ZKG INTERNATIONAL

ISSN: 2366-1313

AN EFFICIENT BREAST CANCER DIAGNOSIS USING MACHINE LEARNING

K. NAGESWARA RAO[1], Y. GEETHA KRISHNA[2], A. AKSHITHA[3], CH.SAI SIVA KUMAR[4], D.KOUSHIK RAGHAV[5].

[1] Associate Professor, CSE,Chalapathi Institute of Technology,Guntur, India

[2]UG Student,CSE,Chalapathi Institute of Technology,Guntur, India

[3]UG Student,CSE,Chalapathi Institute of Technology,Guntur, India

[4]UG Student,CSE,Chalapathi Institute of Technology,Guntur, India

[5]UG Student,CSE,Chalapathi Institute of Technology,Guntur, India

ABSTRACT: Breast Cancer is one of the most exquisite and internecine disease among all of the diseases in medical science. It is one of the crucial reasons of death among the females all over the world. We present a novel modality for the diagnosis of breast cancer and introduces with the Support Vector Machine and K Nearest Neighbors which are the supervised machine learning techniques for breast cancer detection by training its attributes.

The proposed system uses confusion matrix to get an accurate outcome. The breast cancer termed as Wisconsin breast cancer diagnosis data set is taken from UCI machine learning repository. The performance of the proposed system is appraised considering accuracy, sensitivity, specificity, false discovery rate, false omission rate and Matthews correlation coefficient. The approach provides better result both for training and testing. Furthermore, the techniques achieved the accuracy of 96.57% and 96.14% by Support Vector Machine and K-Nearest Neighbors individually along with the specificity of 96.65% and 97.31% in testing phase.

1. INTRODUCTION

A recent statistic from World Cancer Research Fund International (WCRFI) shows that breast cancer is the most widely recognized cancer in ladies around the world. There are almost 1.7 million new cases which are detected in 2012 that signifies around 12 percent of altogether new cancer cases and 25 percent of all cancers in ladies. Breast cancer outperformed the position of fifth for the reason of death in ladies. In numerous nations with advanced technology in medical science, the 5-year survival rate of initial phase breast cancer is 80–90%, dropping to 24% for breast cancer analyzed at progressive phase. Breast cancer cells are found in the tissues of the breast. In very recent years there are numerous modalities have been developed for the diagnosis of breast cancer.

In biopsy testing, the biopsy is occupied from the tissues of the breast. The test provides higher accurate result but the procedure to take the biopsy from breast is very painful and pathetic. So, the most of the patients are not intrigued with this test. A mammogram is the most widely used technique for the detection of breast cancer which provides the 2D projection images of the breast. There are two kinds of mammography techniques that are widely used. These are the screen-film mammography (SFM) and digital mammography (DM). SFM is utilized in asymptomatic ladies' breast i.e., problem free breast where it receipts two sights of both breasts. The time duration of screening mammograms (conventional

ISSN: 2366-1313

mammography) is around 20 minutes. It cannot detect benign cancer properly. Digital Mammography overcomes the problem of screening mammograms. It is related with computer system i.e.; the data of digital mammography are kept in a computer. In DM, the images are taken and image processing methods are applied to improving the quality of the image. DM performs better in case of misdiagnosed cancer samples. Magnetic Reasoning Imaging (MRI) is another most common technique for the diagnosis of breast cancer. Although MRI is a very complex test, sometimes it miscues some cancer whereas mammograms may detect. MRI is used for the ladies who have attacked in breast cancer to define the real size of the breast and find another disease in the breast. It provides an excellent result for 3D images and displays the dynamic functionality. MRI is done with the help of contrast-enhanced imaging. Although several techniques have been introduced, none of the techniques are able to provide an accurate and reliable outcome. All of the modalities are involved with doctors or physicians or other medical staffs. So, a system which can operate without any medical equipment's and medical staffs may lead to an appropriate solution.

We introduce an innovative modality to classify the input attributes according to the presence or absence of benign or malignant types of breast cancer. We have used two supervised machine learning techniques termed as Support Vector Machine and K-Nearest Neighbours which are related with learning computations that investigate data utilized for regression analysis and classification to identify the breast cancer.

