PREDICTING THE RISK OF BREAST CANCER DISEASE USING MACHINE LEARNING AND DIAGNOSIS RESULTS

Gouthami Yanamanagandla[1], Dr. Vicky Nair[2]

[1]M.Tech Scholar, Department of Computer Science and Engineering, TKR College of Engineering and Technology, Hyderabad, India

[2]Professor, Department of Computer Science and Engineering, TKR College of Engineering and Technology, Hyderabad, India

***Abstract*— Breast cancer, a disease or tumor which takes place withinside the breast tissue. This cancer is the most mutual form of breast most cancer mainly present in women and a few chances in men across the world. This is the major cause of women dying in the present generation. This paper will give a relative evaluation of machine learning models, deep learning techniques and data mining strategies that are used to predict this cancer. Most of the scientists kept their energy, work, time and efforts in the prognosis and diagnosis of types of breast most cancers. Each procedure has a different rate of accuracies which vary with the situation, the tools and data sets used. The primary focus here is to relatively examine the different existing machine learning models & data mining strategies to discover the maximum suitable technique that assists the data set with proper accuracy of prediction.**

***Keywords*— Machine learning techniques, breast cancer, K-nearest neighbors, malignant, benign, data mining techniques, types of breast cancers, diagnosis, Support Vector Machine.**

1. INTRODUCTION

Breast most cancers is the maximum deadly and diverse disease in the current generation which causes the dying of women present on the sphere. Breast most cancers is the 2nd biggest disorder which is accountable for women dying. Numerous machine learning models and data mining strategies are present for detecting the breast cancer predictions. The percentage of occurring the breast cancer in women when compared to men is nothing. It doesn't mean they do not have this cancer but when compared it is nothing. Discovering the maximum appropriate and suitable technique for predicting the breast most cancers is the critical task. This cancer is initiated over malignant tumors, while the boom of the mobileular goes out of the control. A large amount of fat and fibrous tissue of the chest will begin to grow and that turns into the reason of this cancer.

The affected breast cells will blowout through lumps causing cancers at different phases. The affected breast cells when unfold all over the breast part, different types of breast cancers take place. The most found breast cancers are Ductal Carcinoma in Situ (DCIS), Invasive Ductal Carcinoma (IDC), Mixed Tumor Breast Cancer (MTBC), Lobular Breast Cancer (LBC), Mucinous Breast Cancer (MBC), Inflammatory Breast Cancer (IBC). The first and foremost cancer is DCIS. DCIS cancer is also called as non-invasive carcinoma and this cancer happens when the irregular cells unfold outside of the breast part. And the second one is IDC; this type of cancer takes place when the irregular cells unfold to all breast tissues. IDC typically occurs in men and it is also referred to as invasive ductal carcinoma. The third form of cancer is MTBC, lobular cells and abnormal duct cell causes this cancer. MTBC has

Vol 12, Issue 07, JULY /2021

ISSN NO:0377-9254

another name invasive breast cancer. The 4th one is LBC; it happens inside lobule. LBC increases the chance of other different invasive cancers. The Fifth kind is MBC, also called colloid breast cancer. MBC is due to infiltrating ductal cells. This cancer happens when affected tissues unfold across duct. The last but not least form of cancer is IBC. IBC begins to appear when lymphatic vessels are blocked in ruptured cells and it leads to reddening and swelling of chest part. IBC is a quick developing breast cancer.

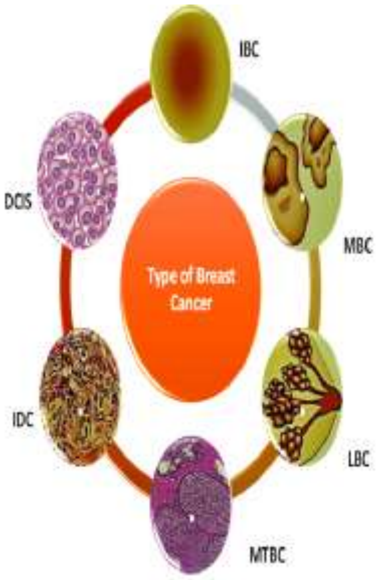


Fig.1 Variety forms of cancers

Symptoms

- ✓ New breast lump
- ✓ Partial thickening or swelling of breast tissue
- ✓ Breast skin irritation
- ✓ Breast skin depression
- ✓ Red or scaly skin in the breast area
- ✓ Breast pain
- ✓ Secretion of blood instead of milk
- ✓ Pulling the nipple

