International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 02 | Feb 2021 www.irjet.net p-ISSN: 2395-0072

Breast Cancer Detection using Machine Learning Algorithm

Hiba Masood[1]

[1] *Master of Engineering student, Department of Computer Science and Technology, V.V.P.I.E.T. Solapur, Maharashtra, India.*

Abstract - Breast cancer is one of the most widely spreading diseases and the second leading cause of cancer death among women. [3].The survival rate increases on detecting breast cancer early as better treatment can be provided. Data classification using machine learning has been widely used in the diagnosis of breast cancer and for early detection of breast cancer. The aim of this literature review is to focus on the use of machine learning in classification of available data in breast cancer early detection and diagnosis. On reviewing several papers of artificial intelligence it is apparent that there are different techniques available for cancer detection. The objective of this study is to summarize various review and technical articles on diagnosis and prognosis of breast cancer. It gives an overview of the current research being carried out on various breast cancer datasets using the data mining techniques to enhance the breast cancer diagnosis and prognosis.

Key Words: breast cancer, machine learning, artificial neural networks, decision tree, support vector machine-nearest neighbour, healthcare system, Wisconsin breast cancer database

1. INTRODUCTION

Breast cancer (BC) is one of the most common malignancies in women. Early diagnosis of Breast Cancer and metastasis among the patients based on an accurate system can increase survival of the patients to >86%. Breast cancer starts when malignant lumps which are cancerous begin to grow from the breast cells. Doctors may wrongly diagnose benign tumour (which is noncancerous) as malignant tumour. There is need for a computer aided detection (CAD) systems which uses machine learning approach to provide accurate diagnosis of breast cancer. [8]These computers aided detection(CAD) systems can aid in detecting breast cancer at an early stage. When, breast cancer is detected early enough, the survival rate increases because better treatment can be provided [3]. Early detection of cancer is critical to improve breast cancer survival and to reduce the high mortality rate of Breast Cancer. Despite early detection and the advent of new treatments, about 50% of patients will develop distant metastases during their follow-up time. According to WHO, India has approximate 1.5 million patient diagnosed of breast cancer every year, in year 2015 alone500,000women are estimated to die because of this cancer[5],and roughly 9.6 million deaths in 2018 [10]. This gap between incidence and mortality is wide, which means performance improvement is needed on early breast cancer detection. Hence, improvements in existing techniques are required to predict breast cancer at an early stage.
There are two type of breast cancer as follows:

1.1 Benign Tumors: Noncancerous

If the cells are not cancerous, the tumor is benign. It won't invade nearby tissues or spread to other areas of the body (metastasize). A benign tumor is less worrisome unless it is pressing on nearby tissues, nerves, or blood vessels and causing damage. Fibroids in the uterus or lymphomas are examples of benign tumors. Benign tumors may need to be removed by surgery. They can grow very large, sometimes weighing pounds. They can press on vital organs or block channels. Some types of benign tumors such as intestinal polyps are considered precancerous and are removed to prevent them becoming malignant. Benign tumors usually don't recur once removed, but if they do it is usually in the same place. [24]

1.2 Malignant Tumors: Cancerous

Malignant means that the tumor is made of cancer cells and it can invade nearby tissues. Some cancer cells can move into the bloodstream or lymph nodes, where they can spread to other tissues within the body—this is called metastasis. Cancer can occur anywhere in the body including the breast, intestines, lungs, reproductive organs, blood, and skin. Once breast cancer has spread to the lymph nodes, the cancer cells can travel to other areas of the body, like the liver or bones. The breast cancer cells can then form tumors in those locations. A biopsy of these tumors might show characteristics of the original breast cancer tumor. [23]

1.3 Machine Learning

Patient's medical records have a large amount of data such as doctor's examination records, nursing notes, imaging studies, laboratory results, medications and progression notes. Moreover, the data are collected from different sources and the data types cannot be managed and processed easily without computer aided systems.

The basis of Machine learning is it enables the Systems to learn themselves automatically and to improve their performance through experience without any instructions by programmer. The primary objective is to evaluate the performance in classifying data with respect to efficiency and effectiveness of each algorithm in terms of classification, test accuracy, precision, and recall. There is lot of information and data available, which gives opportunity for analyzing processes, to perform research in classification and in data mining fields, to test tools of machine learning and carry on experiments for tuning main methods of supervised learning.

Machine learning may be categorized as follows.

