iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** V **Month of publication:** May 2022

DOI: https://doi.org/10.22214/ijraset.2022.43055

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue V May 2022- Available at www.ijraset.com

Breast Cancer Detection Using Machine Learning Techniques

Sarthak Vyas[1], Abhinav Chauhan[2], Deepak Rana[3], Noman Ansari[4]
[1, 2, 3, 4]*Meerut Institute of Engineering and Technology(MIET), Affl. Dr. A. P. J. Abdul Kalam Technical University(AKTU) , Meerut, Uttar Pradesh 250002, India.*

***Abstract:** Cancer is one of the most prominent cause of fatalities around the world, accounting over 1 crore deaths in past year out of which 22.6% deaths were due to Breast cancer (BC). BC is the most common type of cancer among women, it accounts for 14.7 % of cancer cases in India. Multiple pieces of research have been conducted over the early detection of BC that can help begin treatment on time thus decreasing the mortality rate. Out of the total diagnosed, only about 86% are diagnosed correctly. Biopsy images of cells have a risk of false detection which may endanger the life of a person. There is a dire need of discovering new alternative methods that have an easy implementation with different data sets, are cost-effective, reliable, and safer, that can give an accurate prediction. This paper proposes a model combined with several Machine Learning algorithms(ML) that are Decision Trees, Artificial Neural Network, K-Nearest Neighbour, Support Vector Machine for an effective and accurate breast cancer diagnosis.*

Index Terms - Breast Cancer; Breast Cancer Detection; Machine Learning
Key Words: Breast Cancer (BC); Artificial neural networks (ANN); Wisconsin breast cancer data set.

I. INTRODUCTION

Breast cancer (BC) is one of the most prominent type of cancer among women all around the world, according to a research conducted by World Health Organization (WHO). BC is a leading causes of death among women all around the world. Breast cancer also has an exceedingly high rate of cancer fatalities in India which is around 14% and is the most common cancer among women. BC. affects about 5% of Indian women, but it affects about 12.5 percent of women in Europe and the United States. It confirms that women in Malaysia who have breast cancer present at a later stage of the disease than women in other countries. Breast cancer is in most of the cases easy to diagnose if any particular symptom appears. Some women with breast cancer, on the other hand, face no symptoms. Thus, periodic breast cancer screening is crucial for early detection.

As prognosis is so critical for long-term survival, early detection of breast cancer benefits early treatment and diagnosis. Because cancer can be detected, diagnosed, and treated only if detected early, the chance of death is reduced by early detection. It plays a vital role patient's survival. Delay in diagnosing cancer or detecting it at a later stage may lead to the spreading of disease and complications in treatment.

Cancer-related research done in the past on the effects of a late cancer diagnosis has found that it is very closely linked to the disease progressing to advanced stages, lowering the likelihood of saving the patient's life.

An analysis of 87 researchers found that female breast cancer patients who begin treatment within 90 days after the onset of symptoms had a considerably higher likelihood of surviving than those who wait more than 90 days.

Many earlier studies have found that detecting breast cancer in its early stages and starting the treatment on time increases the chances of survival by preventing malignant (Cancerous) cells from spreading throughout the body.

This paper's main contribution is a evaluation and study of the role of various machine learning approaches in breast cancer early detection.

Artificial intelligence (AI) and Machine Learning together can be implemented to improve breast cancer detection, while also avoiding overtreatment. Nonetheless, merging AI with Machine Learning (ML) approaches helps achieve accurate prediction and decision-making. For e.g., deciding whether or not the patient needs surgery based on the biopsy results for detecting breast cancer.

Mammograms are currently the most utilized test, they can give false positive (high-risk) results, which can lead to unnecessary biopsies and procedures. When surgery is performed to remove malignant cells, it is sometimes discovered that the cells are benign that are non-cancerous. This implies that the patient will be subjected to unnecessary, unpleasant, and a costly surgery.

M.L. Algorithms have a number of benefits, including their ability to perform well on healthcare-related datasets such as pictures, x-rays, and blood samples. Some strategies are better suited to small datasets, while others are best suited to large datasets. Noise can be an issue with some methods.

The research paper is arranged as follows: Section I introduces breast cancer and its diagnosis, Section II Contains a brief note on all the Machine Learning algorithms used for the detection, and a summary of previous work done in this field. IV & V Section concludes the paper.

