A systematic literature survey on breast cancer and advancements in its detection using Machine learning

Dr. Yelepi Usha Rani¹
VNR Vignana Jyothi Institute of Engineering and Technology, Department of Information Technology

Dr. G. Sudhakar²
Assistant Professor(c), School of Information Technology, JNTUH
Department of Computer Science Engineering,

Lakshmi Sowmya Kotturi³
VNR Vignana Jyothi Institute of Engineering and Technology, Department of Information Technology

Abstract: The main objective of this literature survey is to provide insights about various approaches and researches that are going on to develop an accurate CAD system that would help identify breast cancer at early stages consequently helping the chances of survival. This paper also sheds light on the obstacles and challenging issues in developing an efficient system. Also, we will walk through the various existing medical and non-medical procedures to cure cancer like radiation therapy, chemotherapy, nanotechnology, etc. The paper digs deep down about each aspect starting from the growth of cancer cells to its cure-all along the way discussing types and comparing them about how dangerous these could be and different side effects of the treatments in detail [2]. This paper overall presents a complete overview of breast cancer and research done in this context along with future trends about CAD systems and ML techniques used for E-health diagnosis.

Keywords: Breast cancer, Computer-aided diagnosis (CAD), chemotherapy, tumor, Gene Therapy, Support Vector Machine (SVM) Wisconsin Breast Cancer Diagnosis (WBCD), Artificial Neural Network (ANN).

I. INTRODUCTION

Nowadays, Breast cancer is one of the highly occurring cancers and its diagnosis is life-threatening. Every 1 woman out of 8 has developed breast cancer over their lifetime in the US which explains the solemnity. Breast cancer is dreadful cancer and the death rate is also high along with a critical diagnosis process. In the past years along with advancements in technology there have been changes in the way of understanding the disease, many types of research have enabled more effective and less complex procedures to identify as well as cure breast cancer. With more efficient scanning methods and awareness among people, it has become easier to identify tumors and detect whether they are toxic(cancer) or not. Consequently, survival rates for breast cancer have shown improvement notably mostly in the young woman. Further, we would discuss the types of breast cancers, their main causes, various clinical symptoms, and wide approaches both drug-free like surgery and radiation methods and drug treatments including gene therapy, chemotherapy, etc. breast cancer.

Generally, a study has shown that many factors like hormones, lifestyle, place of living, and in some cases, genetics also increase the chances of developing breast cancer. Around 5% of breast cancers occur due to inheritance i.e., due to gene mutations it comes as an encumbrance to the family from ages. Whereas there some factors that are personal like Obesity, aging, hormonal imbalances (mostly postmenstrual) are the other factors that cause breast cancer [1]. Breast cancer is a popularly known dangerous cancer where global statistics state that 10.4% of all cancer is contributed by breast cancer and in 2004. It has caused 519,000 deaths all over the world [2]. Also, most cancer occurs in the duct cells of the breast called duct cancer and other causes of cancer are lobular cancers that occur in the line of lobules and various other tissues like stromal tissues.

Breast cancer refers to cancers that develop from the tissues in the breast, mostly from the inner surface of milk ducts or the section that generates and supplies the milk to the ducts. For very obvious reasons like hormonal fluctuations, breast cancer is almost 100 times more usual in women when compared to men.

Cancer cells are very similar in both structure and formation of compounds like other natural cells in the organism as they are originated from the healthy cells, they have high similarity with minute differences thus yet not completely identical like the DNA and RNA. Thus, it's hard to detect and differentiate cancer cells from other cells in organisms [3]. Cancer cells are formed from normal cells with slight changes or modification/mutation of chromosomes in DNA and RNA. These mutations can occur impulsively or prompt by other factors like exposure to radiation (mobile phones, UV rays), electromagnetic wave (CTC scans, microwaves, X-rays, Gamma-rays, etc.), infinitesimal particles like viruses, bacteria, and fungi, aging of DNA and RNA chromosomes, etc. of these can produce modifications to the DNA/RNA which will originate cancer.
Cancer is also called "Entropic Disease" as it rises entropy of the human being to the cause where it cannot control itself that the exterior intervention is needed to aid the body to return to a normal and enduring entropic state [4].

Cancer originates and develops when the system isn’t functioning normally but instead producing an abnormal number of cells that the system cannot afford to eliminate/kill or control the growth either [4]. The rate at which DNA/RNA mutations occur depends on the conditions like huge exposure to radiation, chemicals (chemists, doctors, engineers, etc.), undernourishment [5], persons with a genetic tendency to mutations [6], and others just with age factor (above 80).

