BREAST CANCER DETECTION USING DIFFERENT CLASSIFIER MODELS

Botchway, Kodjo Opoku

A20338464

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.

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 heterogenic 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 axillary lymph nodes. Improvements in multimodal therapy have increased the likelihood that 70–80% of patients will recover. (Harbeck, et al., 2019)

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.

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 http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29. 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 multilayer feedforward 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.

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.

BACKGROUND

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. (Hameed, Zahia, Garcia-Zapirain, Aguirre, & Vanegas, 2020) 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. (R. Krithiga, 2020)

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. (Y.M. George, 2014). They also claimed that their results were better and that the model was applicable 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 (M. TIWARI, 2020), 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.

APPROACH

When working with the data, the first process was to build some understanding of the problem internally, put together the dataset, clean, and pre-process the data. To do this, we explored the distributions of the models and identified whether the was a need for prior feature engineering to be performed on the independent variables.

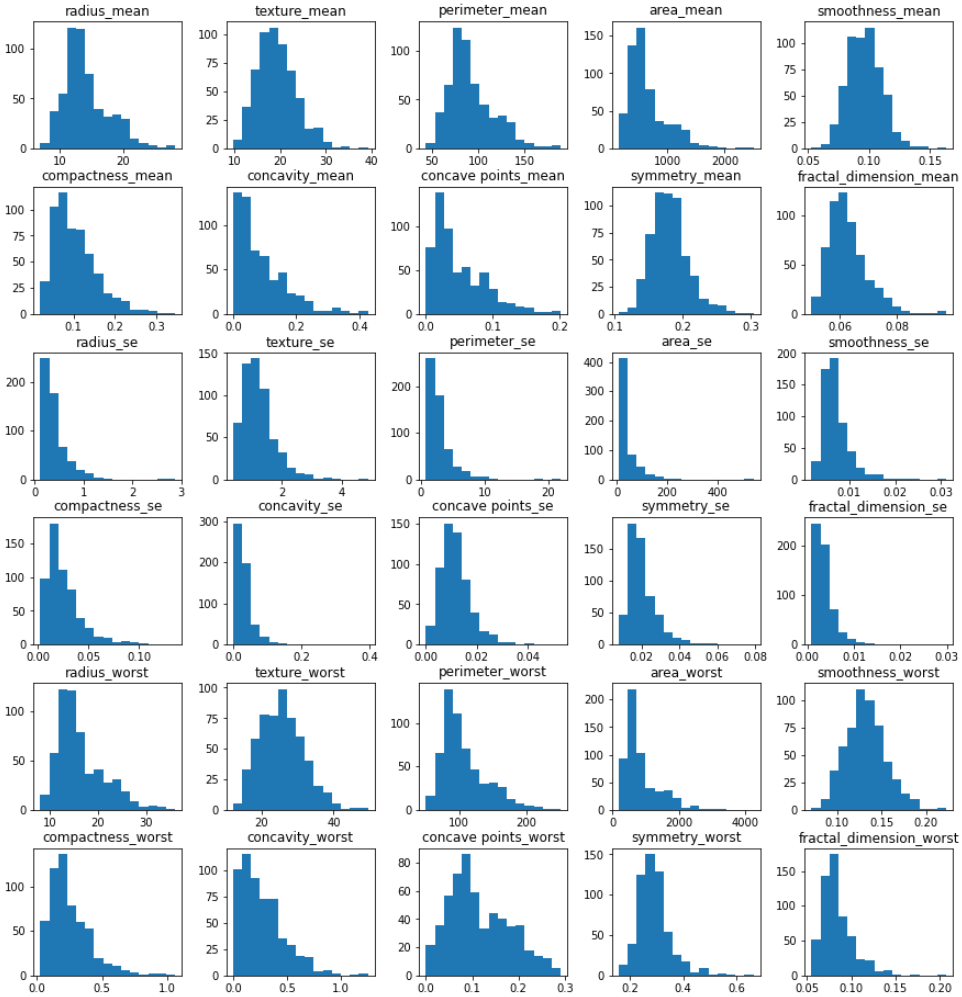

Figure 1: Distributions of variables in the dataset

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, because we did not have outliers that appeared to be influential when inserted in the machine learning algorithm, we went with StandardScalar. The features were normalized and were then split into testing and training data to be used in the section for model deployment.

