Breast Cancer Detection

by Snigdha Sinha

Submission date: 27-Nov-2021 01:08PM (UTC+0530)
Submission ID: 1713522937
File name: Breast_Cancer_Prediction.pdf (103.09K)
Word count: 3106
Character count: 16466

Breast Cancer Detection

Smruthi Gowtham
Department of Computer Science
PES University, Bangalore, India
smruthig01@gmail.com

Vridhi Goyal
Department of Computer Science
PES University, Bangalore, India
goyalvridhi@gmail.com

Snigdha Sinha
Department of Computer Science
PES University, Bangalore, India
snigdhasinha0811@gmail.com

Yashi Chawla
Department of Computer Science
PES University, Bangalore, India
yashichawla1@gmail.com

***Abstract*—Breast cancer is a prevalent incursive cancer among women. It is one of the primary causes of cancer-related death, which can be classified as Malignant or Benign. Breast cancer diagnosis is time consuming and due to its gravity, it is imperative to design a solution to automate the process of identification of the same in its early stages so that it can be treated efficiently. Breast Cancer prediction aims at extracting features from the given samples and predicting it as Benign or Malignant. The dataset chosen is extracted from the Wisconsin Breast Cancer Dataset. This implementation compares six basic models namely, Logistic Regression, SVC, AdaBoost, Neural Network Nearest Neighbours and Random Forest through performance metrics such as accuracy, F1-score and F1-stratified score. The results reveal that Logistic Regression is the best performing model with an F1-score of 0.9788 and F1-stratified score of 0.9822.**

***Index Terms*—Breast Cancer, Prediction, Regression, AdaBoost, neural network, nearest neighbors**

I. Introduction

Breast Cancer originates in the lining cells of the glandular tissue ducts of the breast. Initially, the cancerous growth is limited to the duct where it usually does not show any symptoms and has very little scope for spreading. Gradually, the cancerous cells in stage 0 cancers may advance and invade the neighbouring breast tissue and then progress to the adjacent lymph nodes or other organs in the body within reach. Death due to Breast Cancer is because of widespread metastasis. There are many risk factors when it comes to breast cancer, such as race, age, genes, exercise level, alcohol consumption, etc. and there are multiple types of breast cancer with differing spread, stages, and aggressiveness.

Breast cancer treatment includes surgical removal, radiation therapy, etc, and can be highly effective, if it can be detected at an early stage. Misdiagnoses can lead to improper treatments where people lose their prime time for treatment which is why Breast Cancer prediction has been a principal topic for research in the medical field. Implementation of a prediction model which would help in early detection of breast cancer will be very helpful in increasing the survival rates of the patients diagnosed.

The dataset used in our project, The Wisconsin Breast Cancer Dataset (WBCD), acquired from the repository of UCI Machine learning, is a benchmark dataset. The dataset is distributed over 37.25% cancerous samples and 62.75% non-cancerous samples.

TABLE I
Summary of Wisconsin Breast Cancer Dataset

Attribute Description	Range		
	Mean	***Standard Error***	***Maximum Value***
Radius	6.98 - 28.11	0.11 - 2.87	7.93 - 36.04
Texture	9.71 - 39.28	0.36 - 4.89	12.02 - 49.54
Area	143.50 - 2501.00	6.80 - 542.20	185.20 - 4254.00
Perimeter	43.79 - 188.50	0.76 - 21.98	50.41 - 251.20
Smoothness	0.05 - 0.16	0.00 - 0.03	0.07 - 0.22
Concavity	0.00 - 0.43	0.00 - 0.40	0.00 - 1.25
Concave Points	0.00 - 0.20	0.00 - 0.05	0.00 - 0.29
Compactness	0.02 - 0.35	0.00 - 0.14	0.03 - 1.06
Fractal Dimension	0.05 - 0.10	0.00 - 0.03	0.06 - 0.21
Symmetry	0.11 - 0.30	0.01 - 0.08	0.16 - 0.66

In 2020, over 2.3 million women were diagnosed with breast cancer with a death rate of 29.78 percent. According to a 2020 survey, breast cancer was detected in 7.8 million women in the past 5 years. There are more Disability-Adjusted Life Years (DALYs) [1] lost by women to breast cancer than any other type of cancer.

In this project, we aim to come up with an improvised approach which helps in effective and accurate prediction of malignant and benign breast cancer.