1.1 Prologue about breast cancer

Usually, the cancer disease is named after the location or the body part where it originates thus, breast cancer can be defined as the turbulent growth and proliferation of cells that confine from the breast tissue [7].

The breast anatomy shows up the two main sorts of tissues i.e., glandular tissues and the stromal tissues which are just for support. Glandular tissues are the immense source for Lactation which comes from milk-producing glands called lobules and also consists of the ducts which allow milk to flow out. While stromal tissues encompass sebaceous and fibroconnective tissues of the breast. Along with these tissues the breast also contains lymphatic tissue which is the impervious system tissue that abolishes cellular fluids and any kind of waste material is excreted out of the body [8].

Different types of lumps can originate from different tissues/areas of the breast. Most tumors are generally benign which are non-toxic and easily curable with less death rate. For example, any change in the fibrocystic tissue can lead to a benign tumor which is a non-cancerous condition because of which women may develop lumps, cysts, and tumors with symptoms like piled-up packets of unwanted fluid, bristling, stiffness, fibrosis, and different areas on the breast might show thickening, mammalgia, or shooting pain in the breast [9]. Most cases of breast cancers show up in ducts referred to as ductal cancers. Few originate within the cells that are across the lobes called the lobular cancers, while a few numbers of other cancers develop within the other tissues [10].

Breast cancer is widely divided into Non-Invasive (only Ducts), Invasive (Breakthrough ducts and lobules and widespread). And the widely spread cancers are Lobular, Ductal, infiltrating lobular carcinoma, infiltrating ductal carcinoma[11].

The below figure-(i) gives insight into the various types of breast cancers and the number of cases of each type for a total sample of N.

The classic symptom of carcinoma is generally a tumor identifies within the breast or armpit which can be felt with our hands and sometimes touching it can be painful which is a strong symptom and an immediate doctor consultant is recommended. Doing breast self-exam (BSE) every month helps one to keep track of their breast texture, any abnormal changes in size, periodical changes, and any rashes/prickling along with skin condition. The overall changing features of areas where cancer development is happening shows up like pain in the nipples or any discharge either bloody or clear fluids, swelling/ mass lump inside the breast or within the armpit i.e., inside the lymph nodes, retracted (a project of nipple inwards) nipple, blemished skin on the nipple, continuous shooting pain in the breast or discomfort.

During the Advanced stages of cancer, the symptoms are more vicious like along with lymph in the underarm there might be a pain in the bone (bone metastases), wheezing problems (lung metastases), loss of appetite (liver metastases), abnormal weight loss (liver metastases), hemicrania, pain in nerves or fatigue. Breast cancer is typically diagnosed by the inaction of tumors detected by mammograms or by pulsation [12].

Figure-(i)

Figure-(ii)
Figure-(ii) shows a flow chart of how tumors detected from X-rays are classified into cancerous or non-cancerous by considering parameters like thickness, mass, etc.

When cancer is detected/diagnosed conclusions are made about the stage of cancer-based on the symptoms and severity of the disease to get an insight into how advanced the disease is. Information about the stage of cancer helps doctors decide the methods that are appropriate for diagnosis and therefore the predict the future conditions of the patient [13]. Stages of carcinoma could be defined as a place (not invasive) or invasive which is the most general way of doing it. Stages could also be defined intimately and designated particularly based on the variety (O-IV).

<table>
<thead>
<tr>
<th>STAGE</th>
<th>PRIMARY TUMOR</th>
<th>NODES</th>
<th>METASTATES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage I A</td>
<td>≤ 20mm</td>
<td>none</td>
<td>none</td>
</tr>
<tr>
<td>Stage I B</td>
<td>≤ 20mm</td>
<td>Nodal micrometastases (&gt;0.2mm and &lt;2.0mm)</td>
<td>none</td>
</tr>
<tr>
<td>Stage IIA</td>
<td>≤ 20mm</td>
<td>N1 None</td>
<td>none</td>
</tr>
<tr>
<td></td>
<td>&gt;20 mm to ≤ 50mm</td>
<td>N1 None</td>
<td>none</td>
</tr>
<tr>
<td>Stage IIB</td>
<td>&gt;20 mm to ≤ 50mm</td>
<td>N2 N1 or N2</td>
<td>none</td>
</tr>
<tr>
<td></td>
<td>&gt;50 mm</td>
<td>N0-N2</td>
<td>none</td>
</tr>
<tr>
<td>Stage IIIA</td>
<td>≤ 50mm</td>
<td>N3</td>
<td>none</td>
</tr>
<tr>
<td>Stage IIIB</td>
<td>Extension to chest wall &amp; skin</td>
<td>N0-N2</td>
<td>none</td>
</tr>
<tr>
<td>Stage IIIC</td>
<td>Any size</td>
<td>None</td>
<td>detectable</td>
</tr>
<tr>
<td>Stage IV</td>
<td>Any size</td>
<td>Any involvement</td>
<td>detectable</td>
</tr>
</tbody>
</table>