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, K-Means Clustering, Support Vector Machines (SVM), and, as mentioned, a neural network. These were incorporated through python libraries and packages for model deployment.

INTERMEDIATE RESULTS

The current results show that the models are doing very well on the classification problem. The accuracy across the six models is greater than 95 percent. This primarily would have been good considering the raw accuracy values. 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 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.

The next step after this was to change the default initialized parameters and perform some hyperparameter optimization iterating through the different sequences to obtain the best model for the classification. I would also investigate if performing a Principal Component Analysis on the features to reduce the incorporated number of variables would make any difference in the model.

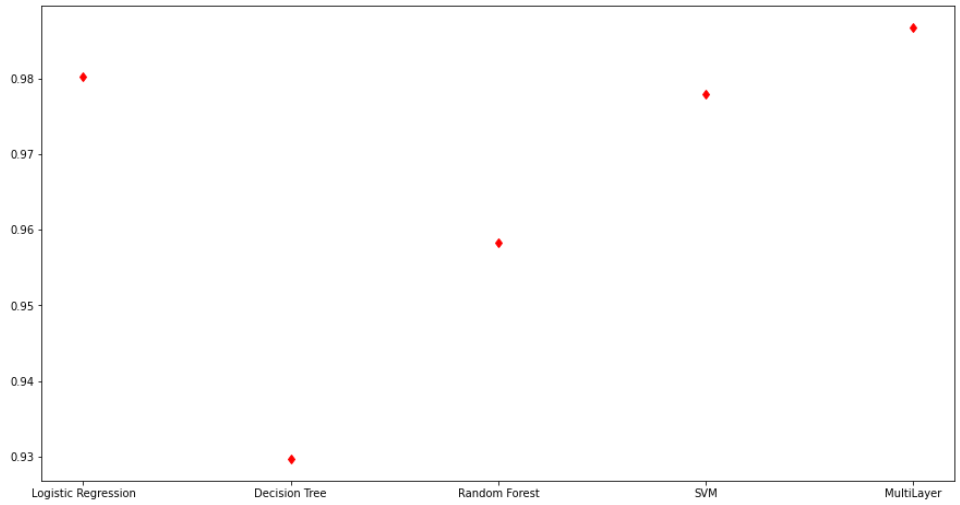

Figure 2: Currrent training accuracy values of the models used for now

References

Breast Cancer. (2022, September 26). Retrieved from cdc.gov: https://www.cdc.gov/cancer/breast/basic_info/what-is-breast-cancer.htm

Hameed, Z., Zahia, S., Garcia-Zapirain, B., Aguirre, J., & Vanegas, A. (2020). Breast cancer histopathology image classification using an ensemble of deep learning models. *Sensors*.

Harbeck, N., Penault- Llorca, F., Cortes, J., Gnant, M., Houssami, N., Poortmans, P., . . . Tsang, J. (2019). Breast Cancer. *Nature Reviews Disease Primers*.

Hickey, M., Peate, M., Saunders, C. M., & M., F. (2009). Breast cancer in young women and its impact on reproductive function. *National Library of Medicine*, 323 - 39.

M. TIWARI, R. B. (2020). Breast cancer prediction using deep learning and machine learning techniques. *SSRN Electron. J.*

R. Krithiga, P. G. (2020). Deep learning based breast cancer detection and classification using fuzzy merging techniques. *Mach. Vis. Appl.*, 1-18.

Y.M. George, H. Z. (2014). Remote computer-aided breast cancer detection and diagnosis system based on cytological images. *IEEE Syst. J.*, 949-964.