BREAST CANCER DETECTION USING DIFFERENT CLASSIFIER MODELS

Anonymous CVPR submission

Paper ID *****

Abstract

Breast cancer is a disease that stems from the uncontrollable increase in the cells situated in the breast region. This cancer occurs most frequently in women and very rarely in men, accounting for a large percentage of cancers affecting people today. Most of the constitutions of breast cancer patients happen to be women, and even though a lot of awareness is being created to shed light on it, there is still so much to be done to increase proactivity. The early detection of abnormally growing or maturing cells is a critical step forward in the treatment process to enable doctors and physicians to save time, reduce the severity, and in extreme cases, prevent fatal endings. This paper builds a classification model for different breast cancer cell samples. The plan for approaching the problem is to categorize and correctly predict the diagnosis of breast cancer data, whether they are malignant or benign. To do this efficiently to the highest degree, we would use different traditional machine learning algorithms and a multilayer neural network to develop models and generate comparisons amongst them through hyperparameter tuning and alterations. Our model's accuracy will help detect cancer cells in patients with the recorded data. The initial results at this stage show that the classifier model, given the dataset's features, does an excellent job in classifying the cells and would help, being used hand in hand with actual experts to help deduce the probability of a patient having breast cancer.

1. Introduction

The most commonly occurring kind of cancer in women is breast cancer, and this is a molecularly diverse illness. The disease exists at such a heterogeneous level that targeting particular cells in an attempt at remedy is seemingly impossible. There are also different types of breast cancer and other effects on the human body. These are dependent on which cells mutate into cancerous cells. (Breast Cancer, 2022) The methods and approaches for therapy have changed over the past decades to accommodate this heterogeneity, with more attention and focus directed towards more biologically targeted medicines and treatment de-escalation to lessen side effects. Early breast cancer is considered treatable if confined within the breast or has only progressed to the auxiliary lymph nodes. Improvements in multimodal therapy have increased the likelihood that 70–80% of patients will recover. [3] Even with this, in retrospect, many people still suffer from the fatality of breast cancer. In many cases, the patients or those affected are either unaware of it or do not get diagnosed when the cancer is in its early stages. This is not specific to those with an idea or a lead that nudges them to get tested but to everyone, especially women, to get checks now and then.

1.1. Problem Statement

The main idea of this project is to use the data from the digitized images of a breast mass to make accurate deployable classifications of the samples to detect breast cancer. The dataset that will be used for the project is obtained from Breast Cancer Dataset. To effectively accomplish this, we will use different machine learning models and weigh and compare the accuracy of the resulting models. Using the features provided in the dataset, the plan is to use three traditional ML classifier models, a multi-layer feed-forward neural network model, and a simple CNN model. The reason for testing these different models is to obtain the model with the highest classification accuracy. We plan to use different hyperparameters to optimize the models and ensure that the model's complexity and efficiency are not compromised.

1.2. Background Preamble

Some work has been done on the detection based on the actual images from patients. This will be the baseline to understand whether the results are positive. This is not particularly necessary for this project, but we wanted some domain knowledge to weigh our results against real-life instances. The results are going to be evaluated both qualitatively and quantitatively for context. Plots will also be made between

the predicted and actual values. We will use different binary evaluation measures for calculating our model performance, including R-square value, misclassification score, precision, and the F1 score.

2. Background/ Related Work

Research classifying the cells in potential breast cancer patients has been ongoing for quite some time. This section identifies the relevant background work related to the project. Much of the available information being collected and improved involves making classifications using the actual scanned images of the tissues. Hameed and his team conducted research directed toward using an ensemble of deep-learning models to classify breast cancer histopathology images. [2] The accuracy in classifying the actual images came to about 95.29%, and they were able to conclude that the experiment's results showed the effectiveness of the approach to solving this problem. Another deep learning approach adopted by Krigitha achieved similarly good results, with values of accuracy, sensitivity, and specificity of 0.986, 0.947, and 0.964. [4] Experiments by Y. M. George and his team also showed that the results of using the four different models, Multilayer perceptron, probabilistic neural network, learning vector quantization, and SVM showed effective and comparable results even in other circumstances. [1] They also claimed that their results were better and that the model applied to multiple problems. Most of these approaches were considered using different deep learning approaches. Other works were also tackled using conventional machine learning algorithms. In the study by TIWARI [6], the classification problem was handled additionally using logistic regression and K-Nearest Neighbour in addition to a neural network model and an SVM model. In this work, the data used for the classification was extracted from image data and not the candid images themselves, similar to the work done in this study.

