is called colloid breast cancer, sometimes referred to as mucinous breast cancer (MBC) [12]. It occurs when abnormal tissues encircle the duct [13]. The most recent type of breast cancer, known as IBC (Inflammatory Breast Cancer), causes breast swelling and reddening. This kind of breast cancer rapidly progresses when lymphatic pathways in broken cells are shut [14]. The technique of removing valuable information from huge databases is known as data mining. Any sort of disease can be identified using data mining techniques. For instance, a variety of cancer disorders, such as prostate cancer, lungs cancer, and leukaemia, can be diagnosed and their prognosis can be predicted using machine learning, statistics, databases, fuzzy sets, data warehouses, and neural networks [15]. The basis of conventional cancer diagnosis methods is the "gold standard" approach, which consists of three steps (clinical examination, radiological imaging, and pathology test) [18]. The typical methodology uses regression to indicate the presence of cancer, whereas the most recent machine learning methodologies and algorithms are based on model creation. The model was developed to forecast unknown data, and it successfully produced the anticipated outcomes during training and testing [19]. The three main approaches on which machine learning is based are preprocessing, features selection or extraction, and classification [20]. Feature extraction, the primary machine learning component, assists in the diagnosis and prognosis of cancer and may be able to differentiate between benign and malignant tumours [21]. Thanks to data mining and machine learning algorithms, we can identify and predict specific types of breast cancer, like the one shown in Figure 1. We can get important information about breast cancer patients by using data mining techniques including classification, regression, and clustering [22]. We can assess the chances of predicting different forms of breast cancer [24] by using the training datasets included in these algorithms [23]. This article is divided into various sections. Section II discusses the primary machine learning methods for breast

cancer prediction. Section III covers the main ensemble methods for predicting breast cancer; Section IV discusses deep learning techniques for diagnosing breast cancer; Section V provides a survey of breast cancer; Section VI reviews various machine learning and deep learning algorithms; Section VII summarises the studies and resources we used in this research; Section VIII provides the discussion; and Section IX brings the study to a close.

A. Literature Review

There are several applications for Machine Learning, the most significant of which is Data Mining. People are often prone to making mistakes during analysis or when trying to establish relationships between multiple features. There have been many attempts by researchers to build an accurate model for predicting the malignancy of a tumor. One such model was built by Roberto Lopez.

His learning algorithm was trained by a dataset from the University of Wisconsin- Madison. His algorithm was trained by approximately 600 patients covering over 10 features with a training accuracy of 71.

Thus the main problem with such training models is the low accuracy achieved due to anomalies in the dataset as well as the gradient descent ending up at the local minima instead of the global minima.

As we increase the number of features to be considered, the complexity of our model increases which may become more difficult to visualize. In this paper, two features have been considered, namely Clump Thickness and Marginal Adhesion. The data for this problem has been taken from the UCI Machine Learning Repository.

Data of over 100 patients was recorded and analyzed by our learning algorithm. We try to estimate the probability and predict whether the patient's tumor is malignant or benign. Several studies have been reported that have focused on breast cancer survivals. These studies have applied different approaches to the given problem and achieved high classification accuracies. Details of some of the previous research works are given in the following:

Liu et al.3 used decision table (DT)-based predictive models for breast cancer survivability, concluding that the survival rate of patients was 86.52 percent. They employed the undersampling C5 technique and bagging algorithm to deal with the imbalanced problem, thus improving the predictive performance on breast cancer.

Tan and Gilbert4 demonstrated the usefulness of employing ensemble methods in classifying microarray data and presented some theoretical explanations on the performance of ensemble methods. As a result, they suggest that ensemble machine learning should be considered for the task of classifying gene expression data for cancerous samples.

Chaurasia and Pal5 compare the performance criterion of supervised learning classifiers, such as Naı¨ve Bayes, SVM-RBF kernel, RBF neural networks, Decision Tree (Dt) (J48), and simple classification and regression tree (CART), to find the best classifier in breast cancer datasets. The experimental result shows that SVM-RBF kernel is more accurate than other classifiers; it scores at the accuracy level of 96.84 percent in the Wisconsin Breast Cancer (original) datasets.

Chaurasia and Pal6 offered three popular data mining algorithms: CART, ID3 (iterative dichotomized 3), and DT for diagnosing heart diseases, and the results presented demonstrated that CART obtained higher accuracy within less time. Chaurasia and Pal7 conducted an experiment to identify the most common data mining algorithms, implemented in modern Medical Diagnosis, and evaluate their performance on several medical datasets.