II. Previous Work

Abien Fred M. Agarap's paper titled 'On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset'[2] compares machine learning (ML) algorithms of Linear Regression, Support Vector Machine (SVM), Softmax Regression, GRU-SVM, Nearest Neighbour search and Multilayer perceptron (MLP). Out of the six models mentioned, MLP emerges as the highest performing model with an accuracy of approximately 99.04%. However, the GRU-SVM also gives commendable results with an accuracy of 93.75%, which can be explained due to the sensitivity of RNNs to weight initialization[3], as they are arbitrarily assigned and due to the non-linearities[4] which were introduced due to the mechanisms of gating.

In the paper titled 'Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction'[5], by Yixuan Li and Zixuan Chen, five ML models, namely Decision Tree, Random Forest, SVM, Neural Network and Logistic Regression have been compared through indicators like accuracy values, AUC values and F-measure metric. The comparison showed that Random Forest performs better than the other four models and yields the highest accuracy score of 96.1% and F-measure metric of 94.1%.

While the two papers succeed in achieving high model accuracy, they do not talk about cross validation techniques which can provide a more accurate model prediction.

However, this approach has been discussed in the study by Haifeng Wang and Sang Won Yoon in their paper 'Breast Cancer Prediction Using Data Mining Method.'[6] It discusses breast cancer feature extraction based on data mining methods and presents a comparison among four models namely, Artificial Neural Network (ANN), AdaBoost Tree, Naive Bayes Classifier and SVM. 10-fold cross-validation method is implemented to approximate the error of these models. It also compares eight hybrid models namely, PCs-SVM, PCi-SVM, PCs-ANN, PCi-ANN, PCs-Naïve, PCi-Naïve, PCsAdaBoost, and PCi-AdaBoost. A series of paired t-test are performed to compare the prediction accuracy for each model and PCi-ANN turned out to be the one with the highest accuracy.

The cross validation technique has also been implemented by Md. Milon Islam et. al in the paper titled 'Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques'[7]. They have used a ten-fold cross-validation technique on the Wisconsin Breast Cancer Dataset (nine-fold for training and one for testing the model) and have performed a comparison between five ML algorithms, namely, SVM, Random Forest, ANN, Logistic Regression and K Nearest Neighbours. The performance of the models were measured using metrics such as accuracy, specificity, F1 score and Matthews Correlation Coefficient. Judging by these metrics, ANN was found to be the best model with an accuracy of 98.57%, a precision of 97.82% and an F1 score of 98.9%.

III. Proposed Solution

This study focuses on binary classification of the given samples of cell nuclei as benign or malignant. For the analysis of our problem statement, six ML models have been used, namely Logistic Regression, Support Vector Classifier (SVC), AdaBoost, Neural Network, K-Neighbours classifier and Random Forest.

A. *Logistic Regression*

Logistic Regression, despite its name is a classification model rather than a regression model. It is used for finding existence of a relationship between a discrete dependent variable and several independent variables. It can be generalized to multi-class classification. Rather than Mean Squared Error (MSE), Logistic Regression uses Maximum Likelihood Estimate (MLE) as the loss function.

B. *Support Vector Classifier*

Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for both regression and classification problems. In SVM, each feature for a particular data point represents one coordinate. The combination of all features results in the data point being plotted as a single point in f-dimensional space, where f is the number of features. For the purpose of classification, a hyperplane is chosen such that it maximises the margin of difference between the classes.

C. *Adaptive Boosting*

Adaptive Boosting, often abbreviated as AdaBoost, is an ensemble learning method. It is one of the earliest boosting algorithms. It combines several weak classifiers into one single strong classifier. The weak learners in AdaBoost are classifiers with a little over fifty percent accuracy. It works by putting more weight on misclassified instances and reducing the weight for correctly classified ones, such that subsequent classifiers focus more on difficult cases.

D. *Neural Network*

A Neural Network is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain. It comprises of layer of nodes which can be classified into three different containers: An input layer, hidden layers, an output layer. Nodes are interconnected and are associated with weights and biases. When the output of a node, post activation function, crosses the threshold value, the node gets activated. Such a node then forwards the data to the next layer. A node which has not been activated will not pass the data on. It can infer non linear relationships between the explanatory and response variables.

E. *K Nearest Neighbours*

K Nearest Neighbours (KNN) is a supervised learning algorithm. It is an instance based algorithm as it does not build a model based purely on the training data and waits for the test instance to do the same. For the given problem statement, KNN has been used for classification purpose. The inductive bias for KNN is that, given a new data point with an unknown class, it tends to belong to the same class as the bulk in its immediate neighbourhood.

F. *Random Forest*

Random Forest is an ensemble learning technique that combines several classifiers to provide solutions to complex problems such as classification. The algorithm comprises of many decision trees. The 'forest' generated is trained through bagging or bootstrap aggregation. It produces a reasonable prediction without hyperparameter tuning. It establishes the outcome based on the prediction of the decision tree while eradicating all it's limitations.