Table-(i)

Table-(i) is a chart showing different stages of cancer with information about tumor size and metastases.

There are many treatments and advancements in the process of curing breast cancer being discovered along with time like Quadrantectomy, Lumpectomy, Radiation Therapy, chemotherapy, Nanotechnology. And recent approaches that are used for curing breast cancer are Gene Therapy, Oncogenes Inactivation, Targeted cell Suicide, and many more but all of them are better when the cancer is detected in its early stages.

II. ADVANCEMENTS IN THE PREDICTION OF CANCER USING ML TECHNIQUES.

Breast Cancer is one of the popularly known dreadful cancers stands second when compared to all other cancers in death rate. There are broadly two types of calcifications micro and macro whereas macro is common in women and is almost found in every alternate woman after crossing the age of 50. They appear as white spots in the X-rays and are mostly non-dangerous. But microcalcifications are complex if they appear as white pinpricks, they are considered as non-cancerous but if they start clustering it is a sign that cancer cells are about to originate. Due to this complex nature of microcalcification, many times radiologists fail to identify and diagnose breast cancer. The former studies and researches lead to the development of computer-aided diagnosis (CAD) systems using ML and advanced technologies like AI which can help the radiologist to identify breast cancer based on its irregularities and diagnose efficiently. Generally, researches have shown that to develop the CAD many robust ML algorithms like Support vector machine (SVM) kernels, Decision Trees were used to classify whether the tumors are cancerous or not.

Fortunately, with advancements in treatments and detection methods of cancer survival rate among breast cancer patients has improved significantly. Another factor that biases the survivability rate is cardiovascular health. Early cancer detection presents a critical factor related to patient survival. But to date, early detection of cancer remains a confront and can’t be possible without the active participation of governments for creating awareness, health institutions using high-end technologies and tools for diagnosis, and patients being on guard and observer any changes in their body with periodical checks. Yet with all the devising steps taken there some risk factors like age, genes, and geographical variations that are uncontrollable, and also disease non-uniformness is complex and hard to identify in particular using the old classic methodologies like X-rays, MRI scans, ultrasounds, and mammography.

But then existing systems are all helpful only once cancer is detected but detection is one of the major hurdles as some type of cancer doesn't show any symptoms or some show only in advance stages. On the other hand, early detection has a promising growth in the rate of curing and survival thus for advance and efficient detection we need to develop a classification system or a CAD (computer aided diagnosis). The generalized steps that are followed for classifying the tumor into benign or malignant are based on the different features of the tumor-like the tumor radii, Clump thickness, Marginal adhesion, mitosis.
Machine learning techniques are widely used these days to analyze and predict or classify and get depth insights about important scientific parameters. For example, in the medical field ML can be used to predict the disease, its progression over some time, extract important information from data collected as an outcome to the research done, even suitable treatments can be chosen based on knowledge obtained efficiently. Data processing is a process of mining useful information from large amounts of data collected as records from the patients in clinics or part of any research. By using statistics and computing algorithms (ML) we can easily analyze and find interesting patterns, valuable information classifies, and even develops systems to predict outcomes. Particularly for cancer by considering features like geographic location, race, gender and hormonal, socio-economic and evaluation of health needs, in turn identifying the type of cancer and its stages. Furthermore, a thorough inspection of patient’s details and records using data pre-processing can open up to unknown facts and unexpected relations can be identified which can help diagnose the exact cause.

