Biomedical & Pharmacology Journal, March 2022. Vol. **15**(1), p. 13-25

A Comprehensive Review Study on: Optimized Data Mining, Machine Learning and Deep Learning Techniques for Breast Cancer Prediction in Big Data Context

Madhu Kirola[1], Minakshi Memoria[1], Ankur Dumka[2], Amrendra Tripathi[3] and Kapil Joshi[1]*

[1]Uttaranchal University, Dehradun, India.
[2]Women's Institute of Technology, Dehradun, India.
[3]UPES, Dehradun, India.
*Corresponding Author E-mail: kapilengg0509@gmail.com

https://dx.doi.org/10.13005/bpj/2339

(Received: 10 July 2021; accepted: 10 January 2022)

In recent years, big data in health care is commonly used for the prediction of diseases. The most common cancer is breast cancer infections of metropolitan Indian women as well as in women worldwide with a broadly factor occurrence among nations and regions. According to WHO, among 14% of all cancer tumours in women breast cancer is well-known cancer in women in India also. Few researches have been done on breast cancer prediction on Big data. Big data is now triggering a revolution in healthcare, resulting in better and more optimized outcomes. Rapid technological advancements have increased data generation; EHR (Electronic Health Record) systems produce a massive amount of patient-level data. In the healthcare industry, applications of big data will help to improve outcomes. However, the traditional prediction models have less efficiency in terms of accuracy and error rate. This review article is about the comparative assessment of complex data mining, machine learning, deep learning models used for identifying breast cancer because accuracy rate of any particular algorithm depends on various factors such as implementation framework, datasets(small or large),types of dataset used(attribute based or image based)etc. Aim of this review article is to help to choose the appropriate breast cancer prediction techniques specifically in the Big data environment to produce effective and efficient result, Because "Early detection is the key to prevention-in case of any cancer".

Keywords: Bigdata; Deep Learning Algorithm; Data Mining Algorithm; DCIS; LCIS; Invasive; Non-Invasive; Machine Learning Algorithms.

Nowadays healthcare data includes free form text such as doctors notes, reports from radiologists, still images such as CAT scans, photos, videos, recorded patient historical data, genomic files, biometric and other scientific data from clinical research and drug production.It also collects data from wearables, medical equipment, respirators, blood pressure monitors, and other linked devices using the Internet of Things (IoT). All this data, in addition to data which exists in independent,standalone systems — EMR, PACS, RTHS, EMPI,LIS, and PMS — is also part of the new data on health care.Big data technology is needed to capture and handle these large quantities of data involved and to provide reliable responses from several reputable sources that represent the

latest medical research. Big Data and advanced analytics provide solutions to some of the key issues facing the healthcare industry today. Digital healthcare demands that medical practitioners have access to the study of all data in its original formats instantly, explicitly and in a natural language. Cancer is the world's second leading cause of death. Breast cancer is the top cancers that affect the Indian population too. Breast cancer diseases is the most common type of cancer detected in women in India. Breast cancer accounts for 2.09 million cases and 627000 deaths globally. It can occur at any age but the incidence rates in India begin to rise in the early thirties and peak at ages 50-64 years. Globocan 2018 (Globocan belongs to IARC - International Agency for Research on Cancer) and NCRP (National Cancer Registry programme, India) represented(data for the year 2018, published on 12 sept 2018[13,14,15].17december 2020,

'Incidence' indicates number of newly diagnosed women with breast cancer that year. Newly diagnosed in India in the year 2018 were 162,468 women with breast cancer[Figure 1]. Breast cancer also accounted for 27.7% of all newly diagnosed women's cancers. That means, in India, that one in four newly diagnosed cancer in women was breast cancer[36].

'Mortality' reflects the number of women who died that year from breast cancer. 87,090 women died in India for the year 2018 from breast cancer [Figure 2]. Breast cancer accounted for approximately 23.5 per cent of all deaths linked to cancer in women in India. Which means that in India's women nearly one in four deaths due to breast cancer[36].

Breast Cancer

Breast cancer is an inflammatory tumour developed in the mammary gland. Cancer begins when the cells start to develop out of control. Breast cancer is a group of diseases where breast tissue cells shift and grow uncontrolled, usually leading to a lump or mass. The majority of breast cancers originate in the milk glands[3]. Breast cancer is diagnosed by mammograms, breast self-examination (BSE), biopsy, and advanced breast tissue testing. Breast cancer care may include surgery, radiation, hormone therapy, chemotherapy and laser therapy.

Breast cancer can spread when cancer cells invade the bloodstream or lymphatic system and travel to other parts of the body. Breast cancer cells usually form a lump or a tumour, which can be seen on an X-ray or felt as a hard mass. Breast cancer risk can be reduced by keeping track of controllable risk factors.Breast cancer is almost

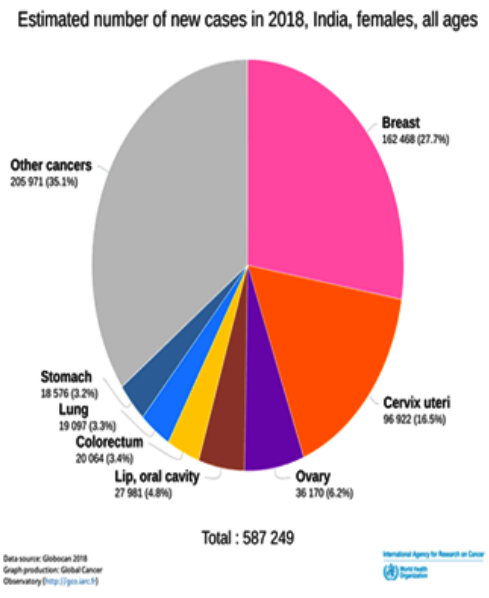

Fig. 1. Breast Cancer Incidence

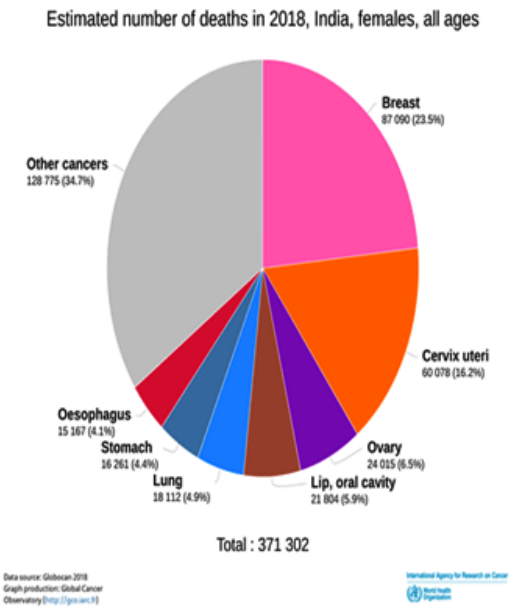

Fig. 2. Breast Cancer Mortality

exclusively a female disease, but it is also very common in men now days.