3. Approach

3.1. Data Preparation

The data preparation section of a research paper is an essential part of the study as it describes the steps taken to collect, clean, and organize the data used in the analysis. This section provides a clear and detailed account of the data sources, preprocessing steps, and any transformations or manipulations performed on the data. As mentioned, the data was obtained from Breast Cancer Dataset. There was no cleaning or manipulation required for this dataset. There was no missing data or data imputation needed in this case. Working with the data, the next process was to build some internal understanding of the problem and get results that best represent the classification problem. To do this, we explored distributions of the model features and identified

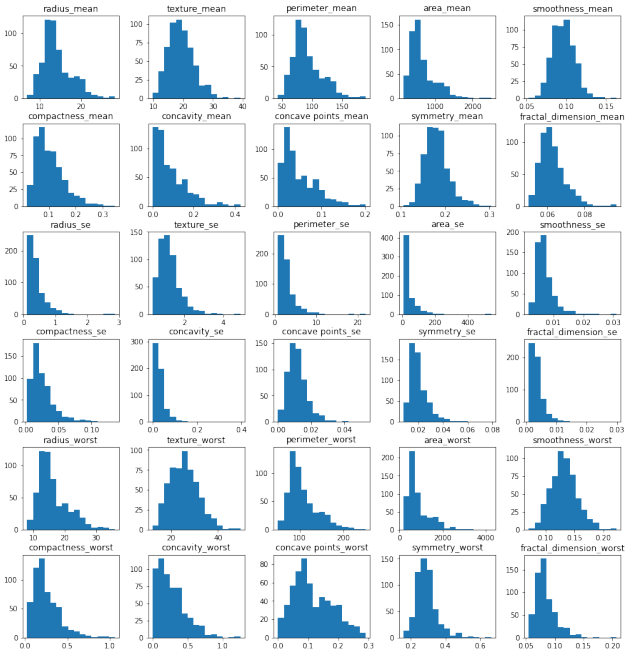


Figure 1. Distributions of variables in the dataset

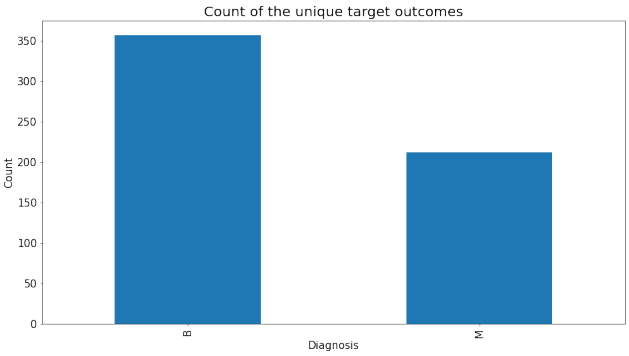


Figure 2. Distribution of the classes under the target variable

whether there was a need for any further feature engineering to be performed on the independent variables.

Since the distributions of the independent variable had different spreads across the values, normalization was prudent to be performed on the data to correctly accord weightage to the independent variables. The "RobustScalar" and the "StandardScalar" sklearn packages were considered for this procedure. Still, since we did not have outliers that appeared to be influential when inserted in the machine learning algorithm, we went with StandardScalar. The target variable, 'Diagnosis,' was also explored in terms of the distribution across the different classes.

Following the breakdown of the counts of the target variable, there was a raised concern about whether this would be an imbalanced dataset. Still, under further investigation

and from different sources, one being [5], a dataset with a distribution of 40:60 would not exactly be considered an imbalanced dataset. With this, I went forward without balancing the data. As an added assurance my results would not be affected, I used the classification report and the confusion matrix as a benchmark for testing my solutions so that I was confident the results and interpretations were not affected and the integrity was maintained. The features of the data were then split into testing and training data to be used in the section for model deployment.

3.2. Machine Learning Algorithms

This study employed five machine learning algorithms with a multilayer feedforward neural network model. The algorithms used at this stage were: Logistic Regression, Random Forest Classification, Decision Tree Classification, Support Vector Machines (SVM), and, as mentioned, a neural network. These were incorporated through python libraries and packages for model deployment. Logistic Regression Logistic regression is a simple yet powerful linear classification algorithm used to predict a binary outcome, such as whether a patient has breast cancer. It works by finding the linear decision boundary that best separates the two classes. Random Forest Classification On the other hand, random forest classification is an ensemble learning method that uses multiple decision trees to make predictions. It works by training multiple decision trees on different data subsets, then averaging their predictions to get a more accurate and stable estimate. This approach is often more accurate than a single decision tree and can also be less prone to overfitting. Decision Tree Classification Decision tree classification is a simple and intuitive algorithm used to predict a target label based on the values of other features in the data. It works by splitting the data into smaller and smaller subsets based on the importance of each feature and then using these splits to make predictions about the target label. This can be efficient for many classification tasks, including breast cancer classification. Support Vector Machines Support vector machines (SVMs) are another popular classification algorithm often used for difficult or complex classification tasks. They work by finding the decision boundary that maximally separates the two classes and can be very effective in high-dimensional data. Neural Network Finally, a neural network is a machine learning algorithm composed of many interconnected processing nodes called neurons. It works by learning the relationships between the input data and the target label and then using those relationships to predict new data. Neural networks are often used for complex classification tasks, such as breast cancer classification, and can achieve very high accuracy when trained on large and representative datasets. These are the models that were going to be employed in this paper.

4. Code Reuse

For this project's scope, a lot of the code used was incorporated and learned throughout my degree program. Other sources of code that were included in the notebook were from the sources listed below:

1. Stack Overflow for solutions for incorrectly running code.
2. Github Help for insights.
3. Towards Data Science to understand some more information about the fundamental uses of the selected model and their interpretations.
4. W3 Schools for some code testing.
5. Course material from CS 5783-65257

Most of the code, however, was coded by me, through and through. I did seek a lot of guidance from other random google sources and random sources, but I wrote the majority of the code.

5. Experiment

5.1. Results

The results of running the models show that they are doing very well on the classification problem in terms of the accuracy of the training and validation sets. To ensure the accuracy values we obtained from the models, I did some k-fold cross-validations and represented the accuracy as a mean of the procedure. I used a value of 10 splits for the k-fold validation and a scoring metric based on 'accuracy.' The accuracy across the six models is greater than 90 percent. This primarily would have been good considering the raw accuracy values. Figure 3 and Figure 4 are depictions of these accuracies and the comparisons among each other. In the training datasets, the best-performing model was the neural network model and the lowest-performing, though still good, was the Decision Tree model. Regarding the validation accuracies, the neural network model performed best; this time, the random forest classifier was the lowest performer. However, since the problem classification is to make sure that we can accurately classify the malignant and benign cases, we would be geared toward getting a higher specificity score, the correctly predicted malignant classes, as a ratio to the overall malignant classes. We would want this value to be as close to one as possible because the idea is to accurately predict the malignant classes of the model instead of getting a high accuracy. For example, we consider outcome one, where all malignant cases are correctly identified as malignant, and some benign cases are considered malignant with an accuracy value of 0.95. The second outcome would have most malignant identified as such and

Model	Training Accuracy
Logistic Regression	0.978019
Decision Tree	0.940628
Random Forest	0.958164
SVM	0.977923
MultiLayer	0.989011

Table 1. Training Accuracy

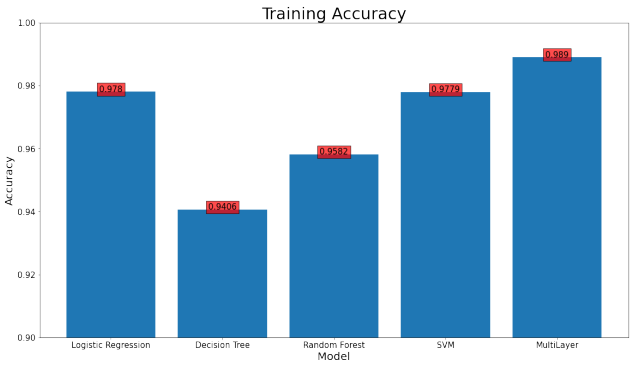


Figure 3. Training accuracy values of the models

Model	Validation Accuracy
Logistic Regression	0.973684
Decision Tree	0.964912
Random Forest	0.956140
SVM	0.973684
MultiLayer	0.982456