1.3.1 Supervised Learning:

Here the input is a labelled set of training data given to the program to learn. The program has to find the group according to the labelled input. Supervised learning is so named because the programme teaches the learning process about what results should be obtained from the training data, as a guide to the algorithms. Supervised techniques processed the datasets based on its previously known output. If the output is continuous, the regression algorithms are considered the best choice. Predicting the occurrence of cancer is not the only interest of the researchers. However, predicting the surviving probability take an important place in the health field. [13, 15].

1.3.2 Unsupervised Learning:

Here there are no label and the learner need to find the pattern and discover the group. These approaches create clusters from raw, unlabelled or unclassified data. These clusters can be used later to develop classification schemes or classifiers. Unsupervised learning, all the data is unlabelled. It should be able to find the internal structure or relationship between different inputs. The most representative technique of unsupervised learning is clustering. [13]. A review of current research reveals that almost all the ML (machine learning) algorithms employed in the Breast Cancer diagnosis and prognosis are supervised. Furthermore, most of these supervised learning algorithms belong to a specific category of classifiers that classify on the basis of conditional probabilities or conditional decisions.

2. LITERATURE REVIEW

Ali et al. work on big health data it was concluded that machine learning algorithm is best suited for detection of specific problem and it is possible to predict individual cancer risk via deep learning based solely on personal health informatics. [2]

Meraryslan Meraliyev and others have worked on problems with breast cancer prediction and worked on solutions, using 5 modelling algorithms with Greedy Search and K-fold Cross-validation. Algorithms like Neural Networks, Decision Tree Classifier, Logistic Regression, K-nearest Neighbour (KNN) and Support Vector Machines (SVM) were considered. The results of modelling showed algorithms like SVM and KNN are the best ones for breast cancer prediction. [3]

Medisetty Hari Krishna and others have analysed the medical data by various data mining and machine learning techniques. They utilized four main algorithms: Support vector classifier, Random Forest, Gradient Boosting, Naive Bayes, Cart Model, Neural Network and Linear Regression algorithm on the Wisconsin Breast Cancer (original) datasets and compared efficiency and effectiveness of those algorithms in terms of accuracy, precision, sensitivity and specificity to find the best classification accuracy. Support vector has proven its efficiency in Breast Cancer prediction and diagnosis; it achieves the best performance in terms of precision and low error rate. [4]

Ali Al Bataineh applied five of the popular nonlinear machine learning algorithms for breast cancer detection in form ofMultilayer Perceptron (MLP), K-Nearest Neighbours (KNN), Classification and Regression Trees (CART), Gaussian Naïve Bayes (NB), Support Vector Machines (SVM) and NB(Gaussian Naive Bayes). The results concluded accuracy of MLP(Multilayer Perceptron) on the training data was 96.70% which is better than the other four algorithms. Results of this study confirmed that the MLP (Multilayer Perceptron) model has the highest performance in terms of accuracy, precision, and recall of 99.12%, 99.00%, and 99.00% respectively. [5]

Abien Fred study on the Wisconsin Diagnostic Dataset suggested, a CV technique such as k-fold cross validation should be employed. The application of such a technique will provide more accurate measure of model prediction

performance and assist in determining the most optimal hyper-parameters for the ML algorithms. [6]

A survey done by B.M. Gayathri and others, assessed performance of different machine learning algorithms such as Support Vector Machine(SVM) and Relevance Vector Machine(RVM). Overall they found many researchers have applied the algorithm of neural networks for predicting cancers, especially the breast cancer. If studies on Relevance Vector Machine (RVM) continue, then it is likely that the use of Relevance Vector Machine (RVM) will become much more useful in diagnosing breast cancer. [7]

Habib Dhahri and other solved the problem of automatic detection of breast cancer using a machine learning algorithm by conducting three different experiments using the breast cancer dataset. In the first test, they proved that the three most popular evolutionary algorithms can achieve the same performance after effective configuration. The second experiment focused on the fact that combining features selection methods improves the accuracy performance. The third experiment, they deduced how to automatically design the machine learning supervised classifier. The proposed model looks naturally suited for control parameter setting of the machine learning algorithms in one side and automated breast cancer diagnosis on the other side. [8]

Youness Khourdifi and Mohamed Bahaj used five learning algorithms: SVM (support vector machine), Random Forest, Naive Bayes, and K-NN (K-Nearest Neighbors), applied to the breast cancer dataset, and tried to compare them according to many criteria: accuracy, turnaround time, sensitivity, and specificity. In their work SVM (support vector machine) has proven its performance on several levels by the lowest error rate, and shortest turnaround time. [9]