Five algorithms were chosen: Naı¨ve Bayes, RBF Network, Simple Logistic, J48 and Decision Tree. For the evaluation two Irvine Machine Learning Chaurasia et al. 121 Repository (UCI- UC) databases were used: heart disease and breast cancer datasets. Several performance metrics were utilized: percent of correct classifications, True/False Positive rates, area under the curve (AUC), precision, recall, F-measure, and a set of errors.

Li et al.8 discovered many diversified and significant rules from high-dimensional profiling data and proposed aggregation of the discriminating power of these rules for reliable predictions. The discovered rules are found to contain low-ranked features; these features are found to be sometimes necessary for classifiers to achieve perfect accuracy. Kaewchinporn et al.9 presented a new classification algorithm tree bagging and weighted clustering (TBWC) combination of decision tree with bagging and clustering. This algorithm is experimented on two medical datasets: cardiotocography 1, cardiotocography 2 and other datasets not related to medical domain. Delen et al.10 had taken 202,932 breast cancer patients records, which then pre-classified into two groups of "survived" (93,273) and "not survived" (109,659). The results of predicting the survivability were in the range of 93

II. Machine Learning Algorithm For Breast Cancer Prediction

We enter a sizable amount of data, the machine learning model examines the data, and using the trained model, we can forecast the future [24], [26], [27]. A method of automatic learning is machine learning [25]. The primary machine learning algorithms for predicting breast cancer are as follows:

1) Artificial Neural Network (ANN): Artificial neural networks are a useful method for data mining [28]. A neural network is composed of three layers: input, hidden, and output. To extract the overly complex patterns, this technique is used [29]. The approach is based on distributed memory [31], network architecture [32]–[34], teamwork, and parallel processing [30].
2) Logistics Regression (LR): The approach uses more dependent variables and is supervised learning. The result of this procedure is a binary number. For a certain set of data, logistic regression [35] can provide a continuous outcome. This approach uses a statistical model with binary variables [32].

3) K-Nearest Neighbor (KNN) : This approach is used to recognise patterns. It is a useful method for forecasting breast cancer. The same amount of time was spent on each class in order to identify the trend. K Nearest Neighbour [36] pulls the relevant highlighted data from a huge dataset. We use feature similarity to categorise a huge dataset [32].
4) Decision Tree (DT) : Decision trees are built using classification and regression models [37]. There are fewer subsets of the data set. These smaller data sets may be used to make predictions with the highest degree of accuracy. Among the decision tree techniques are CART [38], C4.5 [39], C5.0 [40], and conditional trees [32, [41].
5) Naive Bayes Algorithm (NB) : This method makes use of a huge training dataset. The algorithm used to determine probability uses the Bayesian approach [42]. It provides the highest level of accuracy for estimating the input probabilities of noisy data [43]. This classifier compares training datasets and training tuples using analogies [32].
6) Support Vector Machine (SVM) : This supervised learning method addresses both classification and regression concerns [44]. In order to deal with the regression issue, it employs mathematical and theoretical functions. It delivers the highest accuracy rate when making predictions using a large dataset. It is a potent machine learning technique that is based on 3D and 2D modelling [32], [45].
7) Random Forest (RF) :Classification and regression problems are handled using the Random Forest algorithm [46], which is based on supervised learning. It is a machine learning building block that utilises historical datasets to forecast fresh data [32].
8) K Mean Algorithm :Data can be sorted into manageable categories using the clustering technique K mean. Algorithms are used to assess how similar different data points are to one another. Every data point contains the most suitable cluster for analysing a sizable dataset [48].
9) K. Gaussian Mixture Algorithm :The most popular learning technique is unsupervised learning. The term "soft clustering methodology" refers to a technique for calculating the likelihood of different types of clustered data. The implementation of this algorithm is based on expectation maximisation [51].

III. Ensemble Techniques for Breast Cancer Prediction

It is possible to utilise both homogeneous and heterogeneous ensemble techniques. Homogeneous ensemble strategies [52] combine one base method with two or more configuration methods, including bagging and boosting technique, whereas heterogeneous ensemble techniques [53]–[55] mix two or more base methods. The foundation of ensemble approaches is supervised learning, which makes precise predictions based on predetermined hypotheses.

1) Bagging
Evasion attacks focus on exploiting gaps in IDS algorithms, particularly in their ability to detect obfuscated or non-standard traffic. These attacks are meticulously crafted to bypass detection systems by using techniques such as packet fragmentation, payload encryption, and protocol manipulation.
2) Boosting
Boosting is a homogeneous week learner that combines a number of weak classifiers into a single strong classifier [52]. It is based on methods for building up the model step-by-step using certain training data [54], [55].
3) Stacking
Stacking is a heterogeneous [52] weak learner that integrates a variety of machine learning methods for prediction on the same dataset. It integrates the predictions of two or more fundamental models [54, 55].