Types of Breast Cancer

Breast Cancer

The categories of breast cancer can also determine by identifying that whether the cancer has spread or not referred as -Invasive Breast Cancer and Non-Invasive breast Cancer [Figure 3]

Non-Invasive breast cancer originate in the milk glands but does not spread to the rest of the breast tissues. There are two categories of Non-Invasive Breast Cancer are-Ductal carcinoma in situ (DCIS) is a malignant cell contained within breast ducts, detects by mammograms.Lobular Carcinoma In Situ (LCIS) is a malignant cell contained within breast lobule, detected on biopsy done for others indications [Figure 3].

On the other side Invasive Breast cancer refers to any form of breast cancer that has spread into the surrounding breast tissues. Such as- Invasive Ductal Carcinoma(IDC) infiltrates surrounding breast tissues palpable if large or detected on mammograms. The majority 81% of breast cancers cases are invasive type. Invasive Lobular Carcinoma (ILC) infiltrates surrounding breast tissues bilateral diagnosed microscopically[13,14,15]. Inflammatory Carcinoma,IBC (inflammatory breast cancer) is an uncommon and severe kind of breast cancer that manifests as a rash or irritated skin region. It obstructs the lymph veins in the skin of the breasts. A mammogram or ultrasound cannot detect inflammatory breast cancer therefore microscopically it can detect. Paget's disease is a type of breast cancer that is extremely rare. Paget's disease of the breast begins on the nipple and progresses to the black circle of skin surrounding the nipple (areola). Male breast cancer is typically

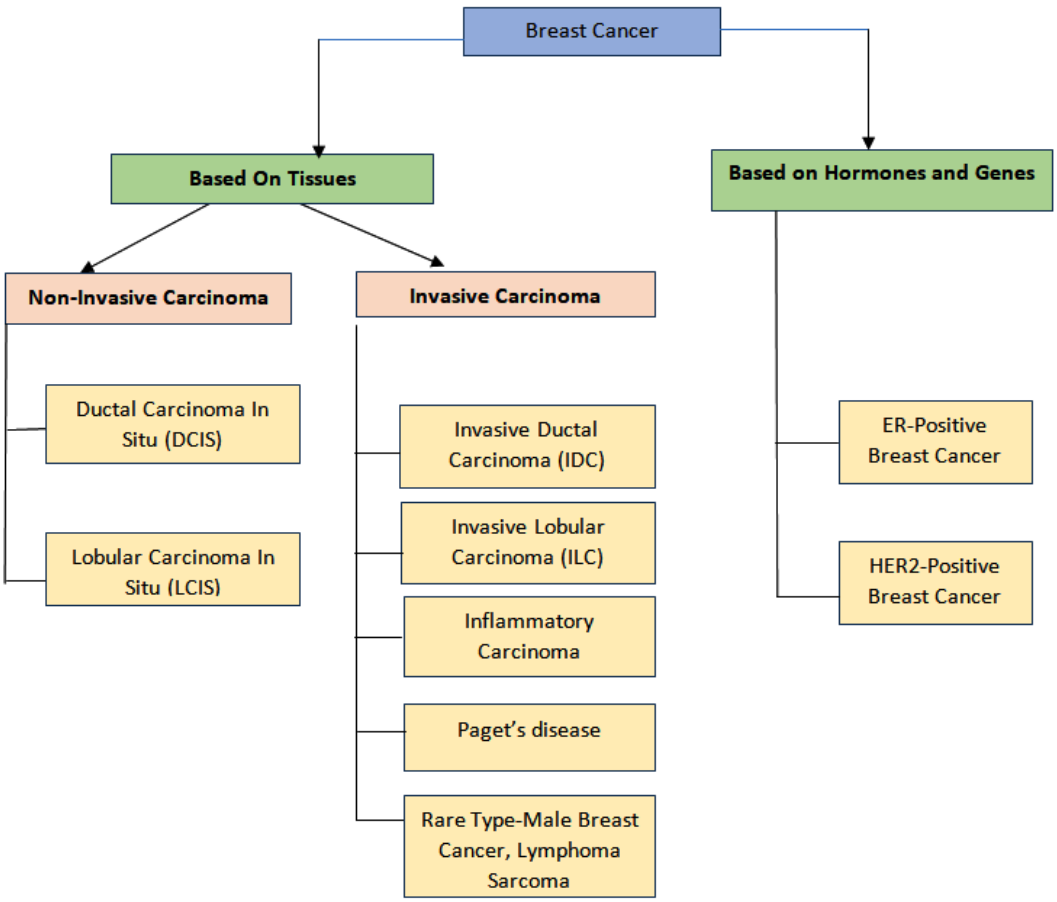

Fig. 3. Classification of Breast cancer

diagnosed as invasive ductal carcinoma at an advanced age [Figure 3].

Breast Cancer Signs and Symptoms

A very common symptom of breast cancer is new lymph node or hard mass around the breast or underarm area. Breast cancer is more likely to be a painless, hard mass with irregular edges, but it may also be soft, delicate.Since breast cancer usually has no symptoms because the tumour is small and easy to treat, screening is important for early detection. Breast cancer symptoms differ from one person to the next. Some people don't show any indications or symptoms at all. Breast cancer can manifest itself in a variety of ways, including:

• Change in Breast Texture

• Detecting a lump, hard knot or area of thickened tissues in the breast or underarm area.

• Skin on the breast, nipple or areola becomes red, scaly or feel warm.

Change in Breast shape or size

• Swelling or shrinking of the breast.

• Breast pain without swelling or Shrinking.

• Recent asymmetry in breast size.

Other Changes

• Tenderness in the Breast area.

• One or both nipple have slightly turned inward or inverted.

• Discharge of clear fluid or bloody fluid from Breast.

Breast Cancer Stages and Survival Rate

According to the stage system of the SEER Committee(Surveillance,E pidemiology,a ndE ndResults) overview[16,27]

• Cells that are abnormal in the duct lining or a part of the breast. Breast cancer is more likely to occur in one or both breasts. At this point, the survival rate is 100%.

• Breast cancer is a form of cancer that affects the tissues of the breast. Tumour is less than an inch in diameter. This stage has a 95 % to 98 % survival rate.

• It is also related with tissues of the breast. Tumour measures less than two inches in diameter. Cancer has the potential to spread to the auxiliary lymph nodes. At this point, the survival rate is 88%.

• Affect tissues of the breast. Tumour has a diameter of more than two inches. Cancer has the potential to spread to the auxiliary lymph nodes. Inflammation, dimpling, or a shift of skin colour are all possibilities. At this point, the survival rate is between 50% and 60%.

• Beyond the breast, cancer has spread to other parts of the body. At this point, the survival rate ranges from 15% to 20%.