Ebru Aydındag Bayrak and others from Department of Computer Engineering Istanbul University Turkey discussed two popular machine learning techniques for Wisconsin Breast Cancer classification. Based on the performance metrics of the applied machine learning techniques, SVM (Sequential Minimal Optimization Algorithm) showed the best performance in the accuracy of 96, 9957 % for the diagnosis and prediction from WBC dataset. [10]

Abdelghani Bellaachia and others interpreted that the preliminary results are promising for the application of the data mining methods into the survivability prediction problem in medical databases. Their approach takes into consideration, the Survival Time Recode (STR), the Vital Status Recode (VSR) and Cause of Death (COD). [11]

Anusha Bharat and others from Institute of Technology Bangalore, India have concluded in their study, each algorithm performs in a different way depending on the dataset and the parameter selection. For overall methodology, KNN technique had given the best results. Naive Bayes and logistic regression had also performed well in diagnosis of breast cancer. The SVM (Support vector machine) used in the analysis was only applicable when the number of class variable was binary. To solve this problem scientist came up with multiclass SVM (Support Vector Machine).[12]

Muhammet Fatih Aslan and others in their work on Breast Cancer Diagnosis by Different Machine Learning Methods Using Blood Analysis. This dataset was different from other datasets in terms of feature type. The significance of data in breast cancer detection was investigated by ML(machine learning) methods. Analysis was performed with four different ML (machine learning) methods. Interfaces for a Artificial Neural Networks (ANN) and were Extreme Learning Machine (ELM) developed. In addition, the hyper parameter values giving the least errors for Artificial Neural Networks (ANN), Extreme Learning Machine (ELM), K-Nearest Neighbor (KNN) and SVM (support vector machines) methods were determined using hyper parameter optimization technique. Accuracy rates and training times were obtained according to these values. The results indicated highest accuracy rate and the lowest training period by Standard Extreme Learning Machine (ELM). They proved, the use of Standard Extreme Learning Machine (ELM) is more advantageous in terms of time when there are a high number of samples.[13]

Joseph A. and others published a review on Applications of Machine Learning in Cancer Prediction and Prognosis, identified number of trends with respect to the types of machine learning methods being used, the types of training data being integrated, the kinds of endpoint predictions being made, the types of cancers being studied and the overall performance of these methods in predicting cancer susceptibility or outcomes. They concluded, if the quality of studies continues to improve, it is likely that the use of machine learning classifier will become much more commonplace in many clinical and hospital settings. [14]

In the review published by Wenbin and others from China have shown for several decades Artificial Neural Networks

(ANN)s have dominated Breast cancer diagnosis and prognosis, recently alternative ML(machine learning) methods are applied to intelligent healthcare systems to provide a variety of options. Lots of algorithms achieved very high accuracy in WBCD (Wisconsin Breast Cancer dataset). Classification accuracy is very important assessment criterion but it is not the only one. ML (machine learning) techniques have shown their remarkable ability to improve classification and prediction accuracy. [15]

Bibhuprasad Sahu, and other published a research article on A Hybrid Approach for Breast Cancer Classification and Diagnosis proposed predictive model for diagnosis of cancer. They incorporated Multivariate statistical and machine learning techniques for better accuracy. They measured performance of different classifier techniques. Their study result reveals Artificial Neural Networks (ANN) plays major factor for detection of cancer diagnosis to save the human life from the dangerous disease. [16]

Smita Jhajharia, and others review article on risk factors, susceptibility, and machine learning techniques for cancer prediction suggested the multitude of various general and miscellaneous risk factors have not been comprehensively taken into account for the modeling of a predictive tool. There is a need for a robust mathematical model incorporating all that have been left. They have highlighted various Absolute Risk Prediction Models like the Gail model and BRCAPRO model. [17]

Madhuri Gupta and others have concluded; it is needed to improve the prediction of breast cancer in order to increase the accuracy of diagnosis. They analyzed four widely used machine learning techniques: Multi-Layer Perceptron (MLP), Support vector machine (SVM), K- Nearest Neighbor (KNN) and Decision Tree (DT). The performance in terms of accuracy, Multi-Layer Perceptron (MLP) is better as compared to other techniques. Multi-Layer Perceptron (MLP) technique also performs better than other techniques when Cross Validation metrics is used in breast cancer prediction. In order to further improve accuracy, Info Gain test, Gain Ratio test and Chi-square tests will be incorporated in future. [18]

Shubham Sharma and others from Amity University Uttar Pradesh did a comparative study of different machine learning algorithms, for the detection of breast cancer. Performance comparison of the machine learning algorithms techniques was carried out using the Wisconsin Diagnosis Breast Cancer data set. It was observed that each of the algorithms had an accuracy of more than 94%, to determine benign tumor or malignant tumor. They concluded that K-Nearest Neighbor (KNN) is the most effective in detection of the breast cancer as it had the best accuracy, precision and F1 score over the other algorithms. [19]