II. METHODOLOGY

A. Boosting

Boosting strategies work in similar soul as stowing techniques: we assemble a group of prototypes which are amassed towards acquiring some solid student whose conduct is superior. Notwithstanding, dissimilar to stowing that targets lessening difference, supporting, a strategy that is used for linking together various weak learners in an exceptionally adaptative manner: each separate weak model in this grouping is fitted giving more reliability to the dataset which was previously missing. Naturally, each latest prototype spotlights one's endeavours especially for troublesome perceptions suitable till the current moment, therefore the user acquire, toward the finish of the interaction, a solid student with lower inclination (regardless of whether we can see that supporting can likewise diminish difference). Helping, such as packing, can be utilized for relapse as well concerning characterization issues.

Being primarily engaged at decreasing predisposition, the base models that are frequently considered for supporting are models with low fluctuation however high inclination. Here, utilizing two significant helping calculations: adaboost and inclination supporting.

These two meta-calculations vary on how they make and total the frail students during the successive interaction. Versatile helping refreshes the loads appended to every one of the preparation dataset perceptions though slope supporting updates the worth of these perceptions. This principal contrast comes from the way the two techniques attempt to take care of the enhancement issue of observing all that model that can be composed as a weighted number of powerless students.

B. Support Vector Machine

The purpose of the calculation by vector machine usually is to find a hyper-plane in a M-multi-layered space (M — the batch of elements) which distinctly characterizes given data of interest.

To quarantine the two given categories of data of interest, there are scores of thinkable hyperplanes that could be harvested. The objective here that we are trying to achieve is to observe a plane that has the most prominent edge, i.e., the most prominent distance between prominent sections of statistics of the two given classes. Expanding the border interval provides some aid so subsequent statistics focal points can be grouped together with greater reliability.

In the SVM calculation, we are hoping to expand the edge between the data of interest and the hyperplane. The misfortune work that augments the edge is pivot misfortune.

C. Random Forest Classifier

Random forest is an administered learning algorithm. It is an assortment of Decision Trees. Decision Tree is various levelled in nature in which nodes address specific circumstances on a specific arrangement of highlights, and branches split the decision towards the leaf nodes. Leaf decides the class marks. Decision Tree can be developed either by utilizing Recursive Partitioning or by Conditional Inference Tree. Recursive Partitioning is the bit-by-bit process by which a Decision Tree built by either parting or not parting every node. We can say that the tree is advanced by parting the source set into subsets in view of a property estimation test. The recursion ends if subset at a node has a gross similar value to the objective variable. Contingent Inference Tree is a factual based approach that involves non parametric tests as dividing models that is rectified for different testing to avoid over fitting. Random Forest is reasonable for high layered information displaying as it can deal with missing qualities, persistent, all out and parallel information however for very informational indexes, the size of the trees can take up a ton of memory. It can will quite often over-fit, so there is a need to tune the hyper-boundaries

RF(Random forest) algo. is utilized at the point of highest quality in the RF model.. RF constructs multiple DT’s utilizing random examples with substitutions to increase the efficiency of DTs. Every constructed tree group their perceptions, and greater part casts a ballot decision is picked. RF is utilized in the solo mode for surveying vicinities among information focuses.

The random forest approach is a bagging method where profound trees, fitted on bootstrap tests, are joined to create a result with lower difference. Be that as it may, random forests additionally utilize one more stunt to make the various fitted trees a piece less related with every others: while developing each tree, rather than just inspecting over the perceptions in the dataset to produce a bootstrap test, we likewise test over highlights and keep just a random subset of them to assemble the tree.

Classification Report

a. Train	precision	recall	f1-score	support
0	1.00	1.00	1.00	252
1	1.00	1.00	1.00	146
accuracy			1.00	398
macro avg	1.00	1.00	1.00	398
weighted avg	1.00	1.00	1.00	398

b. Test	precision	recall	f1-score	support
0	0.90	0.94	0.92	105
1	0.90	0.83	0.87	66
accuracy			0.90	171
macro avg	0.90	0.89	0.89	171
weighted avg	0.90	0.90	0.90	171

D. Ensemble Methods

Ensemble learning is an ML Technique where different models are prepared to overcome a similar issue and obtain to obtain improved results. The fundamental speculation is that when frail models are accurately joined, we can get more precise as well as vigorous models.