Breast Cancer Causes

This is caused because of the progressions or change in DNA of the cells. A portion of the peril factors are kind-hearted condition like hyper plasia increment danger of bosom malignant growth. Having a prior history of malignant growth expands the opportunity of causing disease.

Classification of different Data Mining, Machine Learning and Deep Learning Techniques for Breast Cancer Prediction

Training datasets are being used in data mining techniques. Data mining is a process for detecting common patterns in a data set that are accurate, unique, and useful data.[Figure 4]

An algorithm that learns from data and improves over time is referred to as machine learning.Machine learning is the analysis of an algorithm that can generate data automatically. Machine learning makes use of data mining techniques and another learning algorithm to create models of what is going on behind the scenes of such data in order to predict potential outcomes. Machine learning algorithms [Figure 4] generate models based on the knowledge that describes the relationship between items in data sets in order to predict future outcomes.

When a computer model learn from images, texts, audios based datasets and perform classification tasks directly is known as Deep learning. Deep learning models [Figure 4] can achieve cutting-edge precision, sometimes even outperforming humans. A large quantity of labelled data and a multilayer neural network architecture are used to train models.

Ensemble algorithms are supervised learning techniques work on the base of hypothesis[Figure 4].There are two types of ensemble approaches: homogeneous and heterogeneous. Homogeneous ensemble techniques combine one base method with two or more configuration methods, whereas heterogeneous ensemble techniques mix two or more base methods.

Review of the Literature

In the field of medical data analysis, many

studies on breast cancer have been published, and the majority of them claim to have high classification accuracy.

K. Venkateswara Rao et.al[1] proposed an examination report, utilizing various strategies for features selection used for features extraction with various techniques for features grouping to characterize breast cancer malignant growth. Information on breast cancer disease is taken from the UCI store and analysed by utilizing the WEKA strategy, and proposed methods are applied to precisely classify details. This examination plainly characterizes that the strategy for information digging is viable for anticipating breast cancer disease. The WEKA device is viewed as truly outstanding and most dependable data classification techniques in data mining. Contrasted with different algorithm on the data set for breast cancer malignant growth, SVM gives reliable outcomes. 286 cases and 10 qualities of breast cancer disease have been investigated with 82.53%accuracy rate.

Madhu Kumari et.al[2] proposed a prediction framework plan that could foresee the event of breast cancer malignant growth at a beginning phase by assessing the smallest set of features chose from the clinical dataset. To play out the proposed explore, the Wisconsin breast cancer dataset (WBCD) was utilized. Utilizing characterization exactness that wasacquired by contrasting genuine with anticipated qualities, the capability of the proposed technique is obtained. The result shows that this investigation accomplishes the optimum accuracy of classification 99.28%.

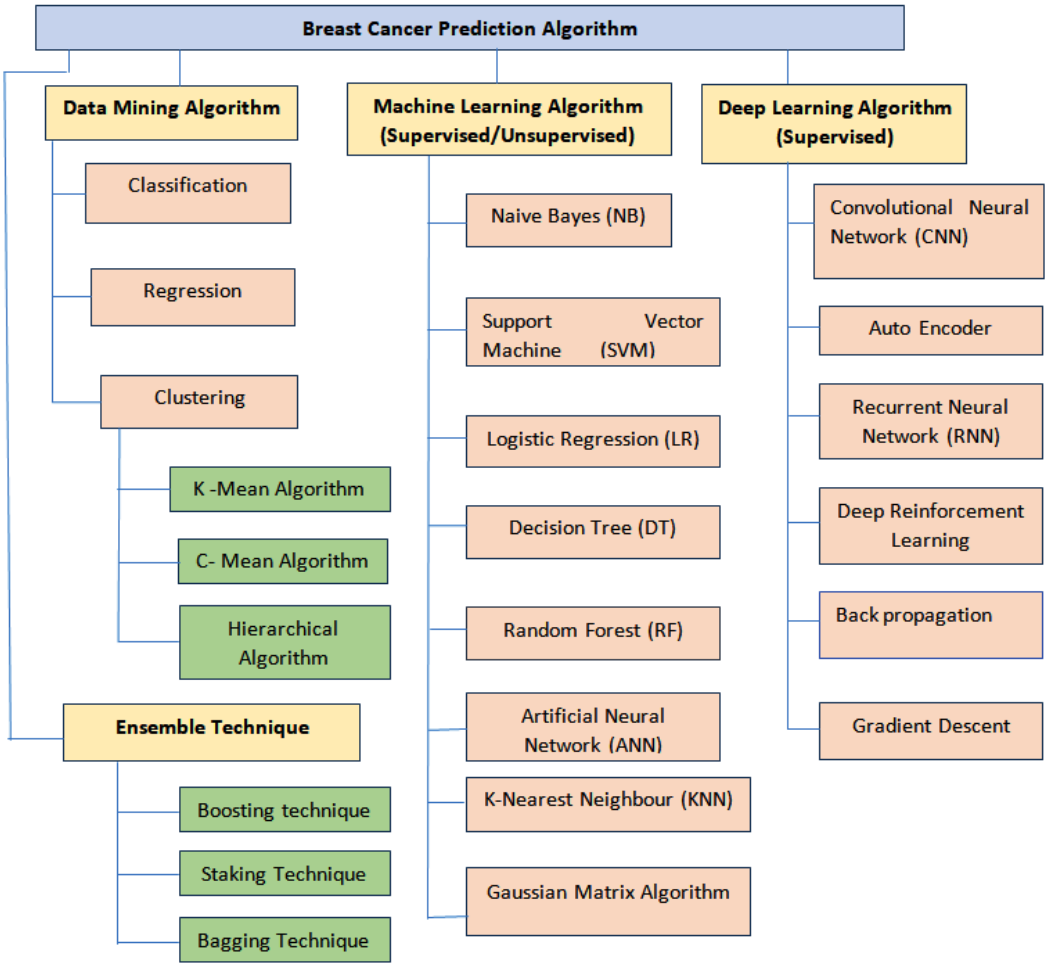

Fig. 4. Different Supervised and Unsupervised Breast Cancer Prediction Techniques

Table 1. Comparative Review of Data mining, Machine Learning and Deep learning Techniques for breast Cancer Prediction