Naresh Khuriwal and Dr Nidhi Mishra proposed the deep learning method, convolutional neural network that is mostly used for classification of images dataset. After the implementation this method they achieved 99.67% accuracy. [20]

Sangeetha D, N. And others published their work in the International Journal for Research in Applied Science & Engineering Technology (IJRASET) in 2018 on Predicting Cancer using Machine Learning Algorithms. [21]

David A. and others have deduced the following conclusion in 2019. They analysed WDBC (Wisconsin Diagnostic Breast Cancer) dataset using dimensionality reduction techniques and three popular ML (machine learning) algorithms to classify malignant and benign tumors. Their experimental work proves that classification performance is dependent on the ML (machine learning) classification technique chosen. The results showed that SVM-LDA (support vector machine-linear discriminant analysis) and ANN-LDA (Artificial Neural Networks-linear discriminant analysis) outperforms the other ML (machine learning) classifier models. This chosen approach showed good and promising results over the validation dataset. It obtained a classification accuracy of 98.82%, sensitivity of 98.41%, specificity of 99.07% and area under the receiver operating characteristic curve of 0.9994. This research work reveals that feature selection and feature extraction can help improve the diagnosis of benign and malignant tumors using machine learning techniques. [22]

Shelly Gupta et al. have observed that the accuracy for the diagnosis analysis of various applied data mining classification techniques is highly acceptable and can help the medical professionals in decision making for early diagnosis and to avoid biopsy. The prognostic problem was mainly analysed under Artificial Neural Networks (ANN) s and its accuracy came higher in comparison to other classification techniques applied for the same. But more efficient models can also be provided for prognosis problem like by inheriting the best features of defined models. Thus the best model can be obtained after building several different types of models, or by trying different technologies and algorithms. [23]

Leili Tapaka, and others have used performance of six machine learning and two classical techniques in predicting survival and metastasis occurrence in patients with breast cancer. Their finding showed that the SVM (support vector machine) and LDA (linear discriminant analysis) were the best models to predict survival in terms of several criteria and the LDA (linear discriminant analysis) was the best technique to predict metastasis among Breast Cancer patients in this study. [24]

Hiba Asria, and others, employed four main algorithms: SVM(support vector machine), Naïve Bayes (NB),K- Nearest Neighbour (KNN) and C4.5 on the Wisconsin Breast Cancer (original) datasets. The study revealed SVM (support vector machine) reaches an accuracy of 97.13% and outperforms all other algorithms. They concluded SVM (support vector machine) has proven its efficiency in Breast Cancer prediction and diagnosis and achieves the best performance in terms of precision and low error rate. [25]

B Nithya, published her research work which shows that machine learning tools and techniques are essential in numerous disease predictions. There are lot of open problems and future challenges in dealing with massive amounts of heterogeneous, distributed, diverse, highly dynamic data sets and increasingly large amounts of unstructured and non-standardized information with respect to varied types of diseases. She recommended efficient machine learning approaches becomes essential in the health care industry to address these challenges. Machine Learning techniques could revolutionize the entire healthcare industry by providing accurate insights and predictions related to symptoms, diagnoses, procedures and medications. [26]

Margaret C.in her review article claimed, neural networks are one of many different computational techniques that may be applied to cancer diagnostics and treatment. It is entirely possible that the next cancer breakthrough may take place in a CPU instead of a test tube in future. The final conclusion was cancer research literature supports the claim that Artificial Neural Networks (ANN)s are effective tools in cancer diagnosis and treatment, and suggests that there is an expanding role for computer technologies in the future of medicine.[27]

Li Shen, and others published the following conclusion in their scientific report. The results demonstrated that deep learning models trained in an end-to-end fashion can be highly accurate and potentially readily transferable across diverse mammography platforms. Deep learning methods have enormous potential to further improve the accuracy of breast cancer detection on screening mammography as the available training datasets and computational resources expand. The end-to-end approach can also be applied to other medical imaging problems where ROI (cancerous region of interest) annotations are scarce. [28]

The work by Anji Reddy and others, have supported the new method DNNS (Deep Neural Network with Support Value) for detecting Breast Cancer. For better performance, efficiency, and quality of images, a normalization process was employed. Experimental results proved that the proposed DNNS (Deep Neural Network with Support Value) is quite better than the existing methods. It is ensured that the proposed algorithm is advantageous in both performance, efficiency and quality of images is crucial in the latest medical systems. [29]

