Artificial Intelligence Based Implementation Approach for Prediction of Breast Cancer Masses

Tirth Vishalbhai Dave
Graduate Student
School of Computer Science and Engineering
Vellore Institute of Technology Chennai
Chennai, India
tirthdave1508@gmail.com

Surendiran Balasubramanian
Associate Professor
Dept. of Computer Science and Engineering
National Institute of Technology Puducherry
Karaikal, India
surendiran@nitpy.ac.in

Vallidevi Krishnamurthy
Associate Professor
School of Computer Science and Engineering
Vellore Institute of Technology Chennai
Chennai, India
vallidevi.k@vit.ac.in

Veeraraghavan Krishnamurthy
Specialist Gastroenterology
NMC Specialty hospital
Al Ain, United Arab Emirates
drveeraraghavangastro@gmail.com

Sangeetha Vilvanathan
Assistant Director, Central Leprosy
Teaching & Research Institute, Ministry of Health & Family Welfare (Govt. of India),
Chengalpattu Tamil Nadu, India 0000-0002-4330-34

ABSTRACT

Breast cancer is the common disease women face in this digital era. Based on the shape, size and density of the mammograms, benign and cancerous masses can be differentiated. Applications of machine learning in breast cancer are explored by focusing on predicting the possibility of a person having breast cancer. A few models are implemented in this chapter and a hybrid model named Voter Model is also implemented to have a better result. On an average the Voter model produces the results with an accuracy of 99.7%.

Keywords— Machine Learning, Breast Cancer Prediction, Voter Model

I. INTRODUCTION

According to a report on breast cancer [1] 'Breast cancer is the most common cancer among women in the United States. The most notable characteristic of the descriptive epidemiology of breast cancer in recent years is perhaps the rapidly increasing incidence rates in developing countries. There are various research works that discusses about the classification of masses as benign or malignant [13]. In this proposed work, a hybrid prediction model named VoterModel for predicting the breast cancer based on statistical values is explained. The statistical estimates for breast cancer in the United States for 2021 are mentioned further. An estimated 268,600 new cases of invasive breast cancer are expected to be diagnosed in women in the U.S. About 62,930 new cases of non-invasive (in situ) breast cancer will be diagnosed. For women in the U.S., breast cancer death rates are higher than those for any other cancer, besides lung cancer. In recent analysis, it has been noted that, across the world, every 3 minutes a woman gets diagnosed with breast cancer. Also, every 13 minutes a woman dies from the same breast cancer disease.

In this chapter, machine learning concepts are used to predict Breast Cancer, using the dataset from The University of Wisconsin Breast CancerDiagnosis Dataset (WBCD)[2] Section 2 describes the breast cancer dataset used and the features of a cancer cell taken into consideration. Section 3 talks about the algorithms use in breast cancer prediction. Section 4 gives an elaborate discussion on the factors taken into consideration while building various models. This Section 4 also gives the accuracies obtained at the end. Followed by this section are the references. The section 5 is the survey based on deep learning based algorithms.

II. DATASET DESCRIPTION

The data collected so far can be classified into two groups: benign and malignant cases; 569 total cases, 357 classified as benign and 212 as malignant. The data being used was found at the UC Irvine Machine Learning Repository. The features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. The following features were considered to build the models. Radius (mean of distances from center to points on the perimeter), texture (standard deviation of gray-scale values), perimeter, area, smoothness (local variation in radius lengths), compactness, concavity (severity of concave portions of the contour), Concave points (number of concave portions of the contour), symmetry, fractal dimension.

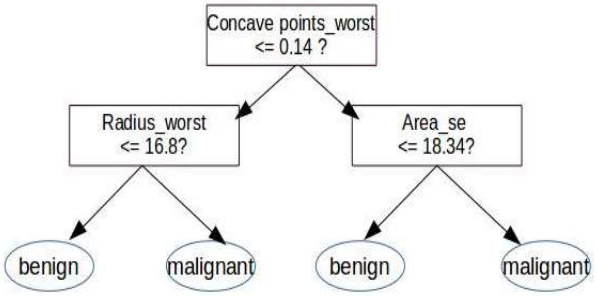

Figure 1: Decision Tree Classifier

III. Machine Learning Based Implementation Details

Python libraries such as Scikit-learn, Matplotlib, Keras were used for the models that were built.