S.No	Author Name	Method/Algorithm Used	Tools	Data Set	Data Type	No of Attributes	Performance
1	K. Venkateswara Rao, L. Mary Gladence , V.Raja Lakshmi[7]	Classification Techniques- SVM, Naive Bayes, Regression , J48, Bagging	Weka	UCI Machine Learning Repository	Numeric Attributes	Dataset has 286 instances and 10 attributes.	SVM with higher accuracy 82.53%
2	Madhu Kumari, Vijendra Singh[2]	Classification –supervised machine learning algorithm-SVM, Linear Regression and KNN	MatLab	Wisconsin breast cancer dataset (WBCD)	Numeric Attributes	Dataset contains 699 instances 9 attributes	KNN classifier with higher accuracy 99.28%
3	Hiba Asria, Hajar Mousannifb, Hassan Al Moatassime, Thomas Noeld[3]	Scale Vector Machine(SVM), NB,C4.5,K-NN	Weka	Wisconsin Breast Cancer (original) datasets	Numeric Attributes	699 instances and 11 attributes	SVM with highest accuracy 97.13%
4	Sara AlGhunaim1 and Heyam Al-Baity[4]	Support Vector machine (SVM), Decision Tree, and Random Forest,	Spark and Weka	University of California, IRVINE Repository, Wisconsin Breast Cancer (original) datasets	Numeric Attributes	683 instances, 254 Samples	SVM classifier in the Spark environment reaches an accuracy of 99.68%
5	M. Supriya1 & A. J. Deepa [5]	Optimized Artificial Neural Network (OANN), Gray Wolf Optimization (GWO) algorithm.	Implemented Using Java	Wisconsin Breast Cancer (original) datasets	Numeric Attriattributes	571 instances and 10 attributes	OANN classifier (with and without feature selection) 96% of accuracy
6		ANN with whale optimization algorithm (WOA), dragonfly algorithm (DA), multiverse optimization (MVA), and grey wolf optimization (GWO)	MatLab Version 2017a	UCI Machine Learning Repository	Numeric Attributes	699 instances and 9 attributes.	ANN with dragonfly algorithm (DA) and grey wolf optimization Algorithm(GWOA) achieve 98% accuracy
7	Md. Milon Islam, Md. Rezwanul Haque, Hasib Iqbal, Md. Munirul Hasan, Mahmudul Hasan, Muhammad Nomani Kabir[1]	Artifcial neural networks (ANNs) SVM, K-nearest neighbors(KNN) Random forests, and Logistic Regression(LR)	Jupyter with python programming language	UCI Machine Learning Repository	Numeric Attributes	629 instances and 9 attributes.	ANN -98.57%
8	Habib Dhahri, Eslam Al Maghayreh, Awais Mahmood, Wail Elkilani,and Mohammed Faisal Nagi [8]	support vector machine (SVM) , Knearest neighbor (KNN) , decision tree (DT), gradient boosting classifier (GB) , random forest (RF) , logistic regression (LR) , AdaBoost classifier (AB), Gaussian Naive Bayes (GNB), and linear discriminant analysis (LDA)	Weka with python	UCI Machine Learning Repository-Wisconsin Breast Cancer dataset	Numeric Attributes	569 instances and 10 attributes	AdaBoost classifier (AB)with Genetic programming-98.24%
9	Walid Cherif[9]	Optimization of K-NN algorithm by clustering	MatLab	UCI dataset	Numeric Attributes	565 Instances and 30 attributes	Optimized of K-NN algorithm with 94% accuracy
10	Hui Huang, Xi'an Feng, Suying Zhou, Jionghui Jiang, Huiling Chen, Yuping Li and Chengye Li[10]	Fruit fly optimization algorithm (FOA) with support vector machine (SVM)	MatLab	Wenzhou People's Hospital China	Numeric Attributes	470 Instances and 14 attributes	LFOA-SVM with higher accuracy rate 93.83%

11	Sapiah Binti Sakri, Nuraini Binti, Abdul Rashid and Zuhaira Muhammad Zain[11]	With Particle swarm optimization as a feature selection, classifiers-naive Bayes, K-nearest neighbour, and fast decision tree techniques.	Weka	Wisconsin Dataset-Breast Cancer Prognostic	Numeric Attributes	198 instances and 34 attributes	Naive Bayes produced better output with and without PSO(particle swarm optimization)
12	Seyed Reza Kamel, Reyhaneh Yaghoub Zadeh and Maryam kheirabadi [12]	Support vector machine (SVM) with Gray Wolf Optimization (GWO)	MatLab	UCI dataset	Numeric Attributes	699 records, 9 features	SVM algorithm and the GWO with 100% accuracy
13	P. Israni [17]	Sopport Vector Machine() SVM), K-nearest neighbour (KNN),Logistic Regression (LR)	MatLab	UCI MACHINE repository	Numeric Attributes	32 attributes	Support Vector Machine-92.78% K-Nearest Neighbour-92.23% Logistic Regression - 92.10%
14	A.A. Bataineh [18]	MLP,KNN,CART, Naive Bayes, SVM	MatLab	UCI Machine repository	Numeric Attributes	32 attributes	MLP -99.12% KNN -95.61% CART-93.85% Naive Bayes -94.73% SVM -98.24%
15	Y. Khourdifi and M. Bahaj[20]	KNN, SVM,RF, Naïve Bayes	Weka	Wisconsin Breast Cancer (original) datasets	Numeric Attributes	11 attributes	KNN-96.1% SVM-97.9% RF-96%, NB- 92.6%
16	A. Reddy, B. Soni, and S. Reddy[24]	DNNS(Deep Neural Network with Support Value)	MatLab	M. G Cancer Hospital & Research Institute, Visakhapatnam, India	Image based data set	Complete Image	DNNS- 97.21%
17	J. Zheng, D. Lin, Z. Gao, S. Wang, M. He, and J. Fan[48]	Deep Learning assisted Efficient Adaboost Algorithm (DLA-EABA)	Weka	https://wiki.cancerimagingarchive.net/	Image based data set	Complete Image	DLA-EABA -97.2%
18	M. H. Memon, J. P. Li, A. U. Haq, M. H. Memon, and W. Zhou[39]	SVM with linear kernel, SVM with RBF-Kernel, SVM with polynomial kernel ,SVM with sigmoid Kernel	Weka	Wisconsin Breast Cancer (original) datasets	Numeric attributes	32 attributes	SVM with linear Kernel-99%, SVM with RBF-Kernel-98%, SVM with polynomial kernel-87%, SVM with sigmoid Kernel-94%
19	E. A. Bayrak, P. Kirci, and T. Ensari[42]	Support Vector Machine, Artificial Neural Network,	Weka	Wisconsin Dataset-Breast Cancer Prognostic	Numeric attributes	11 attributes	Support Vector Machine-96.9% Artificial Neural Network-95.4%
20	S. Bharati, M. A. Rahman, and P. Podder [46]	Naïve Bayes, Random Forest, Logistic Regression, K-Nearest Neighbours (KNN)	Weka	UCI Machine repository	Numeric attributes	10 attributes	Naïve Bayes-71.68% Random Forest-69.5%, Logistic Regression-68.85, Multilayered Perception, Multilayered Perception-64.6%; K-Nearest Neighbours(KNN)-72.37%

Hiba Asria et.al[3] set up a report, on execution examination which is completed on the "Wisconsin Breast Cancer (unique datasets between different machine learning classification algorithm: Support Vector Machine (SVM), Decision Tree (C4.5), Naïve Bayes (NB) and K-Nearest Neighbour (k-NN)"[3]. The primary target is to decide the exactness of the classification algorithm with respect to the adequacy and effectiveness of every classification algorithm as far as exactness, accuracy, affectability and explicitness. Experimental work done on WEKA tool data mining method. Taking everything into account, in Breast Cancer prediction and diagnosis, SVM has demonstrated its adequacy with 97.13% and accomplishes the most elevated outcomes regarding precision and low error rate.