E. Bagging

While preparing a model, regardless assuming we are managing an order or a relapse issue, we get a capacity that takes an information, returns a result and that is characterized as for the preparation dataset. Because of the hypothetical variance of the preparation dataset (we remind that a dataset is a noticed example coming from a genuine obscure fundamental circulation), the fitted model is likewise dependent upon changeability: if another dataset had been noticed, we would have acquired an alternate model

Bagging is then basic: we need to fit a few free models and "normal" their forecasts to acquire a model with a lower variance. We cannot fit completely autonomous models since it would require a lot of information. Also, we depend on the given "inexact properties" of bootstrap tests to fit models that are autonomous.

Classification Report

a. Train	precision	recall	f1-score	support
0	1.00	1.00	1.00	252
1	1.00	1.00	1.00	146
accuracy			1.00	398
macro avg	1.00	1.00	1.00	398
weighted avg	1.00	1.00	1.00	398

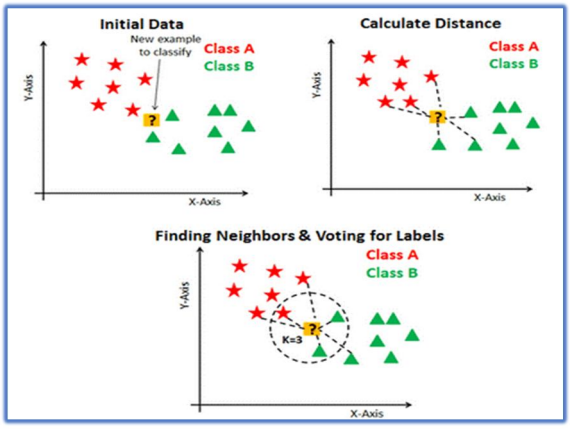
b. Test	precision	recall	f1-score	support
0	0.90	0.94	0.92	105
1	0.90	0.83	0.87	66
accuracy			0.90	171
macro avg	0.90	0.89	0.89	171
weighted avg	0.90	0.90	0.90	171

F. K Nearest Neighbour (KNN)

K-Nearest Neighbour is one of the easiest Machine Learning Algorithms. It can be used for Classification and Regression, but it is used for Classification only. This algorithm accepts the closeness between the new information and accessible cases and put the new case into the classification that is generally like the accessible classifications. KNN addresses both characterization and relapse assignments. In Classification technique, it arranges the articles considering the k closest getting ready models in the component space. The working norm behind KNN is that it accepts that the comparative data centres lie in same environmental components. It lessens the weight of building a model, changing a couple of limits, or building also assumptions. It gets the chance of proximity considering mathematical condition called as Euclidean distance, calculation of distance between two spots in a plane. Assume the two focuses in a plane are A (x0, y0) and B (x1, y1) then, at that point, the Euclidean distance between them is determined as follows:

$$\sqrt{(x0 - x1)^2 + (y0 - y1)^2}$$

Equation for Euclidean Distance



To figure out which of the K cases in the preparation dataset are like another information, a distance measure is utilized. For genuine esteemed input factors, the most well-known distance measure is the Euclidean distance.

Steps to be done during the K-NN calculation are as per the following:

1) Divide the information into preparing and test information.
2) Select a worth K.
3) Determine which distance work is to be utilized.
4) Choose an example from the test information that should be ordered and register the distance to its n preparing tests.
5) Sort the distances got and take the k-closest information tests.
6) Assign the test class to the class in view of the larger part vote of its k neighbours.

Important Tuning Parameters for KNN

- n_neighbours - The number of nearest neighbours K in the K-NN algorithm
- Weights - weight function used in predictions.

Following is the code snippet for the same: -

```
param_grid = {'n_neighbors':[3,4,5,6,7,8,9,10,11,12],
'weights': ['uniform', 'distance']}
knn = GridSearchCV(KNeighborsClassifier(), param_grid = param_grid, cv=5, scoring = 'f1_weighted')
knn.fit(std_data_train, train_y)
```

Final accuracy Score-
Train Data: 0.9849246231155779
Test Data: 0.9239766081871345