IV. Survey on Breast Cancer

China has the largest population in the world. According to a recent organisational study (GLOBOCAN-2018) [65], women encounter breast cancer at a rate of 19.2 percent, compared to men who experience it at a rate of 8.6 percent. 1.2 million People die from this sickness each year. 48,100 cases of DCID cancer in female patients were recorded by the American Cancer Society. A US study from 2019 predicts that 41,760 women and 500 men will die from breast cancer [66]. A US study found that 3.8 million women are still alive but are fighting breast cancer. Incidences of Ductal Carcinoma in Situ (DCIS) breast cancer in US women reached 59,838 in 2019 [67]. Breast cancer has claimed the lives of 458,000 people globally. In 2012, breast cancer claimed the lives of 48 percent of Chinese women, compared to 52 percent worldwide [68]. In 2015, researchers looked at data from 1,517 women to find out the survival and recurrence rates for breast cancer; they found that the mortality rate was 132 and the recurrence rate was 100 [69].

V. Review Of Machine Learning Algorithms For Breast Cancer Prediction

This study's main objective is to assess various machine learning and data mining techniques that have helped with breast cancer prediction. Our primary objective is to identify the most accurate and suitable algorithm for breast cancer prediction. We have reviewed and analyzed earlier studies on breast cancer prediction systems in order to accomplish this. Additionally, articles based on K-nearest neighbor, support vector machines, naïve Bayes, linear regression, linear discriminant analysis, and various ensemble methods (Decision Tree, Random Forest, Boosting, and AdaBoost) were evaluated. The great majority of researchers used ensemble techniques in addition to linear and nonlinear or nonlinear alone. We've separated our review post into sections that contrast and compare each algorithm based on how accurate it is as a consequence. We will next identify the top machine learning technique for predicting breast cancer after that comparison.

VI. Conclusion

In this study, we investigated several data mining, machine learning, and deep learning approaches for breast cancer prediction. Our main objective is to find the best algorithm to more precisely predict the development of breast cancer. The main objective of this article is to summarise all previous work on machine learning techniques for breast cancer prediction. It also provides beginners with all the knowledge they need to comprehend machine learning algorithms and provide the foundation for deep learning. The examination of the many types of breast cancer is the first in this article's review. We looked over fourteen research publications to gain additional knowledge about the main types, signs, and causes of breast cancer. The most significant machine learning, ensemble, and deep learning methodologies were then reviewed. The algorithms that are used to predict breast cancer are considerably enhanced by these techniques. Future development will need to address a few issues that are still present. To deal with the issue of the limited datasets available, researchers might employ a variety of data augmentation techniques. Because it might lead to bias towards either a positive or negative prediction, researchers should consider the problem of the gap between positive and negative data. A crucial issue with an uneven quantity of breast cancer photos against affected patches needs to be solved for accurate breast cancer diagnosis and prognosis.

References

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

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

[3] Keles, M. Kaya, "Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: A Comparative Study."

[4] TehnickiVjesnik - Technical Gazette, vol. 26, no. 1, 2019, p. 149+. [4] V. Chaurasia and S. Pal, "Data Mining Techniques: ToPredict and Resolve Breast Cancer Survivability", IJCSMC, Vol. 3, Issue. 1, January 2014, pg.10 – 22.

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

[6] R. K. Kavitha1, D. D. Rangasamy, "Breast Cancer Survivability Using Adaptive Voting Ensemble Machine Learning Algorithm

[7] 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

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

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

[10] Y. Khourdifi and M. Bahaj, "Feature Selection with Fast Correlation-Based Filter for Breast Cancer Prediction and Classification

[11] Using Machine Learning Algorithms," 2018 International Symposium on Advanced Electrical and Communication Technologies (ISAECT), Rabat, Morocco, 2018, pp. 1-6.

[12] R. M. Mohana, R. DelshiHowsalya Devi, Anita Bai, "Lung Cancer Detection using Nearest Neighbour Classifier", InternationalJournalofRecent Technology and Engineering (IJRTE), Volume-8,Issue-2S11, September 2019

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

[14] Haifeng Wang and Sang Won Yoon, "Breast Cancer Prediction Using Data Mining Method", Proceedings of the 2015 Industrialand Systems Engineering Research Conference, [14] AbdelghaniBellaachia, ErhanGuven, "Predicting Breast CancerSurvivability Using Data Mining Techniques".