Thus, using ML techniques, we can mine the interesting patterns in medical data but it’s a bit hard and complicated process. Thus, with the help of data mining techniques and by using data from the healthcare industry valid pre-cancer detection systems and other various CAD, healthcare-related systems are being designed. Also using the cancer data from the clinics and diagnosis records we can predict the stage of cancer, its severity and also help doctors make the right decisions about the treatment. These days health industry is extensively using technology as a tool for diagnosis. Thus, by using ML algorithms and data processing techniques we can assist to help doctors efficiently and quickly identify the abnormalities and help in the process of diagnosis in real-time thus surviving the purpose of identifying cancer in early stages. The knowledge mining models can be classified into two categories as shown below fig.5: predictive model and descriptive model [14]. Predictive data processing can also be called forecasting or regression where we predict the unknown or missing value or predict outcomes based on the collective relationship between the response variables. They consist of two sets of variables the independent variable (set of features used to predict) and the dependent variable which can be predicted based on its relationship with the independent variables [14]. Descriptive data processing generally uses unsupervised functions which have no class variable. They are used to classify the new records based on training data and also can be used to identify interesting patterns, form rules that humans can interpret [14]. A detailed summary of where the Predictive and Descriptive data processing techniques are used for early cancer detection is presented within the next section. The researchers are classified based on the data processing model used.

This research was done in the year 2019 they mainly focused on DL and ML techniques and tried to build a computer-aided diagnosis system (CAD). Specifically, they searched for different popular websites like google scholar and a center for biomedical literature PubMed they also studied DL techniques over the past 5 years using Multiview mammogram datasets. After a keen comparison between ML and DI techniques, they concluded that DL holds great potential and capability for clinical analysis and efficient diagnosis of breast cancer when compared to existing CAD. From the literature, it is identified that varying breast masses and densities make it hard to detect and classify as cancer or not when compared with the case where calcifications are not when compared with the case where calcifications are used as features to classify. The traditional ML methods provide constricted approaches by considering density types of datasets. When compared with ML the DL methods show promising advancements for cancer diagnosis, there are still problems like data scantiness, integrity issues in data collection, notations used may cause ambiguity, some variables may be categorical which have to be dealt with a separate step called encoding, noise in data, computational expense, but all these problems can be addressed with improved computational infrastructure and enrichment in DL algorithms. They used mammography to capture and detect breast cancer and this method is efficient and safer as...
they expose the breast to a lower dose of radiation when compared to existing methods. Also, there are two different views namely the cranio-caudal (CC) view and the mediolateral oblique (MLO) view. The major hurdle was to remove the PMs (pectoral muscle) which may lead to false conclusions. They also used the segmentation step before modeling to improve the quality of the image.

B. R. Chithrakkannan [18], 2019

In this research, the mammogram technique is used to capture the X-ray images of the breast and is used as input data. The first step is to pre-process input images in which they applied Gaussian Filter and Edge detection techniques to obtain the exact location of tumors and enhance the image quality. Once pre-processing is done, they applied Wavelet Transformation to identify the first-order features, and then by obtaining a grayscale histogram using GLCM second-order features were extracted. The obtained computational parameters are processed by developing a supervised classification model for which a deep neural network with multilayers was used. The final dataset was created at the training phase which was ready to build a model. In the testing phase, the newly acquired images from a patient are given as input to the classifier after completing the image processing steps like Pre-processing, and based on the model-built outcomes are obtained with a certain level of confidence based on the accuracy of the model built. The vital observations are that using a Bilateral filter which has a nonlinear nature helps to protect the keen portion of the image that contributes the useful statistical information. This filter also clears the fringe and noise involved in the image. The bilateral filter is defined as:

$$I_{\text{filtered}}(x) = \frac{1}{W_p} \sum_{x_i \in \Omega} I(x_i) f_g(|I(x_i) - I(x)|) g_s(|x_i - x|)$$

Where the normalization term is

$$W_p = \sum_{x_i \in \Omega} f_g(|I(x_i) - I(x)|) g_s(|x_i - x|)$$

Moving on in the step of feature extracting the following features of GLCM are estimated based on entropy which is used to measure randomness, average (mean) which can show the average value of each feature, Standard deviation which shows how much each value is deviating from the central mean and to classify the interesting regions in the images using texture feature. In this texture feature, they have concluded that using the DNN algorithm would result in better accuracy and sensitivity level when compared to other algorithms that they have studied and implemented. Thus, this shows that DNN can be used for the early detection of cancer with no physical contact. The main advantage to use this method is that it avoids the destruction or any kind of damage to the part of the image being tested. This feature obtained from wavelet transformation and GLCM feature extraction was useful to classify the greatest number of sample images collected accurately. Finally, the results obtained have shown that among all the used methodologies DNN provides a far better classifier rate with a minimum error rate for all test samples. So, when several algorithms like ANN, SVM, KNN are compared DNN provided better accuracy and efficient classification rate that ranged from minimum 87% to 96% as the highest accuracy.