2. LITERATURE SURVEY

"Comparative study of machine learning algorithms for breast cancer detection and diagnosis" by **Dana Bazazeh, R. Shubair – Year – 2016**

Breast cancer is one of the most widespread diseases among women in the UAE and worldwide. Correct and early diagnosis is an extremely important step in rehabilitation and treatment. However, it is not an easy one due to several uncertainties in detection using mammograms. Machine Learning (ML) techniques can be used to develop tools for physicians that can be used as an effective mechanism for early detection and diagnosis of breast cancer which will greatly enhance the survival rate of patients. This paper compares three of the most popular ML techniques commonly used for breast cancer detection and diagnosis, namely Support Vector Machine (SVM), Random Forest (RF) and Bayesian Networks (BN). The Wisconsin original breast cancer data set was used as a training set to evaluate and compare the performance of the three ML classifiers in terms of key parameters such as accuracy, recall, precision and area of ROC. The results obtained in this paper provide an overview of the state of art ML techniques for breast cancer detection.

"Comparison of Machine Learning Algorithms in Breast Cancer Diagnosis Using the Coimbra Dataset" by **Austria, Yvonne Denice; Lalata, Joel Anne ; Maria, Luis Benedict S. Jr. Goh, Jenn Ern ; Goh, Mark Liang Theng ; Vicente, Henrique – Year-2019**

Breast cancer is the second most common type of cancer worldwide, and its diagnosis is still a challenge in the field of medicine. In recent years, machine learning algorithms have been proposed as a powerful tool to improve the accuracy and efficiency of breast cancer diagnosis. This study compares the performance of four machine learning algorithms, namely, k-Nearest Neighbors (k-NN), Decision Tree (DT), Support Vector Machine (SVM), and Random Forest (RF), in the diagnosis of breast cancer using the Coimbra dataset. The Coimbra dataset consists of clinical and

ISSN: 2366-1313

demographic data of patients with breast cancer, collected from a hospital in Coimbra, Portugal. The dataset includes 116 instances with 9 attributes, including age, BMI, and various blood markers. The study used 10-fold cross-validation to evaluate the performance of each algorithm. The results show that all four algorithms achieved high accuracy in the classification task, with SVM achieving the highest accuracy of 97.4%. Furthermore, the study performed feature selection analysis and identified age, glucose, and BMI as the most important features in the classification task. This study demonstrates the effectiveness of machine learning algorithms in breast cancer diagnosis and provides insights into the most important features for the classification task.

3. EXISTING SYSTEM

Traditionally, the diagnosis of breast cancer and the classification of the cancer as malignant or benign was done by various medical procedures like:

• **Breast exam** – The doctor would check the breasts and lymph nodes in the armpits to check if there are any lumps or abnormalities.

• **Mammogram** – These are like X-ray of the breast. They are used to check whether there is breast cancer or not. If any issues are found, the doctor may ask the patient to take a diagnostic mammogram to check for further abnormalities.

• **Breast ultrasound** - Ultrasound uses sound waves to produce images to determine whether a new breast lump is a solid mass or a fluid-filled cyst.

• **Removing a sample of breast cells for testing (biopsy)** – This is probably the only definite way of checking if a patient has breast cancer. The doctor uses a specialized needle device guided by the X-ray or any other test to take samples of tissues from the area to be checked.

• **MRI of the breasts** - An MRI machine uses a magnet and radio waves to create pictures to see the interiors of the breast tissues. Blood tests, CT scans and PET scans are also done to check for breast cancer.

Blood tests, CT scans and PET scans are also done to check for breast cancer.

Disadvantages:

• Time consuming.
• Not completely accurate.
• Very expensive

4. PROPOSED SYSTEM

In the proposed system we plan on using existing data of breast cancer patients which has been collected for a number of years and run different machine learning algorithms on them. These algorithms will analyze the data from the datasets to predict whether the patient has breast cancer or not and it will also tell us if the cancer is malignant or benign. It is done by taking the patient's data and mapping it with the dataset and checking whether there are any patterns found with the data. If a patient has breast cancer, then instead of taking more tests to check whether the cancer is malignant or benign, ML can be used to predict the case based on the huge amount of data on breast cancer.

This proposed system helps the patients as it reduces the amount of money they need to spend just for the diagnosis. Also, if the tumour is benign, then it is not cancerous, and the patient doesn't need to go through any of the other tests. This saves a lot of time as well.

Advantages:

• Reduces costs for medical tests.
• Does not take huge amount of time.
• Intelligent way of using available data.
• Accurate

We have proposed a model using Kernel Support Vector Machine and K-Nearest Neighbors which is implemented in a high configuration computer.

ISSN: 2366-1313

We have utilized 629 (90%) instances of total data for training both in Support Vector Machine and K-Nearest Neighbors individually. The remaining 70 (10%) instances used for testing both in SVM and K-NN individually. The graphical representation of the confusion matrix for each modality is illustrated.