Sara Al-Ghunaim et.al[4] "consider the issue of breast cancer forecast in the , thought about two assorted data context consider majorly two varieties of data- Gene Expression(GE) and DNA methylation (DM). The goal of this work is proportional up the machine learning algorithms which are utilized for characterization by applying each dataset independently and together. For this reason, they picked Apache Spark as a framework and three distinctive classification algorithms, Support Vector Machine (SVM), Decision Tree (DT), and Random Forest(RF), to make nine models that help in anticipating breast cancer disease. A comparative study conducted by utilizing three situations with GE, DM, and GE and DM consolidated, to show which of the three kinds of information would create the best outcome as far as precision and error rate and just as a test correlation performed between two frameworks (Spark and Weka), to show their conduct when managing huge amount of data i.e-Big Data. Where The research results showed that the scaled SVM classifier in the Spark framework beats different classifiers, as it accomplished the most highest elevated precision and the least error rate with the GE dataset. SVM arrives at an exactness of 99.68% and hence beats different classifiers on both Spark and Weka environment".

M.Supriya et.al[5] "proposed a breast cancer prediction framework using Optimized Artificial Neural Network (OANN). Fundamentally, the unprocessed breast cancer data are viewed as the input. The large amount of data big data (BD) stockpiling contains some rehashed data. Secondarily, such rehashed information are disposed of by using Hadoop Map-Reduce. In the ensuing stage, the data are pre-processed

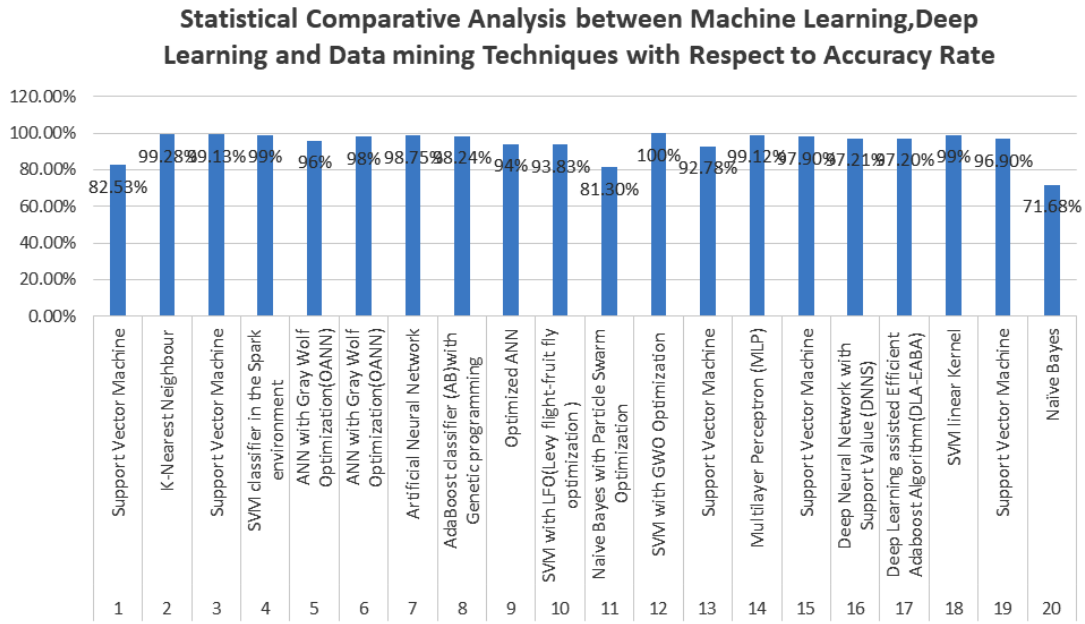

Fig. 5. Statistical Comparative Analysis of Data mining, Machine Learning andDeep learning Techniques(w.r.t Accuracy level) for breast Cancer Prediction in last Five years (2016-2020)

using replacing of missing attributes (RMA) and normalization procedures. Therefore, the features are picked by using Modified Dragonfly algorithm (MDF). At that point, the chosefeatures are inputted for classification. Here, it grouped the features using OANN. Optimization is done by utilizing the Gray Wolf Optimization (GWO) algorithm. Experiential results are appeared differently in relation to winning IWDT (Improved Weighted-Decision Tree) in regard of exactness, recall, precision, and ROC. The proposed OANN classifier (with and without features selection) accomplishes over 96% of accuracy for each data. The ROC performance of the proposed OANN accomplishes more prominent outcomes when contrasted with the existing one".

A.M.Hemeida, November 2019[6]" addresses execution of transformative optimization algorithm for mining two popular data indexes in machine learning by carrying out four diverse streamlining methods of optimization. The chose data indexes utilized for assessing the proposed optimization algorithm are Iris dataset and Breast Cancer dataset. In the order issue of this paper, the neural organization (NN) is utilized with four streamlining optimization procedures, which are Whale Optimization Algorithm(WOA), Dragonfly Algorithm (DA), Multiverse Optimization (MVA), and Gray Wolf Optimization (GWO). Diverse control boundaries were considered for precise decisions of the proposed optimization procedures. The comparative investigation demonstrates that, the GWO, and MVO give precise outcomes over both WO, and DA regarding convergence, runtime, classification rate, and MSE. Hybrid algorithms consisting of two diverse optimization strategies can be considered for future investigation for data mining tasks".

Md. Milon Islam et.al[7]" works on Support Vector Machine (SVM), K-Nearest Neighbours, Random Forest, Artificial Neural Networks (ANNs), and Logistic Regression are five supervised machine learning approaches that have been distinguished (LR). The UCI repository provides links to the Wisconsin Breast Cancer dataset. The yield of the test is surveyed in terms of precision, sensitivity, specificities, accuracy, negative predictive value, false negative rate, false positive rate, F1 score, and Matthews Correlation Coefficient.The findings shows that ANNs have the highest Precision, Accuracy, and F1 score of 98.57%, 97.82 %, and 0.9890, respectively, while SVM has the second highest Precision, Accuracy, and F1 score of 97.14 %, 95.65%, and 0.9777 respectively".