Table 2. Validation Accuracy

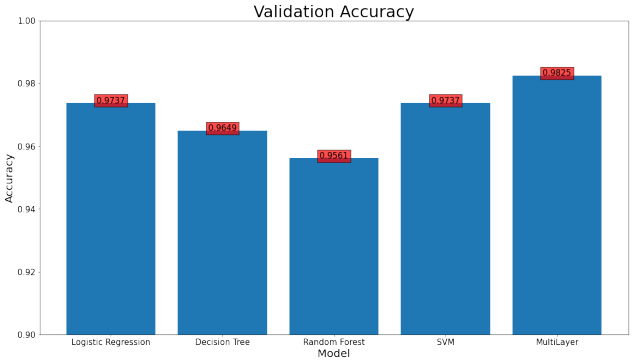
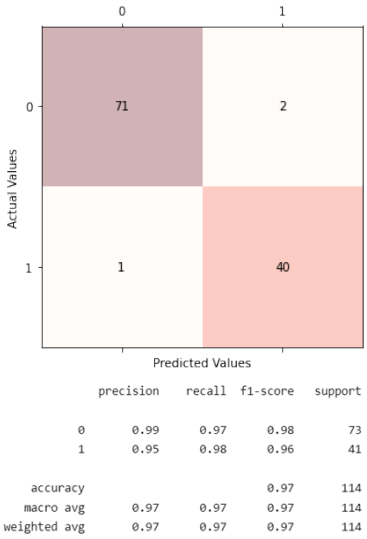


Figure 4. Validation accuracy values of the models

just a few malignant classed as benign with an accuracy of 0.98. For our problem statement solution, we would prefer outcome one because, even with the lower accuracy, it is doing better in grouping all the cases we should be concerned about than just getting a higher accuracy.

5.2. Discussion

For the interpretation of the results from the experiment, I decided to go into further detail in evaluating the model based on the information provided in the confusion matrix. As explained before, I was more focused on the false negative classifications before the accuracy values for this project's scope. Again, this is because, in a scope such as this, it would not be an issue if a benign breast cancer case was mistaken for a malignant one. Here, it would mean the patients will be subject to further tests and screenings even though it was probably unnecessary. That is acceptable. However, we would not want a malignant case to be classified as benign because that would be overlooked and potentially harmful and fatal to the patient. Considering this, we used the recall values for all the models for the first evaluation of the best model to use.



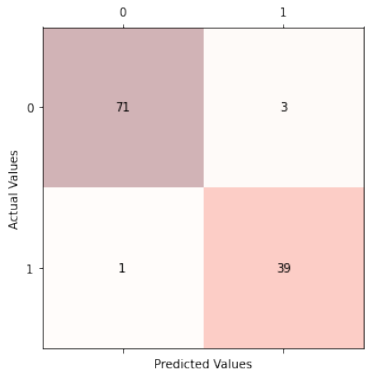
	precision	recall	f1-score	support
0	0.99	0.97	0.98	73
1	0.95	0.98	0.96	41
accuracy			0.97	114
macro avg	0.97	0.97	0.97	114
weighted avg	0.97	0.97	0.97	114

Figure 5. Confusion Matrix and Accuracy values for Logistic Regression

The results eventually directed towards the neural network model and the random forest models having the highest recall score for the malignant cases. This was the initial evaluation. Since we had more than one model with a recall score of 1, I decided to go with the model with the highest accuracy between the two chosen ones. That ended up being the Multilayer Feedforward Neural Network model.

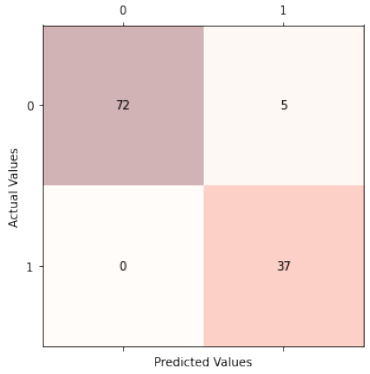
6. Conclusion

After experiments with multiple machine learning models, it was found that a neural network model outperformed all other models in terms of accuracy and recall for breast cancer classification. The neural network demonstrated a high-performance level and effectively accurately classified breast cancer cases. This suggests that neural networks may



	precision	recall	f1-score	support
0	0.99	0.97	0.98	73
1	0.95	0.98	0.96	41
accuracy			0.97	114
macro avg	0.97	0.97	0.97	114
weighted avg	0.97	0.97	0.97	114

Figure 6. Confusion Matrix and Accuracy values for Decision Tree



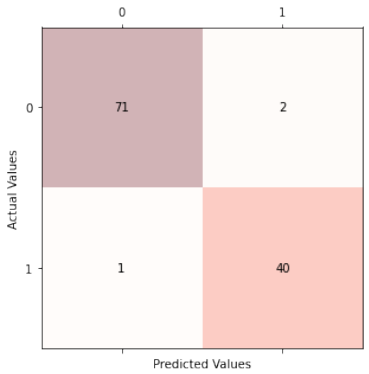
	precision	recall	f1-score	support
0	1.00	0.94	0.97	77
1	0.88	1.00	0.94	37
accuracy			0.96	114
macro avg	0.94	0.97	0.95	114
weighted avg	0.96	0.96	0.96	114

Figure 7. Confusion Matrix and Accuracy values for Random Forest

be a promising approach for breast cancer classification in the future. As a future project scope, I would want to explore different tuning of the hyperparameters in the machine learning model to determine if changing some of the default variables would lead to better accuracy metric values.

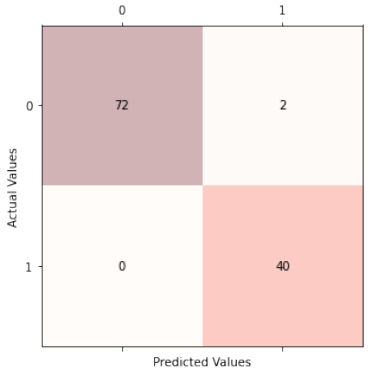
7. Division of Labor

I completed the project scope and the work done for this paper, so there is no division of labor.



	precision	recall	f1-score	support
0	0.99	0.97	0.98	73
1	0.95	0.98	0.96	41
accuracy			0.97	114
macro avg	0.97	0.97	0.97	114
weighted avg	0.97	0.97	0.97	114

Figure 8. Confusion Matrix and Accuracy values for SVM Classifier



	precision	recall	f1-score	support
0	1.00	0.97	0.99	74
1	0.95	1.00	0.98	40
accuracy			0.98	114
macro avg	0.98	0.99	0.98	114
weighted avg	0.98	0.98	0.98	114

Figure 9. Confusion Matrix and Accuracy values for Neural Net Model

References

[1] Y. M. George, H. H. Zayed, M. I. Roushdy, and B. M. Elbagoury. Remote computer-aided breast cancer detection and diagnosis system based on cytological images. *IEEE Syst. J.*, pages 949–964, 2014. 2

[2] Z. Hammed, S. Zahia, B. Garcia-Zapirain, J. J. Aguirre, and A. M. Vanegas. Breast cancer histopathology image classification using an ensemble of deep learning models. *Sensors*, 2020. 2

[3] Nadia Harbeck, Frédérique Penault-Llorca, Javier Cortes, Michael Gnant, Nehmat Houssami, Philip Poortmans, Kathryn Ruddy, and Janice Tsang. Breast cancer. *Nature Re-*

views Disease Primers, 2019. 1

[4] R. Krigitha and P. Geetha. Deep learning based breast cancer detection and classification using fuzzy merging techniques. *Mach. Vis. Appl.*, pages 1–18, 2020. 2

[5] Saikat Mazumder. 5 techniques to handle imbalanced data for a classification problem, June 2004. Analytics Vidhya. 3

[6] M. TIWARI, R. Bharuka, P. Shah, and R. Lokare. Breast cancer prediction using deep learning and machine learning techniques. *SSRN Electron. J.*, 2020. 2