5. SYSTEM ARCHITECTURE

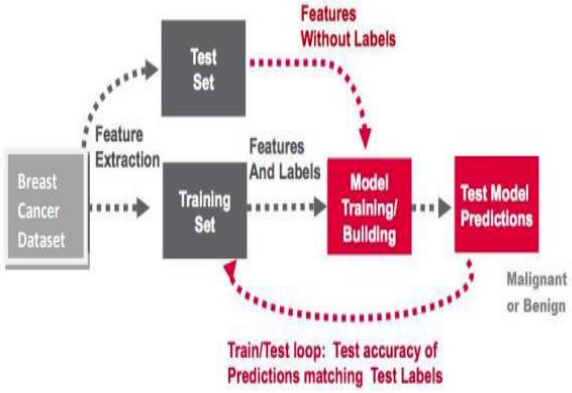

Fig 3: System Architecture

6. IMPLEMENTATION

The steps followed to do this project are:

1. Collection of datasets.
2. Understanding features of dataset.
3. Pre-processing the data.
4. Split data into training dataset and testing dataset.
5. Apply ML algorithms to dataset to predict the breast cancer.
6. Improving results.

Collection of datasets

The Wisconsin Diagnostic Breast Cancer (WDBC) dataset which can be found in University of California, Irvin's Machine Learning Dataset Repository will be used for this project. All ML algorithms will be performed on this dataset. The features which are in the dataset were computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. These features describe different characteristics of the cell nuclei found in the image. There are 569 data points in the dataset: 212 for Malignant and 357 for Benign.

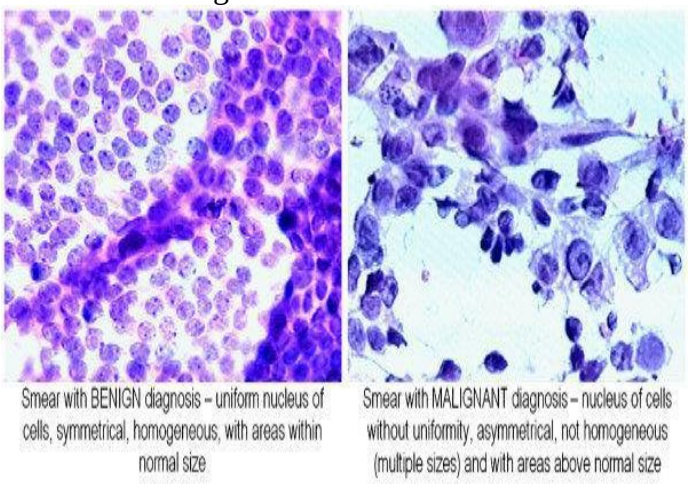

Fig 4 Digitized images of FNA (a) Benign (b) Malignant

The dataset contains these features:

- Radius
- Texture
- Perimeter
- Area
- Smoothness
- Compactness
- Concavity
- Concave Points
- Symmetry
- Fractal Dimension

ISSN: 2366-1313

Radius	Mean of distances from center to points on the perimeter
Texture	Standard deviation of gray-scale values
Perimeter	The total distance between the snake points constitutes the nuclear perimeter.
Area	Number of pixel on the interior of the snake and adding one-half of the pixel in the perimeter
Smoothness	Local variation in radius length, quantified by measuring the difference between the length of a radial line and the mean length of lines surrounding it.
Compactness	Perimeter ^2 / area
Concavity	Severity of concave portions of the contour
Concave points	Number of concave portions of the contour
Symmetry	The length difference between lines perpendicular to the major axis to the cell boundary in both directions.
Fractal dimension	Coastline approximation. A higher value corresponds to a less regular contour and thus to a higher probability of malignancy

Fig : Description Of Features Used In The Dataset

Each of these features have information on mean, standard error, and "worst" or largest (mean of the three largest values) computed. Hence, the dataset has a total of 30 features.

The breast cancer named as Wisconsin Breast Cancer (WBC) data set is retrieved from UCI machine learning repository dataset. This dataset comprises of 699 instances, where the cases are labelled as either benign or malignant and 458 (65.50%) of the cases are benign and 241 (34.50%) are malignant. The dataset is partitioned into two classes 2 and 4, where 2 denotes the benign class and 4 denotes the malignant class. The dataset has 11 features that are Clump Thickness(x1), Uniformity of Cell Size(x2), Uniformity of Cell Shape(x3), Marginal Adhesion(x4), Single Epithelial Cell Size(x5), Bare Nuclei(x6), Bland Chromatin(x7), Normal Nuclei(x8), Mitoses(x9) except sample code number and class.