III. DECISION TREE

Decision Trees (DTs) are one of the most helpful administered learning calculations out there. Rather than solo realizing (where there is no result variable to direct the growing experience and information is investigated by calculations to track down designs), in administered learning your current information is as of now marked and you know which conduct you need to foresee in the new information you get. This is the sort of calculations that independent vehicles use to perceive walkers and items, or associations exploit to gauge client's lifetime esteem and their beat rates. Decision Trees are Machine Learning calculations that continuously partition informational collections into more modest information bunches considering an unmistakable element, until they arrive at sets that are sufficiently little to be portrayed by some mark. They expect that you have information that is marked (labelled with at least one names, like the plant name in pictures of plants), so they attempt to name new information considering that information.
Decision Trees calculations are amazing to tackle arrangement (where machines sort information into classes, like regardless of whether an email is spam) and relapse (where machines anticipate values, like a property cost) issues. Relapse Trees are utilized when the reliant variable is consistent or quantitative (for example if we have any desire to gauge the likelihood that a client will default on an advance), and Classification Trees are utilized when the reliant variable is clear cut or subjective (for example to assess the blood classification of an individual). The significance of DTs depends on the way that they have bunches of utilizations. Being one of the involved calculations in Machine Learning, they are applied to various functionalities in a few enterprises

Important Tuning Parameters for DT

- criterion - measure for quality of a split
- max_depth - The maximum depth of the tree.
- max_leaf_nodes - Number of features to take for the best split
- min_sample_leaf - The minimum samples required to become a leaf node. This may have effect of smoothing the model.
- min_sample_split - The minimum number of required samples to split internal node.

Following is the code snippet for the same: -

```
param_grid = {'max_depth': np.arange(3, 5),
          'max_features': np.arange(3,5)}
 tree = GridSearchCV(DecisionTreeClassifier(), param_grid, cv = 5)
 tree.fit( train_x, train_y )
```

The overall accuracy score for the Train Data is: 0.957286432160804
The overall accuracy score for the Test Data is: 0.8947368421052632

Classification Report

a. Train

	precision	recall	f1-score	support
0	0.96	0.97	0.97	252
1	0.94	0.94	0.94	146
accuracy			0.96	398
macro avg	0.95	0.95	0.95	398
weighted avg	0.96	0.96	0.96	398

b.Test

	precision	recall	f1-score	support
0	0.88	0.95	0.92	105
1	0.91	0.80	0.85	66
accuracy			**0.89**	171
macro avg	0.90	0.88	0.89	171
weighted avg	0.90	0.89	0.89	171