Habib Dhahri, Eslam et.al[8] " focused on Genetic programming and machine learning algorithms, with the aim of developing a framework that can reliably distinguish between benign and malignant breast tumours. The aim of this research was to develop the learning algorithm. The best features and perfect parameter values of machine learning classifiers are selected using a genetic programming technique. Sensitivity, specificity, precision, accuracy, and roc curves were used to test the proposed method's efficiency. The study shows that by combining feature pre-processing methods and classifier algorithms, genetic programming can automatically find the best model".

Walid CHERIF[9] gives a "new solution to speed up KNN algorithm based on clustering and attributes filtering to optimize K-Nearest Neighbours algorithm (KNN) performance and to accelerate its process . Therefore the paper's filtering to optimizecontributions are threefold: firstly ,the clustering of class cases, secondly, the identification of the most significant attributes, and third is the assessment of similarities by coefficients of reliability. Classification results indicate that the proposed algorithm outperforms KNN, NB, SVM on the considered dataset with an f-measure slightly exceeding 94%".

Hui Huang et.al[10] " established anenhanced machine learning framework to diagnose the breast cancer. The centre of this framework is the adoption of the Levy Flight (LF) Strategy (LFOA) enhanced fruit fly optimization algorithm (FOA) to optimise two main support vector machine (SVM) parameters and to construct LFOA-based SVM (LFOA-SVM) to diagnose breast cancer. In terms of different performance metrics, the experimental results show that the suggested LFOA-SVM approach can beat other counterparts. The proposed method has achieved a classification accuracy of 93.83%, sensitivity of 91.22%, specificity of 96.53% and MCC of 0.8799 for breast cancer diagnosis based on the high-level features".

Sapiah Binti Sakri et.al[11] suggested "improving the efficiency of most classification algorithms by using techniques for feature selection

to minimise the number of features. Compared to other features, certain characteristics are more significant and affect the results of the classification algorithms, The objective of this research is to compare the accuracy of a few existing data mining algorithms in predicting breast cancer recurrence. It incorporates a particle swarm optimization as a feature selection in three well-known classifiers: Naive Bayes, K-nearest neighbour, and fast decision tree learner, with the goal of boosting the prediction model's accuracy..With and without PSO (Particle swarm optimization), naive Bayes produced better performance, whereas when used with PSO, the other two methods improved".

Compared to other models for classification, Seyed Reza Kamel et.al[12] uses "data mining as a blend of Gray Wolf Optimization (GWO) feature selection process and support vector machine (SVM) to improve the accuracy of breast cancer diagnosis compared to previous methods, a new technique introduced with high precision. The approach proposed had a better ability to detect breast cancer comparison to prior approaches. Experimental results are gathered Using the MATLAB and UCI datasets. The best results are obtained from a fusion of the SVM algorithm and the GWO to select the subset of suitable features. The accuracy, sensitivity and specificity were 100%, 100% and 100% compared to the other algorithms".

P. Israni[17] "provides research workUsing Support Vector Machine (SVM) with 10-fold cross validation-an efficient BCD model for detecting breast cancer. When there are multiple input features for cancer detection, the problem becomes more complicated. To reduce the feature space from a higher to a lower dimension, Principal Component Analysis (PCA) is utilized. The PCA improves the model's accuracy, according to the results of the experiment. Other supervised learning algorithms such as Decision trees (DT), Random Forest, k-Nearest Neighbours (k-NN), Stochastic Gradient Descent (SGD), AdaBoost, Neural Network (NN), and Nave Bayes are compared to the suggested BCD model. F1 measure, ROC curve, Accuracy, Lift curve, and Calibration Plot are among the evaluation metrics that show that the proposed BCD model outperforms and provides the highest accuracy among the other examined methods.The accuracy of the proposed BCD model is 98.1 % and AUC is 0.995, which are the highest among the other implemented models".

A. A. Bataineh[18] "studies compares the performance of five nonlinear machine learning algorithms: Multilayer Perceptron (MLP), K-Nearest Neighbours (KNN), Classification and Regression Trees (CART), Gaussian Nave Bayes (NB), and Support Vector Machines (SVM). The major goal is to assess each algorithm's efficiency and efficacy in terms of classification test accuracy, precision, and recall when it comes to classifying data. MLP has a training data accuracy of 96.70 %, which is higher than the other four algorithms. Following the estimation, the predictive models' performance is tested using the k-fold cross-validation procedure on unknown data in terms of accuracy, precision, and recall. The MLP model had the highest accuracy, precision, and recall of 99.12%, 99.00 %, and 99.00 %, respectively, according to the findings of this research.Wisconsin Breast Cancer Diagnostic (WBCD) dataset were used for the study."

A.Reddy, S. Reddy, and B. Soni[24] "The authors present the innovative DNNS Breast Cancer Detection Method. Unlike other methods, the proposed solution is based on a deep neural network's Support value. A normalizing procedure has been used to improve the performance, efficiency, and quality of photographs. Experimental results show that the proposed DNNS outperforms the existing one methods."

Zheng, Jing & Lin, Denan & Gao, Zhongjun & Wang, Shuang & He, Mingjie & Fan, Jipeng[49]" proposed With modern computing approaches, a mathematically proposed Deep Learning assisted Efficient Adaboost Algorithm (DLA-EABA) for breast cancer diagnosis has been developed. Tumor classification methods employing transfers, in addition to typical computer vision methodologies, are being actively researched through the use of deep convolutional neural networks(CNNs). This work focuses on finding the best way by integrating various machine learning methodologies with methods for choosing and extracting features, as well as evaluating their output using classification and segmentation algorithms. When compared to other current systems, the experimental findings demonstrate that the high accuracy level of 97.2%, sensitivity of 98.3%, and specificity of 96.5%."

Comparative Analysis between techniques used for Breast Cancer Prediction (Year 2016 to 2020)

The above survey provides the detailed description of classification of breast cancer using various machine learning, data mining, as well as deep learning techniques on the basis of Algorithm/ method used for prediction, tools, data set, data type,number of attributes considered for the study as depicted in Table 1.

Result Analysis

According to the study, traditional data mining and machine learning approaches have limited use, whereas hybridization of machine learning techniques with optimization techniques as well as hybridization of deep learning methods with optimization methods have a lot of potential for clinical analysis and boosting the diagnostic capacity of existing computer-based application systems like SVM WITH Gary Wolf optimization[12] with 100% accuracy, and ANN with Dragon Fly Optimization Algorithm[6] with 98% accuracy rate as shown in Figure[5].This statistical [Figure 5] and comparative analysis [Table 1] also shows that very few studies of breast cancer prediction are based on mammograms/images.The availability of datasets is a major barrier in using machine learning and deep learning techniques to predict breast cancer because for computational measurements, each method requires a considerable amount of training data.In this paper, we provide an overview of data mining, machine learning, and deep learning methodologies, with an emphasis on the accuracy rate of breast cancer prediction. We looked for publications in data mining, machine learning, and deep learning techniques in the field of medical data analysis and searched BMC Bioinformatics, Biomed, Google Scholar, IEEE, Science-Direct, Springer, and Web of Science databases, as well as Research Gate, where multi-view mammography based data set /numeric attributes based data set used for research study.

CONCLUSION

When the objective is to obtain more efficient trends and knowledge that allow improved analysis, decision making, and process automation, analysing large sets of data is difficult. Unfortunately, conventional approaches to using machine-learning algorithms were unable to meet the modern challenges of big data, especially scalability.For breast cancer prediction, numerous data mining ,machine learning and deep learning algorithms are evaluated. The primary goal of this review article is to identify existing machine learning and deep learning based research for breast cancer prediction and to determine the most appropriate approach for predicting the incidence rate.It has been observed that there is still a lot of work has to be done in the future. Because Big data is currently causing a revolution in healthcare. Since today's digital healthcare needs intelligent integration and aggregation of accessible patient information and computer data, structured, semi-structured, and unstructured, in their original formats, there is a need to manage this vast amount of data. Second, due to the small dataset availability, very few research studies are focused on breast cancer images. Therefore a model can be proposed for the prediction of breast cancer from the histopathological images based data sets on Big data. Initially, Hadoop architecture can be generated topreserve the data samples in order to envisage the work on Big data after that optimized Convolutional neural network algorithm can be implemented for prediction.

REFERENCES

1. Md. Milon Islam, Md. Rezwanul Haque, Hasib Iqbal, Md. Munirul Hasan, Mahmudul Hasan,Muhammad Nomani Kabir" Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques". SN COMPUT. SCI. **1,** © Springer Nature Singapore Pte Ltd ,2020,Art.no-290.
2. Madhu Kumari , Vijendra Singh" Breast Cancer Prediction system" International Conference on Computational Intelligence and Data Science (ICCIDS 2018) Procedia Computer Science 132, Elsevier, 2018: 371–376.
3. Hiba Asria, Hajar Mousannifb, Hassan Al Moatassime C, Thomas Noeld" Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis" The 6th International Symposium on Frontiers in Ambient and Mobile Systems (FAMS 2016) Procedia Computer Science 83 , Elsevier, 2016,pp.1064 – 1069.
4. S. Alghunaim and H. H. Al-Baity, "On the Scalability of Machine-Learning Algorithms for Breast Cancer Prediction in Big Data Context," in *IEEE Access*, 2019, pp.91535-91546.
5. M. Supriya1 & A. J. Deepa2"A novel approach

for breast cancer prediction using optimized ANN classifier based on big data environment" Health Care Management Science,Springer, Science+Business Media, LLC, part of Springer Nature ,2019,pp.414-426.

6. A.M. Hemeida a, Salem Alkhalaf b , A. Mady c , E.A. Mahmoud c , M.E. Hussein c , Ayman M. Baha Eldin d "Implementation of nature-inspired optimization algorithms in some data mining tasks" Published by Elsevier B.V. on behalf of Faculty of Engineering, Ain Shams University. *Ain Shams Engineering Journal,*: 309-318 (2020).
7. K. Venkateswara Rao, L. Mary Gladence, V. Raja Lakshmi" Research of Feature Selection Methods to Predict Breast Cancer" *International Journal of Recent Technology and Engineering,* 2356-2367 (2019).
8. Habib Dhahri, Eslam Al Maghayreh, Awais Mahmood, Wail Elkilani, Mohammed Faisal Nagi, "Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms", *Journal of Healthcare Engineering,* Article Id-4253641, 11 pages (2019).
9. Walid Cherif,"Optimization of K-NN algorithm by clustering and reliability coefficients: application to breast-cancer diagnosis", *Elsevier Procedia Computer Science,* 293-299 (2018).
10. Hui Huang, Xi'an Feng, Suying Zhou, Jionghui Jiang, Huiling Chen, Yuping Li and Chengye Li"A new fruit fly optimization algorithm enhanced support vector machine for diagnosis of breast cancer based on high-level features" *BMC Bioinformatics*, **20**; Art,No-290 (2019).
11. Sapiah Binti Sakri, Nuraini Binti Abdul Rashid, and Zuhaira Muhammad Zain" Particle Swarm Optimization Feature Selection for Breast Cancer Recurrence Prediction" Special section on Big Data Learning and Discovery, IEEE Access, 29637-29647 (2018).
12. Kamel, S.R., YaghoubZadeh, R. & Kheirabadi, M. "Improving the performance of support-vector machine by selecting the best features by Gray Wolf algorithm to increase the accuracy of diagnosis of breast cancer". *J Big Data,* **6**: Art. No-90 (2019).
13. American Cancer Society. 2018. Global Cancer: Facts & Figures, 4th edition, pp-12-15.
14. India against cancer 2019, "Breast Cancer", National Institute of Cancer Prevention and Research, viewed 12 November 2019.
15. American Cancer Society. Breast Cancer Facts & Figures 2019-2020. Atlanta: American Cancer Society, Inc. 2019**.**
16. P. Mekha and N. Teeyasuksaet, ''Deep learning algorithms for predicting breast cancer based on tumor cells,'' in Proc. Joint Int. Conf. Digit. Arts, Media Technol. With ECTI Northern Sect. *Conf. Electr., Electron., Comput. Telecommun. Eng.* (ECTI DAMT-NCON): 343–346 (2019).
17. P. Israni, ''Breast cancer diagnosis (BCD) model using machine learning,'' *Int. J. Innov. Technol. Exploring Eng.* 4456–4463 (2019).
18. A. A. Bataineh, ''A comparative analysis of nonlinear machine learning algorithms for breast cancer detection,'' *Int. J. Mach. Learn. Comput,* 248–254 (2019).
19. M. K. Keles, ''Breast cancer prediction and detection using data mining classification algorithms: A comparative study,'' Tehnièki Vjesnik, pp. 149–155 (2019).
20. Y. Khourdifi and M. Bahaj, "Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification," 2018 International Conference on Electronics, Control*, Optimization and Computer Science (ICECOCS),* 2018, Corpus ID: 58013185**.**
21. Y. Lu, J.-Y. Li, Y.-T. Su, and A.-A. Liu, "A review of breast cancer detection in medical images," in *Proc. IEEE Vis. Commun. Image Process. (VCIP),* 1–4 (2018).
22. R. Hou, M. A. Mazurowski, L. J. Grimm, J. R. Marks, L. M. King, C. C. Maley, E.-S.-S. Hwang, and J. Y. Lo, ''Prediction of upstaged ductal carcinoma in situ using forced labeling and domain adaptation,'' *IEEE Trans. Biomed. Eng.,* 1565–1572 (2020).
23. A. Memis, N. Ozdemir, M. Parildar, E. E. Ustun, and Y. Erhan, ''Mucinous (colloid) breast cancer: Mammographic and US features with histologic correlation,'' *Eur. J. Radiol.,* 39–43 (2000).
24. A. Reddy, B. Soni, and S. Reddy, "Breast cancer detection by leveraging machine learning," *ICT Express,* 320-324 (2020).
25. Z. Salod and Y. Singh, ''Comparison of the performance of machine learning algorithms in breast cancer screening and detection: A protocol,'' *J. Public Health Res.,* 1677 (2019).
26. S. Eltalhi and H. Kutrani, ''Breast cancer diagnosis and prediction using machine learning and data mining techniques: A review,'' *IOSR J. Dental Med. Sci.,* 85–94 (2019).
27. M. D. Ganggayah, N. A. Taib, Y. C. Har, P. Lio, and S. K. Dhillon, ''Predicting factors for survival of breast cancer patients using machine learning techniques,'' *BMC Med. Inform. Decis. Making,* Art.No-48 (2019).
28. A. A. Ibrahim, A. I. Hashad, and N. E. M. Shawky, ''A comparison of open source data mining tools for breast cancer classification,'' in Handbook of Research on Machine Learning Innovations and Trends. Hershey, PA, USA: IGI

Global,: 636–651 (2017).
29. M. Hosni, I. Abnane, A. Idri, J. M. C. de Gea, and J. L. Fernández Alemán, ''Reviewing ensemble classification methods in breast cancer,'' *Comput. Methods Programs Biomed.* 89–112 (2019).
30. M. Abdar and V. Makarenkov, ''CWV-BANN-SVM ensemble learning classifier for an accurate diagnosis of breast cancer,'' *Measurement,* :557–570 (2019).
31. S. P. Rajamohana, A. Dharani, P. Anushree, B. Santhiya, and K. Umamaheswari, ''Machine learning techniques for healthcare applications: Early autism detection using ensemble approach and breast cancer prediction using SMO and IBK,'' in Cognitive Social Mining Applications in Data Analytics and Forensics. Hershey, PA, USA: *IGI Global,* 2019,: 236–251.
32. M. Togacar and B. Ergen, ''Deep learning approach for classification of breast cancer,'' in Proc. Int. Conf. Artif. Intell. Data Process. (IDAP),: 1–5 (2018).
33. M. Tiwari, R. Bharuka, P. Shah, and R. Lokare, ''Breast cancer prediction using deep learning and machine learning techniques,'' SSRN, New York, NY, USA, *ech. Rep*. 3558786 (2020).
34. D. Selvathi and A. A. Poornila, ''Deep learning techniques for breast cancer detection using medical image analysis,'' in Biologically Rationalized Computing Techniques for Image Processing Applications. Cham, Switzerland: Springer, 159–186 (2018).
35. G. Hamed, M. A. E.-R. Marey, S. E.-S. Amin, and M. F. Tolba, ''Deep learning in breast cancer detection and classification,'' in Proc. Joint Eur.-US Workshop Appl. Invariance Comput. Vis. Cham, Switzerland: Springer, 322–333 (2020).
36. F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, ''Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,'' CA, *Cancer J. Clin.,* 394–424 (2018).
37. S. Khalil, L. Hatch, C. R. Price, S. H. Palakurty, E. Simoneit, A. Radisic, A. Pargas, I. Shetty, M. Lyman, P. Couchot, R. Roetzheim, L. Guerra, and E. Gonzalez, ''Addressing breast cancer screening disparities among uninsured and insured patients: A student-run free clinic initiative,'' J. Community Health, 2019, pp. 1–5, Oct.
38. C. Siotos, A. Naska, R. J. Bello, A. Uzosike, P. Orfanos, D. M. Euhus, M. A. Manahan, C. M. Cooney, P. Lagiou, and G. D. Rosson, ''Survival and disease recurrence rates among breast cancer patients following mastectomy with or without breast reconstruction,'' *Plastic Reconstructive Surg.,* 169e–177e (2019).
39. M. H. Memon, J. P. Li, A. U. Haq, M. H. Memon, and W. Zhou, ''Breast cancer detection in the IOT health environment using modified recursive feature selection,'' *Wireless Commun. Mobile Comput.,* 1–19 (2019).
40. A. A. Said, L. A. Abd-Elmegid, S. Kholeif, and A. Abdelsamie, ''Classification based on clustering model for predicting main outcomes of breast cancer using hyper-parameters optimization,'' *Int. J. Adv. Comput. Sci. Appl*. 268–273 (2018).
41. A. Bharat, N. Pooja, and R. A. Reddy, ''Using machine learning algorithms for breast cancer risk prediction and diagnosis,'' in Proc. 3rd Int. *Conf. Circuits, Control, Commun. Comput*. (IC): 1–4 (2018).
42. E. A. Bayrak, P. Kirci, and T. Ensari, ''Comparison of machine learning methods for breast cancer diagnosis,'' in *Proc. Sci. Meeting Elect.-Electron. Biomed. Eng. Comput. Sci. (EBBT),:* 1–3 (2019).
43. M. Abdar, M. Zomorodi-Moghadam, X. Zhou, R. Gururajan, X. Tao, P. D. Barua, and R. Gururajan, ''A new nested ensemble technique for automated diagnosis of breast cancer,'' *Pattern Recognit. Lett.,:* 123–131 (2020).
44. D. A. Omondiagbe, S. Veeramani, and A. S. Sidhu, ''Machine learning classification techniques for breast cancer diagnosis,'' *IOP Conf. Ser., Mater. Sci. Eng.,* **495**: Art. no. 012033 (2019).
45. S. N. Singh and S. Thakral, ''Using data mining tools for breast cancer prediction and analysis,'' in Proc. 4th *Int. Conf. Comput. Commun. Automat. (ICCCA),* 1–4 (2018).
46. S. Bharati, M. A. Rahman, and P. Podder, ''Breast cancer prediction applying different classification algorithm with comparative analysis using WEKA,'' in Proc. 4th Int. Conf. Electr. Eng. Inf. Commun. Technol. (iCEEiCT), 581–584 (2018).
47. L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, ''Deep learning to improve breast cancer detection on screening mammography,'' *Sci. Rep.,* 1–12 (2019).
48. Zheng, Jing & Lin, Denan & Gao, Zhongjun & Wang, Shuang & He, Mingjie & Fan, Jipeng ''Deep Learning Assisted Efficient AdaBoost Algorithm for Breast Cancer Detection and Early Diagnosis''. *IEEE Access.* 2020. PP. 1-1.
49. M. S. Yarabarla, L. K. Ravi, and A. Sivasangari, ''Breast cancer prediction via machine learning,'' in Proc. 3rd *Int. Conf. Trends Electron. Informat. (ICOEI),* 121–124 (2019).
50. U. Ojha and S. Goel, ''A study on prediction of breast cancer recurrence using data mining techniques,'' in Proc. 7th Int. *Conf. Cloud Computer, Data Sci. Eng.-Confluence*, 527–530 (2017).