C. Anji Reddy Vaka [19], 2020

Generally, studies have shown that hormones, lifestyle, geographic location, socioeconomic factors, diet, and environmental factors have a direct impact on the chances of developing breast cancer. Around 5% of BC cases were due to the inheritance of gene mutations. Other factors like overweight, aging, hormonal imbalances (mostly postmenstrual) also may increase the probability of developing BC.

In this study, they considered a dataset from the M. G Cancer Hospital & Research Institute, Visakhapatnam, India. The dataset has around 8009 breast tissue images using histopathology. The sample images were collected from six-eight three patients at various magnification levels for a great annotated image. Here they use the SVM Classifier technique before which they apply RFE for feature selection. RFE is a technique that eliminates irrelevant features in a recursive manner and pruning is based on specific attributes. So, as they implement SVM-RFE together it eliminates features with low weightage when compared to their features and this process repeats in iterations. Then they use AdaBoost which is the best technique to estimate the accuracy of a classifier as it works by uniting several ailing classifiers. Even a bi-cluster technique can be used for better assessment of accuracy obtained by a classifier. In the phase of training, each subject is assigned weights, and decisions are made based on “weighted majority voting”. RNNs are the type of NN's which use sequential dimension for dimensionality reduction but are mostly not suitable when data is time-series data. RNN's differ from NN's as they predict the activation points in n-th step as they process the data in n-th step.

CNN doesn't go well with the spatial data that is comprised in the image pixels and so they use “discrete convolution”. To meet those constraints and accordingly suitable Gray images are to be taken as input. Coming to HA-BiRNN it has two layers which are for encoding one encodes the sentences and the other is for words respectively. Along with this, based on punctuation, grammar, content sentence level, and word-level attention are also considered.

But all these methods have limitations like the Naive Bayes Classifier results in average outcomes when the training data isn't represented in proper format and also it assumes that all variables are independent of the class variable. Whereas SVM cannot tackle large datasets ad works at a very high computational expense. When the dataset has skewness and imbalanced distribution techniques like Bi-clustering and Ada boost give biased outcomes and lead to false classifications. RCNN requires high computational time to train the data. HA-BiRNN gives incorrect scores when ultrasound images of the breast were given as input.

Thus, they introduce a new version of DNNS which stated better efficiency and performance also maintaining the quality of images which is crucial when it comes to the medical dataset.
They initially used a Gaussian filter to remove noise in the dataset then calculated the Entropy value for feature extraction and finally, Histo-sigmoid fuzzy clustering is applied. Finally, DNNS is used but altered to improve efficiency in classifying also took steps to improve the quality of the image as high-quality input is the key to get good results.

The support value-based normalization is calculated using:

$$SN = supportvalue \times \frac{y - y_{\min}}{y_{\max} - y_{\min}}$$

Where $y_{\min}$ and $y_{\max}$ are the maximum values in image $y$, where $SN$ is the support value-based normalized image. Thus, DNNS recorded the highest accuracy when compared with the results of other algorithms.

D. Esmat Mirbagheri [20], 2020

This paper focuses on the data collected from different sectors like clinical, research, and non-clinical and integrating them thus enhancing the amount and quality of data available for research and provide intuitive solutions accurately and efficiently. They state that clinical care and clinical research lacks behind and need a bridge for more amount of data. As medical data is considered to be sensitive and needs to be maintained confidentially acquiring it and also integrating it into a single system is a complex problem. Their main goal was to integrate breast cancer data from different sources in different formats and find the common data elements that help in the collection, processing, and sharing of data for research. As part of this research, they almost studied all researches done from 2007 to 2009. And after examining 25 studies it showed that 52% of research on breast cancer is being carried out in the United States the researchers urged during the period 2013-2015. The outcome of this research was to find the CDE's so the data can be shared and integrated between different environments by maintaining synchronization without any anomalies.

The ambiguities disrupt and conflicts raised were settled down by conducting formal meetings between the two research teams. Further, the according to the degree of agreement approximated by using the kappa coefficient ($\kappa$) was 0.85 which is statistically significant at $P < 0.001$. As part of these reviews, the sources were chosen as per the Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow chart. By following the study selection steps based on the relevancy of citations around 396 studies were opted from the whole database among which 184 were duplicate studies. From the remaining 212 also there were some exclusions done as there were articles that had access only to abstract, some had matching titles but the context of the paper was different. After a precise and lengthy procedure, the teams have jotted down a final list of CDE's that can be used as a primary standard list of features while collecting and integrating data for breast cancer from various sources. To look at the final CDE's found in three different sectors as mentioned above refer to [15].

E. Ganesh N. Sharma[21], 2010

This paper addresses the various types of causes, and dominant symptoms and also various clinical and non-clinical methods of curing breast cancer. They discussed how cancer is dangerous and global statistics about breast cancer stating that 10.4% of all cancer is contributed by breast cancer and in 2004 it has caused 519,000 deaths all over the world. In detail, the cell cycle, the different types of tissues, and how the tumor grows at different parts of the breast were discussed. Also, mostly cancer occurs in the duct cells of the breast called duct cancer and other causes of cancer are lobular cancers that occur in the line of lobules and various other tissues like stromal tissues. In the following research work, they talk about types of cancers that occur frequently and also less commonly occurring. Moving on they listed the various possible causes of cancer like genetic, hormonal, lifestyle, dietary, environmental causes, etc. Further, they presented a chart about the stages of cancer [16].

Then list the existing medical and non-medical procedures to cure cancer like radiation therapy, chemotherapy, nanotechnology, etc. The paper digs deep down about each aspect starting from the growth of cancer cells to their cure all along the way discussing types and comparing them about how dangerous these could be and different side effects of the treatments in detail.

F. Ebru Aydındag Bayrak [22], 2019

In this research after comparing with many ML techniques for prediction of cancer at early stages both SVM and ANN showed better results when worked with Wisconsin dataset which was released in 1992 with 699 instances. To evaluate the classification and compare various techniques used many performance metrics like accuracy, precision, ROC curve, the recall was taken into count.

In SVM the classes are separated with an area of problem space between the two class labels i.e., hyperplane which consists of extremely important features from all classes know as support vectors are used for classifying the lumps as cancerous or non-cancerous by considering tumor size, patients age.

Whereas, ANN is used to train with efficient non-linear functions. In biological neuron systems like the human brain process systems where many nodes are connected forming a network. It is a complex network of artificial neurons. Based on the input/output characteristics a function is generated. The function generates the weights of each input which indeed computes an output if it goes beyond a given threshold value. The output can be the input to other neurons in the network. This process is repeated and, in each iteration, a new output is generated.

For modeling, they split the data in the ratio 66% training data and 33% testing data. Also, to evaluate the model they used a 10 k-cross-fold validation technique. They also used the WEKA tool and applied SVN and ANN on the same
dataset to compare and validate their results. In the experiment conducted it is found that SVM resulted in 97.13 % accuracy with a low error rate.

**G. Lal Hussain [23], 2018**

This research uses machine learning methods such as SVM kernel, Navies Bayes method, and decision tree to classify the mammogram image samples as cancerous or non-cancerous. The two main indicators in a mammogram for the context of breast cancer have been assessed by micro-calcifications and the other is masses. The recognition and microscopic identification of low-contrast images due to the large scattering of the shape requires a multi-dimensional feature extraction strategy. To address this problem various features based on texture, morphology, EFD is extracted. Also, SIFT feature was used to make mammograms resistant to brightness, scaling, change in angle, etc. MATLAB version 2015 developed breast cancer radiographs. They used 10k-fold cross-validation by dividing the image database into train and test data. The performance of each model was assessed and compared using performance metrics like sensitivity, precision, recall, confusion matrix, specificity, and ROC curve and its area.

**H. Md. Imran Hossain Showrov[24], 2019**

This research used the three most popular supervised techniques SVM, naive Bayes, and ANN for classifying the records of the Wisconsin dataset released in 1992 with 699 instances and 9 attribute values and a single class attribute with two values benign and malignant.

As part of this research, they implemented algorithms like Linear SVM, RBF SVM, Sigmoid SVM, polynomial SVM, Feed Forward NN, RBF NN. They conduct a comparison between all these models and based on accuracy they decide the best model.

Before even consulting a doctor, a woman can predict it is a serious condition based on the symptoms that are dominating and periodic checks can also help detect cancer in early stages. The motive of this research is to build the best classifier that would help detect breast cancer efficiently. But because of so many moral, social issues in health care, getting the greatest possible precision is needed to take the best action for those affected. Patient as soon as possible. Depending on the characteristics of the data, the different types of algorithms show different results for the same data set (WBCD). Here they trained SVM, ANN, and different parts of Naive Bayes based on a specific training data set (WBCD). Some might give excellent results while some algorithms ended up giving average results. Though Feed Forward NN comes at a cost high computational time it can efficiently deal with large datasets. SVM is suitable for large or small amounts of data whose amount of data is small. Linear SVM is well suited for this amount of data because of the small amount of data and the small dimensionality. The data set can be presented in two forms one using statistics and the other is analytical. Using accuracy as the performance metric SVM has given 96.72% and works best n WBCD. The RBF NN listed second by resulting in a 95.88% accuracy.

**I. Amna Ali [25], 2010**

In recent years, researchers started extensively using machine learning techniques for cancer prediction. ML techniques help the Practitioners make wise decisions about the diagnosis of the patient and give accurate results about stages of cancer and precisely define the critical cases. In addition, as it helps early detection of cancer patients can be treated with minimal and mostly painless techniques avoiding complex procedures like surgeries and gene therapy. In this, they proposed a merger to create a composite classifier to predict breast cancer patient's chances of survival. Here the individual support vectors are combined to provide a combined classification architecture improving the chances for the prognosis of breast cancer. In this research, they implemented a technique to merge the classification results of 2 individual SVM-based classifiers by FLS type 2 interval for accurate prediction of breast cancer. This method implements two versions of SVM algorithms ic-SVM and nu-SVM with RBF kernel to the SEER dataset. Depending on the precision value obtained from each SVM-based classifier the distance information between data can be estimated. Once the hyperplane point is identified the individual support vectors are dumped into FLS type -2 so they combinedly form a composite classifier. When compared to individual SVM classifiers the composite classifier has given better results and deals with data more stably and works more generalized way.

**J. P. Sathiyanarayanan [26], 2019**

The early detection of breast cancer plays a major role in reducing the death rate which is one of the major challenges in the medical field today. Among many types of cancers, women are mostly exposed to breast cancer which is one of the dreadful cancers which high complexations and death rate. There are two forms of breast cancer: Malignant, benign type. Breast cancer is developed due to mutations in DNA and can be identified by symptoms like shooting pain in breasts, abnormal changes in size, skin color that might turn red and itchy with pricking pain, and the overall appearance of the breast. Many types of research have been done in the past in this field using ML techniques like SVM, KNN, MLP but lagged as they can't be used in real-time because they didn't reach the needs. In this paper, the idea was to use a decision tree which is a supervised learning technique for detecting breast cancer. And by observing the results when compared to all other algorithms decision tree has given the highest accuracy of 99% and well classified the records with the least error rate.
CONCLUSION
Breast cancer is prevailing rapidly due to life style and high exposures to radiation emitted from devices and having hormonal imbalances along with inherited toxic genes. Unfortunately, even the death rate is high which can be put a bay by detecting cancer at early stages which is one of the major hurdles in the medical industry. Though many advancements have made diagnosis of cancer efficient they would be more useful when tumors are detected in early stages. Also, on other hand it helps the doctors make right decisions about treatment by giving insights about patient and finding interesting patterns in their records creating links and gives better understanding about the patient’s state and how to rationally further proceed. All of this is possible with a CAD system which needs to meet the standards in order to bring it into use practically in real time. In above section of the paper, we studied and compared various researches where they used ML, AI, ANN techniques to develop models that can predict breast cancer and classify it as cancerous or non-cancerous based on various parameters like tumor size, metastases, texture, no. of nodes. Though there are many researches going on and methods being proposed yet there is open space for developing hybrid models and efficient algorithms, also collecting and integrating quality data along with huge amounts remains an open end. Many more advancements and strong communication between clinics, researchers and medical industry is needed to accumulate both quality and quantity data. And classification of breast cancer is an interesting area to experiment with and researches have a great scope to propose new and efficient methods for proper and non-painful and less complex diagnosis of BC making the world better place to live.

The table below gives a list of previous studies that worked on improving the standards and efficiency of early detection of breast cancer. All of them mostly used ML techniques while some have chosen deep learning and AI methods to address the same challenge and create a CAD system with better accuracy so that it can be adapted into hospitals and practitioners can rely on it for making better decisions for diagnosis.
<table>
<thead>
<tr>
<th>Reference</th>
<th>DM Model</th>
<th>MLT</th>
<th>Problem Statement</th>
<th>Performance Metric</th>
<th>DATASETS</th>
<th>Phase of care</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alireza Osareh [27], 2010</td>
<td>Descriptive</td>
<td>SVM, KNN &amp; probabilistic neural networks</td>
<td>Distinguish the type of tumor into benign or malignant using PCA and SNR feature ranking</td>
<td>Accuracy: 98.80% and 96.33%</td>
<td>1) FNAC breast lesions 2) gene microarrays [28]</td>
<td>primary prevention and detection</td>
</tr>
<tr>
<td>Murat Karabatak [29], 2015</td>
<td>Descriptive</td>
<td>Naive Bayesian</td>
<td>Customize NB algorithm for effective detection of breast cancer and experimented with 5-fold cross-validation</td>
<td>Sensitivity: 99.11%, specificity: 98.25% and the accuracy value: 98.54%</td>
<td>Wisconsin breast cancer database</td>
<td>primary prevention and detection</td>
</tr>
<tr>
<td>Vikas Chaurasia[30], 2017</td>
<td>Descriptive</td>
<td>Sequential Minimal Optimization (SMO), IBK, BF Tree.</td>
<td>To compare the accomplishment of different classification techniques using WEKA software</td>
<td>Accuracy: 96.2%, 95.90%, 95.46%</td>
<td>Wisconsin breast cancer dataset</td>
<td>primary prevention and detection</td>
</tr>
<tr>
<td>Valérie Bourdès[31], 2010</td>
<td>ANN</td>
<td>Multi-layer perception, Logistic regression</td>
<td>Comparing a neural network (NNs) of multilayer perceptron with typical logistic regression (LR) to identify key variables the effect cause of cancer</td>
<td>ROC Curves</td>
<td>Collected database from 1996-2006, French center and by the end of March 2006, 4070 records were stored on to the database, by examining 3929 patients by taking 220 explanatory variables into account</td>
<td>primary prevention and detection</td>
</tr>
<tr>
<td>Ya-Qin Liu [32], 2009</td>
<td>Descriptive</td>
<td>Bagging algorithm-- Decision Tree</td>
<td>proposing a method to predict using a decision tree for 5 years survivability based on imbalanced data</td>
<td>10 k-fold cross-validation, AUC under ROC curve, accuracy, specificity, and sensitivity</td>
<td>SEER breast cancer datasets</td>
<td>primary prevention and detection</td>
</tr>
<tr>
<td>Y.Ireaneus Anna Rajani [33], 2009</td>
<td>Descriptive</td>
<td>SVM classifier</td>
<td>Detecting tumors from suspicious areas extract features, finally model with SVM.</td>
<td>sensitivity of 88.75%.</td>
<td>mini MIAS database (75 mammographic)</td>
<td>primary prevention and detection</td>
</tr>
<tr>
<td>Tim W. Nattkemper [34], 2005</td>
<td>Descriptive</td>
<td>K-means, SOM, SVM, decision trees, KNN</td>
<td>using supervised and unsupervised machine learning to analyze and extract features of a tumor and classify it based on key features as benign/malignant.</td>
<td>ROC CURVES AUC -0.88 SVM, 74% of DT</td>
<td>Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) time-course data.</td>
<td>primary prevention and detection</td>
</tr>
<tr>
<td>Layla Abdel-Ilah[35], 2017</td>
<td>Descriptive</td>
<td>neural networks</td>
<td>Improve accuracy to an acceptable level using ANN</td>
<td>TANSIG activation function</td>
<td>Wisconsin breast cancer database</td>
<td>primary prevention and detection</td>
</tr>
<tr>
<td>Lei Liu [36], 2018</td>
<td>Regression</td>
<td>Logistic regression</td>
<td>improve accuracy by removing the internal subjective human factors</td>
<td>accuracy 96.5%</td>
<td>Wisconsin breast cancer database</td>
<td>primary prevention and detection</td>
</tr>
</tbody>
</table>

**Table 2: Overview of Literature studies**
REFERENCES


[16] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3255438/bin/1APTR-1-109-e005.png


[33] Early Detection of Breast Cancer using SVM Classifier Technique, Y. Irenao Anus Anna Rajani, S.Thamarai Selvi

[34] Tim W. Nattkemper, Bert Arminich, Oliver Lichte, Wiebke Timm, Andreas Degenhard, Linda Pointon, Carmel Hayes, Martin O. Leach, Evaluation of radiological features for breast tumor classification in clinical screening with machine learning methods, Artificial Intelligence in Medicine