A. Decision Tree Model

A decision tree [3] is a flowchart-like structure in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from the root to a leaf represent classification rules. Fig. 1 shows a part of the decision tree classifier built for breast cancer. The series of questions are answered by test data and accordingly a particular branch of the tree is chosen to proceed. The final leaf node arrived at; represents the class the sample belongs to. Hence each sample is classified as benign or malignant.

B. Random Forest

A random forest [4] is a collection of several decision trees. This provides a more stable and accurate prediction. Random Forest adds additional randomness to the model while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model. Decision trees might suffer from over-fitting [5]. Random Forest prevents overfitting by creating random subsets of the features and building smaller trees using these subsets. Figure 2 depicts three decision trees. Tree 1 and Tree 3 classify a test sample as malignant, while Tree 2 classifies as benign. Random forest decides based on most votes, hence predicts the sample to be malignant.

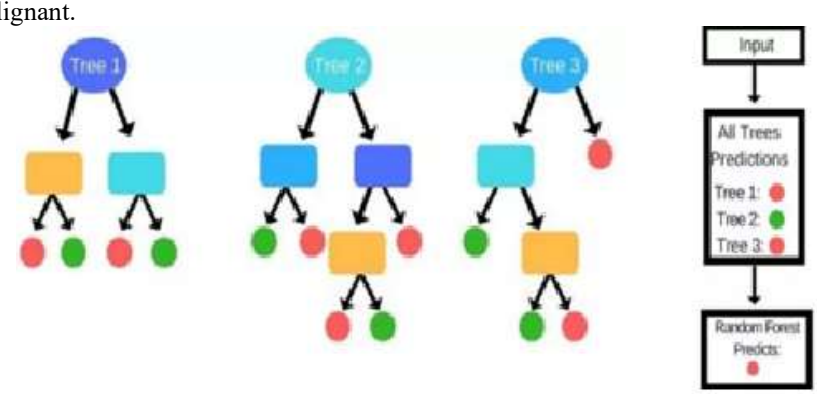

Figure 2: Random Forest Classifier

Over-fitting refers to a model that fits the training data too well. This occurs when a model learns every detail and noise in the training data to the extent that it shows very poor performance on unseen data. This is because the noise or random fluctuations in the training data is picked up and learned as concepts by the model. But these

concepts might not apply to new data and the model is no longer able to generalize.

C. Extra Tree

Extra tree classifiers obtained by randomizing the random forest further. Each tree is trained using the whole learning sample (rather than a bootstrap sample), and the top-down splitting in the tree learner is randomized. Instead of computing the locally optimal cut-point for each feature under consideration, a random cut-point is selected. Then, of all the randomly generated splits, the split that yields the highest score is chosen to split the node.

D. Support Vector Machine

Each data item in the support vector machine [6] is plotted as a point in n-dimensional space (n is number of features) with the value of each feature being the value of a particular coordinate. Classification is performed by finding the hyper-plane that differentiates the two classes very well. Linear kernel finds a linear hyperplane to classify the samples.