Risk Factors

- Genetic Mutations
- Taking drugs
- Not being physically active
- Being overweight
- Taking hormonal pills
- Not having pregnancy
- Drinking alcohol
- Using tobacco
- Older age
- Not having healthy diet
- Diarrhea
- Vomiting
- Hair loss
- Appetite loss
- Lifestyle is not good
- Medical treatment not taken in correct form

2. LITERATURE SURVEY

At present, the current system we are following is we need to go to the physician whenever there is a problem in breast. The physician first manually checks the patient asks details about it as how is he feeling, what are the symptoms he is facing. After taking all information from patient he comes to the conclusion and make decisions like whether to go for scanning of breast or not. If scanning is required, the patient should take the scanning and again come to physician to know the problem he/she is facing. The physician then sees the reports tells the patient about the problem she is facing. As there is no cure to get rid of this cancer and the only way is to take care and manage using treatment. As there is no correct treatment found after occurring the cancer so we should take precautions about this cancer in order not to happen. For this we have know some information about this cancer. Like why it happens what would be the reasons, risk factors of occurring these factors. If these things are known to us, we can hundred percent prevent this cancer from not happening.

Priyanka Gandhi and Professor Shalini L from the analysis of machine learning procedures at VIT University for predicting the breast cancer, vellore. This article explores ML technology to improve diagnostic accuracy. Compare CART, Random Forest, KNearest Neighbors, and other methods. The dataset used comes from the Machine Learning Library at the University of California at Irvine. It is found that the performance of the KNN algorithm is much better than other technologies compared with it. The most accurate model is KNearest Neighbor. Classification models such as random forest (RF) and boosted trees show like accuracy. Consequently, the utmost precise classifier can be used to detect tumors, therefore to find a cure as soon as possible.

3. PROPOSED SYSTEM

In this, firstly we take data from the patients of affected persons and make a dataset. We store this data firstly in an XL sheets containing name of the person, age, gender, email-id, contact number, address, type of cancer, symptoms, risk factors, causes, stage, diagnosis, treatment, reviewed doctor name and status of patient presently. And store these data in a database. And we write programs for the models we are using. And the information stored in database is taken by these models when predicting the risk and we have connectivity programs for these two things. After that predictor called Service provider can do following predictions from the dataset present with him. He can view all details of the patients of having the cancer. He can also have the patients list by taking into consideration of different factors like by symptom, risk factor, diagnosis, type of cancer, causes etc. Here we can predict such things like patient who is facing particular symptom may have high chances of occurring this cancer by the model. It is performed by each and every model and collects the accuracies results for it. The model which predicts highest accuracy is taken for further predictions. We can have ratios for type of breast cancer a person is facing. And also, ratios for symptoms, risk factors and causes so that we can predict which one is very dangerous. We can also predict patients who is having high ages can have this cancer.

The algorithms compared and used here are KNN, DT, SVM, RF, Naïve bayes and ANN. And this projected system mainly does classification first that means they train the data and then test the data and compare with the algorithms present with each and every algorithm used and then the model that fives best accuracy is taken into consideration for further purposes. The description of the proposed model: -

1. Patients make appointments via the website provided.
2. The respective patient sees the physician by consulting him/her in a hospital.
3. The doctor first manually checks the patient and if there is any problem, then he goes for scanning.
4. Scanning gives the image/x-ray of the affected patient breast.
5. If a lump is found, a biopsy should be performed.
6. After biopsy, the details are given to system by physician.
7. After that, model detects the stage of cancer.
8. The report is passed to the patient to the account.

3.2 CLASSIFICATION

The classification procedure is a supervised procedure which uses to classify new observation categories constructed upon training information. In classification, the model acquires from a given set of data or observation and then categorizes the innovative observation into multiple classes. There are 4 types of classification. They are Binary classification, multi-class classification, multi-label classification and imbalanced classification. Unlike Regression, the outcome for classification is a variable not a value. It takes labelled input data which means each input has its corresponding output. In the classification procedure, the discrete output function (y) is assigned to the input variable (x).

y = f (x), where y = categorial output.

3.2 MACHINE LEARNING

Machine learning is a subgroup of AI. This is the study of how to make the behavior and decision-making of machines more humane, so that they can study and progress their individual plans. It can be made by minimal manual involvement, that is, nope clear software design is required. The learning process is computerized, progressed based on reports of the machines during progression. Virtuous pleasant statistics is sustained to systems, and exceptional procedures are used to construct machine learning fashions in order to instruct systems in these statistics or data. Desire of procedure relies upon form of information, kind of activity which wishes to remain computerized. A typical Machine Learning process undergoes 3 stages that is training, validation and testing. After training, in validation we need to measure the error and manage the noise and then testing.

4. ALORITHMS

4.1 K NEAREST NEIGHBOR

It is the simplest supervised algorithm. It is used for regression and classification, but mainly for classification challenges. This is a non-parametric procedure, because it doesn't make assumptions about the basic data. It is a lazy learning process, as it won't learn immediately from training set, but as an alternative holds data set, performs operations on data set during classification. The KNN algorithm only stores the dataset throughout the training phase, after getting innovative data, it organizes it into categories which are very alike to innovative data. It measures similarity between new case and the obtainable case and categorizes cases built upon likeness.

Working:

- Load training data
- Choose k(any integer) no of neighbors
- Calculate Euclidean distance for k no of neighbors
- Basing the Euclidean distance value, arrange them
- Pick upper k rows from arranged order
- Allot a class to test point based upon frequent class occurring

4.2 SUPPORT VECTOR MACHINE

SVM is the flexible supervised ML procedure. It is mainly useful in regression and classification challenges but primarily for classification purposes. They have an extra-ordinary ability to handle multiple continuous and categorial variables. Comparing with other algorithms they have unique way of representing and implementing. They perform classification by selecting a hyperplane which maximizes the margin among 2 classes. The vector which describes the hyperplane is support vector. This algorithm picks the acute vectors that assist in increasing hyperplane. These extreme instances called support vectors, hence consequently called the procedure as Support Vector Machine.

Steps:

- Load information set
- Explore the data
- Preprocess the data
- Divide data
- Divide information into 2 sets; training set and testing set
- Algorithm has to be trained
- Make some predictions
- Evaluate the result

4.3 ARTIFICIAL NEURAL NETWORK

ANN is a computational procedure. Artificial neural networks are constructed identical to a humane mind, the neural nodes are linked to one another identical to a network. Artificial Neural Network (ANN) is an element of AI designed to stimulate the functions of the humane mind. The processing unit forms an artificial neural network, which in turn is composed of input and output. It is capable of doing pattern recognition. It is also an information processing technique and compute values form inputs. Artificial Neural network contains 3 layers; -input, output and the hidden layer.

Algorithm for the neural network:

- Weight's initialization.
- Multiply the weight by the input and add it.
- Compare the result with the threshold to calculate the output.
- Update the weight.
- Repeat.

4.4 NAÏVE BAYES

Naïve Bayes is the supervised technique which is built upon Bayes theorem. This algorithm is useful for solving classification challenges and mainly in text classification. Naive Bayes is one of the classifiers named probabilistic classifier, as it makes predictions by considering object probability.

Steps in Naïve Bayes:

- Data preprocessing
- Naive Bayesian fitting training set
- Test result prediction
- Results test accuracy (confusion matrix creation)
- Test set result display.

4.5 DECISION TREE

Decision Tree is one of the supervised techniques. It is useful for doing classification and regression challenges. And primarily used to solve problems of classification. DT is one of the classifiers and DT is also called as tree-structure in which the characteristics of the data set is given by internal node, the branches represent decision rules, and the outcome is given by leaf node.

It contains mainly 2 nodes: -decision node & leaf node where the decision node is used in taking decisions and it has several branches as options and the leaf node is considered as the result node where no further decisions made on this node. It is a top-down technique also. It contains 2 types of decision tress one is Continuous variable decision tree and second is Categorial variable decision tree.

Steps in Decision Tree:

- Data Preprocessing Stage
- Fit Decision Tree Algorithm to Training Set
- Predicting Test Result
- Test Result Accuracy (Confusion Matrix Creation)
- Display Test Set Result.

4.6 RANDOM FOREST

Random Forest is the prevalent supervised technique. It is useful for mainly doing classification challenges and also regression challenges. RF is one of the classifiers which holds multiple decision trees in each subset of a assumed data set, and compute the typical value that improves prediction accuracy for the data set. Random forest does not depend on decision trees. Instead, it gets predictions from every tree and then forecasts the last result which is built upon polls of majority estimates. The more trees in the forest, the higher the accuracy and avoid over-fitting problems. It is based on ensemble technique concept, which combines multiple classifiers to solve complex problem and improves model performance.

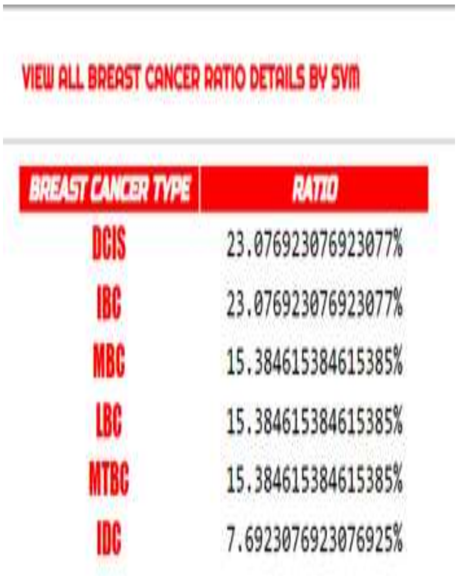
Steps in Random Forest:

- Select a random sample from a given data set.
- Algorithm will create a decision tree for each sample and you will get the result of the prediction for each decision tree.
- Next, you will vote on each prediction result.
- Finally, choose the prediction result with the most votes as the final result.

5. RESULTS AND DISCUSSION

This study summarizes distinctive machine learning models and data mining strategies in predicting the risk for breast most cancers. It gives comparative precis of learning strategies for the breast most cancer predictions. This evaluation aims to study which capabilities are maximum beneficial in predicting malignant or benign cancer and additionally form of cancer and their accuracies for it. This also allows in taking accurate prognosis remedy for the type of cancer. Performance for the proposed technique is given by considering the actual classification value and the predicted classification value. Ratio for predicting this cancer is calculated by Standardized detection ratio.Accuracy for the proposed system is evaluated by using confusion matrix acquired for the predictor/classifier taken.

The following figures depicts the ratios and accuracies calculated for different symptoms, causes, risk factors, stages and type of distinctive breast cancers taken for a dataset. The ratios and accuracies that are calculated by SVM algorithm, K-Nearest Neighbors Algorithm are given below. And also, a pie-chart is drawn for the patients having highest occurrence of this breast cancer risk.



VIEW ALL BREAST CANCER RATIO DETAILS BY SVM

BREAST CANCER TYPE	RATIO
DCIS	23.076923076923077%
IBC	23.076923076923077%
MBC	15.384615384615385%
LBC	15.384615384615385%
MTBC	15.384615384615385%
IDC	7.6923076923076925%

Fig.2 SVM Ratios for occurrence of most type of breast cancer

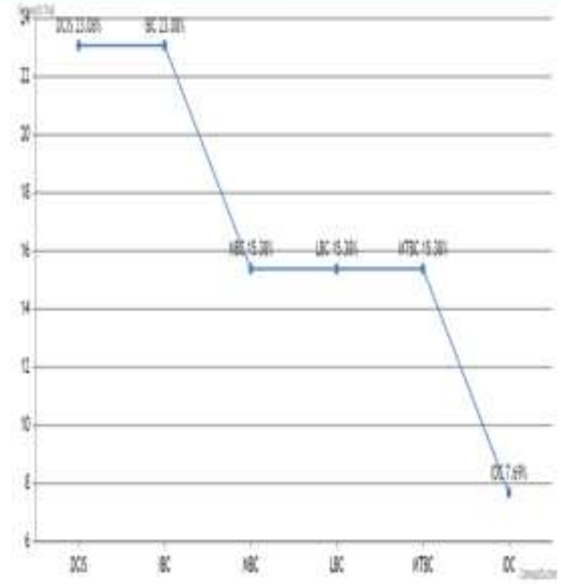


Fig.3 Accuracy results for different types of cancers

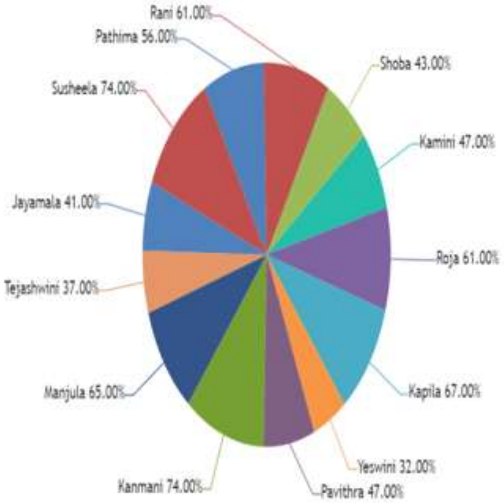
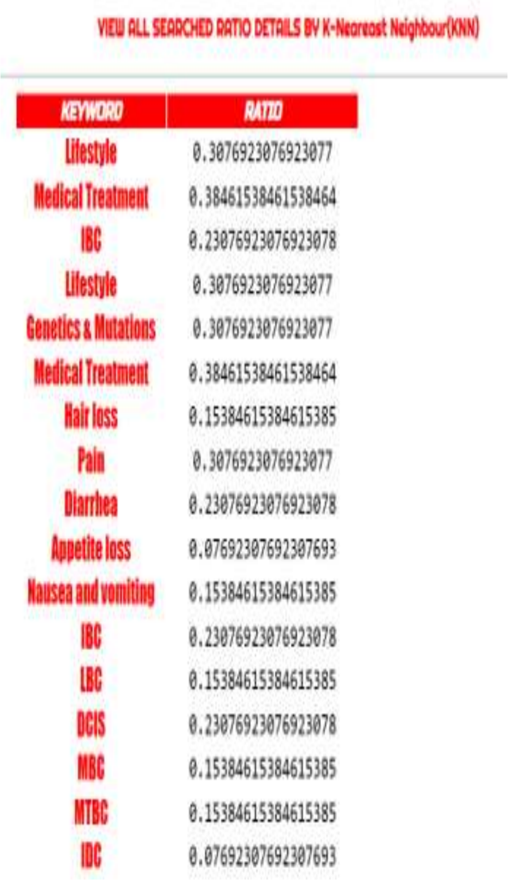


Fig.4 Accuracy results of patients having breast cancer

Vol 12, Issue 07, JULY /2021

ISSN NO:0377-9254



VIEW ALL SEARCHED RATIO DETAILS BY K-Nearest Neighbour(KNN)

KEYWORD	RATIO
Lifestyle	0.3076923076923077
Medical Treatment	0.38461538461538464
IBC	0.23076923076923078
Lifestyle	0.3076923076923077
Genetics & Mutations	0.3076923076923077
Medical Treatment	0.38461538461538464
Hair loss	0.15384615384615385
Pain	0.3076923076923077
Diarrhea	0.23076923076923078
Appetite loss	0.07692307692307693
Nausea and vomiting	0.15384615384615385
IBC	0.23076923076923078
LBC	0.15384615384615385
DCIS	0.23076923076923078
MBC	0.15384615384615385
MTBC	0.15384615384615385
IDC	0.07692307692307693

Fig.5 Ratios for Symptoms,risk factors and typesand causes of breast cancer

6. CONCLUSION

Breast most cancers if predicted at an early degree will save the lives of people both men and women. Finding the cause of the disease first is an important factor. So here we are taking all the machine learning models/techniques present and comparing one with other for detecting, the best one that gives good range of accuracy is considered as the best technique. We have taken the different types of datasets and also predicted various all factors one which each other. The suitable algorithm is able to classify whether the disease is in first stage, middle stage or advanced stage. It also tells or classifies cancers into malignant tumor or benign tumor. Early treatment is the best treatment for any kind of disease/cancer. Machine learning algorithms used for these types of research helps to reduce manual mistakes, lowers human mistakes and advances the existing system present in many ways.

In the future scope, work has to be still continued because still there are many issues to be solved by the researchers. The difficulty of inequality of good and negative records needs to be taken into consideration with the assistance of scientists, as it is able to result in biasness in the direction of good or bad prediction. The other crucial difficulty to resolve is the unfair quantity of different forms of cancer photos towards exaggerated portions in proper analyzing and predicting the breast most cancers.

REFERENCES

[1] Y.-S. Sun, Z. Zhao, Z.-N. Yang, F. Xu, H.-J. Lu, Z.-Y. Zhu, W. Shi, J. Jiang, P.-P. Yao, and H.-P. Zhu, ‘‘Risk factors and preventions of breast cancer,’’

[2] Priyanka Gandhi and Professor Shalini L “The analysis of machine learning procedures”VIT University,2018,pp.1-5.

[3] Y. Khourdifi and M. Bahaj, ‘‘Applyingmachine learning algorithms for breast cancer prediction and classification,’’ in Proc. Int. Conf. Electron., Control, Optim. Comput. Sci. (ICECOCS), Dec. 2018, pp. 1–5.

[4] ”Ultrasound characterisation of breast masses”, The Indian journal of radiology imaging by S. Gokhale., Vol. 19, pp. 242-249, 2009

[6]“Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis” by Wenbin Yue , Zidong Wang, 9 May 2018

[7] Y. Lu, J.-Y. Li, Y.-T. Su, and A.-A. Liu, ‘‘A review of breast cancer detection in medical images,’’ in Proc. IEEE Vis. Commun. Image Process. (VCIP), Dec. 2018, pp. 1–4.

[8] A R. Chaudhury, R. Iyer, K. K. Iychettira, and A. Sreedevi, “Diagnosis of invasive ductal carcinoma using image processing technique,” in 2011 International Conference on Image Information Processing, pp. 1– 6, IEEE.