IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 30, 2020, accepted August 9, 2020, date of publication August 14, 2020, date of current version August 26, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3016715

Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis

NOREEN FATIMA[1], LI LIU[1], SHA HONG[1], AND HAROON AHMED[2], (Student Member, IEEE)
[1]School of Big Data and Software Engineering, Chongqing University, Chongqing 400044, China
[2]School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China

Corresponding author: Li Liu (dcsliuli@cqu.edu.cn)

This work was supported by National Natural Science Foundation of China (Grant 61977012, 61977054), the Central Universities in China (Grant 2019CDJGFDSJ001, XDJK2019B023), the Chongqing Provincial Human Resource and Social Security Department (Grant cx2017092).

ABSTRACT Breast cancer is type of tumor that occurs in the tissues of the breast. It is most common type of cancer found in women around the world and it is among the leading causes of deaths in women. This article presents the comparative analysis of machine learning, deep learning and data mining techniques being used for the prediction of breast cancer. Many researchers have put their efforts on breast cancer diagnoses and prognoses, every technique has different accuracy rate and it varies for different situations, tools and datasets being used. Our main focus is to comparatively analyze different existing Machine Learning and Data Mining techniques in order to find out the most appropriate method that will support the large dataset with good accuracy of prediction. The main purpose of this review is to highlight all the previous studies of machine learning algorithms that are being used for breast cancer prediction and this article provides the all necessary information to the beginners who want to analyze the machine learning algorithms to gain the base of deep learning.

INDEX TERMS Machine learning, breast cancer prediction, deep learning, data mining, ensemble techniques.

I. INTRODUCTION

Breast cancer is one of the most lethal and heterogeneous disease in this present era that causes the death of enormous number of women all over the world. It is the second largest disease that is responsible of women death [1]. There are various machine learning [2] and data mining algorithms that are being used for the prediction of breast cancer. Finding the most suitable and appropriate algorithm for the prediction of breast cancer is one of the important task. Breast cancer is originated through malignant tumors, when the growth of the cell got out of control [3]. A lot of fatty and fibrous tissues of the breast start abnormal growth that becomes the cause of breast cancer. The cancer cells spread throughout the tumors that cause different stages of cancer. There are different types of breast cancer [4] which occurs when affected cells and tissues spread throughout the body. Ductal Carcinoma in Situ (DCIS) is type of the breast cancer that occurs when abnormal cells spread outside the breast it is also known as the non-invasive cancer [5]. The second type is Invasive Ductal Carcinoma (IDC) [6] and it is also known as infiltrative ductal carcinoma [7]. This type of the cancer occurs when the abnormal cells of breast spread over all the breast tissues and IDC cancer is usually found in men [8]. Mixed Tumors Breast Cancer (MTBC) is the third type of breast cancer and it is also known as invasive mammary breast cancer [9]. Abnormal duct cell and lobular cell causes such kind of cancer [10]. Fourth type of cancer is Lobular Breast Cancer (LBC) [11] which occurs inside the lobule. It increases the chances of other invasive cancers. Mucinous Breast Cancer (MBC) [12] is the fifth type that occurs because of invasive ductal cells, it is also known as colloid breast cancer. It occurs when the abnormal tissues spread around the duct [13]. Inflammatory Breast Cancer (IBC) is last type that causes swelling and

The associate editor coordinating the review of this manuscript and approving it for publication was Md. Moinul Hossain.

reddening of breast. It is a fast growing breast cancer, when the lymph vessels block in break cell, this type of cancer starts to appear [14].

Data mining is a process of discovering the useful information from a big dataset, data mining techniques and functions help to discover any kind of disease, data mining techniques such as machine learning, statistics, database, fuzzy set, data warehouse and neural network help in diagnosis and prognosis of different cancer diseases [15] such as prostate cancer, lungs cancer [16] and leukaemia [17]. Traditional methodology of cancer detection is based on ''the gold standard'' method that consists of three tests: clinical examination, radiological imaging and pathology test [18]. This conventional method indicates the presence of cancer and it is based on regression process while the new machine learning techniques and algorithms are based on model design. Model is designed for the prediction of unseen data and provides the good expected result in their training and testing stages [19]. Machine learning process is based on three main strategies that consists of preprocessing, features selection or extraction and classification [20]. Feature extraction is the main part of machine learning process and actually helps in diagnosis and prognosis of cancer, this process can elaborate the cancer set in to benign and malignant tumors [21].

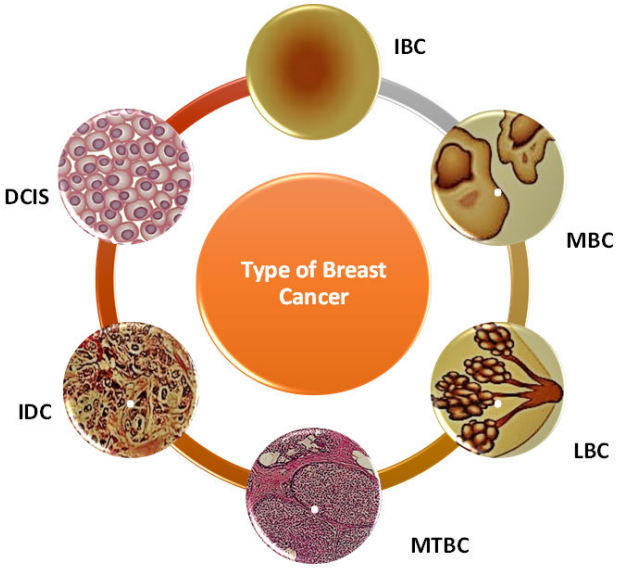

FIGURE 1. Demonstration of major types of Breast Cancer.

Data mining and machine learning algorithms help us for diagnoses and predictions of such types of breast cancer as shown in Figure 1. Data mining techniques [22] such as classification, regression and clustering help us to get the meaningful information about the breast cancer patients. These algorithms [23] consist of training dataset, with the help of these datasets we can find chances of prediction of different kinds of breast cancer [24]. This article is divided into different sections. Section II is about the major machine learning algorithms that are being used for breast cancer prediction, section III is about the major ensemble techniques being used for the prediction of breast cancer, section IV is about the deep learning techniques for breast cancer diagnosis, section V is the survey on breast cancer, section VI is review of different machine learning and deep learning algorithms, section VII is about the study selection and materials that we have used in this research, section VIII provides the discussion and section IX provides the conclusion of this review article.

II. MACHINE LEARNING ALGORITHMS FOR BREAST CANCER PREDICTION

Machine learning is an automatic learning method [25], the algorithms are designed to learn from past dataset, we input a large number of data, machine learning model analyze that data and on the basis of that train model we can make a prediction about future [24], [26], [27]. For breast cancer predictions, major machine learning algorithms are as follow:

A. ARTIFICIAL NEURAL NETWORK (ANN)

Artificial Neural Network [28] is a common algorithm for data mining process. Neural network consists of input layer, hidden layer and output layer. This technique is used to extract the pattern that is too complex [29]. Algorithm is based on parallel processing [30], distributed memory [31], collective solution and network architecture [32]–[34].

B. LOGISTICS REGRESSION (LR)

It is a supervised learning algorithm that includes more dependent variables. The response of this algorithm is in the binary form. Logistics regression [35] can provide the continuous outcome of a specific data. This algorithm consists of statistical model with binary variables [32].

C. K-NEAREST NEIGHBOR (KNN)

This algorithm is used in pattern recognition. It is a good approach for breast cancer prediction. In order to recognize the pattern, each class has given an equal importance. K Nearest Neighbor [36] extract the similar featured data from a large dataset. On the basis of features similarity we classify a big dataset [32].

D. DECISION TREE (DT)

Decision tree [37] is based on classification and regression model. Dataset is divided into smaller number of subsets. These smaller set of data can make prediction with the highest level of precision. Decision tree method includes CART [38], C4.5 [39], C5.0 [40] and conditional tree [32], [41].

E. NAIVE BAYES ALGORITHM (NB)

This model is used to make an assumption of large training dataset. The algorithm is used to calculate the probability through Bayesian method [42]. It provides the highest accuracy while calculating the probabilities of noisy data that is used as an input [43]. It is an analogy classifier that is used for comparing training dataset with training tuple [32].

F. SUPPORT VECTOR MACHINE (SVM)

It is a supervised learning algorithm which is used for both classification and regression problems [44]. It consists of theoretical and numeric functions to solve the regression problem. It provides the highest accuracy rate while doing prediction of large dataset. It is a strong machine learning technique that is based on 3D and 2D modelling [32], [45].

G. RANDOM FOREST (RF)

Random Forest algorithm [46] is based on supervised learning [47] that is used to solve both classification and regression problems. It is a building block of machine learning that is used for prediction of new data on the basis of previous dataset [32].

H. K MEAN ALGORITHM

K mean is clustering algorithm that provides the partition of data in the form of small clusters. Algorithm is used to find out the similarity between different data points. Data points exactly consist of at least one cluster that is most suitable for the evaluation of big dataset [48].

I. C MEAN ALGORITHM

Clusters are identified on the similarity basis. Cluster that consist of similar data point belongs to one single family. In C mean algorithm each data point belongs to one single cluster. It is mostly used in medical images segmentation and disease prediction [49].

J. HIERARCHICAL ALGORITHM

Hierarchical algorithm mostly provides the evaluation of raw data in the form of matrix. Each cluster is separated from other clusters in the form of hierarchy. Every single cluster consists of similar data points. Probabilistic model is used to measure the distance between each cluster [50].

K. GAUSSIAN MIXTURE ALGORITHM

It is most popular technique of unsupervised learning. It is known as soft clustering technique which is used to compute the probability of different types of clustered data. The implementation of this algorithm is based on expectation maximization [51].

III. ENSEMBLE TECHNIQUES FOR BREAST CANCER PREDICTION

Ensemble techniques are considered as homogeneous and heterogeneous; homogenous ensemble techniques [52] are the combination of one base method and two or more configuration methods such as bagging and boosting technique while the heterogeneous is used to combine two or more base methods, and ensemble technique is based on supervised learning that provides the good prediction on the basis of some hypothesis [53]–[55].

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

Stacking is heterogeneous [52] weak learner that combines the different machine learning algorithms for prediction on same dataset. It consists of two or more base models and merges the prediction of base model [54], [55].

IV. DEEP LEARNING TECHNIQUES FOR BREAST CANCER PREDICTION

Deep learning is broader form of ANN (Artificial Neural Network). Deep learning algorithms consist of multiple layers architecture. These algorithms are used to process the large number of natural data and have ability to recognize all the data from different categories [56]. We mostly apply the unsupervised deep learning techniques [57] when we have huge amount of unlabeled data [58], [59].

A. AUTO ENCODER

Auto encoder basically consists of encoder that followed the decoder, encoder usually transfer the input in the form of variables like x, y and decoder takes that input and try to get back all the original input. The main purpose of auto encoder is to learn from a big dataset, by training its network that ignore the irrelevant signals such as noise [58], [60], [61].