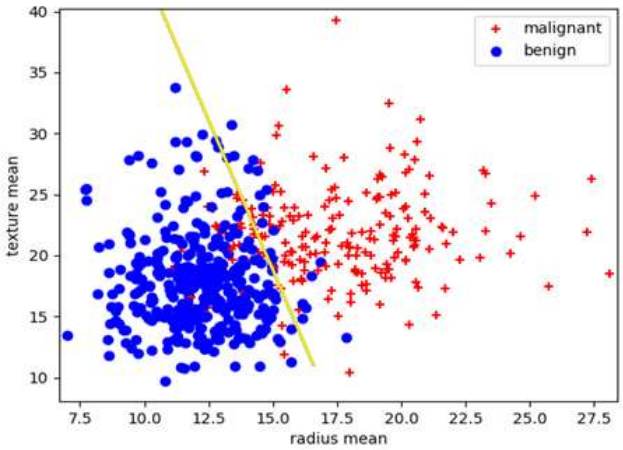

Figure 3: Support Vector Machine Classifier

Figure 3 shows an SVM classifier trained with features radius and texture. The red '+' represent malignant samples, while the blue circles represent benign samples. SVM classifier identifies the best hyperplane that classifies the data into their classes. This is represented by the yellow line. The model should consider accuracy as well as aim to maximize margin from samples to prevent overfitting.

E. Logistic Regression

Logistic regression algorithm uses a linear equation with independent predictors to predict a value. The predicted value can be anywhere between negative infinity to positive infinity. Logistic regression produces an output which is a class variable, i.e 0-no, 1-yes. Squashing of output of the linear equation into a range of [0,1] is done. In Fig. 4, X axis represents the feature mean radius used in breast cancer prediction. The blue curve represents the sigmoid function which is used to squash the predicted value between 0 and 1. All samples below this curve are benign, and the ones above, are malignant.

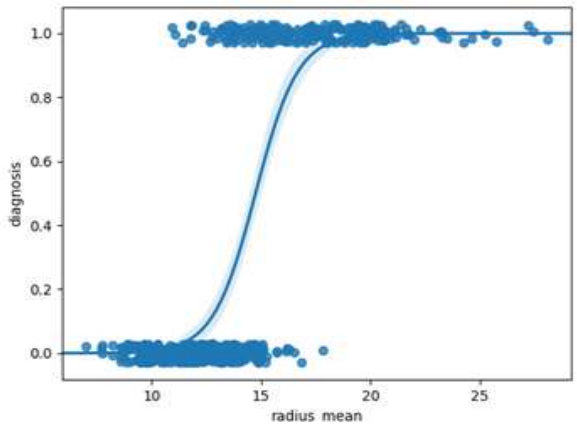

Figure 4: Logistic Regression Classifier

F. Naïve Bayes

Naive Bayes [7]is a classification technique based on Bayes Theorem [8] with an assumption of independence among predictors. Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature, hence the name Naive. P(Ck | X) shows probability of a sample belonging to class malignant or benign. P(Ck) represents this probability, while P(X) represents the probability of a feature occurring, eg. mean symmetry being<=0.15.

$$\text{Equation 1: } P(Ck \mid X) = \frac{P(Ck)P(X|Ck)}{P(X)}$$

Equation 1 is the equation given by Bayes theorem.

- P(Ck| X) is the posterior probability of class c(target) given predictor x(attributes).
- P(Ck) is prior probability of class.
- P(X | Ck) is the likelihood which is the probability of predictor given class.
- P(X) is the prior probability of predictor.

G. Artificial Neural Networks

The neural network [9] captures information from the outcomes of previous data between cases. During training, the network is provided the results of previous cases as input along with the features. The neural network has an advantage over other methods in that it is also able to take features of all cases involved as inputs. Therefore, it can draw on the outcomes of previous training examples. The neural network used for the dataset under consideration is also shown here.

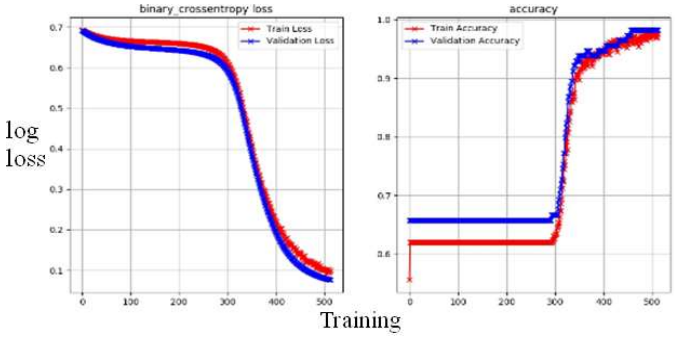

Figure 5: ANN classifier

Figure 5 describes how the loss (which should ideally be 0) decreases with each training iteration. The train and validation accuracy show a drastic increase as training progresses.