IV. Workflow

The Wisconsin Breast cancer Dataset contains 569 instances and 32 features that provide accurate information regarding the occurrence of breast cancer. The attributes of the dataset represent the characteristics of the cell nuclei present in the digitised image of a fine needle aspirate (FNA) of a breast mass.

A. *Data Cleaning*

During the process of data cleaning, the dataset (WBCD) was tested for inconsistencies, duplicates and null values and was found to be a clean dataset.

B. *Data Preprocessing*

For the types of models that have been trained for the given problem statement, it has been found that building models from the raw data without preprocessing yields poor results. Hence, multiple combinations of standardization, normalization and Principal Component Analysis (PCA) have been implemented as a part of data preprocessing. Furthermore, it was found that unstandardized data also yields relatively poor results. Hence the various combinations used were as follows:

TABLE II
Combination of Preprocessing Techniques

Standardize	***Normalize***	***PCA***
True	True	15 components
True	True	False
True	False	15 components
True	False	False

C. *Feature Correlation and Variance*

To reduce redundancy in the number of attributes chosen, the correlation between the attributes was studied and the results were observed. A series of input-target correlations were run that were indicative of what factors most influence a tumor to be malignant or benign. It was observed that attributes Radius and Concavity showed a high positive correlation to the target variable diagnosis. Perimeter_mean and radius_mean had a correlation coefficient of one which indicates a high linear correlation. (ADD PCA HERE)

D. *Evaluation Metrics and Cross Validation*

For our analysis, different evaluation metrics such as accuracy, F1 score and F1-stratified score have been proposed. For the purpose of testing the performance of the classifiers on varied sets of data, K-Fold cross-validation technique has been adopted across the entire dataset. This helps to verify that introduction of high bias does not make the model prone to underfitting. For the purpose of ensuring equal distribution of malignant and benign cases, both stratified and non stratified splitting has been used.

E. *Classification Models*

For the purpose of observing how different classifiers solve this issue, classifiers working on different approaches were proposed - Logistic Regression, Support Vector Classifier (SVC), AdaBoost, Neural Network and K-Neighbours classifier. GridSearch was performed with different hyperparameter values on all the classifiers used, in order to obtain a model with optimal performance.

1) Logistic Regression: Logistic Regression was chosen as the first classifier model. A cross validated Grid Search was used for the purpose of tuning the model. This enabled it to try out various combinations of parameters and allowed it to choose the optimal model.

TABLE III
Possible Parameters for Grid Search

Attributes	***Possible Values***
Penalty	l1, l2, elasticnet
C	0.5, 0.4, 0.3, 0.2, 0.1, 1, 10
Tol	1e-4, 1e-3, 1e-2, 1e-1, 1
Solver	"liblinear", "lbfgs"
Dual	True, False

TABLE IV
Chosen Parameters after Grid Search

Attributes	***Chosen Values***
Penalty	l2
C	0.4
Tol	1
Solver	"liblinear"
Dual	True

2) SVC: The next chosen classifier was SVC. GridSearch was again deployed to tune each hyperparameter to choose the best set which yields the optimum results.

3) Adaboost: The third chosen classifier was AdaBoost. The best hyperparameters were chosen by performing a grid search on the possible values for each parameter while varying the depth of the decision tree.

TABLE V
POSSIBLE PARAMETERS FOR GRID SEARCH

Attributes	*Possible Values*
kernel	”linear”, ”rbf”, ”poly”, ”sigmoid”
C	0.5, 0.4, 0.3, 0.2, 0.1, 1, 10
degree	2, 3, 4, 5
gamma	scaled, auto
shrinking	True, False
tol	1e-4, 1e-3, 1e-2, 1e-1, 1
coef0	0.5, 0.4, 0.3, 0.2, 0.0, 0.1, 1, 10

TABLE VI
CHOSEN PARAMETERS AFTER GRID SEARCH

Attributes	*Chosen Values*
kernel	”poly”
C	0.1
degree	3
gamma	auto
shrinking	True
tol	1e-1
coef0	10

TABLE IX
CHOSEN PARAMETERS AFTER GRID SEARCH

Attributes	*Chosen Values*
Layer Dimensions	128, 64, 1
Optimizer	Adam
Loss	binary cross entropy
Metric	accuracy
Epochs	200

TABLE X
POSSIBLE PARAMETERS FOR GRID SEARCH

Attributes	*Possible Values*
n_neighbors	2, 5, 10, 15, 20, 50
weights	distance, uniform
algorithm	brute, auto, kd tree, ball tree
leaf_size	1, 2, 4, 6, 8, 10, 20, 30, 40, 50
p	1, 2, 3
metric	euclidean, manhattan, minkowski, mahalanobis, cosine

4) Neural Network: Neural Network is the fourth classifier that was used. To tune the model, the number of layers and the neurons per layer were modified continuously till the best result was obtained. ReLu was chosen as the activation function for every layer except for the output layer which was sigmoid and Adam was chosen as the optimizer. As the classification problem is a binary one, the loss function chosen was binary cross entropy.

5) KNN: The next chosen classifier model was K Nearest Neighbours. The model was tuned by varying the value of K and the chosen distance metric to determine the best result. To improve the model results, weights were also introduced. The grid search algorithm was used to implement the same.

6) Random Forest: For the final model, the Random Forest classifier was trained. The Random Forest, being a powerful model in itself, does not require a lot of fine tuning to achieve high accuracy. Each decision tree in the forest was tuned based on it’s parameters. Grid search was deployed in order to choose the optimal set of hyperparameters which yielded the best outcome.

V. EXPERIMENTAL RESULTS

The six classifiers trained for the purpose of solving the aforementioned problem were Logistic Regression, SVC, AdaBoost, Neural Network, KNN and Random Forest. For each of these models, four possible pipelines of preprocessing techniques have been used, as mentioned in section IV B.
The metrics used for evaluating each of these models for all the different preprocessing pipelines are F1 score, stratified F1 score, accuracy and precision. The following is the summary of the F1 scores obtained.
—- FORM BIG TABLE —-

TABLE VII
POSSIBLE PARAMETERS FOR GRID SEARCH

Attributes	*Possible Values*
n$_{e}stimators$	10, 50, 100, 250, 500, 1000
learning$_{r}ate$	0.1, 0.3, 0.5, 1, 5, 10
algorithm	”SAMME”, ”SAMME.R”
base$_{e}stimator$	None, RandomForestClassifier

TABLE VIII
CHOSEN PARAMETERS AFTER GRID SEARCH

Attributes	*Chosen Values*
kernel	”poly”
C	0.1
degree	3
gamma	auto
shrinking	True
tol	1e-1
coef0	10

TABLE XI
CHOSEN PARAMETERS AFTER GRID SEARCH

Attributes	*Chosen Values*
weights	uniform
n neighbors	5
leaf size	1
p	1
metric	manhattan
algorithm	auto

TABLE XII
POSSIBLE PARAMETERS FOR GRID SEARCH

Attributes	*Possible Values*
criterion	entropy, gini
max features	None, auto, log2
bootstrap	False, True
min samples leaf	1, 2, 4, 6, 8, 10
min samples split	1, 2, 4, 6, 8, 10
n_estimators	10, 50, 100, 250, 500, 1000

TABLE XIII
CHOSEN PARAMETERS AFTER GRID SEARCH

Attributes	***Chosen Values***
bootstrap	False
n estimators	250
min samples leaf	2
criterion	entropy
min samples split	10
max features	auto

The best result was yielded by the Logistic Regression Classifier on a preprocessing pipeline consisting of standardization and PCA reducing the feature set to 15 components. The F1 and F1-stratified scores obtained for the above model were 0.9822 and 0.9788 respectively. —– PUT ACCURACY PLOT OF TOP MODELS—-

CONCLUSION

——-write about contribution——-

ACKNOWLEDGMENT

We wish to acknowledge Dr. Gowri Srinivasa, our Data Analytics Professor, for giving us the opportunity to work on this project which helped to enhance our knowledge. [SEEMS INCOMPLETE, PLEASE ADD SOMETHING.] We would also like to thank the Teaching Assistants who supported us throughout the course of this project.

REFERENCES

[1] WHO Fact Sheet on Breast Cancer. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

[2] Abien Fred Agarap. 2017. A Neural Network Architecture Combining Gated Recurrent Unit (GRU) and Support Vector Machine (SVM) for Intrusion Detection in Network Traffic Data. arXiv preprint arXiv:1709.03082 (2017).

[3] Abdulrahman Alalshekmubarak and Leslie S Smith. 2013. A novel approach combining recurrent neural network and support vector machines for time series classification. In Innovations in Information Technology (IIT), 2013 9th International Conference on. IEEE, 42–47

[4] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014).

[5] Yixuan Li, Zixuan Chen. Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction. Applied and Computational Mathematics. Vol. 7, No. 4, 2018, pp. 212-216. doi: 10.11648/j.acm.20180704.15

[6] Wang, Haifeng Yoon, Sang Won. (2015). Breast Cancer Prediction Using Data Mining Method.