S Kesavan, and others published a literature review in 2020 related to the mass detection of breast cancer. Recent networks namely Convolution Neural Networks were used in their paper for improving the efficiency of mass detection of breast cancer. This proposes novel bosom malignant growth location and order strategy which utilizes district based and surface based highlights for bosom disease portrayal and arrangement. [30]

3. OVERVIEW OF RELATED WORK

It is evident from the literature review that Machine learning technique are widely used for diagnosed of breast cancer using classification. Machine learning (ML) techniques offer various probabilistic and statistical methods that allow intelligent systems to learn from reoccurring past experiences to detect and identify patterns from a dataset. The researchers focused on studies using artificial neural networks (ANNs), support vector machines (SVMs), decision trees (DTs), and k-nearest neighbours (k-NNs) techniques. And they also used the Wisconsin breast cancer database. The researchers provided a clear and intuitive catalogue of information. The researchers believed that many algorithms have achieved very high accuracy using the Wisconsin breast cancer database (WBCD), but the development of improved algorithms is still necessary. Among the better designed and validated studies, it is clear that machine learning methods can be used to substantially (15-25%) improve the accuracy of predicting cancer susceptibility, recurrence and mortality. At a more fundamental level, it is also evident that machine

learning is also helping to improve our basic understanding of cancer development and progression. [14]

After reviewing the current literature it is apparent there is a need to formulate a system for early detection of breast cancer to improve the prognosis. We are working on project which will help diagnosis of breast cancer in early stage by using mammogram images. We will be analysing mammogram images using different machine learning techniques and applying artificial intelligence algorithm to detect breast cancer in early stage.

4. PROPOSED SYSTEM

Worldwide 14 million patients are diagnosed with cancer per year by pathologist. That's million people who will face years of uncertainty. Pathologists have been performing cancer diagnosis and prognosis for decades. The problem comes in a next part according to the Oslo university hospital, the accuracy of detection is 60% of pathologists. In this system machine learning (ML) algorithms and artificial intelligence (AI) based computer system are used to improve the accuracy rate of detection and diagnosis of the breast cancer at early stage. This project lays foundation in making detection of the cancer automated so that more and more people can get it diagnosed at an early stage and gets cured.

4.1 Implementation:

The signs of detection are Masses and micro calcification clusters which are important in early detection of breast cancer. Micro calcification is nothing but tiny mineral deposits within the breast tissue. They look similar to small white colored spots. They may or may not be caused by cancer. Masses can be many things, including cysts (fluid-filled sacs) and non-cancerous solid tumors, but they could also be cancerous. The difficulty in cancer detection is that the abnormalities from normal breast tissues are hard to read because of their subtle appearance and ambiguous margins.

4.2 Methodology:

In this project we will be using adaptive mean filter to remove noise from images, since it is better among all the spatial filters andit also distinguishes fine details from noise. The Adaptive Median Filter performs spatial processing to determine which pixels in an image have been affected by impulse noise. The Adaptive Median Filter classifies pixels as noise by comparing each pixel in the image to its surrounding neighbor pixels. A pixel that is different from a majority of its neighbors, not structurally aligned with those pixels to which it is similar, is labeled as impulse noise. These noise pixels are then replaced by the median pixel value of the pixels in the neighborhood that have passed the noise labeling test.

Initially, we will be converting the image into gray scale image using rgb2gray () function then applying adaptive mean filtering to the resulting image and then converted the image into unsigned integer 8 using unit8 () function.
In this way the image will be preprocessed image. Then GMM segmentation (Gaussian Mixture Model) is done on the preprocessed image with number of regions 2 and number of GMM components 2 and maximum number iterations 10. Then we can perform k-means segmentationwith k=2. Then Implement HMRF-EM (Hidden Markov Random Field Model) and get its Expectation-Maximization Algorithm.

4.3 Machine Learning Types:

The supervised learning algorithms are required for training a set of medical image to identify a specific breast cancer tumor type. This will be labeled depending on pathologist result.