H. Voter Model

The No Free Lunch Theorem [10] states that any one algorithm that searches for an optimal cost or fitness solution is not universally superior to any other algorithm. In essence, different algorithms prove to be more effective for different data sets. Thus, instead of relying on a single algorithm completely, VoterModel algorithm relies equally on all of them.

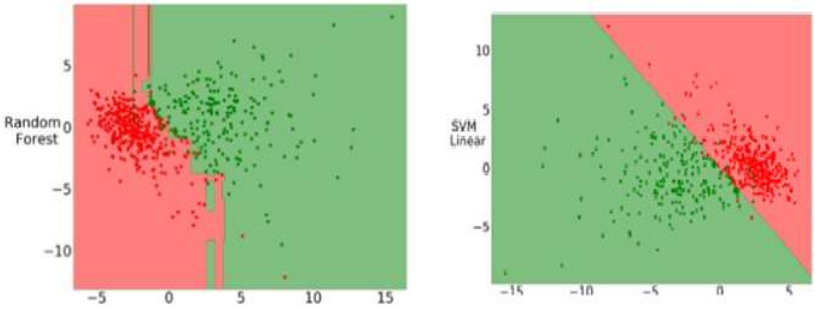

Figure 6: Fitting Curves of Random Forest and SVM Classifier

This can be explained by the Figure 6, which show the fitting curves of Random Forest and SVM classifiers. The features were reduced into 2 columns. The points (which indicates the samples) and curves (which indicates the boundaries) were plotted on the graph. Over fitting of boundaries to accommodate the points is clearly visible.

Voter Model Algorithm:

```
Initialize votes for "benign" and "malignant" to 0
Train the data with the models under consideration.
Use the trained model to classify test data as "benign" or "malignant."
If prediction is "malignant":
        Increment votes of "malignant" by 1
Else:
        Increment votes of "benign" by 1
If "malignant" has higher vote count:
        Test data is considered as "malignant"
Else:   Test data is considered as "benign"
```

Voter Model considers any machine learning model. Every model vote whether a test data is to be classified as benign or malignant. Based on most votes, a sample is classified as either benign or malignant using the model proposed. This can reduce over fitting as it prevents complete dependence on a single classifier. This has been proven based on the accuracy achieved by this model in comparison to the other models considered earlier.

IV. DISCUSSION

In the support vector machine model, proper parameter selection plays an important role in obtaining a correct classification. The linear kernel function is used to separate both the classes. Gamma should not be too high, as this can cause over-fitting.

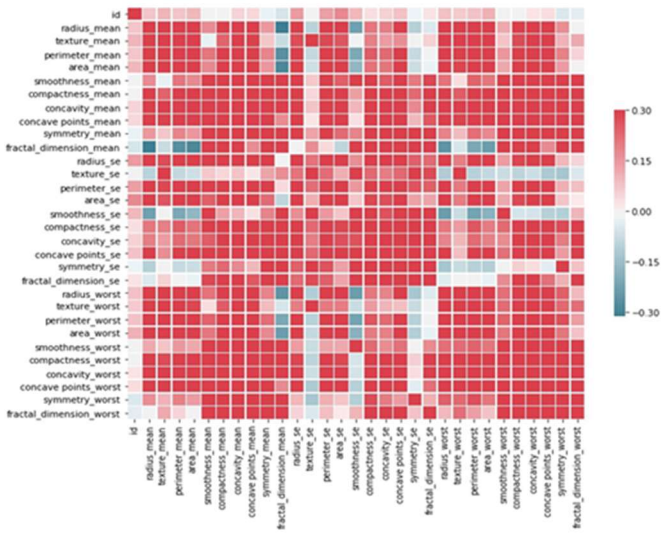

Figure 7: Correlation Matrix

The value is independent of how the remaining probability is split between incorrect classes. Cross-entropy loss uses a log function to measure the performance of a classification model whose output is a probability value between 0 and 1. Cross-entropy loss decreases as the predicted probability becomes closer to the actual label. Eg, predicting a probability of 0.015 when the actual observation label is 1 is bad and results in a high loss.