The Performance Measure Indices

The performance of machine learning techniques is measured in terms of some performance measure indices. A confusion matrix for actual and predicted class is formed comprising of TP, FP, TN, and FN to evaluate the parameter. The significance of the terms is given below. The performance of the proposed system is measured by the following formulas:

$$Accuracy\ (Acc) = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (2)$$

$$Sensitivity\ (Sen) = \frac{TP}{(TP+FN)} \quad (3)$$

$$Specificity\ (Spec) = \frac{TN}{(TN+FP)} \quad (4)$$

$$False\ Discovery\ Rate\ (FDR) = \frac{FP}{(FP+TP)} \quad (5)$$

$$False\ Omission\ Rate\ (FOR) = \frac{FN}{(FN+TN)} \quad (6)$$

$$Matthews\ Correlation\ Coefficient\ (MCC) = \frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (7)$$

Pre-processing the data.

Pre-processing the data is an important step in data analysis and machine learning. Here are some common pre-processing steps:

1. Data cleaning: Removing or correcting missing, duplicate, or incorrect data.
2. Data normalization: Scaling data to a consistent range, such as between 0 and 1 or -1 and 1, to improve the performance of some machine learning algorithms.
3. Data encoding: Converting categorical data into numerical values that can be used by machine learning algorithms.
4. Feature selection: Choosing the most relevant features or variables to include in the analysis or model.

ISSN: 2366-1313

5. Feature extraction: Creating new features or variables from the existing ones to improve the performance of the analysis or model.
6. Data transformation: Applying mathematical or statistical operations to the data, such as log transformation or standardization, to improve the performance of the analysis or model.
7. Data reduction: Reducing the size of the data by selecting a representative subset of the observations or variables, to improve the performance of the analysis or model.

Split data into training dataset and testing dataset.

Splitting the data into training and testing datasets is a crucial step in machine learning to evaluate the performance of the model. The training dataset is used to train the model, while the testing dataset is used to evaluate how well the model generalizes to new, unseen data.

Here are the general steps to split the data into training and testing datasets:

1. First, shuffle the data randomly to remove any inherent ordering or biases in the data.
2. Decide on a ratio for splitting the data into training and testing datasets. A common ratio is 80:20, where 80% of the data is used for training and 20% is used for testing.
3. Use a function or library to split the data into training and testing datasets. For example, in Python, you can use the train_ test_ split function from the scikit-learn library to split the data.
4. Once the data is split, you can use the training dataset to train the model and the testing dataset to evaluate its performance.

It's important to note that the split should be random to avoid any bias, and the same split should be used consistently for all evaluations of the model to ensure comparability.

Apply ML algorithms to dataset to predict the breast cancer.

There are several machine learning algorithms that can be applied to predict breast cancer based on a dataset of relevant features. The algorithms used in this project are.

1. Support Vector Machines (SVM): A supervised learning algorithm that can be used for classification or regression. SVM tries to find a hyper plane in a high-dimensional space that separates the classes.
2. k-Nearest Neighbors (k-NN): A non-parametric algorithm that can be used for classification or regression. k-NN tries to classify a new data point based on the class of its k nearest neighbors in the training dataset.

Improving results.

To improve the results of a machine learning model for predicting breast cancer, here are some possible strategies:

Feature engineering: The quality and relevance of the features used as input to the model can significantly impact the model's performance. Feature engineering involves selecting the most relevant features, creating new features, and transforming existing features to improve their relevance and effectiveness in predicting the target variable.

Hyper parameter tuning: Most machine learning algorithms have several hyper parameters that can be adjusted to optimize performance. Hyper parameter tuning involves experimenting with different values of the hyper parameters to find the combination that produces the best results.

Ensemble methods: Combining the predictions of multiple models, either through bagging, boosting, or stacking techniques, can often improve the performance of the final model.

Data augmentation: Increasing the size and diversity of the training data can help the model learn more patterns and generalize better to new data. Techniques such as oversampling, under sampling, and synthetic

ISSN: 2366-1313

data generation can be used to augment the training data.

7. SCREEN SHORT

```
Accuracy score of train KNN
96.25550660792952
Accuracy score of test KNN
97.55102040816327
[[160   4]
 [  2  79]]
Accuracy score of train KNN
96.25550660792952
Accuracy score of test KNN
97.55102040816327
[[160   4]
 [  2  79]]
Accuracy score of train KNN
96.25550660792952
Accuracy score of test KNN
97.55102040816327
[[160   4]
 [  2  79]]
Accuracy score of train KNN
96.25550660792952
Accuracy score of test KNN
97.55102040816327
[[160   4]
 [  3  78]]
Accuracy score of train KNN
96.0352422907489
Accuracy score of test KNN
97.14285714285714
[[160   4]
 [  4  77]]
Accuracy score of train SVM
96.47577092511013
Accuracy score of test SVM
96.73469387755102
```

8. CONCLUSION

Breast cancer diagnosis is very significant in the area of Medicare and Biomedical. In this project we focused on building a classifier which aims at predicting the most severe cancer known as breast cancer. Breast cancer is a remarkably risky disease that causes a lot of death for numerous ladies all over the world. So, early detection of this cancer can save a lot of valuable life.

We proposed a model that predict the breast cancer based on Support Vector Machine and K-Nearest Neighbours. The SVM has been implemented by the Python to be the most effective in classifying the diagnostic data set into the two classes in view of the seriousness of the cancer. We end up with an accuracy of 99.68% in SVM in training phase. The proposed model will be very helpful for the medical staffs as well as general people.

The classifier obtained by supervised machine learning techniques will be very supportive in the
field of medical disorders and proper diagnosing.

FUTURE SCOPE

Our study proposed a breast cancer diagnosis model based on SVM and KNN. Our study proposed model outperformed other models .Finally ,our study is expected to improve health care systems and help reduce the breast cancer risks for individuals. Our study focused on a small set of population; thus the result may not be generalized for wider cases. A future study should consider other clinical datasets, diagnosis model and feature selection methods..

In the future, we can also use a dataset to predict the re-occurrence of breast cancer after a surgery or chemotherapy session. Artificial Neural Networks can be applied to make the diagnosis better and smarter. Accuracy can be increased by selecting better features.

REFERENCES

1. Breast cancer statistics. [Online]. Available: http://www.wcrf.org/int/cancer-facts-figures/dataspecific-
cancers/breast-cancer-statistics, accessed on: Aug. 25, 2017.

2. Arbab Masood Ahmad, Gul Muhammad, Khan, S. Ali Mahmud, Julian F. Miller, “Breast Cancer Detection Using Cartesian Genetic Programming evolved Artificial Neural Networks,” Philadelphia, Pennsylvania, USA, GECCO’12, July 7–11, 2012.

3. Ahmad Taher Azar, Shaimaa Ahmed El-Said, “Probabilistic neural network for breast cancer
classification,” Neural Computing and Applications, Springer, vol. 23, pp.1737-1751, 2013.

ISSN: 2366-1313

4. Warner E, Messersmith H, Causer P et al, "Systematic review: using magnetic resonance imaging to screen women at high risk for breast cancer," Annals of Internal Medicine,148(9):671–679,06 May, 2008.
5. Emina Alicˇkovic´, Abdul hamit Subasi, "Breast cancer diagnosis using GA feature selection and Rotation Forest" Neural Computing and Applications, Springer, Volume 28, issue 4, pp 753–763, April 2017.
6. Fadzil Ahmad, Nor Ashidi Mat Isa, Zakaria Hussain, Siti Noraini Sulaimon, "A genetic algorithm-based multi-objective optimization of an artificial neural network classifier for breast cancer diagnosis", Neural Computing and Applications, Springer, Volume 23, Issue 5, pp 1427–1435 ,October 2013.
7. M. K. Hasan, M. M. Islam and M. M. A. Hashem, "Mathematical model development to detect breast cancer using multigene genetic programming," 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), Dhaka, pp. 574-579, 2016.
8. H. AttyaLafta, N. Khadim Ayoob and A. A. Hussein, "Breast cancer diagnosis using genetic algorithm for training feed forward back propagation," 2017 Annual Conference on New Trends in Information & Communications Technology Applications (NTICT), Baghdad, pp. 144- 149, 2017.
9. D. Bazazeh and R. Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), Ras Al Khaimah, pp. 1-4, 2016.
10. Gareth James and Daniela Witten and Trevor Hastie and Robert Tibshirani, An Introduction to Statistical Learning, 1st ed., 2013.
11. Breast Cancer Wisconsin (Original) Data Set, [Online]. Available: https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer wisconsin/breastcancer-wisconsin.data, accessed on: Aug. 25, 2017.
12. Ahmad Taher Azar, Shaimaa Ahmed, El-Said ,"Performance analysis of support vector machines classifiers in breast cancer mammography recognition" Neural Computing and Applications, Springer, Volume 24, Issue 5, pp 1163–1177, April 2014.
13. Austria, Y.D.; Lalata, J.-A.; Maria, L.B.S., Jr
14. Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science, 83, 1064-1069.
15. Han, H., Wang, W., & Mao, B. (2017). Improved hybrid model combining ICA feature extraction and ANN classification for breast cancer diagnosis. Computer Methods and Programs in Biomedicine, 140, 105-113.