B. SPARSE AUTO ENCODERS

Sparse Auto Encoders automatically learn from unlabeled data. The Sparse Auto Encoder is basically a feed forward and back propagation algorithm with a regular auto encoder [57]. A sparse auto encoder can handle the sparsity regularizer. Sparsity regularizer provides the sparsity of output from the hidden layer of neural network [58], [60], [61].

C. STACKED SPARSE AUTO ENCODER (SSAE)

When the basic layer of Stacked Sparse Auto Encoder (SSAE) [57] are combined together it will construct the stretched sparse, the output of the first layer is merged with the output of the second layer, each stacked spare consists of hidden layers which are based on classifier and provide the output [58], [60], [61].

D. CONVOLUTIONAL NEURAL NETWORK

CNN can analyze the cancer dataset in the form of images, during the preprocessing phase it uses CovNet to analyze

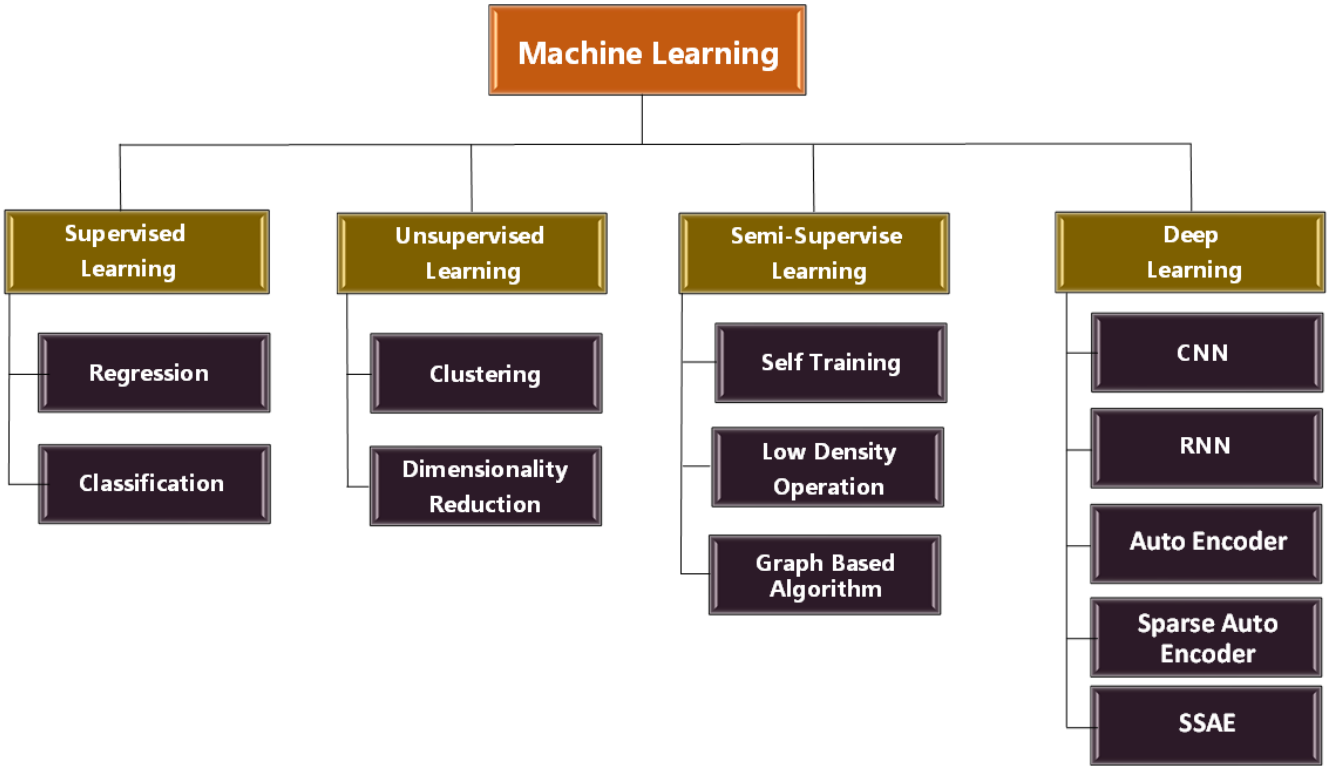

FIGURE 2. Classification of Algorithms in Machine Learning [62].

the different set of data and using some filters the CovNet can capture the different dimensions of images. Layers are divided in to pooling layer, convolutional layer, classification layer and fully contacted layer. CNN is the combination of all these layers [60], [63].

E. RECURRENT NEURAL NETWORK

Recurrent Neural Network (RNN) is a class of neural network, that consists of some hidden states, which uses the output of previous state as an input of next state. It can process a sequence of inputs that uses the same parameters at each layer which reduces the complexity of that parameters more accurately than the other neural networks but it can not process a large number of sequence of inputs through ReLU and Tanh activation functions [58], [64].

V. SURVEY ON BREAST CANCER

China is most populated country around the globe. According to the recent report of organization (GLOBOCAN-2018) the ratio of breast cancer in males is 8.6% while in females is 19.2% [65]. 1.2 million people dying each year from this disease. American Cancer Society diagnosed 48,100 cases of DCID cancer which were found in women. US 2019 report shows that 500 men and 41,760 women are expected to die because of breast cancer [66]. US report shows that the women that are alive but suffering with breast cancer are 3.8 million. In US women 59,838 Ductal Carcinoma in Situ (DCIS) breast cancer cases were found in 2019 [67]. Overall breast cancer deaths are 458,000. In 2012 the death ratio of Chinese people from breast cancer was 48%, while the death ratio all over world in 2012 was 52% [68]. To check the breast cancer survival and recurrence rate, data of 1,517 women was analyzed in 2015, the breast cancer recurrences rate was 100 and the death rate was 132 [69].

VI. REVIEW OF MACHINE LEARNING ALGORITHMS FOR BREAST CANCER PREDICTION

The main purpose of this research is to review different machine learning and data mining algorithms that helped people for the prediction of breast cancer. Our main focus is to find out the most accurate and suitable algorithm for breast cancer prediction. For this, we have reviewed and analyze the past studies of breast cancer prediction algorithms also reviewed the research papers that are based on linear (Linear Regression, Logistic Regression, Linear Discriminant Analysis), nonlinear (Classification and Regression Tree, Naive Bayes, K-Nearest Neighbor, Support Vector Machine) and some ensemble algorithms (Decision Tree, Random Forest, Boosting and AdaBoost). Most of the researcher used combination of linear and nonlinear or combination of nonlinear and ensemble algorithms. So, for this we categorize our review paper in different sections that will provide the comparative analysis of each algorithm on the basis of their accuracy rate. After that comparison we will highlight the most suitable machine learning algorithm for breast cancer prediction.

A. NONLINEAR ALGORITHMS

For the prediction of breast cancer, nonlinear algorithms such as Random Forest, Naive Bayes, Support Vector Machine and K Nearest Neighbor were comparatively used by authors for the prediction of breast cancer. Authors used the Bioinformatics and Medical Science classification technique. This technique was based on the selection of best classifier, comparison of data mining algorithm was performed to choose the most suitable algorithm for the prediction. After the comparative analysis of four classification techniques, authors found that Support Vector Machine (SVM) was more suitable than the other algorithms and it provided 97.9% accuracy [2].

Data mining classification algorithms which include Bagging Algorithm, IBk (Instance based learning with some parameters), Random Committee Algorithm, Random Forest Algorithm, Simple Classification and Regression Tree (Simple CART Algorithm) were used for prediction and detection of breast cancer. Antenna dataset was used to measure the accuracy of each algorithm. Result was analyzed on various Weka types like Bayes, Function, Meta, Lazy, Trees etc. After the analysis, authors came to know that Random Forest Algorithm provided higher accuracy level than other algorithms and was found to be most suitable algorithm for breast cancer predictions. The accuracy rate of Random Forest algorithm was 92.2% while the accuracy of Bagging, IBk and Random Committee Algorithm were found to be 90.9%, 90% and 90.9% respectively [70].

Gene Expression (GE) and DNA methylation data was used for breast cancer prediction. Support Vector Machine (SVM), Decision Tree and Random Forest algorithm were used to classify nine models for the prediction of cancer. Authors comparatively analyzed this dataset on two data mining tools: Weka and Spark, in order to show the accuracy and error rate of algorithms. Authors filtered these two datasets (GE and DM) in order to get the common genes with the main purpose to identify the presence of tumors. After the comparative analysis of these algorithms on two different tools, the accuracy of SVM was found to be higher than the others algorithms, which was 99.68 % on Spark and 98.03 % on Weka tool [71].

Naive Bayes, K Nearest Neighbors and J48 algorithm were used for the prediction of 9 different types of cancer including breast cancer. Dataset was collected with the help of different doctors and experts which consists of 61 attributes and 1059 records, on the basis of training set of data, authors initially compared the symptoms to test result whether it is true or false and if the symptoms got matched then it means that result is true. Through this process authors predicted the different types of breast cancer and also they classified each algorithm on the basis of their accuracy rate. During the process of breast cancer detection the accuracy rate of NB and KNN was found to be higher than J48 decision tree classifier, their accuracies were 98.2%, 98.8% and 98.5% respectively [72].

For the detection of breast cancer, authors used the Support Vector Machine (SVM) which is recursive feature elimination technique with predictive machine learning model. The aim was to the select correct features from a dataset of benign and malignant people. Authors used the dataset from Wisconsin Diagnostics Breast Cancer (WDBC) database. Recursive feature elimination technique was used for the evaluation of SVM algorithm. Performance matrix was designed to check the accuracy rate of predictive model SVM (Support Vector Machine) on different types of kernel. Support Vector Machine provided the 99% accuracy on linear kernel, 98% on RBF kernel, 97% accuracy on polynomial kernel and 84% accuracy on sigmoid kernel [73].

Classification model on dataset was applied for the prediction of breast cancer, dataset was combined in the form of cluster, and each cluster consisted of similar functionality data. To increase the accuracy rate of classification model, authors used another technique which is known as Hyper Parameter Optimization. Dataset was collected from ''National Cancer Institute'' of Egypt, main purpose was to predict the breast cancer in Egyptian people. Hyper Parameter Optimization (HPO) technique was used to get the higher accuracy rate of prediction. Authors initially collected the dataset from NCI center and then applied clustering approach to combine the similar pattern, after that features selection method was applied to select some relevant feature for the prediction process. Decision Tree model was used to categorize each data and hyper parameter optimization technique was applied to check the presence of breast cancer [74].

Support Vector Machine, Decision Tree, Naive Bayes and K Nearest Neighbor was used for breast cancer risk prediction and diagnosis. Dataset was collected from Wisconsin Diagnostics Breast Cancer (WDBC). Experiment was conducted on Weka tool to evaluate the predictive models and authors used the K Fold cross validation technique by splitting the data into training and testing sets. The accuracy of SVM was achieved as 97.13% with 0.07s execution time, which was higher than the other algorithms but the execution time of SVM was higher than the KNN algorithm [75].

B. ENSEMBLE ALGORITHM

Support Vector Machine, Nearest Neighbor Algorithm, Logistics Regression and Naive Bayes was used to observe the breast cancer dataset that was categorized as malignant or benign. Support Vector Machine technique was implemented on two Support Vector Machine kernels, one was linear kernel and other one was Gaussian kernel, Nearest Neighbor algorithm was implemented through Manhattan distance and Euclidean distance. Normal distribution and Kernel distribution methods were used for the implementation of Naive Bayes algorithm. The analysis of all techniques had done by getting the data from UCI depository WDBC and WPBC. Authors used MATLAB tool to correctly classify data with respect to their accuracy. The WPBC dataset consisted of 34 attributes which were used for the prediction of breast cancer while WDBC dataset consisted of 32 attributes which

were used for the diagnosis of breast cancer. K Nearest Neighbors Algorithm was found to be most appropriate algorithm for breast cancer diagnosis and prognosis [76].

Dimensionality reduction technique for feature selection and extraction was used for breast cancer dataset analysis. Authors applied three machine learning techniques including Support Vector Machine (SVM), K Nearest Neighbor (KNN) algorithm and Logistics Regression (LR) on breast cancer dataset. Analysis of these algorithms was observed on the basis of accuracy, precision and sensitivity. The technique Logistics Regression which is based on logistics function was used to calculate the outcome of independent variable. Authors analyzed the KNN algorithm using Euclidean distance. The value of K varies for different datasets. Dataset was collected from UCI depository. Spyder tool was used to measure the accuracy of each algorithm. The accuracy of SVM was found to be 92.78% [77].

Data of breast cancer patients was collected from ICBC (Iranian Centre for Breast Cancer) and compared using three machine learning techniques. Authors analyzed the result of Decision Tree (C4.5), Artificial Neural Network and Support Vector Machine in terms of their accuracy, sensitivity and specificity level. Authors focused on multi-layer perceptron (MLP) model to find the ANN algorithm accuracy. The result showed that Support Vector Machine was the best machine learning prediction algorithm for breast cancer prediction with 95.7% accuracy [78].

Comparison of two algorithms: Support Vector Machine and Artificial Neural Network was performed for breast cancer diagnoses. Support Vector Machine algorithm was used as a pattern recognition of Wisconsin Breast Cancer dataset based on the breast cancer patient's age and the size of the tumors. Support Vector Machine classified the tumor as benign or malignant while Artificial Neural Network was used for the modelling of nonlinear function, K Fold validation technique was applied on both algorithms. Authors measured the validation process on the basis of accuracy. The accuracy of SVM was found to be higher than Artificial Neural Network. The accuracy rate of SVM was 96.9% while the accuracy rate of ANN was 95.4% [79].

Ensemble classifier such as confidence weighted voting method (CWV) combines boosting ANNs (BANN) with SVM used for the diagnosis of breast cancer. Authors applied machine learning algorithms: ANN and SVM for correct diagnosis of breast cancer. Breast cancer detection approach was done through confidence weight voting method and boosting ensemble technique, performance matrix was designed to evaluate these models, authors came to know that the CWV-BANNSVM model was good for the detection of breast cancer and this model had provided the good accuracy results by splitting the data into 50% training and 50% testing test. Dataset was collected from UCI repository university of California named as Irvine. Model was proposed to select the useful information from raw data that was actually consisted of 669 records after applying the ensemble method with machine learning algorithms SVM and ANN, the accuracy level that was achieved through CWV-BANN was higher [53].

Sequential Minimal Optimization (SMO) and K Nearest Neighbours Classifier (IBk) were used for the prediction of breast cancer with some ensemble approaches. Authors used the data that was consisted of 683 instances and 9 input attributes. Experiment was done through Weka data mining tool, K fold cross validation was used for the evaluation of accuracy of algorithm. The accuracy rate of SMO was found to be 96.19% and IBk algorithm accuracy was found to be 95.90% [54].

For the automated diagnosis of breast cancer, authors used the nested ensemble techniques with classifiers, dataset was collected from Irvine UCI repository that had 32 tumors features and 569 subjects. In order to analyze the classifier for the prediction of tumors, authors applied K fold cross validation, evaluation was done through simple ensemble method and provided the comparative analysis of Naive Bayes classifier with Bayes Net. The accuracy of Bayes Net algorithm was found to be 95.25% which was higher than the Naive Bayes Algorithm. Authors also compared the NB and Byes Net with meta classifier, the performance of 3rd meta classifier with SV-Naive Bayes was higher than the others meta classifiers [80].

C. LINEAR AND NONLINEAR ALGORITHM

Features selection and features extraction techniques on Artificial Neural Network (ANN), Support Vector Machine (SVM) and Naive Bayes (NB) were applied for the prediction of breast cancer. Dataset of patients was collected from Wisconsin Diagnostic Breast Cancer. Feature selection is a selection of sub features from a huge dataset that helps in computation process. Authors comparatively analyzed each technique with different type of features selection such as coloration based feature selection (CFS), Linear Discriminant Analysis and Recursive Feature Elimination (RFE). After the comparative analysis with different features selection methods, authors came to know that the accuracy rate of Artificial Neural Network was higher than the other algorithms. The accuracy of Support Vector Machine was 96.4%, Artificial Neural Network was 97.0% and Naive Bayes accuracy was 91% [81].

Data mining tools were used for the prediction of breast cancer, main purpose was to classify Naive Bayes (NB) algorithm, Bayesian Logistic Regression, Simple CART and J48 on the basis of some parameters. Dataset was collected from Wisconsin Breast Cancer (WBCO). Decision Tree algorithm was used to split whole data into subsets while J48 was based on decision node, these nodes guess the expected data from whole set of data. The main purpose of the author was to find out the best classifier on the basis of their Kappa Statistics, Error Rate and their accuracy parameters. The accuracy rate defined the percentage of correctively predicted data. Weka tools was used to analyze all these classifiers. After that analysis, authors came to know that Simple CART is more

appropriate than the other algorithms and it provided the 98.13% accuracy [82].

Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbors (KNN) and Classification And Regression Tree (CART) algorithms were used for breast cancer prediction. Authors compared the data mining algorithms to achieve the best accuracy with less error rate. Analysis was performed using the Weka libraries on the original dataset of Wisconsin Breast Cancer. Dataset consisted of 357 benign patient and 212 records of malignant patients, 80% of data was divided in to train set while 20% data was used for testing, to avoid the overfitting and underfitting issues authors used the cross validation, regularization and dropout techniques. After analysis the accuracy rate of Support Vector Machine (SVM) was found to be higher than the other data mining algorithms, that provided the 99.1% accuracy [83].

For the prediction of breast cancer, Greedy Algorithm was proposed with some constraints search. Constrained Search Sequential Floating Forward Search (CSSFFS) is a feature selection algorithm with Support Vector Machine whose main purpose is to extract some relevant features from a large set of data and it also removes irrelevant features. Dataset for this experiment was collected from machine learning depository WDBC. CSSFFS was used as hybrid algorithm with Support Vector Machine (SVM), authors used the K Fold cross validation technique on different algorithms, whose main purpose was to extract the irrelevant features and calculate the accuracy rate. Through CSSFFS algorithm, 15 features were selected and CSSFFS increases the accuracy rate of other machine learning algorithms. The accuracy rate of RBF network was 93.6%, Naive Bayes was 92.6%, J48 accuracy rate was 92.9% and simple CART algorithm accuracy rate was 92.9% [84].

To get the useful information about the tumors, authors used the hybrid technique by combining the K Mean and SVM algorithm, this hybrid method provided the 97.38% accuracy after the K Fold cross validation process. Dataset was collected from University of California through WDBC center. Six tumor features were selected from 32 original features. The objective of using hybrid method K-SVM was to predict the tumors whether it is malignant or benign. K-SVM algorithm provided the better performance with minimum error rate [85].

D. NONLINEAR AND ENSEMBLE ALGORITHM

Decision tree, Naive Bayes and K Nearest Neighbor was applied comparatively on dataset for breast cancer prediction. Authors used the Wisconsin original dataset that was collected from UCI machine learning repository. The dataset used had 10 attributes with 458 benign and 241 malignant patients, three major matrices were designed on the basis of two classes: actual healthy and actual not healthy to predict the sensitivity of data. To analyze the performance of each algorithm Weka tool was used, after the comparative analysis of each technique authors came to know that the accuracy of Naive Bayes algorithm was 95.99% which was higher than the accuracy of decision tree and K Nearest Neighbor (KNN) [86].

Authors applied Naive Bayes classifiers, Support Vector Machine, K Star, Decision tree and ANN to analyze the patients dataset. The analysis of all algorithms was performed using Weka (Data mining tool). SMO was implemented on RBF kernel. This algorithm normalized all the attributes. KNN was implemented using MLP (Multi-Layer Perceptron). MLP had input layer, a hidden layer and output layer and K Star algorithm was used to determine the similarity of data. After the comparative analysis of all these algorithms on the dataset of university of medical centre, Institute of Oncology, authors came to know that the accuracy of J48 Decision tree was 75.52% which was higher than the all other algorithms [87].

Different machine learning algorithms including Decision tree J48, Neural Network, Logistic Regression (LR), Support Vector Machine (SVM) and K Nearest Neighbor (KNN) were used for the prediction of breast cancer. Authors collected the dataset from Wisconsin Breast Cancer (WBC). In order to compare the accuracy of algorithms, Weka tool was used that is based on classification, association, clustering and visualization of data. The accuracy rate of SVM algorithm was found to be 97.59% which was higher than the all other techniques. Authors evaluated that SVM (Support Vector Machine) was more suitable algorithm for breast cancer prediction [88].

Comparison of five nonlinear machine learning algorithms including Multi-Layer Perceptron (MLP), K Nearest Neighbors (KNN), Classification And Regression Tree (CART), Support Vector Machines (SVM) and Gaussian Naive Bayes was done for breast cancer detection. Main objective of the author was to compare the efficiency and effectiveness of algorithms for breast cancer detection. Author also measured the accuracy of each algorithm separately. Analysis was performed on Wisconsin breast cancer diagnostics dataset (WBCD). Author used the K Fold validation method to predict the accuracy of each algorithm. The accuracy of MLP was found to be 96.70% which was higher than the KNN, CART and NB algorithm [89].

Surveillance Epidemiology and End Result (SEER) database was used to analyze the breast cancer survivability rate. SEER data is more reliable for prediction of different stages of breast cancer. Authors used three data mining techniques including Naive Bayes, back propagate neural network and C4.5 decision tree algorithm. Comparison was done through data mining tool Weka. Decision tree C4.5 algorithm was found to be more appropriate algorithm for breast cancer but it does not include the record of missing data. Survivability of cancer patients was calculated using Weka toolkit that showed the graphical representation of tumor size and its rank, tumor size was higher than its rank. The accuracy of C4.5 algorithm was found to be 86.7% which was higher accuracy than the other algorithms [90].

Data mining pre-processing and classification algorithms was used to detect the breast cancer. After classification

of dataset, authors choose two data mining algorithms: J48 Decision Tree and ZeroR. ZeroR classifier was based on frequent analysis of data that was analyzed using the frequency table while J48 machine learning algorithm was applied to predict the value of dependent variable from a dataset of independent variables. Authors used the pathology report to analyze the attributes. On the basis of dataset patterns, author selected some important attributes to predict the occurrence of breast cancer [91].

ADTree, J48 and CART algorithm was used to analyze the Breast Cancer dataset of Indian cancer centre Adyar, Chennai and took digital images in the form of DICOM (Digital Imaging and Communication in Medicine). The dataset that authors used was in the form of CSV and authors applied three different data mining algorithms to check the accuracy level. Authors came to know that CART algorithm is more suitable for breast cancer analysis than others, because the accuracy level of CART was 98.50% while the other algorithms Adtree and J48 accuracy rate was 97.70% and 98.10% respectively [92].

Breast cancer is most common disease in Nigerian women, while in Nigeria there is no prediction and detection of this heterogeneous disease. Authors collected the breast cancer dataset from LASUTH, cancer registry, Nigeria. Dataset had 17 different breast cancer attributes. Authors used Naive Bayes and J48 decision tree algorithm, Naive Bayes probabilistic model is used to handle the number of classes based on probabilistic theory. In J48 decision tree, top to down greedy search was applied on training dataset. Authors came to know that the accuracy level of Naive Bayes was 82.6%, while the accuracy of J48 Decision tree was 94.2% that shows that J48 is most suitable algorithm for breast cancer prediction and detection [93].

Researchers provided the comparative analysis of Naive Bayes, Random Forest, Logistics Regression, Multi-Layer Perceptron and K Nearest Neighbors for the breast cancer prediction. Evaluation of all these algorithms was performed in terms of Kappa Statistics analysis, TP rate, FP rate and precision of each algorithm. Dataset of Breast Cancer patients was collected from UCI machine learning repository. From dataset, 10 different attributes were collected to predict the breast cancer. Each algorithm was applied on dataset to analyze the accuracy of each algorithm. The accuracy of K Nearest Neighbors, Naive Bayes and Random Forest was 72.3%, 71.6% and 69.5% respectively while Logistics Regression and Multi-Layer Perceptron classified instances accuracy was 68.8% and 64.6% respectively [94].

E. DEEP LEARNING ALGORITHM

To predict the breast cancer on tumor cells, authors used the deep learning technique with different activation functions: Tanh, Rectifier, Maxout and Exprectifier, to provide the comparative analysis with machine learning algorithms such as Naive Bayes, Decision Tree, Support Vector Machine (SVM) and Random Forest. Wisconsin dataset was used and it had 457 benign class tumors and 241 malignant class tumors. After the comparative analysis of Decision Tree, Naive Bayes, Random Forest and Support Vector Machine (SVM), authors came to know that the accuracy rate of the algorithm using Exponential Rectifier Linear Unit (ELU) activation function was found to be highest with 96.99% accuracy [95].

A model was proposed to predict the reparative appearance of breast cancer. This model consisted of two main algorithms: Extreme Learning Machine and Bat algorithm. Bat algorithm was used to create the biases and random weights. The dataset was collected from Wisconsin Breast Cancer Prognostic. The dataset was analyzed on MATLAB tool. The implementation was performed by collecting the relevant attributes from a big dataset. For attributes selection, coefficient correlation method was used and after that Bat algorithm and Extreme Learning parameters were applied to check the recurrent and non-recurrent of breast cancer. Deep learning activation functions such as sigmoid, sine and tanh were used to check the testing accuracy on different training stages. Tanh activation function provided the good accuracy than the other activation functions that was 93.75% [96].

Deep learning techniques such as Stack Sparse Auto Encoder (SSAE), Sparse Auto Encoder (SAE) and Convolutional Neural Network (CNN) were used for the error free detection of breast cancer using mammograms. Preprocessing involves the noise removal, background removal and artifact suppression. The next step is ROI segmentation that was applied for the detection of tumor by removing the pectoral muscle. The last step was cancer detection for that process, authors provided the input generation, then they construct the deep neural network and after training and testing phase final input was generated. Dataset had 322 digitized images which were actually mammograms. Confusion matrix was designed to analyze the accuracy, sensitivity and precision using MIAS database. The accuracy of SSAE, SAE and CNN was 98.9%, 98.5% and 97% respectively. Stack Sparse Auto Encoder provided the good accuracy for the detection of breast cancer in early stages [60].

By using end to end training approach authors developed an end to end training for the detection of breast cancer with the help of deep learning algorithm. Analysis was done on Linux workstation, the method was carried out by developing the match and whole images classification on CBIS-DDSM. Confusion matrix was designed to construct the Resnet 50 and VGG16 patch classifiers. Deep learning methodology was used to analyze the cancer patients images, learning efficiency was analyzed through different training set and visualization of images was improved by adding more and more patches around the ROI and in the background [97].

To enhance the cancer diagnosis, authors applied the unsupervised and deep learning method. Authors initially reduced the dimensionality of features by using PCA method and then they applied the PCA to represent features as a compressed structure which were actually encoded through some sample set and randomly selected gene expression. Sparse Auto Encoder technique was applied on multilayers. Authors

used the learning classifier known as Softmax regression to analyze the results [63].

For the diagnosis, prognosis and prediction of breast cancer, authors modularized images into MRI, Digital Images and Ultrasound. Dataset was collected from online open source platform wiki, authors applied deep learning algorithms: Autoencoders, CNN and LSTM to achieve the higher level of accuracy, authors proposed the high level learning model Adaboost (DLA-EABS) for final predictions that provided the good accuracy with maximum survival rate. The accuracy of this model was 97.2% [98].

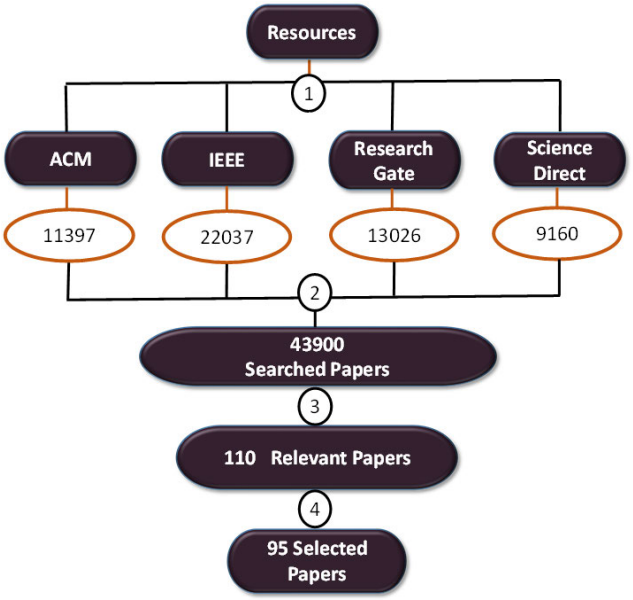

FIGURE 3. Paper selection process.

VII. OVERVIEW OF STUDY SELECTION

For the review of breast cancer predictions, the number of papers that we have considered for study selection at each stage are shown in Figure 3. Total number of papers we have found using keyword search were 43,900 that we have got from different platforms like ACM, IEEE, Research Gate and Science Direct. Our search query was focused on four keywords: machine learning, deep learning, data mining and breast cancer prognosis or diagnosis.

Our main aim was to focus on papers that included prediction of breast cancer using machine learning and deep learning techniques. We have applied inclusion and exclusion criteria for the selection of relevant material. So, after removing all the duplicate papers, we have selected 110 papers for deep study that was purely related to our research. After thorough study of these papers we have shortlisted 95 papers and finally considered for this review. We have considered three channels for our research that are journals, conferences and books. We have reviewed total 64 journal papers, 31 conference papers and 7 books.

Table 1 demonstrates the selection of paper according to different years, table is divided into journal and conference papers that we have reviewed. The last column in Table 1 is

TABLE 1. Year wise number of journal and conference papers.

Year	Journal Papers	Conference Papers	Total
Before 2011	13	4	17
2011	1	2	3
2012	3	0	3
2013	3	3	6
2014	2	0	2
2015	7	4	11
2016	6	1	7
2017	1	3	4
2018	4	7	11
2019	19	6	25
2020	5	1	6
2011-2020	51	27	78

about the total number of papers in one specific year that we have selected for our study. Majority of papers that we have selected for our review analysis are from the last 10 years i.e. 2011-2020 with the maximum number of papers from last year (2019) that give us the information about the breast cancer prediction using machine learning and major deep learning techniques in the present era.

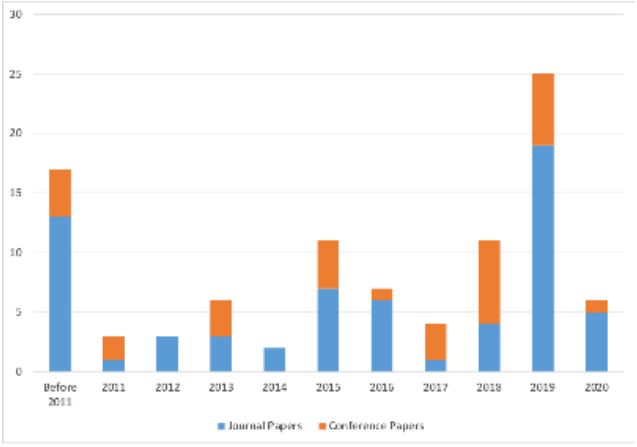

FIGURE 4. Number of papers per year.

Figure 4 demonstrates the per year frequency of selected research articles in bar plot, it is clearly shown that our main focus is to collect the most recent research papers about the breast cancer prediction. The graph highlights the research papers that we have studied for our breast cancer review. The graph is year based; first we combined number of papers before 2011 then graph is started from the year 2011 and each year shows the number of journal and conference papers that are separately highlighted with two different colors in the graph. We can see that in our review we have maximum number of research papers from last year (2019), in order to find the most recent and appropriate techniques or methods for breast cancer prediction.

VIII. DISCUSSION

This research summarizes different machine learning, deep learning and data mining algorithms for the prediction of breast cancer. Table 2 provides comparative summary of machine learning techniques for breast cancer prediction on the basis of tools, data sources, data type, data

TABLE 2. Comparative review of machine learning techniques for breast cancer prediction.

Algorithm	Tool	Data set	Number of Attributes	Data type	Pre Processing	Data Processing Method	Evaluation Method	Validation Technique	Accuracy	R#
Support-Vector Machine(SVM) K-Nearest Neighbor (KNN) Logistics Regression(LR) Naive Bayes(NB)	MATLAB	UCI depository WDBC WPBC	32WDBC 34WPBC	Numeric	Discrete value	Feature selection	Function graph(LR) Normal distribution Kernel distribution (NB)	SVM Dynamic KNN Euclidean LR generalize NB Kernel	93% 95% 90% 92%	[76]
Support-Vector Machine(SVM) K Nearest Neighbor (KNN) Logistics Regression(LR)	Spyder	UCI depository	32 Attributes	Numeric	Discrete value	Feature selection Dimen-sionally Reduction	Principle Component Analysis (PCA)	SVM-PCA testing KNN Euclidean LR Logistic Function	92.78% 92.23% 92.10%	[77]
Decision Tree(C4.5) Artificial Neural Network(ANN) Support-Vector Machine (SVM)	Weka	Iranian Center ICBC	22 Attributes	Numeric	Mixed value	Data Cleaning and Prepa-ration	Expectation Maxi-mization Algorithm (EMA)	10 Fold Cross Validation	93.6% 94.7% 95.7%	[78]
MLP K-Nearest Neighbor(KNN) CART Naive Bayes SVM	Mathe-matical	UCI depository WDBC	32 Attributes	Images	Binary Val-ues	Standardize rescaling method	Binary Classi-fication Accuracy Method	10 Fold Cross Validation	99.12% 95.61% 93.85% 94.73% 98.24%	[89]
C4.5 Support-Vector Machine CART Naive Bayes(NB) K-Nearest Neighbor(KNN)	Weka	Wisconsin Breast Cancer	11 Attributes	Numeric	Discrete Values	Kappa statistics Mean absolute error	Confusion Matrix	10 Fold Cross Validation	95.13% 97.13% 95.99% 95.27%	[99]
Naive Bayes Artificial Neural Network CART	Weka	SEER Database	151,886 records	Numeric	Mixed Values	Features Extraction	Confusion Matrix	10 Fold Cross Validation	84.5% 86.5% 86.7%	[90]
Random Forest Algorithm Random commit-tee and bagging Simple CART Tree Base IBk	Weka	Antenna Database	4 Attributes	Numeric	Discrete Values	Features Extraction	Binary Classi-fication Accuracy Method	10 Fold Cross Validation	92.2% 90.9% 90.1% 90% 90%	[70]
J48 Decision Tree ZeroR	Weka	Pathology Report	10 Attributes	Numeric	Continuous Values	Kappa statistics Mean absolute error	Confusion Matrix	Cross Validation	95.4% 1%	[91]
J48 Algorithm Adtree(Alternating Decision Tree) CART	Weka	India Cancer Institute Adyar Chennai	9 Attributes	Images	Discrete Values	Features Selection	Binary Classification Accuracy Method	Weighted average of parameters	98.10% 97.70% 98.50%	[92]
Naive Bayes J48 Decision Tree	Weka	LASUTH cancer Dataset, Nigeria	17 Attributes	Text	Continuous Values	Features Extraction	Confusion Matrix	Performance Evaluation Model	82.6% 94.2%	[93]
Naive Bayes Decision Tree K-Nearest Neigh-bor(KNN)	Weka	Wisconsin Hospital	9 Attributes	Numeric	Mixed Values	Features Selection	Confusion Matrix	Performance Evaluation Model	95% 94.99% 95.94%	[86]
Naive Bayes Classifier SMO Decision Tree KStar	Weka	University of medical center Institute of Oncol-ogy	10 Attributes	Numeric	Discrete Values	Kappa statistics Mean absolute error	Confusion Matrix	Performance Evaluation Model	71.67% 69.58% 75.52% 73.52%	[87]

TABLE 2. *(Continued.)* **Comparative review of machine learning techniques for breast cancer prediction.**

Algorithm	Tool	Data set	Number of Attributes	Data type	Pre Processing	Data Processing Method	Evaluation Method	Validation Technique	Accuracy	R#
J48 Naive Bayes MLP Logistic Regression Support Vector Machine K-Nearest Neighbor (KNN)	Weka	Wisconsin Breast Cancer (WBC) Data Center	9 Attributes	Numeric	Mixed Values	Features Selection	Confusion Matrix for comparatively analyzed algorithms	10 Fold Cross Validation	95.59% 96.79% 94.78% 96.79% 97.59% 95.19%	[88]
Artificial Neural Network Support Vector Machine	Weka	Wisconsin Hospital	11 Attributes	Numeric	Discrete Values	Common Featured Values Selection Method	Performance Matrix	Minimal Optimization (SMO) Lib SVM	95.4% 96.9%	[79]
Naive Bayes Random Forest Logistics Regression Multilayered Perception K-Nearest Neighbors	Weka	UCI Machine Learning Repository USA	10 Attributes	Numeric	Discrete Values	Kappa statistics	Binary Classification Accuracy Method	MCC and ROC area	71.6% 69.5% 68.8% 64.6% 72.37%	[94]
Support-Vector Machine(SVM) Algorithm Decision Tree Random Forest Algorithm	Weka Spark Weka Spark Weka Spark	University of California Irvine repository WDBC	254 samples	Numeric	Mixed Values	Features Selection	Gene Expression and DNA Methylation Techniques for dataset evaluation	10 Fold Cross Validation	98.03% 99.68% 95.09% 98.80% 96.07% 98.09%	[71]
Naive Bayes(NB) K-Nearest Neighbors (KNN) J48	Weka	Collected from Doctor and cancer Experts	61 Attributes	Numeric	Discrete Values	Features Selection	Confusion Matrix	10 Fold Cross Validation	98% 98% 97%	[72]
Simple CART RBF Network Naïve Bayes J48	Weka	Wisconsin Diagnostic Breast Cancer (WDBC)	32 Attributes	Images	Discrete Values	Features Selection And Extraction	Constrained Search Sequential Floating Forward Search (CSSFFS)	10 Fold Cross Validation	92.8% 93.6% 92.6% 92.9%	[84]
SVM (Linear kernel) SVM (RBF-Kernel) SVM (Polynomial Kernel) SVM (Sigmoid Kernel)	Weka	Wisconsin Diagnostic Breast Cancer (WDBC)	32 Attributes	Numeric	Mixed Values	Data Selection Recursive Features Elimination	Confusion Matrix	Predictive Model	99% 98% 87% 94%	[73]
Naive Bayes Bagging Algorithm Regression Algorithm SVM J48	Weka	Wisconsin Breast Cancer (WDBC)	11 Attributes	Numeric	Mixed Values	Features Selection	Performance Classifiers	Precision ROC area F-measure	72.7% 65.3% 60.03% 82.53% 79.8%	[85]
Naive Bayes Bayesian Logistic Regression Simple CART J48	Weka	Wisconsin Breast Cancer (WDBC)	10 Attributes	Numeric	Discrete Values	Kappa statistics Mean absolute error	Performance Matrix	Error Comparison	95.26% 65.42% 98.13% 97.27%	[82]
K-Nearest Neighbors(KNN) Support Vector Machine(SVM) Random Forest Naive Bayes	Weka	Wisconsin Breast Cancer (WDBC)	11 Attributes	Numeric	Discrete Values	Root mean and relative Squared error	Confusion Matrix	10 Fold Cross Validation	96.1% 97.9% 96% 92.6%	[2]

TABLE 3. Comparative Analysis of Major Machine Learning Techniques (on the basis of accuracy level).

Ref.	SVM	KNN	LR	NB	C4.5	MLP	CART	ANN	TB	J48	DT
[76]	93%	95.68%	90%	92.1%	N/A	N/A	N/A	N/A	N/A	N/A	N/A
[77]	99.4%	98.5%	N/A	98.3%	N/A	N/A	N/A	N/A	N/A	N/A	97.9%
[78]	95.7%	N/A	N/A	N/A	93.6%	94.7%	N/A	N/A	N/A	N/A	N/A
[89]	98.24%	95.61%	N/A	94.73%	N/A	99.12%	93.85%	N/A	N/A	N/A	N/A
[99]	97.13%	95.27%	N/A	95.99%	95.13%	N/A	N/A	N/A	N/A	N/A	N/A
[90]	N/A	N/A	84.5%	86.7%	N/A	86.5%	N/A	N/A	N/A	N/A	N/A
[92]	N/A	N/A	N/A	N/A	N/A	N/A	98.50%	N/A	97.70%	98.10%	N/A
[86]	N/A	94.99%	N/A	95.9%	N/A	N/A	N/A	N/A	N/A	N/A	95.99
[79]	96.99%	N/A	N/A	N/A	N/A	95.4%	N/A	N/A	N/A	N/A	N/A
[94]	N/A	72.37%	68.8%	71.6%	N/A	64.6%	N/A	N/A	N/A	N/A	N/A
[88]	97.59%	95.19%	96.79%	96.79%	N/A	94.78%	N/A	N/A	N/A	95.59%	N/A
[71]	99.68%	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	98.80%
[72]	N/A	N/A	98.80%	N/A	98.89%	N/A	N/A	N/A	N/A	98.5%	N/A
[82]	N/A	N/A	65.42%	95.26%	N/A	N/A	98.13%	N/A	N/A	97.27%	N/A
[2]	97.9%	96.1%	N/A	92.6%	N/A	N/A	N/A	N/A	N/A	N/A	N/A

TABLE 4. Comprehensive review of major machine techniques (in terms of breast cancer prediction).

Technique	Description	Benefits	Limitations	Year & Reference
Computer Aided Diagnoses System (CAD) for breast cancer prediction.	Comparative analysis of algorithms like K nearest neighbors, support vector machine, random forest and gradient boosting was performed.	Both classification and regression method random forest algorithm provided the highest accuracy,it is a mixture of many train models that provides the predictions about different training classifiers. Hybrid method was developed to performed the accurate computation on UCI online dataset that provide the mode accuracy results	Expected probabilities of occurrence and non-occurrence are calculated through K fold cross validation. Which is more expensive task. Data pre-processing stage took too much time, because it was converted raw data into the valuable form, also the number of patient that are already mentioned in a list were not be considered	2019 [100]
Performance Comparison of classification algorithm on Weka and Spark	Classification models support vector machine, decision tree and random forest was considered for the evaluation of tree types of data that consists of DM, GE and combination of both.	Support vector machine based on parallel computation, have strength to analyze the multiple data at same time, it provide the highest accuracy rate on two different tool Weka and spark Error rate and computation time of SVM is lower than the decision tree and random forest.	Gene Expression data collection is one of the difficult task. To achieve the good result of accuracy, precision and sensitivity of data, large number of samples was needed for computations.	2019 [71]
Non Linear machine learning algorithm comparison	Comparison of MLP with nonlinear machine learning techniques K Nearest Neighbor, CART, Naïve Bayes, Support Vector Machine.	When dataset are linearly separable it provides good accuracy level MLP is consists of different layers each layer perform one single task separately, so the computation of this algorithm was faster enough.	User was responsible to set the hidden layers for MLP algorithm. Setting some value sometimes provided under fitting and sometimes over fitting results.Without 10 fold cross validation, it is impossible to predict the accuracy rate from train data models.	2019 [89]
Comparison of SVM and ANN for breast cancer prediction	Evaluation of SVM and ANN was done through performance metrics such as accuracy, precision, recall and ROC area.	After comparison most suitable technique for the prediction of breast cancer was found SVM because the classes are separated through hyper line that provide the more accuracy result than ANN.	Expected probabilities of occurrence and non-occurrence are calculated through K fold cross validation. Which is more expensive task.	2019 [79]
Optimization of algorithms through Genetics programming technique.	Data was in the form of digitize images, features selection and extraction method were applied to get meaningful information.	Comparative analysis of different machine algorithm was performed after, selecting some feature through polynomial features operator. Extra tree classifier obtained the highest accuracy than other algorithms	It took too much time during the evaluation process and model training.GP algorithm was designed to solve the hyper parameter problem but this algorithm process time was too slow.	2019 [101]
Comparative analysis of Data Mining Classifier for cancer prediction and detection.	Classification algorithm random forest, bagging algorithm, random committee, simple CART and IBK was analyzed through k fold cross validation.	Random forest provided the highest accuracy during evaluation, this algorithm require less efforts. Random forest algorithm do not require the standardization and normalization of data also can handle nonlinear data more efficiently.	Separate model was designed to check that whether there is a tumor or not. This model took too much processing time. K fold cross validation technique are applied for n number of iteration, just to get the desire result, each iteration took too much time.	2019 [70]

pre-processing method, data evaluation method, validation method and accuracy level of each algorithm in different situations. Table 3 summarizes the accuracy level of some important machine learning techniques. Table 4 summarizes the advantages and disadvantages of some important research studies that have been reviewed. There are three major techniques for the prediction of breast cancer: Machine Learning Techniques, Ensemble Techniques and Deep Learning

TABLE 4. *(Continued.)* **Comprehensive review of major machine techniques (in terms of breast cancer prediction).**

Technique	Description	Benefits	Limitations	Year & Reference
Prediction of breast cancer using Naive Bayes, KNN and J48.	Dataset was divided into two parts one is training data and another one is testing data, 10 fold cross validation was applied for the evaluation algorithms.	Most suitable technique for the prediction of cancer dataset Classified the data according to the similarity of each instances Provide the good accuracy for both training data and testing data	Testing phase is slow and also take too much time Difficult to choose require K value. To predict about the new data K-nearest only find the nearest neighbor from training data.	2019 [72]
Recursive Features Selection for Breast Cancer Detection.	Analysis of SVM algorithm was performed on different kernels such as Linear, RBF, Polynomial and Sigmoid. Accuracy rate of SVM on linear kernel was higher than the others.	SVM linear kernel provide the highest accuracy while the selection of appropriate features for breast cancer prediction. Predicted model and features selection technique was designed for computation of large dataset, that provide the good accuracy.	Computation time was increases while the extraction of irrelevant features. Error rate of SVM linear kernel was higher than others, while the computation process was slower.	2019 [73]
Breast cancer diagnoses through classification techniques.	Comparison of Machine Learning techniques done through dimensionally reduction technique using Linear discriminant analysis LND.	Classification model was built through training dataset, this phase took consume too much time during preprocessing. Features selection and extraction help to identify the presences of tumor also improve the classification of benign and malignant patients. Features extraction decrease the data storages issue efficiently.	R programming language is used for the implementation, this language consist of lots of packages, processing is lesser than the other languages. Evaluation phase took too much time because of CFS, LDA, PCA method.	2019 [81]
Breast Cancer risk prediction and Diagnosis.	Performance metrics evaluate C4.5,SVM,NB and KNN in terms of their accuracy, precision and sensitivity. SVM provide the highest accuracy result than others	ROC curve provide the good evaluation of each algorithm. Prediction of correctively classified instances rate higher through SVM algorithm. Also this algorithm provided lower error rate value.	Processing time of SVM was 0.007 while KNN was 0.01s.Model was designed to train data for the evaluation of correctly and in correctively classify the instances that was difficult and complex task.	2018 [75]
Best machine learning for breast cancer prediction.	Dataset was divided into two .K fold validation technique was applied before features selection and extraction method. SVM provide the more accuracy for prediction	Predictive model was designed, SVM provided the 99.7% accuracy for benign class and 94.6% for malignant class. Turnaround time and error rate of SVM is lesser than other algorithm.	Good and appropriate selection of method is important for evaluation of machine learning algorithm Confusion matrix was design for expected class result, matrix correctively predict the instances but with prediction time was maximum	2018 [2]
Hyper parameter Optimization for Breast Cancer Prediction.	HPO technique through clustering method was used to get the best suitable prediction algorithm for Breast cancer	Hyper parameter through clustering method provided the highest accuracy. Hyper parameter handled both categorical and continues type of data more effectively.	selected features also provided some redundant data. BCOAP model consists of too many phases each phase took lost if time for the evaluation of breast cancer data	2018 [74]
Artificial Neural Network for breast cancer.	Back propagation algorithm used for prediction through ANN algorithm, Every hidden layer provided the different accuracy during evaluation	Multi-layered Neural network created weight arbitrary that provided the Mean Square Error whose rate is too less. Feed forward algorithm help to reduce the error through weight modification.	Require high processing and time for large number of data, that affect the overall accuracy of data To achieve the good accuracy, precision and sensitivity of data, large number of sample are needed for computations.	2018 [29]
Comparison of data mining for breast cancer classification.	Fusion classifier which was the combination of more than two classifier was designed to evaluate the algorithm on different data mining tool.	Single classification provided the highest accuracy than the fusion classification. WPBC, WBC, LBCD Dataset was provided the better level accuracy during the evaluation of different algorithm when the confusion Metrix was design.	Weka tool provided the best accuracy for WPBC and WBC dataset but the accuracy level was not good for LBCD dataset.	2017 [43]
Breast Cancer Prediction using Data Mining techniques.	Confusion matrix was designed for the comparative analysis of classification algorithm and clustering algorithm.	Classification algorithm C4.5 and SVM provided the better result than the other algorithm. EM was also founded the most appropriate clustering algorithm for breast cancer.	Finding the effect algorithm that predict the accruing and recurring of diseases is one of the most difficult task.	2017 [102]
Breast Cancer diagnoses and prognoses through machine algorithms.	Time build model was designed for the performance analysis of different classifiers	Evaluation of each classifier through confusion matrix shows that accuracy rate of SVM is higher than the other SVM provided the less error rate for prognoses of breast cancer	the process time of SVM was higher than the KNN algorithm but KNN was a lazy learner method that had not provided the good accuracy result	2016 [99]
Analysis of Breast cancer through Classification Algorithm	Performance of classification algorithm was analyzed in terms of their accuracy, sensitivity and precision.	Comparison of each classification algorithm was done through the evaluation of weighted average values. CART algorithm provided the better accuracy for prognoses of breast cancer with minimum time.	Model was design to comparatively analyzed the data mining decision tree algorithm J48,CART,ADtree.Evaluation phase took too much time.	2016 [92]
Data Mining classification techniques for risk prediction of breast cancer	Naive Bayes and J48 comparatively analyzed through performance Matrix. Accuracy rate of J48 algorithm was found to be higher than the Naive Bayes	Naive Bayes provided the less error rate while computations. Number of the attribute was increases while increases the sample size of data that provided the good accuracy results.	Expression rule was designed to show the best attributes for breast cancer prediction but the evaluation process was too complicated	2015 [93]

TABLE 5. Number of papers for major techniques.

Techniques	References	No. of Papers
Machine Learning Techniques	[24]–[35], [37]–[51]	27
Ensemble Techniques	[52]–[55]	4
Deep Learning Techniques	[56]–[61], [63], [64]	8

Techniques. Table 5 summarizes the number of papers that have been reviewed related to each major technique. The review of five different kinds of algorithms are provided including Non-Linear Algorithms, Ensemble Algorithms, Deep Learning Algorithms, combination of Linear and Non-Linear Algorithms and combination of Non-Linear and Ensemble Algorithms. Table 6 summarizes the number of papers based on type of algorithm that have been studied and analyzed for breast cancer predictions.

Each technique is suitable under different conditions and on different type of dataset, after the comparative analysis of these algorithms we came to know that machine learning algorithm SVM is the most suitable algorithm for prediction of breast cancer, different researcher [2], [74], [76], [79], [82], [85]–[89], [91], [92] has provided the analysis of prediction algorithms by using the dataset from Wisconsin Diagnostic Breast Cancer (WDBC), the analysis shows that each time the accuracy of SVM algorithm is higher than the other machine learning algorithms. Researcher Sara Al Ghunaim *et al.* provided the comparative result of machine learning algorithms on two different tools: Weka and Spark, the accuracies of these algorithms show that Support Vector Machine is most suitable method for prediction of cancer, the accuracy of SVM on Weka was 98.03% while on spark tool was 99.68% which is the highest accuracy from all the other machine learning algorithms [71]. For the prediction of breast cancer, author [60] used the deep learning techniques, dataset is collected from MIAS database, deep learning algorithms: CNN, SAE and SSAE were used for analysis but the highest accuracy was 98.9%. Different researchers [2], [52]–[54], [60], [63], [70]–[97], [99]–[101] provided the total 24 different algorithms that we have reviewed for breast cancer predictions. There are different methods of machine learning and deep learning that are currently being used for cancer prediction. The accuracies of these algorithms still vary for different datasets. Therefore, we are still looking for some advanced level models and techniques including deep learning and machine learning algorithms that can provide the best accuracy of predictions of breast cancer and can be generalized for any type of dataset being used.

For the prediction of breast cancer through machine learning and deep learning techniques the major challenge is the availability of datasets. Each algorithm requires a large amount of training data for its computational measurements, however many researchers are now putting their efforts to get the datasets of cancer patients in the form of medical images, these images contain the confidential information about the cancer patients and many of these datasets are open source and available in the form of raw images. To handle the issue of limited dataset, many researchers are now using the data augmentation schemes, that consist of some key features

TABLE 6. Number of papers for algorithm types.

Algorithms	References	No. of Papers
Non-Linear Algorithms	[2], [70]–[75]	7
Ensemble Algorithms	[53], [54], [76]–[80]	7
Linear and Non-Linear Algorithms	[81]–[85]	5
Non-Linear and Ensemble Algorithms	[86]–[94]	9
Deep Learning Algorithms	[60], [63], [95]–[98]	6

including cropping, filtering, rotating, cleaning etc. and this technique helps us to get more available dataset of patients.

IX. CONCLUSION

In this article we have reviewed different machine learning, deep learning and data mining algorithms for the prediction of breast cancer. Our main focus is to find out the most suitable algorithm that can predict the occurrences of breast cancer more effectively. The main purpose of this review is to highlight all the previous studies of machine learning algorithms that are being used for breast cancer prediction, this article provides the all necessary information to the beginners who want to analyze the machine learning algorithms to gain the base of deep learning. The review of this article is started from the types of breast cancer, fourteen research papers have been reviewed to get some knowledge about the major types, symptoms and causes of breast cancer. After that, the review of major machine learning techniques, ensemble techniques and deep learning techniques has been provided and these techniques deeply elaborate algorithms that are being used for the predictions of breast cancer. In the future work there are still some issues that needed to be solved. Researchers can solve the issue of limited available dataset by using some data augmentation techniques. The issue of inequality of positive and negative data should be considered by researchers as it can lead to biasness towards positive or negative prediction. Another important issue that needed to be solved is imbalanced number of breast cancer images against affected patches for correct diagnosis and prediction of breast cancer.

REFERENCES

[1] Y.-S. Sun, Z. Zhao, Z.-N. Yang, F. Xu, H.-J. Lu, Z.-Y. Zhu, W. Shi, J. Jiang, P.-P. Yao, and H.-P. Zhu, "Risk factors and preventions of breast cancer," *Int. J. Biol. Sci.*, vol. 13, no. 11, p. 1387, 2017.

[2] Y. Khourdifi and M. Bahaj, "Applying best machine learning algorithms for breast cancer prediction and classification," in *Proc. Int. Conf. Electron., Control, Optim. Comput. Sci. (ICECOCS)*, Dec. 2018, pp. 1–5.

[3] 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)*, Dec. 2018, pp. 1–4.

[4] F. K. Ahmad and N. Yusoff, "Classifying breast cancer types based on fine needle aspiration biopsy data using random forest classifier," in *Proc. 13th Int. Conf. Intellient Syst. Design Appl.*, Dec. 2013, pp. 121–125.

[5] 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.*, vol. 67, no. 6, pp. 1565–1572, Jun. 2020.

[6] A. R. Chaudhury, R. Iyer, K. K. Iychettira, and A. Sreedevi, "Diagnosis of invasive ductal carcinoma using image processing techniques," in *Proc. Int. Conf. Image Inf. Process.*, Nov. 2011, pp. 1–6.

[7] S. Pervez and H. Khan, "Infiltrating ductal carcinoma breast with central necrosis closely mimicking ductal carcinoma *in situ* (comedo type): A case series," *J. Med. Case Rep.*, vol. 1, no. 1, p. 83, Dec. 2007.
[8] D. L. Page, W. D. Dupont, L. W. Rogers, and M. Landenberger, "Intraductal carcinoma of the breast: Follow-up after biopsy only," *Cancers*, vol. 49, no. 4, pp. 751–758, 1982.
[9] A. B. Tuck, F. P. O'Malley, H. Singhal, and K. S. Tonkin, "Osteopontin and p53 expression are associated with tumor progression in a case of synchronous, bilateral, invasive mammary carcinomas," *Arch. Pathol. Lab. Med.*, vol. 121, no. 6, p. 578, 1997.
[10] B. Lee, K. Kim, J. Y. Choi, D. H. Suh, J. H. No, H.-Y. Lee, K.-Y. Eom, H. Kim, S. I. Hwang, H. J. Lee, and Y. B. Kim, "Efficacy of the multidisciplinary tumor board conference in gynecologic oncology: A prospective study," *Medicine*, vol. 96, no. 48, p. e8089, Dec. 2017.
[11] S. Masciari, N. Larsson, J. Senz, N. Boyd, P. Kaurah, M. J. Kandel, L. N. Harris, H. C. Pinheiro, A. Troussard, P. Miron, N. Tung, C. Oliveira, L. Collins, S. Schnitt, J. E. Garber, and D. Huntsman, "Germline E-cadherin mutations in familial lobular breast cancer," *J. Med. Genet.*, vol. 44, no. 11, pp. 726–731, Aug. 2007.
[12] 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.*, vol. 35, no. 1, pp. 39–43, Jul. 2000.
[13] A. Gradilone, G. Naso, C. Raimondi, E. Cortesi, O. Gandini, B. Vincenzi, R. Saltarelli, E. Chiapparino, F. Spremberg, M. Cristofanilli, L. Frati, A. M. Aglianò, and P. Gazzaniga, "Circulating tumor cells (CTCs) in metastatic breast cancer (MBC): Prognosis, drug resistance and phenotypic characterization," *Ann. Oncol.*, vol. 22, no. 1, pp. 86–92, Jan. 2011.
[14] F. M. Robertson, M. Bondy, W. Yang, H. Yamauchi, S. Wiggins, S. Kamrudin, S. Krishnamurthy, H. Le-Petross, L. Bidaut, A. N. Player, S. H. Barsky, W. A. Woodward, T. Buchholz, A. Lucci, N. Ueno, and M. Cristofanilli, "Inflammatory breast cancer: The disease, the biology, the treatment," *CA, Cancer J. Clin.*, vol. 60, no. 6, pp. 351–375, 2010.
[15] M. K. Gupta and P. Chandra, "A comprehensive survey of data mining," *Int. J. Inf. Technol.*, pp. 1–15, Feb. 2020, doi: 10.1007/s41870-020-00427-7.
[16] D. Delen, "Analysis of cancer data: A data mining approach," *Expert Syst.*, vol. 26, no. 1, pp. 100–112, Feb. 2009.
[17] M. Shahbaz, S. Faruq, M. Shaheen, and S. A. Masood, "Cancer diagnosis using data mining technology," *Life Sci. J.*, vol. 9, no. 1, pp. 308–313, 2012.
[18] A. Reddy, B. Soni, and S. Reddy, "Breast cancer detection by leveraging machine learning," *ICT Express*, 2020, doi: 10.1016/j.icte.2020.04.009.
[19] 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.*, vol. 8, no. 3, p. 1677, Dec. 2019.
[20] S. Eltalhi and H. Kutrani, "Breast cancer diagnosis and prediction using machine learning and data mining techniques: A review," *IOSR J. Dental Med. Sci.*, vol. 18, no. 4, pp. 85–94, Apr. 2019.
[21] I. H. Witten and E. Frank, "Data mining: Practical machine learning tools and techniques with Java implementations," *ACM SIGMOD Rec.*, vol. 31, no. 1, pp. 76–77, Mar. 2005.
[22] D. L. Olson and D. Delen, *Advanced Data Mining Techniques*. Springer, 2008.
[23] L. Li, Y. Wu, Y. Ou, Q. Li, Y. Zhou, and D. Chen, "Research on machine learning algorithms and feature extraction for time series," in *Proc. IEEE 28th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Oct. 2017, pp. 1–5.
[24] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
[25] L. Tuggener, M. Amirian, K. Rombach, S. Lorwald, A. Varlet, C. Westermann, and T. Stadelmann, "Automated machine learning in practice: State of the art and recent results," in *Proc. 6th Swiss Conf. Data Sci. (SDS)*, Jun. 2019, pp. 31–36.
[26] A. Dey, "Machine learning algorithms: A review," *Int. J. Comput. Sci. Inf. Technol.*, vol. 7, no. 3, pp. 1174–1179, 2016.
[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*, vol. 19, no. 1, 2019, Art. no. 48.
[28] Y. Uzun and G. Tezel, "Rule learning with machine learning algorithms and artificial neural networks," *J. Seljuk Univ. Natural Appl. Sci.*, vol. 1, no. 2, pp. 1–11, 2012.
[29] P. Singhal and S. Pareek, "Artificial neural network for prediction of breast cancer," in *Proc. 2nd Int. Conf. I-SMAC (IoT Social, Mobile, Anal. Cloud)(I-SMAC)*, 2018, pp. 464–468.
[30] Q. Dai, S.-H. Xu, and X. Li, "Parallel process neural networks and its application in the predication of sunspot number series," in *Proc. 5th Int. Conf. Natural Comput.*, vol. 1, 2009, pp. 237–241.
[31] W. K. Tsai, A. Parlos, and B. Fernandez, "ASDM—A novel neural network model based on sparse distributed memory," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 1990, pp. 771–776.
[32] H. Tran, "A survey of machine learning and data mining techniques used in multimedia system," Dept. Comput. Sci., Univ. Texas Dallas Richardson, Richardson, TX, USA, Tech. Rep., Sep. 2019.
[33] S. D. Borde and K. R. Joshi, "Enhanced signal detection slgorithm using trained neural network for cognitive radio receiver," *Int. J. Electr. Comput. Eng.*, vol. 9, no. 1, p. 323, Feb. 2019.
[34] C. Prasetyo, A. Kardiana, and R. Yuliwulandari, "Breast cancer diagnosis using artificial neural networks with extreme learning techniques," *Int. J. Adv. Res. Artif. Intell.*, vol. 3, no. 7, pp. 10–14, 2014.
[35] C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," *J. Educ. Res.*, vol. 96, no. 1, pp. 3–14, 2002.
[36] S. B. Imandoust and M. Bolandraftar, "Application of k-nearest neighbor (KNN) approach for predicting economic events: Theoretical background," *Int. J. Eng. Res. Appl.*, vol. 3, pp. 605–610, Sep. 2013.
[37] H. Sharma and S. Kumar, "A survey on decision tree algorithms of classification in data mining," *Int. J. Sci. Res.*, vol. 5, no. 4, pp. 2094–2097, 2016.
[38] A. M. Mahmood, M. Imran, N. Satuluri, M. R. Kuppa, and V. Rajesh, "An improved cart decision tree for datasets with irrelevant feature," in *Proc. Int. Conf. Swarm, Evol., Memetic Comput.* Berlin, Germany: Springer, 2011, pp. 539–549.
[39] E. Budiman, Haviluddin, A. H. Kridalaksana, M. Wati, and Purnawansyah, "Performance of decision tree C4.5 algorithm in student academic evaluation," in *Proc. Int. Conf. Comput. Sci. Technol.*, 2017, pp. 380–389.
[40] R. Pandya and J. Pandya, "C5.0 algorithm to improved decision tree with feature selection and reduced error pruning," *Int. J. Comput. Appl.*, vol. 117, no. 16, pp. 18–21, May 2015.
[41] Y.-Y. Song and L. Ying, "Decision tree methods: Applications for classification and prediction," *Shanghai Arch. Psychiatry*, vol. 27, no. 2, p. 130, 2015.
[42] W. Wu, S. Nagarajan, and Z. Chen, "Bayesian machine learning: EEGMEG signal processing measurements," *IEEE Signal Process. Mag.*, vol. 33, no. 1, pp. 14–36, Jan. 2015.
[43] 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, 2017, pp. 636–651.
[44] T. Evgeniou and M. Pontil, "Support vector machines: Theory and applications," in *Advanced Course on Artificial Intelligence*. Berlin, Germany: Springer, 2005, pp. 249–257.
[45] Y. Yang, J. Li, and Y. Yang, "The research of the fast SVM classifier method," in *Proc. 12th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process. (ICCWAMTIP)*, Dec. 2015, pp. 121–124.
[46] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[47] T. O. Ayodele, "Types of machine learning algorithms," *New Adv. Mach. Learn.*, vol. 3, pp. 19–48, Feb. 2010.
[48] Y. Li and H. Wu, "A clustering method based on K-means algorithm," *Phys. Procedia*, vol. 25, pp. 1104–1109, Jan. 2012.
[49] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Comput. Geosci.*, vol. 10, nos. 2–3, pp. 191–203, Jan. 1984.
[50] S. Patel, S. Sihmar, and A. Jatain, "A study of hierarchical clustering algorithms," in *Proc. 2nd Int. Conf. Comput. Sustain. Global Develop. (INDIACom)*, Mar. 2015, pp. 537–541.
[51] J. Zhang, X. Hong, S.-U. Guan, X. Zhao, H. Xin, and N. Xue, "Maximum Gaussian mixture model for classification," in *Proc. 8th Int. Conf. Inf. Technol. Med. Educ. (ITME)*, Dec. 2016, pp. 587–591.
[52] 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.*, vol. 177, pp. 89–112, Aug. 2019.
[53] M. Abdar and V. Makarenkov, "CWV-BANN-SVM ensemble learning classifier for an accurate diagnosis of breast cancer," *Measurement*, vol. 146, pp. 557–570, Nov. 2019.

[54] 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, pp. 236–251.

[55] M. S. Bala and G. R. Lakshmi, "Efficient ensemble classifiers for prediction of breast cancer," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 6, no. 3, pp. 1–5, 2016.

[56] M. Togacar and B. Ergen, "Deep learning approach for classification of breast cancer," in *Proc. Int. Conf. Artif. Intell. Data Process. (IDAP)*, Sep. 2018, pp. 1–5.

[57] Z. Hu, J. Tang, Z. Wang, K. Zhang, L. Zhang, and Q. Sun, "Deep learning for image-based cancer detection and diagnosis: A survey," *Pattern Recognit.*, vol. 83, pp. 134–149, Nov. 2018.

[58] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[59] M. Tiwari, R. Bharuka, P. Shah, and R. Lokare, "Breast cancer prediction using deep learning and machine learning techniques," SSRN, New York, NY, USA, Tech. Rep. 3558786, 2020.

[60] 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, 2018, pp. 159–186.

[61] K. Munir, H. Elahi, A. Ayub, F. Frezza, and A. Rizzi, "Cancer diagnosis using deep learning: A bibliographic review," *Cancers*, vol. 11, no. 9, p. 1235, Aug. 2019.

[62] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE Access*, vol. 7, pp. 19143–19165, 2019.

[63] R. Fakoor, F. Ladhak, A. Nazi, and M. Huber, "Using deep learning to enhance cancer diagnosis and classification," in *Proc. Int. Conf. Mach. Learn.*, New York, NY, USA, vol. 28, 2013, pp. 1–7.

[64] 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, 2020, pp. 322–333.

[65] 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.*, vol. 68, no. 6, pp. 394–424, Nov. 2018.

[66] 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*, vol. 45, pp. 1–5, Oct. 2019.

[67] A. Jemal, E. Ward, and M. J. Thun, "Recent trends in breast cancer incidence rates by age and tumor characteristics among U.S. women," *Breast Cancer Res.*, vol. 9, no. 3, p. R28, Jun. 2007.

[68] A. K. Dubey, U. Gupta, and S. Jain, "A survey on breast cancer scenario and prediction strategy," in *Proc. 3rd Int. Conf. Frontiers Intell. Comput., Theory Appl. (FICTA)*. Cham, Switzerland: Springer, 2014, pp. 367–375.

[69] 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.*, vol. 144, no. 2, p. 169e–177e, 2019.

[70] M. K. Keles, "Breast cancer prediction and detection using data mining classification algorithms: A comparative study," *Tehni ki Vjesnik*, vol. 26, no. 1, pp. 149–155, 2019.

[71] S. Alghunaim and H. H. Al-Baity, "On the scalability of machine-learning algorithms for breast cancer prediction in big data context," *IEEE Access*, vol. 7, pp. 91535–91546, 2019.

[72] S. K. Maliha, R. R. Ema, S. K. Ghosh, H. Ahmed, M. R. J. Mollick, and T. Islam, "Cancer disease prediction using naive Bayes,K-nearest neighbor and J48 algorithm," in *Proc. 10th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2019, pp. 1–7.

[73] 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.*, vol. 2019, pp. 1–19, Nov. 2019.

[74] 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.*, vol. 9, no. 12, pp. 268–273, 2018.

[75] 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)*, Oct. 2018, pp. 1–4.

[76] M. Rana, P. Chandorkar, A. Dsouza, and N. Kazi, "Breast cancer diagnosis and recurrence prediction using machine learning techniques," *Int. J. Res. Eng. Technol.*, vol. 4, no. 4, pp. 1163–2319, 2015.

[77] P. Israni, "Breast cancer diagnosis (BCD) model using machine learning," *Int. J. Innov. Technol. Exploring Eng.*, vol. 8, no. 10, pp. 4456–4463, Aug. 2019.

[78] L. G. Ahmad, A. T. Eshlaghy, A. Poorebrahimi, M. Ebrahimi, and A. R. Razavi, "Using three machine learning techniques for predicting breast cancer recurrence," *J. Health Med. Inform.*, vol. 4, no. 124, p. 3, 2013.

[79] 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)*, Apr. 2019, pp. 1–3.

[80] 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.*, vol. 132, pp. 123–131, Apr. 2020.

[81] D. A. Omondiagbe, S. Veeramani, and A. S. Sidhu, "Machine learning classification techniques for breast cancer diagnosis," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 495, Jun. 2019, Art. no. 012033.

[82] 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)*, Dec. 2018, pp. 1–4.

[83] M. J. Zaki and W. Meira, Jr., *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2019.

[84] S. Aruna and S. Rajagopalan, "A novel SVM based CSSFFS feature selection algorithm for detecting breast cancer," *Int. J. Comput. Appl.*, vol. 31, no. 8, pp. 1–7, 2011.

[85] B. Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1476–1482, Mar. 2014.

[86] C. Shah and A. G. Jivani, "Comparison of data mining classification algorithms for breast cancer prediction," in *Proc. 4th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2013, pp. 1–4.

[87] L. S. Jamil, "Data analysis based on data mining algorithms using weka workbench," *Int. J. Eng. Sci. Res. Technol.*, vol. 5, no. 8, pp. 262–267, 2016.

[88] G. R. Kumar, G. Ramachandra, and K. Nagamani, "An efficient prediction of breast cancer data using data mining techniques," *Int. J. Innov. Eng. Technol.*, vol. 2, no. 4, p. 139, 2013.

[89] A. A. Bataineh, "A comparative analysis of nonlinear machine learning algorithms for breast cancer detection," *Int. J. Mach. Learn. Comput.*, vol. 9, no. 3, pp. 248–254, Jun. 2019.

[90] A. Bellaachia and E. Guven, "Predicting breast cancer survivability using data mining techniques," in *Proc. SIAM Int. Conf. Data Mining*, vol. 58, 2006, pp. 10–110.

[91] J. Talukdar and S. K. Kalita, "Detection of breast cancer using data mining tool (weka)," *Int. J. Sci. Eng. Res.*, vol. 6, no. 11, p. 1124, 2015.

[92] B. Padmapriya and T. Velmurugan, "Classification algorithm based analysis of breast cancer data," *Int. J. Data Mining Techn. Appl.*, vol. 5, no. 1, pp. 43–49, Jun. 2016.

[93] K. Williams, P. A. Idowu, J. A. Balogun, and A. I. Oluwaranti, "Breast cancer risk prediction using data mining classification techniques," *Trans. Netw. Commun.*, vol. 3, no. 2, p. 1, Apr. 2015.

[94] 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)*, Sep. 2018, pp. 581–584.

[95] 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)*, Jan. 2019, pp. 343–346.

[96] Doreswamy and M. U. Salma, "BAT-ELM: A bio inspired model for prediction of breast cancer data," in *Proc. Int. Conf. Appl. Theor. Comput. Commun. Technol. (iCATccT)*, Oct. 2015, pp. 501–506.

[97] 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.*, vol. 9, no. 1, pp. 1–12, Dec. 2019.

[98] J. Zheng, D. Lin, Z. Gao, S. Wang, M. He, and J. Fan, "Deep learning assisted efficient AdaBoost algorithm for breast cancer detection and early diagnosis," *IEEE Access*, vol. 8, pp. 96946–96954, 2020.

[99] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," *Procedia Comput. Sci.*, vol. 83, pp. 1064–1069, Jan. 2016.
[100] M. S. Yarabarla, L. K. Ravi, and A. Sivasangari, "Breast cancer prediction via machine learning," in *Proc. 3rd Int. Conf. Trends Electron. Informat. (ICOEI)*, Apr. 2019, pp. 121–124.
[101] H. Dhahri, E. Al Maghayreh, A. Mahmood, W. Elkilani, and M. Faisal Nagi, "Automated breast cancer diagnosis based on machine learning algorithms," *J. Healthcare Eng.*, vol. 2019, pp. 1–11, Nov. 2019.
[102] U. Ojha and S. Goel, "A study on prediction of breast cancer recurrence using data mining techniques," in *Proc. 7th Int. Conf. Cloud Comput., Data Sci. Eng.-Confluence*, Jan. 2017, pp. 527–530.

NOREEN FATIMA received the B.E. degree in software engineering from COMSATS University Islamabad, Wah Campus, Pakistan, in 2017. She is currently pursuing the M.E. degree in software engineering with Chongqing University, China. Her research interests include data mining, machine learning, deep learning, and their application in biomedical and health informatics.

LI LIU received the Ph.D. degree in computer science from the Université Paris-Sud XI, in 2008. He has served as an Associate Professor with Lanzhou University, China, and a Senior Research Fellow with the School of Computing, National University of Singapore. He is currently an Associate Professor with Chongqing University. He aims to contribute in interdisciplinary research of computer science and human related disciplines. He has published widely in conferences and journals with more than 100 peer-reviewed publications. He has been the Principal Investigator of several funded projects from government and industry. His research interests include pattern recognition, data analysis, and their applications on human behaviors.

SHA HONG received the M.E. degree in electrical control and automatics from Chongqing University, in 1997, and the Ph.D. degree in mechanical engineering in 2001. He is currently an Associate Professor with Chongqing University. His research interests include image processing and e-commerce.

HAROON AHMED (Student Member, IEEE) received the B.E. degree in electronics engineering from COMSATS University Islamabad, Abbottabad Campus, Pakistan, in 2015, and the M.E. degree in electronics and communication engineering from Chongqing University, Chongqing, China, in 2019, where he is currently pursuing the Ph.D. degree in information and communication engineering. His research interests include antenna design, MIMO antennas and systems, machine learning, deep learning, and their application in 5G communications.