The data collected was plotted into a correlation matrix, a table showing correlation coefficients between variables, as shown in Fig 7. Each cell in the table shows the corrélation between two variables which helps to decide the features that can be used for training the model.

The correlation matrix helps to determine the correlated features, some of which are seen listed below.
Some positively correlated features identified are:

- Perimeter Mean and Radius Worst
- Area Mean and Radius Worst
- Texture Mean and Texture Worst
- Area Mean and Area Worst

From the correlation matrix, it was understood that radius, area, and perimeter essentially contain redundant information, which describes the physical appearance of a cell. Since area and perimeter are derived from radius, it is safe to discard both those columns. All the 'worst' columns can be discarded since they are a subset of the 'mean' columns.

For the random forest classifier and extra tree classifier, both the criteria- namely, Gini, as given by Equation 2, and entropy impurities, given by Equation 3, were implemented. Although both are often interchangeably used, for the Wisconsin Breast Cancer Diagnosis Dataset considered, entropy shows slightly better results. Gini prevents miscalculation, while entropy is used for exploratory analysis and can handle missing values. Entropy is apt for attributes that occur in classes.

Gini impurity:

Equation 2: Gini(E) =1 -$\sum_{j=1}^{c} p_j^2$

where, P_j is the fraction of items labeled as class j.

Entropy:

Equation 3: H(E) = -$\sum_{j=1}^{c} p_j \, log p_j$

where C is the number of classes

Table 1: Accurasies Obtained

Model	Columns	
	Criterion	Accuracy
Random Forest	Entropy	0.991
Random Forest	Gini	0.982
Extra Tree	Entropy	0.991
Extra Tree	Gini	0.982
Support Vector Machine	Linear Kernel	0.973
Logistic Regression	-	0.964
Naïve Bayes	-	0.956
Artificial Neural Network	-	0.999
Voter Model	-	0.997

Table 1 shows the accuracies obtained for various models, based on the different criteria considered.

V. SURVEY ON DEEP LEARNING METHODS

Utilizing the Xception deep learning model, [16] Yadavendra et all were able to attain exceptional results, with precision, recall, and F1 measures all reaching a commendable 0.90 under the same testing conditions. As a result, it was evident that the Xception method stands out as the superior choice among the various methods considered for classifying breast cancer tumors, demonstrating consistently high performance across these critical evaluation criteria. This signifies its robustness and effectiveness in accurately identifying and classifying such tumors, making it a preferred option for this task.

[17] Zheng, J et all introduced an innovative approach to breast cancer detection and early diagnosis by combining deep learning with the AdaBoost algorithm. They utilized the AdaBoost algorithm to create an ensemble classifier for the final prediction function. The results from evaluation tests demonstrated that proposed method exhibited superior predictive capabilities compared to other classifiers, with the deep-learning classifier standing out. Their analysis underscored the significant potential for rapid generalization and an efficiency boost in result prediction, driven by the neural network's automatic result derivation. Leveraging insights from the Convolutional Neural Network deep learning model, their DLA-EABA method contributed to enhancing system performance. They customized deep learning techniques to suit the unique attributes of each dataset, resulting in a tailored model for each one. The DLA-EABS method they put forth demonstrated remarkable accuracy in detecting breast cancer masses and subsequently improving patient survival rates. When benchmarked against existing methods, their approach consistently outperformed them in terms of performance.

REFERENCES

[1] Kelsey, J.L. and Horn-Ross, P.L., 1993. Breast cancer: magnitude of the problem and descriptive epidemiology. Epidemiologic reviews, 15(1), pp.7-16.