These labels also known as ground truths; can be specific or general as needed to answer the questions. The machine learning algorithms used here is exposed to enough of these labeled data to allow them to answer questions of interest. Because of the large number of well labeled images required to train models, creating these data sets is often laborious and expensive.[26]

Unsupervised learning algorithms, used here to cluster the data that have similar characteristics, and the unlabeled data can be exposed to the algorithms with the goal of generating labels that will meaningfully organize data.
Machine learning algorithms are used to analyses any data set to extract data driven rules from data set. Furthermore, adding imaging of the targeted image of the body helps the doctor and physician to diagnosis. The artificial intelligence (AI) techniques is used to capture the photography of the targeted body parts which can be captured more reliably.[26]

Semi Supervised Learning Is used for the same applications as supervised learning. But it uses both labeled and unlabeled data for training. Early examples of this include identifying a person's face on a web cam. [26]

Reinforcement Learning (RL) In this type of learning, machine is trained to take specific decisions based on the business requirement with the objective to maximize the efficiency (performance).

This continual learning process ensures less participation of human expertise and saves more time. Reinforcement Learning is often used for robotics, gaming and navigation. With reinforcement learning, the algorithm discovers through trial and error which actions yield the greatest rewards. [26]
The principle of SVM (support vector machine) starts from solving linear separable problems then extends to deal with nonlinear problems.

The way of solving nonlinear problems is to map training samples from the original finite dimensional space to a higher dimensional space to realize linear reparability. SVM (support vector machine) is one of the most popular approaches in Breast Cancer diagnosis and prognosis. [22]

4.3.1. Least square support vector machine (LS-SVM):

LS-SVM is a modification to the Support Vector Machine (SVM) model with a least squares loss function and equality constraints, where
The dual solution can be found by solving a linear system instead of quadratic programming problem. [1, 3, 9, 24].

4.3.2. AdaBoost (AD):

AdaBoost belongs to the machine learning techniques family and can be considered as a meta-algorithm that improves the performance together with other learning techniques. In a classification problem, AdaBoost focuses on the sequentially applying weak classifiers. [24, 29]

4.3.3. Probabilistic Neural Network:

PNN was proposed by Specht in 1988. It is designed to improve the performance of conventional neural networks in which long computation times are required. PNN replaces the sigmoid activation function often used in neural networks with a statistically derived exponential function. The PNN is an extension of what is probably the simplest possible classifier i.e., find the training sample closest to the test sample and assign it the same class. A single PNN is capable of handling multiclass problem.[22]

4.3.4. K-Nearest Neighbor:

KNN classifier is one of the simplest and oldest methods for performing general, nonparametric classification.
In this model, the distances between the test sample and all the other samples in the training set is first measured. Then, the class of the test sample is assigned according to a simple majority vote over the labels of its KNN (K -nearest neighbors). In this classification, the number of neighbors, i.e. K needs to be pre-defined. A reasonable and practical approach would be to use trial and error to identify K such that it gives the lowest misclassification error rate.[3,5,9,12]

4.3.5. Artificial neural networks:

Artificial Neural Networks (ANN) or Neural Networks (NN) is composed of numerous processing elements that are highly connected and analogous to synapses. The activation of artificial neuron is controlled by calculating inputs and weight using a mathematical equation. The back propagation of errors from the output layer to the hidden layer is the main idea of this algorithm.[1,3,5,11].

4.4 Steps of Machine Learning Algorithm

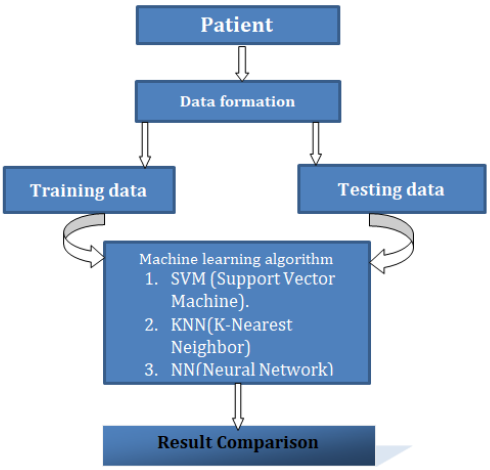

Fig -1: Step of machine learning

4.4.1 Data collection and pre-processing

The first step is collection of data from the retrospective cohort of the patient who have been diagnosed with breast cancer. Data pre-processing is performed to improve the quality of a dataset to get a clean data which can be useful for modelling.
Data cleaning involves removing noise and inconsistencies which are present in the data, thereby improving the quality

of the data. During the data pre-processing stage, the data is partitioned into the training dataset and validation dataset.

A. The training dataset is used in training the machine learning model.
B. The validation dataset is used during the prediction stage.