[7] Islam, M.M., Haque, M.R., Iqbal, H. et al. Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques. SN COMPUT. SCI. 1, 290 (2020).

Breast Cancer Detection

ORIGINALITY REPORT

18%	10%	14%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Haifeng Wang, Bichen Zheng, Sang Won Yoon, Hoo Sang Ko. "A support vector machine-based ensemble algorithm for breast cancer diagnosis", European Journal of Operational Research, 2018 Publication	2%
2	Ameya Rajendra Bhamare, Aditeya Baral, Saarthak Agarwal. "Analysis of Kepler Objects of Interest using Machine Learning for Exoplanet Identification", 2021 International Conference on Intelligent Technologies (CONIT), 2021 Publication	2%
3	Abhishek Narayanan, Anmol Garg, Isha Arora, Tulika Sureka, Manjula Sridhar, H.B. Prasad. "IronSense: Towards the Identification of Fake User-Profiles on Twitter Using Machine Learning", 2018 Fourteenth International Conference on Information Processing (ICINPRO), 2018 Publication	1%

4	Submitted to National University of Ireland, Galway Student Paper	1%
5	link.springer.com Internet Source	1%
6	"Identification of Bio-Markers for Breast Cancer Detection through Data Mining Methods", International Journal of Recent Technology and Engineering, 2019 Publication	1%
7	cran.r-project.org Internet Source	1%
8	edoc.pub Internet Source	1%
9	researchr.org Internet Source	1%
10	Submitted to University of Western Ontario Student Paper	1%
11	apps.dtic.mil Internet Source	1%
12	Tsehay Admassu Assegie, Sushma S. J.. "A Support Vector Machine and Decision Tree Based Breast Cancer Prediction", International Journal of Engineering and Advanced Technology, 2020 Publication	1%

13	Submitted to Queen Mary and Westfield College Student Paper	1 %
14	"ROADS DATA CONFLATION USING UPDATE HIGH RESOLUTION SATELLITE IMAGES", 'Copernicus GmbH' Internet Source	<1 %
15	"An Efficient Data Mining Techniques - Multi-Objective KNN Algorithm to Predict Breast Cancer", International Journal of Recent Technology and Engineering, 2019 Publication	<1 %
16	machinelearningprojects.net Internet Source	<1 %
17	www.arabnews.com Internet Source	<1 %
18	Submitted to Universiti Teknologi MARA Student Paper	<1 %
19	dokumen.pub Internet Source	<1 %
20	Kaushal Sharma, Ajit Kembhavi, Aniruddha Kembhavi, T Sivarani, Sheelu Abraham, Kaustubh Vaghmare. "Application of convolutional neural networks for stellar spectral classification", Monthly Notices of the Royal Astronomical Society, 2020 Publication	<1 %

21	Qing Zhang, Hong Yu. "Computational Approaches for Predicting Biomedical Research Collaborations", PLoS ONE, 2014 Publication	<1 %
22	github.com Internet Source	<1 %
23	mediatum.ub.tum.de Internet Source	<1 %
24	thefreecoursesite.com Internet Source	<1 %
25	Meng Meng, Andreas Rau, Hita Mahardhika. "Public transport travel time perception: Effects of socioeconomic characteristics, trip characteristics and facility usage", Transportation Research Part A: Policy and Practice, 2018 Publication	<1 %
26	Nguyen Thanh Duc, Yong-Moon Lee, Jae Hyun Park, Boreom Lee. "An Ensemble Deep Learning for Automatic Prediction of Papillary Thyroid Carcinoma Using Fine Needle Aspiration Cytology", Expert Systems with Applications, 2021 Publication	<1 %
27	accentsjournals.org Internet Source	<1 %

28	openaccess.uoc.edu Internet Source	<1 %
29	www.ijrte.org Internet Source	<1 %
30	www.mdpi.com Internet Source	<1 %
31	"Advances in Artificial Intelligence and Data Engineering", Springer Science and Business Media LLC, 2021 Publication	<1 %
32	"Artificial Intelligence and Machine Learning in Healthcare", Springer Science and Business Media LLC, 2021 Publication	<1 %
33	Abdulhamit Subasi. "Machine learning techniques", Elsevier BV, 2020 Publication	<1 %
34	Vincent Peter C. Magboo, Ma. Sheila A. Magboo. "Machine Learning Classifiers on Breast Cancer Recurrences", Procedia Computer Science, 2021 Publication	<1 %

Exclude quotes	On	Exclude matches	< 5 words
Exclude bibliography	On		