[2] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[3] Jerez-Aragonés, J.M., Gómez-Ruiz, J.A., Ramos-Jiménez, G., Muñoz-Pérez, J. and Alba-Conejo, E., 2003. A combined neural network and decision trees model for prognosis of breast cancer relapse. Artificial intelligence in medicine, 27(1), pp.45-63.

[4] Nguyen, C., Wang, Y. and Nguyen, H.N., 2013. Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. Journal of Biomedical Science and Engineering, 6(05), p.551.

[5] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1), pp.1929-1958.

[6] Akay, M.F., 2009. Support vector machines combined with feature selection for breast cancer diagnosis. Expert systems with applications, 36(2), pp.3240-3247.

[7] Rish, I., 2001, August. An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46).

[8] Swinburne, R., 2004. Bayes' Theorem. https://philpapers.org/rec/SWIBT-2

[9] Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C. and Meltzer, P.S., 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature medicine, 7(6), p.673.

[10] Ho, Y.C. and Pepyne, D.L., 2001. Simple explanation of the no free lunch theorem of optimization. In Proceedings of the 40th IEEE Conference on Decision and Control (Cat. No. 01CH37228) (Vol. 5, pp. 4409-4414). IEEE.

[11] Saputro, D.R.S. and Widyaningsih, P., 2017, August. Limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method for the parameter estimation on geographically weighted ordinal logistic regression model (GWOLR). In AIP Conference Proceedings (Vol. 1868, No. 1, p. 040009). AIP Publishing.

[12] https://towardsdatascience.com/building-a-simple-machine-learning-model-on-breast-cancer-data-eca4b3b99fa3

[13] Surendiran, B., and A. Vadivel. "Classifying Benign and Malignant Masses Using Statistical Measures." ICTACT Journal on Image and Video Processing 9102 (2011): 319-326.

[14] https://seer.cancer.gov/statfacts/html/breast.html {accessed on 30th March, 2021)

[15] Kalafi, E. Y., Nor, N. A. M., Taib, N. A., Ganggayah, M. D., Town, C., & Dhillon, S. K. (2019). Machine learning and deep learning approaches in breast cancer survival prediction using clinical data. Folia biologica, 65(5/6), 212-220.

[16] Yadavendra, & Chand, S. (2020). A comparative study of breast cancer tumor classification by classical machine learning methods and deep learning method. Machine Vision and Applications, 31(6), 46.

[17] Zheng, J., Lin, D., Gao, Z., Wang, S., He, M., & Fan, J. (2020). Deep learning assisted efficient AdaBoost algorithm for breast cancer detection and early diagnosis. IEEE Access, 8, 96946-96954.

[18] Sun, D., Wang, M., & Li, A. (2018). A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. IEEE/ACM transactions on computational biology and bioinformatics, 16(3), 841-850.

[19] Botlagunta, M., Botlagunta, M. D., Myneni, M. B., Lakshmi, D., Nayyar, A., Gullapalli, J. S., & Shah, M. A. (2023). Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms. Scientific Reports, 13(1), 485.

[20] Surendiran, B., Ramanathan, P., & Vadivel, A. (2015). Effect of BIRADS shape descriptors on breast cancer analysis. International Journal of Medical Engineering and Informatics, 7(1), 65-79.

[21] Shanmuga Priya, S., Saran Raj, S., Surendiran, B., & Arulmurugaselvi, N. (2020). Brain tumour detection in MRI using deep learning. In Evolution in Computational Intelligence: Frontiers in Intelligent Computing: Theory and Applications (FICTA 2020), Volume 1 (pp. 395-403). Singapore: Springer Singapore.

[22] Surendiran, B., & Vadivel, A. (2011). Classifying Benign and Malignant Masses Using Statistical Measures. ICTACT Journal on Image and Video Processing, 9102, 319-326.

[23] Sowrirajan, S. R., & Balasubramanian, S. Brain Tumor Classification Using Machine Learning and Deep Learning Algorithms.