4.4.2. Feature Selection and extraction:

Feature extraction on the other hand, reduces the number of dimensions by transforming features in high dimensional space to fewer dimensions. The feature selection techniques used generally are CFS (Correlation-based Feature Selection) and RFE (Recursive feature elimination) methods. The most well-known feature extraction technique is principal component analysis (PCA) and LDA (linear discriminant analysis). Generally, algorithms for feature selection can be categorized into below classes:

a. Wrapper Methods: wrapper feature selection methods combine random features subset to train the model. As, the lower error rate combination kept and the higher error rate combination features removed.
b. Filter Methods: filter feature selection method selects features by applying score for each feature. Filter feature selection could be used as pre-processing step for wrapper method in large dataset.
c. Embedded Methods: embedded feature selection method selects the features during the model construction.

NodeMcu (ESP8266) is an open source firmware that provides the flexibility to build the IoT based application. NodeMcu has gained its popularity due to its low cost and Wi-Fi enabled features. Thus making the device to operate much faster and making it as a first choice for IoT applications.

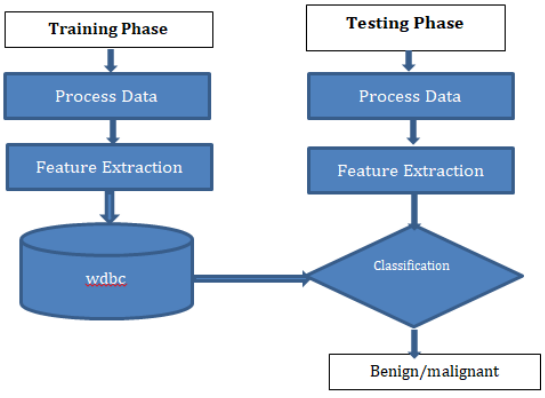

Fig -2: *Training and testing phase*

4.4.3. Classification

Classification-based algorithm and feature selection/extraction technique is good in terms of predicting the prognosis and increases the accuracy. Artificial Neural Network (ANN), Standard Extreme Learning Machine (ELM), Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Classification and Regression Decision Tree (CART), Gaussian Nave Bayes (NB), Multilayer Perceptron (MLP).

5. CONCLUSION

In this paper, we proposed a system that can detect breast cancer and how machine learning algorithm (ML) can improve the early detection and diagnosis of breast cancer. According to many different research papers, support vector machine (SVM) is one of the most powerful machines learning (ML) algorithm that is able to model the human understanding of classifying data. It can find the relationship between data and segregates them accordingly. Here we try to propose the best (accuracy) results for diagnosis and classification in breast cancer.

REFERENCES

[1] S. Willium,"Network Security and Communication", IEEE Transaction, Vol.**31,** Issue.**4**, pp.**123-141**, **2012**. For Journal

[2] R. Solanki, "Principle of Data Mining", McGraw-Hill Publication, **India**, pp. 386-398, 1998. For Book

[3] M. Mohammad, "Performance Impact of Addressing Modes on Encryption Algorithms", In the Proceedings of the 2001 IEEE International Conference on Computer Design (ICCD 2001), Indore, **USA**, pp.**542-545**, **2001.** For Conference

[4] S.K. Sharma, "Performance Analysis of Reactive and Proactive Routing Protocols for Mobile Ad-hoc – Networks", International Journal of Scientific Research in Network Security and Communication, Vol.**1,** No.**5**, pp.**1-4**, **2013**.

[5] S.L. Mewada, "Exploration of Efficient Symmetric AES Algorithm", International Journa of Computer Sciences and Engineering, Vol.**4**, Issue.**11**,pp.**111-117**, **2015**.

[6] A. Mardin, T. Anwar, B. Anwer, "Image Compression: Combination of Discrete Transformation and Matrix Reduction", International Journal of Computer Sciences and Engineering, Vol.**5,** Issue.**1**, pp.**1-6**, **2017**.

[7] H.R. Singh, "Randomly Generated Algorithms and Dynamic Connections", International Journal of

Scientific Research in Network Security and Communication, Vol.**2,** Issue.**1**, pp.**231-238**, **2014.**

[8] Thomas L., "A Scheme to Eliminate Redundant Rebroadcast and Reduce Transmission Delay Using Binary Exponential Algorithm in Ad-Hoc Wireless Networks", International Journal of Computer Sciences and Engineering, Vol.**3,** Issue.**8**, pp.**1-6**, **2017**.

[9] C.T. Lee, A. Girgensohn, J. Zhang, "Browsers to support awareness and Social Interaction," Computer Graphics and Applications, Journal of IEEE Access , Vol.**24,** Issue.**10**, pp.**66-75**, **2012.** doi: 10.1109/MCG.2004.24

[10] Lin C., Lee B.,"Exploration of Routing Protocols in Wireless Mesh Network", In the Proceedings of the 2015 IEEE Symposium on Colossal Big Data Analysis and Networking Security, Canada, pp.**111-117**, **2015**.

[11] S. Tamilarasan, P.K. Sharma, "A Survey on Dynamic Resource Allocation in MIMO Heterogeneous Cognitive Radio Networks based on Priority Scheduling",

[12] International Journal of Computer Sciences and Engineering, Vol.**5,** No.**1**,pp.**53-59**, **2017**. Egyptian Computer Science Journal Vol. 42 No.3 May 2018 ISSN-1110-2586 -29- Intelligent Prediction of Breast Cancer: A Comparative Study Merhan A. Abd-Elrazek 1 , Ahmed A. Othman 2 , Mohamed H. Abd Elaziz3 and Mohamed N. Abd-Elwhab4

[13] Jun Deng, PhD Professor Department of Therapeutic Radiology Yale University School of Medicine November 4, 2017, Ohio River Valley Chapter Fall Symposium, Indianapolis, ININternational Journal of Advances in Science Engineering and Technology,

[14] ISSN: 2321-9009 Volume-5, Issue-3, Jul.-2017 http://iraj.in Choosing Best Machine Learning Algorithm for Breast Cancer Prediction 50 CHOOSING BEST MACHINE LEARNING ALGORITHM FOR BREAST CANCER PREDICTION 1MERARYSLAN MERALIYEV, 2MEIRAMBEK ZHAPAROV, 3KAMALKHAN ARTYKBAYEV

[15] Medisetty Hari krishna, 2 Dr. Kunjam. Nageswara Rao 1 M. Tech Student, ,2Professor 1,2Department of Computer Science and Systems Engineering 1,2Andhra University College of Engineering, Visakhapatnam, AP, India

[16] International Journal of Distributed and Parallel Systems (IJDPS) Vol.4, No.3, May 2013.

[17] Hindawi Journal of Healthcare Engineering Volume 2019, Article ID 4253641, https://doi.org/10.1155/2019/4253641

[18] Ikram GAGAOUA3 Tolga ENSARø 4 Computer Engineering, Istanbul University, Istanbul, Turkey. Meriem AMRANE1 Saliha OUKID2 Computer Science Department, LRDSI Laboratory, University of Blida 1, Blida, Algeria

[19] Abdelghani Bellaachia, Erhan Guven Department of Computer Science The George Washington University Washington DC 200522018,

[20] IEEE Third International Conference on Circuits, Control, Communication and Computing

[21] International Journal of Intelligent Systems and Applications in Engineering Advanced Technology and ScienceISSN:2147-67992147-6799 www.atscience.org/IJISAE.

[22] Joseph A. Cruz, David S. Wishart Departments of Biological Science and Computing Science, University of Alberta Edmonton, AB, Canada T6G 2E8

[23] Department of Computer Science, Brunel University London, Uxbridge, Middlesex UB8 3PH, UK; Wenbin.Yue@brunel.ac.uk(W.Y.); Annette.Payne@brunel.ac.uk(A.P.); Xiaohui.Liu@brunel.ac.uk (X.L.) 2 School of Mathematics, Southeast University, Nanjing 210096, China; hwchen_seu@163.com * Correspondence: Zidong.Wang@brunel.ac.uk Received: 2 February 2018; Accepted: 23 April 2018; Published: 9 May 2018.

[24] Bibhuprasad Sahu1, *, Sachi Nandan Mohanty2 and Saroj Kumar Rout2 1Research Scholar, North Orissa University, Baripada, Odisha. 2Gandhi Institute for Technology, Bhubaneswar, Odisha.

[25] International Journal of Advances in Science Engineering and Technology, ISSN: 2321-9009 Volume-5, Issue-3, Jul.-2017 http://iraj.in

[26] International Journal of Machine Learning and Computing, Vol. 9, No. 3, June 2019.Ali Al Bataineh.

[27] International Journal for Research in Applied Science & Engineering Technology (IJRASET)

[28] ISSN: 0193-4120Page No. 6667- 6670Volume 82 Page Number: 6667 - 6670 Publication Issue:January-February2020.UG Scholar,Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Thandalam, Chennai, Tamilnadu

[29] Procedia Computer Science 83 (2016) 1064 – 1069 1877-0509 © 2016 The Authors.

[30] Volume 6, Issue 6, June 2016 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering Research Paper Available online at: www.ijarcsse.com

[31] Clinical Epidemiology and Global Health Volume 7, Issue 3, September 2019, Pages 293-299 Published by Elsevier B.V. This is an open access article undertheCCBY-NC-NDlicense.