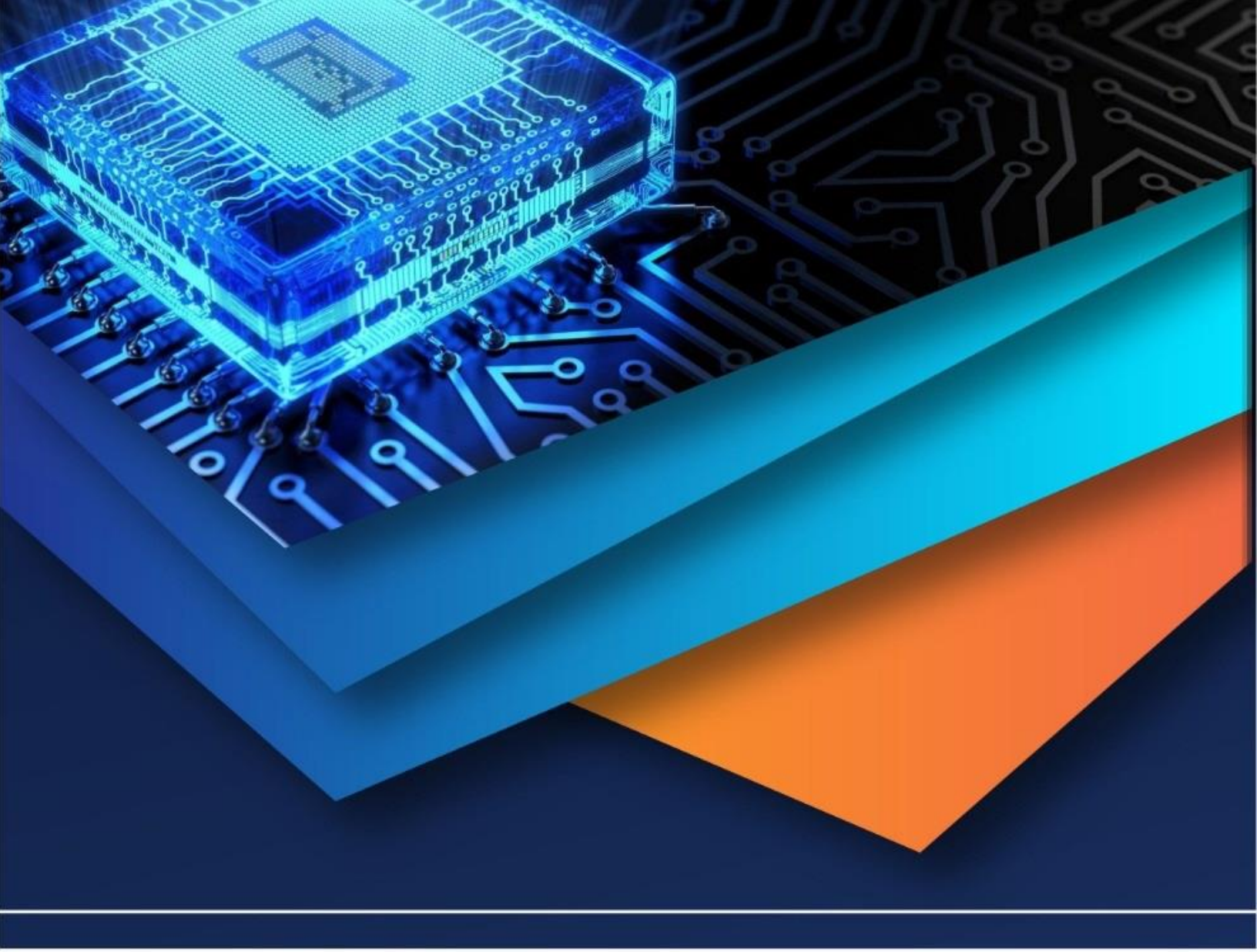
IV. CONCLUSION

Our work is principally centred around the advancement of predictive models to accomplish great precision in foreseeing legitimate disease results utilizing supervised machine learning techniques. The analysis of the outcomes connotes that the mix of multidimensional data along with different classification, feature selection, and dimensionality reduction techniques can give favourable tools for inference in this domain. Further research in this field ought to be done for the better execution of the grouping procedures so it can foresee more factors.

The dataset contains 32 characteristics features that contributes to bring down the multi-dimensional large dataset to a only a few necessary dimensions. The relative multitude of the three applied algorithms the K-Nearest-Neighbour, the Support Vector Machine (SVM), and Logistic Regression, SupportVector provides the most noteworthy exactness of 92.7% when contrasted with different calculations. In this way, we suggest that SVM is the most appropriate calculation for determining the expectation of Breast Cancer Occurrence with complex datasets.

BIBLIOGRAPHY

[1] P. Boix-Montesinos, M.J. Vicent,, A. Armiñán, M. Orzáez, P.M. Soriano-Teruel. The past, the present, and the future of breast cancer models for nanomedicine development Adv. Drug Deliv. Rev., 173 (2021), pp. 306-330
[2] S.V, U.R. Acharya, J.H. Tan,. Sree, , E.Y.K. Ng. Thermography based BCD using texture features and SVM J. Med. Syst., 36 (3) (2012), pp. 1503-1510, 10.1007/s10916-010-9611-z
[3] Nidhal Kamel Taha El-Omari, Vincent O. Efficient Feature Selection and ML Algorithm for Accurate Diagnostics. https://ojs.bilpublishing.com/index.php/jcs.
[4] Yubiao Jin , Lingling Zhuang , Xing Sun. Evaluation of whole axillary status with lymphatic contrast-enhanced ultrasound in patients with breast cancer. 10.1007/s00330-021-08100-8
[5] S.T. Selvi , J. Dheeba. Swarm optimized neural network system for classification of microcalcification in mammogram. 10.1007/s10916-011-9781-3
[6] M.A. T.-S. Kim Al-Antari. Evaluation of deep learning detection and classification towards computer-aided diagnosis of breast lesions
[7] M.M. Freire,F. Soares,, J. Seabra. Classification of breast masses on contrast-enhanced magnetic resonance images through log detrended fluctuation cumulant-based multifractal analysis 10.1109/JSYST.2013.2284101,IEEE Syst. J, 8 (2014).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY