International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 06 | June 2021 www.irjet.net p-ISSN: 2395-0072

A Comprehensive Survey on Predictive Analysis of Breast Cancer

Mrs. Thilagavathi N[1], Samayly S[2], Priyadharshini A[3], Yogeshwari K[4]

[1]*Assistant Professor, Department of Information Technology, Sri Manakula Vinayagar Engineering College, Puducherry-605107*
[2,3,4]*UG Students, Department of Information Technology, Sri Manakula Vinayagar Engineering College, Puducherry-605107*

Abstract - *One of the foremost common types of cancer is breast cancer and early prediction and diagnosis avoid the rising number of deaths. There are several types of research about predicting the type of breast tumors. The main focus is to comparatively analyze different existing techniques to find out the most appropriate method which supports a large amount of dataset with a good and accurate prediction. Earlier methods were used to predict breast cancer diagnoses like data mining techniques, machine learning techniques, and the hybrid form of data mining and machine learning systems with a comparison of their accuracy. The proposed solution is based on a deep learning technique that uses a faster RNN algorithm that achieves a higher accuracy rate. It establishes a better and accurate classification model that makes use of invaluable information in clinical data and medical imaging which improves the performance.*

Key Words: **RCNN, Breast cancer prediction, Deep learning, Data mining, ML.**

1. INTRODUCTION

According to World Health Organization (WHO), twenty-five percent of females are diagnosed with breast cancer at some stage in their life. Breast cancer cells are a kind of tumor that can be seen on an x-ray[1]. It spreads when the cancer cells move into the lymph system or blood and are carried to other parts of the body. This results in changes and mutations in DNA which causes life loss. Breast cancer recurrence or recurrent breast cancer is the one that comes back after the initial treatment or the one that comes in the same or opposite breast when cancer couldn't be detected[2]. There are numerous reasons for the cause of death in breast cancer which depend on the specifications of the patient. In order to address this loss, there are different machine learning [3] and data mining algorithms[4],[5] that are being used for the prediction of breast cancer. Also there are several studies that have been conducted on breast cancer to accurately predicting its rate of occurrence. Prediction of breast cancer cells in the earlier stage can reduce the risk of death. It also has used various earlier techniques such as ultrasound, mammography [6], CT, and MRI for breast cancer. Systematic transrectal ultrasound (breast)-a guided biopsy is considered as the standard method for the definitive diagnosis of breast cancer. The current standard biopsy technique can miss up to 30% of cancers with significant sampling error. Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, machine learning, data mining algorithms[7], ensemble techniques[8], Genetic Algorithm[9] etc., to help health care pathologists with improved accuracy in the diagnosis and prediction of breast cancer. One of the important tasks in the prediction of breast cancer is finding the most appropriate and suitable algorithm [10]. In recent years, there is a significant increase in advancement in the technology of artificial intelligence, especially in deep learning. This study reinforces the importance of early diagnosis with high accuracy in women with breast cancer.

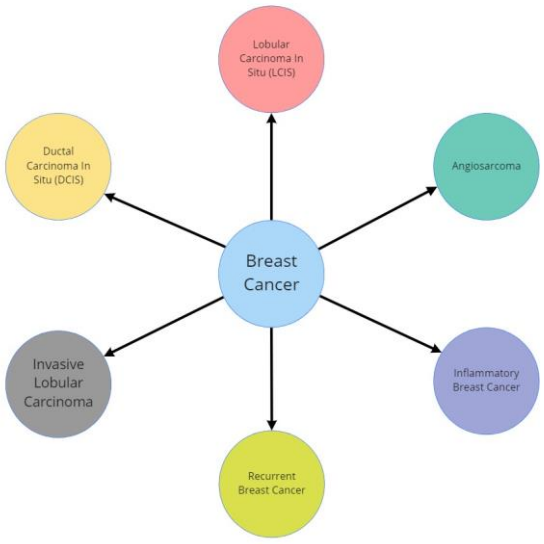

Figure1: TYPES OF BREAST CANCER

Machine learning [11] is one of the most promising breast cancer detection techniques that provides potential support for medical doctors to evaluate faster and more accurately the tumor type. Artificial intelligence and machine learning provide a cheap, easy, fast, and accurate method for tumor kind detection. It is an intelligent system provided with enough information so that it can assist medical studies in the classification and evaluation of the infection type.

Recurrent neural networks [12] are widely used nowadays in different scientific and medical fields. It presents a powerful tool that can work and learn in the same way human brains do. They are able to learn using examples and generalize patterns.. The development of digital processing technologies is increasing the power of the recurrent neural networks. Fast computers with enough memories can perform very complex recognition and classification tasks in an accurate and fast RNN.

Data mining is a technical process of finding patterns and repetitions in large datasets and by which those consistent patterns are identified, sorted, and organized. Data mining techniques[13] have been developed and applied in data mining projects, particularly classification, association, clustering, prediction, sequential models, and decision trees.

This work proposes the use of Recurrent neural networks in the classification of histo- pathological breast images. The classification will be based on the type of cancer tumor in the first stage into benign and malignant cancer [14]. The next step will be used to find out the type of the benign and malign cancer. For this reason, database of pathological images is collected and going to be used with our system. The database is divided into two parts, benign and malignant cancer. Each one of these is also divided into four sub divisions dependent on the type of tumor. Images were gathered using a microscope with different zooming settings.

Deep Learning is a subset of Artificial Intelligence [15]. It helps us to extract the information from the layers present in its architecture. It is used in various applications like Stock Analysis, Fraud Detection, Self-driving cars, Healthcare like cancer prediction and image analysis, etc. They are classified into Supervised, Semi-Supervised and Unsupervised learning types [16]. It helps us to improve the efficiency of predictions. Models can be trained on a large amount of dataset and the more data the model becomes better. Deep learning [17] is classified as stacked sparse auto encoder (SSAE), Convolutional neural network, auto encoders, sparsed auto encoder and so on[18]. Some of the pre-trained model are Alex Net, Google Net, ResNet and widely used dataset for training and testing are Mammogram image, SEER, UCI, WBCD[19].

2. MACHINE LEARNING ALGORITHMS FOR BREAST CANCER PREDICTION:

Machine learning is classified into three types:

- Supervised learning
- Unsupervised learning
- Reinforcement learning

2.1 SUPERVISED LEARNING:

A supervised learning [20] algorithm learns from labelled training data that helps to predict the outcomes for unforeseen data. Datasets are usually partitioned into training and testing phase. Until the model achieves a desired level of accuracy the training process continues on the training data. . Examples of Supervised Learning: Gaussian mixture algorithm, Regression, Decision Tree, Random Forest, KNN, Logistic Regression etc.

2.2 UNSUPERVISED LEARNING:

Unsupervised machine learning algorithms produce patterns from a dataset without processing to labeled or known outcomes. It is best used when we do not have data on the desired outcome. Apriori algorithm, K-means are some of the examples of Unsupervised Learning

2.3 REINFORCEMENT LEARNING:

Reinforcement learning (RL) is one of the type of machine learning in which intelligent agents shall take actions in an environment to increase the process of increasing reward. In reinforcement learning, the robot or agent learns from punishments or rewards based on error and trial. Therefore, it is a highly autonomous and effective learning eventhough the learning is generally very slow.

A. 2.1.1 SUPERVISED LEARNING TECHNIQUES:

GAUSSIAN MIXTURE ALGORITHM:

Gaussian mixture regressors (GMR) [21] were proposed by Ghahramani and Jordan. Supervised Learning Gaussian mixture model (SLGMM) is one the most famous algorithm that improves the recognition accuracy of the GMM. It is also

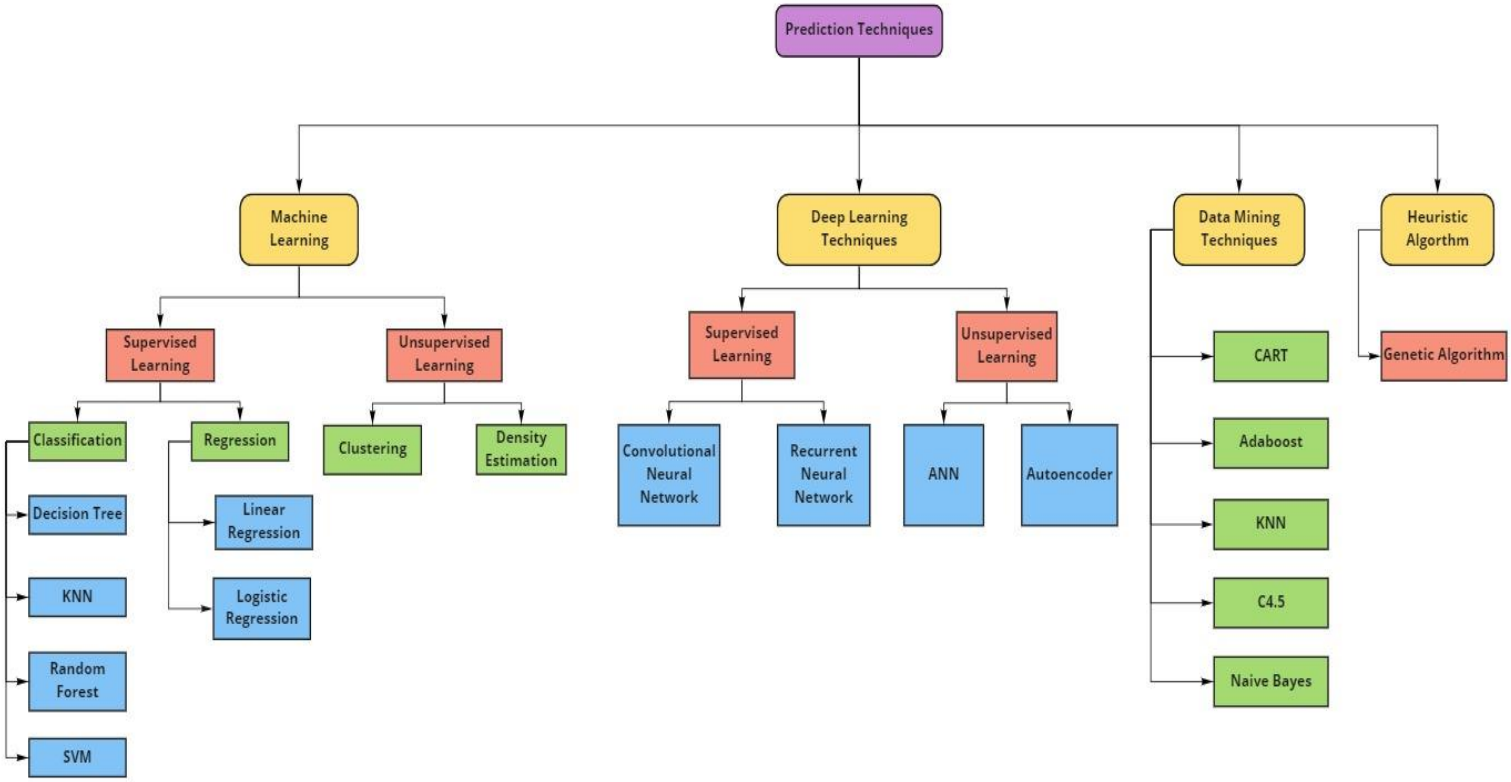

Figure2: CLASSIFICATION OF PREDICTION TECHNIQUES:

known as the soft clustering technique which is used t compute the probability of different types of clustered data. This algorithm is implemented based on expectatio maximization. In the earlier stage, GMR made use of CART or MARS as feature selection tools.

A weighted sum of M component densities is the Gaussian mixture density.

$$p(u|\lambda) = PM\ i = 1\ \alpha i\ bi(u),$$

Where, αi are the mixture weights (αi ≥ 0, i = 1, . . . , M, and PM i =1 αi = 1), and bi(u) are K-variate Gaussian densities with mean vector μi and covariance matrix Σi. The parameter list,

λ= {λ1. . . λM}, defines a particular Gaussian mixture density, where λi = {μi , Σi , αi}.

The advantage of Gaussian mixture regressors is performing model optimization and automatic feature selection. Maximum Gaussian mixture model-based classifier works well in all of the tested real-life datasets and continues to work well when the dimension of the data is high.

RANDOM FOREST:

Random forest is a supervised learning algorithm. It is also one of the widely used algorithms, because of its simplicity and diversity. The forest" it produce, is an ensemble [22] of decision trees, which is trained with the “bagging” technique. The major advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems [23].

DECISION TREE:

Decision trees are powerful classification algorithms that are becoming increasingly popular with the development of data mining in the field of information systems. Quinlan’s ID3, C4, C5, and CART are some of the popular decision tree algorithms [24]. They are considered as one of the easiest algorithms to understand by end-user in Data mining. It shows an effective suggestion among the dataset attributes and represents it in an easier way to understand.

SUPPORT VECTOR MACHINE:

Support Vector Machine [25] is one of the widely used supervised machine learning classification techniques that is applied in the field of cancer prediction, identification, and prognosis. SVM functions by choosing the critical samples from every category. These samples are called support vectors. By generating a linear function that divides them broadly as possible using these support vectors these classes are separated.

Linear kernels outperform polynomial and radial basis kernels with SVM. It achieves the highest AUC value of 0.9944[26] SVM shows as a very promising classifier for breast cancer prediction. SVM is broadly applied in many fields, such as digit recognition, cancer classification, handwritten recognition, face detection, time series forecasting, etc. In binary classification, the training set is classified into:

$$T = \{(ai\ bi), I = 1..,N, ai\ Rm, bi\ \{1,-1\}\}$$

Where ai denotes an M dimension feature vector of the ith case, and bi is a class identifier.

C MEAN ALGORITHM:

C means is a clustering algorithm that provides the division of data in the form of small clusters. The cluster that consists of similar data points belongs to one single class . In C mean algorithm each data point belongs to one single cluster. It is mostly used in medical images [27] segmentation and disease prediction.

LOGISTIC REGRESSION:

Logistic Regression is one of the supervised algorithms used for predicting discrete-valued outputs. Logistic Regression is an extension of the linear regression model for classification. Simple Logistic Regression yields deep predictions and obtains the best model yielding high and accurate results followed by other methods. It models the possibility for classification problems with two possible outcomes. It can also be extended from binary classification to multi-class classification. Then it is called Multinomial Regression. the logistic regression model uses the logistic function to squeeze the output of a linear equation between 0 and 1. The logisti function is defined as:

$$\text{logistic}(\eta) = \frac{1}{1 + \exp(-\eta)}$$

2.2.1 UNSUPERVISED LEARNING TECHNIQUES:

K-MEANS CLUSTERING:

The k-Means clustering algorithm [29] is a classical unsupervised learning method. Unlike supervised learning, clustering is an unsupervised learning method because there is not a need to compare the output of the clustering algorithm to the specific labels to evaluate its performance. It allows grouping the data according to the existing equivalence among them in k clusters that is given as input to the algorithm. The algorithm takes n observations as input and an integer m. The output is a division of the n observations into m sets such that each observation belongs to the cluster with the nearest mean and each data point belongs to only one group. The aim of this algorithm is to minimize the objective function that is know as the squared error function.

K-MEANS MODE:

This algorithm can deal with both continuous(numeric) and categorical data. A mode is a vector of elements that reduces the dissimilarities between the vector itself and each object of the data. Each and every cluster center is an array of means and modes for numeric and categorical characteristics respectively. The means and modes are calculated for each cluster and then each point is proceeded or moved to the cluster with minimum distance.

The Prediction of the disease in a cluster is determined by the number of datasets that have the disease divided by the total number of points in the clusters. Some other algorithms like k-Mode and k-Means [30] achieved accuracy rate only up to 65%.It results that the k-Means-Mode algorithm is better at clustering data than the other two (ie) k-Means and k-Mode algorithms.

3. HEURISTIC TECHNIQUE FOR BREAST CANCER PREDICTION:

GENETIC ALGORITHMS:

Genetic algorithms were invented by J. HOLLAND in the early 1970s. Genetic algorithms[31] are dynamic heuristic techniques that's major aim is to find the smallest set of

genes that ensure it provides a highly accurate classification of cancers cells from microarray data that is done by using a supervised machine learning algorithm. It also provides an effective method that leads to accurate cancer classification using expressions of only a very few genes. GA is proposed to solve the weight optimization problem. It generates high-quality optimization solutions and search problems that depend on mutation and crossover. The GA-based weighted average method excels to the classical weighted average method by an accuracy of 1.13%. Through this divide and conquer approach we have obtained an 84% of accuracy rate.

4. DATA MINING TECHNIQUES FOR BREAST CANCER PREDICTION:

CART (Classification And Regression Trees):

CART (Classification and Regression Trees) was developed by Breiman. CART is based on a binary-tree structure, in which each parent node is split into two children nodes according to a simple yes/no question about the value of a predictor variable. For attribute selection measures to build a decision tree, CART uses Gini Index. It is also based on Hunt's algorithm. Predictions are made with CART [33] by traversing the binary tree given a new input record. Both the categorical and continuous attributes are handled by CART to build a decision tree. It handles missing values. To remove the unreliable branches from the decision tree CART uses cost complexity pruning which improves the accuracy. For many important algorithms like random forest, bagged decision trees, and boosted decision trees CART acts as the foundation. The feature used to split the root node receives overall importance.[34]

K-NN(K-Nearest Neighbor):

K-Nearest Neighbor algorithm was formulated by Fix and Hodges. It is a lazy, non-parametric, machine learning algorithm that can be used to solve both classification and regression problems. The KNN algorithm [36] works by first using a programmed distance function to identify the nearest neighbors of a specific object. The shortest distance of the objects is calculated using overlapping attributes such as patient age, margin, mass shape or density. KNN captures the idea of similarity or closeness with some mathematical strategies to calculate the distance between points on a graph. It is simple, easy to implement and versatile.

ADABOOST:

AdaBoost (Adaptive Boosting)[35] is a very popular boosting technique that combines multiple "weak classifiers" into a single "strong classifier". To improve the overall learning algorithm performance, continuous adjustment of weight importance for each weak classifier model according to the training process should be carried out. Low error rate, performing well in the low noise data set. The advantage of this algorithm is that it requires fewer input parameters and needs little prior knowledge about the weak learner

$$St{+}1\ (i) = St\ (i)\ Zt \quad \{\ -\alpha t\ if\ ht\ (xi\ ) = yi$$

$$e\ \alpha t\ if\ ht\ (xi\ ) \neq yi$$

$$= St\ (i)\ \exp\ (-\alpha t yiht\ (\ ))$$

$$\_\_\_\_\_ \quad \{\ e\ -\alpha t\ if\ ht\ (xi\ ) = yi$$

$$Zt \qquad\qquad e\ \alpha t\ if\ ht\ (xi\ ) \neq yi$$

Where $St\ (i)$ is a distribution for $i\ th$ instance on the $t\ th$ iteration; ht is a base classifier, usually, it is a weak classifier; ϵt is a weighted error, which is calculated by the classification result; αt is an important factor that is assigned to ht based on weighted error

C4.5 Algorithm:

C4.5 [37] is one of the most important Data Mining algorithms that is used to generate a decision tree which is an expansion of ID3 (Iterative Dichotomiser 3) calculation. It enhances the ID3 algorithm by managing both continuous and discrete properties and missing values. The decision trees created by C4.5 that is used for grouping are referred to as a statistical classifier. As it is a supervised learning algorithm it requires a set of training examples so that it can produces a decision trees from a set of training data . It can act and work with both Continuous and Discrete Data. The major advantage of C4.5 over other decision tree systems is that it can handle the issue of incomplete data very well.

NAÏVE BAYES:

Naive Bayes[38] can work efficiently as a single algorithm but is not a single algorithm . The Naïve Bayes method is based on the famous Bayesian approach which is a clear, simple and fast classifier. It is a group of classification algorithms that is combined together that is provided with a

labeled training dataset to construct the tables. It is a model that is simple and easy to build, with no complicated iterative parameter estimation which makes it specifically useful for very large datasets. In spite of its simplicity, the Naive Bayesian classifier performs well and is widely used because it often outperforms more sophisticated classification methods. A naive Bayes classifier is a simple possibility classifier based on applying Bayes' theorem with naive independence assumptions.

5. DEEP LEARNING TECHNIQUES FOR BREAST CANCER PREDICTION:

AUTO ENCODER:

Autoencoder[39] belongs to the unsupervised learning class of neural networks. It is . The main objective of auto encoder is to learn or take in from a large dataset, by training its network that ignores the unwanted signals such as noise. They learn from a lower dimensionality feature representation input data. The training phase consist of two stages: coding and decoding. In the first stage, the input A is encoded by a representation J with some weight matrix YA,J and bias BA,J:

$$J=\sigma\ (YA,JA+BA)$$

Where σ is a sigmoid function also known activation function. An Auto encoder consists of three layers (ie) Encoder, Code and Decoder. The encoder layer encodes the input image as a compressed representation in a minimized dimension. It is then fed to Code layer which represents the compressed input which is fed to the decoder and the final layer Decoder decodes the encoded image back to the original dimension.

SPARSE AUTO ENCODERS:

Sparse Auto Encoders automatically learns features from unlabeled data. It is type of auto encoder that uses sparsity to achieve an information bottleneck. It is basically a feed-forward and backpropagation algorithm that consists of a regular auto encoder. A sparse auto encoder [40] can handl the sparsity regularize which provides the sparsity of output from the hidden layer of the neural network. Experience with different levels of sparsity shows an inverse relationship between the level of sparsity and the relationship nature that is captured in the training data. Stacking of n auto encoders into n hidden layers using unsupervised learning with the supervised method makes the basic structure of the stacked auto encoders (SAEs)[41]

RECURRENT NEURAL NETWORK:

Recurrent neural networks (RNN) are a type of neural networks that are useful in modeling sequence data. It is evolved from feedforward networks, RNNs exhibit similar behavior to how human brains function. The main feature of a neural network is known as a node. The node receives its signals from the other connected nodes or even from another sensor or source. RNN nodes are more dominant than other models in terms of prediction since these models uses backpropagation. Recurrent neural network are used with convolutional layers to extend the essential pixel neighborhood. The major advantage of RNN is that It can model with the sequencing of data so that every sample can be presumed to be dependent on previous ones. Since RNNs deal with sequential data, they are well used for the health informatics fields where large amounts of sequential data are available to process.[42]

ARTIFICIAL NEURAL NETWORK (ANN):

Artificial neural network (ANN) is a mathematical and computational model that consists of many processing elements. Processing units consist of inputs and outputs. ANN learns from the input and produce the desired output. Artificial neural networks have been successfully used in medical science It is a biological oriented network used to predict the breast cancer [43]. It is based on a supervised procedure and comprise three layers: input, hidden, and output .This algorithm is based on parallel processing, distributed memory, and network architecture. ANN models provide some advantages over regression-based models including its capacity to deal with noisy data.

6. LITERATURE SURVEY:

Several studies have been conducted on breast cancer prediction and diagnosis which resulted in different accuracy rates. They used multi techniques and algorithm for the prediction and analysis of breast cancer.

MACHINE LEARNING:

In [1] *Wang, D Zhang and Y.-H Huang (2018)* et al.focused on developing a technique which gives minimum error to increase accuracy. It used Four machine learning algorithm SVM, Logistic Regression, Random Forest and KNN which predict the breast cancer and the outcome of each technique have been compared in this paper using different datasets. It was executed within a simulation environment and conducted

in the JUPYTER platform for analysis of Breast cancer. It achieved an Accuracy of 94.4 % in the prediction of breast cancer.

In [2] Akbugday performed classification on Breast Cancer Wisconsin dataset by using two major machine learning algorithms KNN and SVM. This paper proposed a solution by using K-Nearest Neighbor, SVM for classification of breast cancer and used a WEKA tool to separate the trainings samples. It achieved overall of 93.85%. accuracy rate.

In [3], P.Ramachandran proposed a solution to build a cancer risk prediction system. It is studied that the approach used data mining techniques such as classification, clustering and prediction to identify potential cancer patients. The gathered data is preprocessed, fed into the database and classified to yield significant patterns using decision tree algorithm. K-means clustering algorithm is used to divide and separate cancer and non-cancer patient datasets. It used Limitations include cost constraints.It provides an accuracy rate of 95%.

In [4],*Ahmad LG et* proposed on the basis of predicting with least error rate and highest accuracy when compared with the accuracy of the decision tree model. The support vector machine have been built using polynomial kernel. The performances of the other models have been evaluated using statistical measures, gain and Roc charts. Support vector machine model outperformed the other models on the prediction of the severity of breast masses. Limitations include important variables such as DNA index were not included because of their unavailability. It provides a accuracy rate of 92%.

DEEP LEARNING:

In [5] Dayong Wang, Aditya Khosla presented a deep learning-based system for the automated detection of metastatic cancer from whole slide images of sentinel lymph nodes This method significantly improves the accuracy of pathological diagnoses. It uses a state-of-the-art deep learning model and careful design of post-processing methods for the slide-based classification and lesion-based detection tasks. It uses the Camelyon16 dataset which consists of a total of 400 whole slide images (WSIs) split into 270 for training and 130 for testing. The paper comprises two major phases namely slide-based evaluation and Lesion-based evaluation which results in identifying metastatic breast cancer. It reduces the error rate and accurately provides the result.

In [6] Ainuddin Wahid, Ghulam Murtaza, Liyana Shuiband Abdul Wahab focused on breast cancer classification by using medical imaging modalities through state-of-the-art deep neural network approaches. It is expected to maximize the procedural decision analysis in five aspects, such as types of imaging modalities, datasets and their categories, pre-processing techniques, types of deep neural network, and performance metrics used for breast cancer classification. It achieves an accuracy rate upto 90%.

In [7] Khuriwal. Proposed a paper on the study of diagnosing and detecting Breast cancer in the early stage. It proposed a solution for an accurate and efficient diagnosis system over a short period of time that processes a large amount of data. It describes all algorithms and techniques for image pre-processing and post progressing. NAÏVE BAYES AND SVM algorithms were used for classifying and processing. It used image processing methods for segmentation and filter images for breast cancer diagnosis. The prediction was done by using the MIAS dataset and many filter methods were for image quality improvement. It achieved an accuracy rate of 92%.

In [8] Khuriwal ,N & Mishra, N proposed a Deep Leaning algorithm neural network for diagnosing breast cancer. They used only 12 features for the diagnosis of cancer It made use of the Wisconsin Breast Cancer dataset. It describes the use of deep learning technology for the diagnosis of breast cancer using the UCI Dataset. Because almost all deep learning techniques used for large datasets, Medical Diagnosis and NLP. It focuses on dividing the process into three parts like dataset gathering, Pre-processing algorithm for filter data, and generating the model. It uses pre-processing algorithms like Label Encoder, Normalizer, and Standard Scaler for the scaled datasets. It resulted in 97% accuracy rate.

In [9] Y. J. Tan, K. S. Sim, and F. F. Ting aimed to speed up the diagnosis process by assisting specialists to diagnose and classification breast cancer. It uses a Convolutional Neural Network algorithm, a deep learning technique. The paper proposes breast cancer detection by using CNN Network which has been successfully developed and tested with 322 mammogram images. This method provides a fast diagnosis time and high accuracy system.

DATA MINING:

In [10] Vikas Chaurasia , Saurabh Pa.This presented a paper on the diagnosis system for detecting breast cancer survivability. It uses three popular data mining methods like RepTree, RBF Network and Simple Logistic. These techniques were proposed to obtain fast automatic

diagnostic systems for other diseases.It demonstrated that the Simple Logistic can be used for reducing the dimension of feature space. In addition 10-fold cross validation method was applied for testing phase. It resulted in accuracy of 74.47%

In [11], Haifeng Wang and Sang Won Yoon identified an accurate model for the prediction of breast cancer based on various clinical records. Support vector machine (SVM), artificial neural network (ANN), Naive Bayes classifier, and AdaBoost tree are the four data mining models are applied in this paper. In order to reduce the feature space, the principal component analysis (PCA) method is used. Two widely used datasets Wisconsin Breast Cancer Database (1991) and Wisconsin Diagnostic Breast Cancer (1995)are used for analysis and Prediction of the breast cancer. A 10-fold cross-validation method is implemented to estimate the test error of each and every model. It produces an accuracy rate of 94.82%. The limitation is that since PCA is a linear method some techniques cannot be applied.

In [12] Pallavi Mirajkar1, Dr. G. Prasanna Lakshmi. Proposed an integrated system that is based on combination of various data mining techniques such as analytical hierarchy process, rule based association, classification etc. that is usefull in predicting the patient's disease status. For perfect prediction of cancer disease, the outputs of each algorithm is integrated and compared. The proposed algorithm learns the probability of an object with certain features belonging to a particular group/class. As the user enters into the cancer prediction system, they have to answer the questions, Then the prediction system calculate the risk score. Based on the predicted risk values the range of risk will be assigned. The result can be shown to the user through data base.

7. DISCUSSION:

This research summarizes different machine learning, deep learning and data mining algorithms for the prediction of breast cancer. Breast cancer survival time prediction studies based on ML models occupy a significant part of the contemporary research in this area but their accuracy rate and performance is low when compared to deep learning techniques.

Table 1 provides comparative summary of different machine learning, deep learning and data mining algorithms for breast cancer prediction on the basis of datasets, accuracy level of each algorithm in different situations and the advantages and disadvantages of some important research studies.

Table 1: COMPARISON ON ANALYSIS OF DIFFERENT TECHNIQUES OF BREAST CANCER:

SNO	TITLE	DATASET USED	TECHNIQUE USED	ADVANTAGES	LIMITATION	ACCURACY
1.	Wang et al. *"Breast Cancer Prediction Using Machine Learning" (2018)*	Electronic health records	Logistic regression	5-year survivability prediction using logistic regression	If the number of observations is lesser than the number of features, Logistic Regression should not be used, otherwise, it may lead to overfitting.	93%
2.	Akbugday. *"Classification of Breast Cancer Data Using Machine Learning Algorithms," 2019*	Breast Cancer Wisconsin dataset	KNN and SVM	Optimal k-Value for a k-NN classifier, g kNN is a lightweight, lazy learning algorithm with very short build times.	Accuracy depends on the quality of the data. With large data, the prediction stage might be slow. Sensitive to the scale of the data and irrelevant features.	94%
3.	KELES M. Kaya. *"Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms (2019)*	Wisconsin Dataset	RANDOM FOREST	Each dataset is generated with displacement from the original dataset. Then, trees are developed using a random selection feature, but are not pruned.	It requires much computational power as well as resources as it builds numerous trees to combine their outputs.	94%
4.	Khuriwal *"A Review on Breast Cancer Diagnosis in Mammography Images Using Deep Learning Techniques",(2018)*	Haberman's Survival dataset	NAÏVE BAYES AND SVM	Helps in marginalizing the hyper-parameters and differentiating classes.	It is used only for Bi-classification. It does not support multi-class classification.	92%

5.	Naresh Khuriwal *" Breast Cancer Diagnosis Using Deep Learning Algorithm"*	UCI	CNN	Efficient at Delivering High-quality Results.	Limitation in choosing the size of image. Costly implementation	95%
6.	Shravya *"Prediction of Breast Cancer Using Supervised Machine Learning Techniques"* *(2019)*	UCI repository	SVM	Hyperplane separates two classes which helps in higher accuracy.	SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation	94%
7.	Haifeng Wang and Sang Won Yoon. *"Breast Cancer Prediction Using Data Mining Method" (2015)*	Wisconsin Breast Cancer Database (1991) and Wisconsin Diagnostic Breast Cancer	Support vector machine (SVM), artificial neural network (ANN), Naive Bayes classifier and AdaBoost.	10-fold cross-validation method is implemented to estimate the test error of each and every model.	The limitation is that since PCA is a linear method some techniques cannot be applied.	94.82%.
8.	Chauhan, P., & Swami, A. *"Breast Cancer Prediction Using Genetic Algorithm Based Ensemble Approach* (2018).	UCI machinery	Ensemble, Genetic Algorithm	Ensemble method overcame the limitations of the classical weighted average method. It is very cost-effective for breast cancer prediction	Genetic algorithms do not scale well with complexity.	85%
9.	*"Breast Cancer Prediction with K-Nearest Neighbor Algorithm using Different Distance Measurements "(2018)* by Victoria Rodriguez ,Karan Sharma and Dana Walker.	UCI	KNN	Manhattan distance measurement providesaccuracy, precision, recall, and specificity.	It includes only six measurements in distance function	81.67%.

10	Abdelghani Bellaachia, Erhan Guven, *"Predicting Breast Cancer Survivability Using Data Mining Techniques"*	SEER database	Naïve Bayes	C4.5 algorithm has a much better performance than the other two techniques.	Prediction of survivability is not accurate.	85%
11	Tanzila Saba. *"Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges (2020)"*	FFDM dataset,WBCD dataset,PH2 dataset and BRATS2016 dataset	CNN and SVM	The review has presented six types of cancers Additionally, presented four significant stages of automated cancer diagnosis using benchmark datasets.	Accuracy for each cancer category is far from maturity.	-
12.	Kibeom et. *al "Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm,2017 "*	Gene Expression DatasetCollection	C4.5, Bagging and Adaboost Decision trees	Ensemble Method helps to combine multiple learners.	Genetic algorithms do not scale well with complexity.	92%
13	Ghulam Murtaza, Liyana Shuib, Ainuddin Wahid Abdul Wahab *"Deep Learning-based Breast Cancer Classification through Medical Imaging Modalities: State of the Art and Research Challenges" (2019)*	Wisconsin Dataset	Medical Imaging Modalities	DBT shows a significantly higher rate of screen-detected cancer compared with DM screening	Micro-classification are very small,isolated with various sizes, shapes, dispersed, looks similar to their surroundings; thus, they cannot be identified in mammograms from high-frequency noise	93%
14.	Joseph A. Cruz and David S. Wishart *"Applications of Machine Learning in cancer prediction and prognosis Cancer informatics"*	Pubmed (biomedical literature), the Science Citation Index	SVM, NAÏVEBAYES	Helps to form a decision boundary and helps in classification.	It does not support multi-class classification	94.3%

15	P.Ramachandran *"novel multi layered method combining clustering and decision tree techniques (2014)"*	Wisconsin Dataset	Clustering and Decision tree.	Helps doctors or patient to decide in a short time whether the person is suffering from disease and is generic to all types of disease.	It doesn't perform well when we have large data set because **the** required training time is higher. High cost	94%
16	Pallavi Mirajkar1, Dr. G. Prasanna Lakshmi *"An Integrated Cancer Prediction System Using Data Mining Techniques (2018)"*	UCI	SVM and bootstrap	The prediction system calculate the risk score which increases the performance.	The SVM assumed the independence of the predictor variables.	92%
17	"Prediction of Early Breast Cancer Metastasis from DNA Microarray Data Using High-Dimensional Cox Regression Models"	Desmedt's dataset.van de Vijver's dataset	High-Dimensional Cox Regression Models"	Due to large receptive field, it can extract overall observations of the objects in an image and captures more semantic level information.	The use of simple hashing method cannot provide rich enough information to map the features. Hence effects the representation performance.	89%
18	*"A Review on a Deep Learning Perspective in Brain Cancer Classification (2019) "* by Gopal S Tandel, M. Biswas,	Brats datset	Deep Learning Neural Network	Number of hidden layers	Requires high processing and time for large number of data which affects theoverall accuracy of data	94%
19	SA Medjahed, TA Saadi, A Benyettou *"Breast cancer diagnosis by using k-nearest neighbor with different distances and classification rules"* International Journal of Computer Applications, 2013	Wisconsin breast cancer dataset	Decision Trees	Helps in splitting 96.1 %	It gives low prediction accuracy for a dataset as compared to other machine learning algorithms	79%

8. CONCLUSION

In this article we appraised different machine learning, deep learning and data mining algorithms for the prediction of breast cancer. Our major focus is to find out the most suitable algorithm that can predict the occurrences and types of breast cancer more effectively. Deep learning (DNN) models accept lots of data in different formats. It is a great tool to be used in cancer prognosis prediction since patient's health data contain multi-source data. Existing solutions has not provided an accurate result in prediction of type of breast cancer. Our system uses Recurrent Neural Network (RNN) algorithm along with LSTM and DNN as feature extractor which resulted at high accuracy rate in prediction the type of cancer.

REFERENCES

1 Wang, D Zhang and Y.-H Huang et al ," Breast Cancer Prediction Using Machine Learning " (2018).

2. Akbugday ,"Classification of Breast Cancer Data Using Machine Learning Algorithms," (2019).

3 Y. Khourdifi and M. Bahaj, "Applying best machine learning algorithms for breast cancer prediction and classification,'(2018)

4 Ahmad LG et al applied the "SVM classification model to predict breast cancer recurrence (2016)".

5 . Dayong Wang, Aditya Khosla, "Deep Learning for Identifying Metastatic Breast Cancer(2016)"

6 Khuriwal ,"A Review on Breast Cancer Diagnosis in Mammography Images Using Deep Learning Techniques", (2018).

7 Y. J. Tan, K. S. Sim, and F. F. Ting "Breast Cancer detection Using Convolutional Neural Networks for Mammogram Imaging System (2017)"

8 "Deep Learning-based Breast Cancer Classification through Medical Imaging Modalities: State of the Art and Research Challenges", (2019) by Ainuddin Wahid, Ghulam Murtaza, Liyana Shuib, Abdul Wahab.

9 Vikas Chaurasia , Saurabh Pa ,"Data mining techniques: To predict and resolve breast cancer survivability" (2014)

10 Pallavi Mirajkar1, Dr. G. Prasanna Lakshmi "An Integrated Cancer Prediction System Using Data Mining Techniques (2018)"

11 "Breast Cancer Prediction Using Data Mining Method" (2015) by Haifeng Wang and Sang Won Yoon.

12. Dayong Wang, Aditya Khosla, "Deep Learning for Identifying Metastatic Breast Cancer(2016)"

13. Y. Khourdifi and M. Bahaj, "Applying best machine learning algorithms for breast cancer prediction and classification,'(2018)

14. V. Chaurasia and S. Pal, "Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability", 2014.

15. Abdelghani Bellaachia, Erhan Guven, "Predicting Breast Cancer Survivability Using Data Mining Techniques"

16. Khuriwal ,"A Review on Breast Cancer Diagnosis in Mammography Images Using Deep Learning Techniques", (2018).

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

18. N. Khuriwal and N. Mishra, "Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm," 2018

19. Chauhan, P., & Swami, A,"Breast Cancer Prediction Using Genetic Algorithm Based Ensemble Approach." (2018).

20 L. G. Ahmad, A. T. Eshlaghy, A. Poorebrahimi, M. Ebrahimi, and A. R. Razavi, "Using three machine learning techniques for predicting breast cancer recurrence,'(2013).

21. Akbugday ,"Classification of Breast Cancer Data Using Machine Learning Algorithms," (2019).

22 C. Prasetyo, A. Kardiana, and R. Yuliwulandari, "Breast cancer diagnosis using artificial neural networks with extreme learning techniques", 2014

23. Erhan Guven, Abdelghani Bellaachia, "Predicting Breast Cancer and analysis Using Data Mining Techniques" (2016)

24 . H. Witten and E. Frank, "Data mining: Practical machine learning tools and techniques with Java implementations, Mar. 2005

25. Ch. Shravya, K. Pravalika, Shaik Subhani, "Prediction of Breast Cancer Using Supervised Machine Learning Techniques", April 2019.

26. G. Hamed, M. A. E.-R. Marey, S. E.-S. Amin, and M. F. Tolba, "Deep learning in breast cancer detection and classification.

27."A semi-supervised deep learning method based on stacked sparse auto-encoder for cancer prediction using RNA-seq data ", JunWu[b] ,ZongliL, [c]XiaodongZhao[d] (2018)

28. "Breast Cancer Prediction Using Deep Learning and Machine Learning Techniques", Monika Tiwari Rashi Bharuka.

29. T. O. Ayodele, ''Types of machine learning algorithms,'' New Adv. MachLearn., vol. 3, pp. 19–48, Feb. 2010.

30. J. Zhang, X. Hong, S.-U. Guan, X. Zhao, H. Xin, and N. Xue, ''Maximum Gaussian mixture model for classification'', Dec. 2016.

31. M. S. Bala and G. R. Lakshmi, ''Efficient ensemble classifiers for prediction of breast cancer'', 2016

32. F. K. Ahmad and N. Yusoff, ''Classifying breast cancer types based on fine needle aspiration biopsy data using random forest classifier,''Dec.2013.

33. 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, 2018.

34 T. Evgeniou and M. Pontil, ''Support vector machines: Theory andapplications,'' in Advanced Course on Artificial Intelligence. Berlin,Germany: Springer, 2005, pp. 249–257.

35. L. G. Ahmad, A. T. Eshlaghy, A. Poorebrahimi, M. Ebrahimi, andA. R. Razavi, ''Using three machine learning techniques for predictingbreast cancer recurrence,'' 2013.

36. 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.

37. Y. Li and H. Wu, ''A clustering method based on K-means algorithm,'' Jan. 201

38. 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.,Mar. 2014.

39. Chauhan, P., & Swami, A,"Breast Cancer Prediction Using Genetic Algorithm Based Ensemble Approach." (2018).

40.Vikas Chaurasia , Saurabh Pa ,"Data mining techniques: To predict and resolve breast cancer survivability" (2014)

41. 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

42.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.

43. 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.

44. Victoria Rodriguez ,Karan Sharma and Dana Walker, "Breast Cancer Prediction with K-Nearest Neighbor Algorithm using Different Distance Measurements (2018)"

45. R. Pandya and J. Pandya, ''C5.0 algorithm to improved decision tree with feature selection and reduced error pruning,'' Int. J. Comput. Appl.,May 2015.

46. Breast Cancer Prediction using Naïve Bayes Classifier Megha Rathi* , Arun Kumar Singh*

47.Y. LeCun, Y. Bengio, and G. Hinton, ''Deep learning'', 2015.

48. R. Fakoor, F. Ladhak, A. Nazi, and M. Huber, ''Using deep learning to enhance cancer diagnosis and classification,'' 2013.

49. 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.

50.G. Hamed, M. A. E.-R. Marey, S. E.-S. Amin, and M. F.Tolba, ''Deep learning in breast cancer detection and classification,''2015