Classification Techniques in Machine Learning for Breast Cancer Prediction

Aashi Agarwal

Abstract*—Worldwide, female breast cancer is the most diagnosed cancer, with an estimated 2.3 million new cases (11.7%) every year. Breast cancer caused 685,000 deaths globally in 2020 [1]. Early detection is vital for effective treatment of breast cancer, and the survival rate for localized breast cancer is about 99% [2]. Screening mammograms are the most common way of detection, but still, 1 in every 8 mammogram misses to identify breast cancers. This paper focuses on using machine learning classification algorithms to help the diagnosis of breast cancer. The Wisconsin Breast Cancer Diagnostic dataset is used from the UCI machine learning repository. The dataset contains various characteristics of individual breast cancer cells obtained from a minimally invasive fine needle aspirate and classifies the cancer as Malignant or Benign. We train and test models for Logistic Regression, Logistic Regression with stochastic gradient descent, Support Vector Machine, Random Forest, Decision Tree, Boosted Decision Tree, and Neural Network with this dataset and compare their F1 score, false negatives, and false positives. Our objective is to create a model that has a very high recall, even at a slight expense of precision. A comparison is then made amongst the various techniques for the compute cost of training / inferencing and the performance of the model with respect to characteristics of the dataset. This paper then discusses how machine learning is being used currently for breast cancer prediction, and how we can make it better in the future to help early detection of breast cancer in every part of the world.

***Index Terms*— breast cancer, classification models, logistic regression, support vector machine, decision tree, neural network.**

INTRODUCTION

Worldwide, breast cancer is the most diagnosed cancer, with an estimated 2.3 million new cases (11.7%) [3]. By 2040, the burden from breast cancer is predicted to increase to over 3 million new cases and 1 million deaths every year [4]. It's paradoxical given the disease is 99% curable if detected at an early stage [2]. Global efforts are needed to counteract its growing burden, especially in developing countries where incidence is rising rapidly, and mortality rates are much higher than developed countries due to early detection. This difference is starkly seen by the higher breast cancer incidence rate of 12.5% in North America and lower mortality rate of 7.1% as compared to lower incidence rate of 7% in Southeastern Asia but much higher mortality rate of 8.6% [4].

Using machine learning techniques to aid the prediction of breast cancer and make it available to the masses in an easy-to-use manner can improve the rate of early detection and thereby reduce the mortality rate.

ABOUT THE DATASET

We will use the Wisconsin Breast Cancer Diagnostic dataset from the UCI Machine Learning Repository [5].

The creators of the database documented 30 characteristics of individual breast cancer cells obtained from a minimally invasive fine needle aspirate (FNA). FNA is a kind of biopsy in which a needle attached to a syringe is used to withdraw a small number of cells from a suspicious area [6]. Digitized images were created of these cells and an image analysis software called Xcyt was then used on these images. Based on a curve-fitting algorithm, this software determined the boundaries of the cell nuclei from a digitized 640×400, 8-bit-per-pixel grayscale image of the FNA. The boundaries were used to define 30 features that describe the characteristics of the cell nuclei present in the scanned images [6].

The 30 features are the mean, standard error, and worst case of the following features -

- radius
- texture
- perimeter
- area
- smoothness
- compactness
- concavity
- concave_points
- symmetry
- fractal_dimension

TECHNIQUES USED FOR CLASSIFICATION

Classification is a supervised machine learning predictive modeling problem where a class label is predicted for a given example of input data. Within Classification, binary classification refers to those classification tasks that have two class labels. There are several algorithms available to handle a classification problem with a binary response variable, in this case representing the malignant or benign diagnosis for breast cancer. We will explore the following algorithms in this paper [7] –

Aashi Agarwal, Mission San Jose High School, Fremont, California, USA.

Logistic Regression: Logistic regression is a simple and interpretable algorithm that models the probability of a data point belonging to one of the two classes. It uses the logistic function to transform a linear combination of input features into a probability score.

Logistic Regression with stochastic gradient descent: This optimizes the logistic regression model's parameters by updating them incrementally using small random subsets of the training data, making it computationally efficient and suitable for large datasets and online learning scenarios. The algorithm iteratively adjusts the model to minimize the loss function, converging towards an optimal solution for classification tasks.

Support Vector Machines (SVM): SVM is a powerful algorithm that finds a hyperplane that maximizes the margin between the two classes. This works well with datasets that may have outliers. It is effective in both linear and nonlinear classification tasks, thanks to kernel functions.

Decision Trees: Decision Trees recursively split data into subsets based on the most significant input features. They use a tree-like graph of decisions to make predictions, starting from the root node and branching down to leaf nodes representing class labels or numerical values. Decision Trees are interpretable, versatile, and widely employed for classification and regression tasks in various fields due to their simplicity and effectiveness.

Random Forest: It is an ensemble learning method, consisting of multiple decision trees. It operates by constructing a multitude of decision trees during training and outputs the mean prediction (regression) of the individual trees. By combining predictions from various trees, it enhances accuracy and reduces overfitting, making it a powerful and widely used algorithm in data analysis and predictive modeling.

Gradient Boosted Tree: Gradient Boosted Trees (GBT) is an ensemble learning method that combines the predictions of multiple decision trees sequentially. It builds trees iteratively, with each tree correcting the errors made by the previous ones, optimizing a specified loss function. GBT is powerful for regression and classification tasks, offering high accuracy and handling complex patterns in data.

Neural Networks: Deep learning models, particularly feedforward neural networks and convolutional neural networks (CNNs), can be used for binary classification. They have shown remarkable success in a wide range of applications but require substantial amounts of data and computational resources.

PERFORMANCE METRICS

When comparing machine learning models for breast cancer prediction, several performance factors should be considered to determine which model is most suitable for the task. These factors help in assessing the predictive accuracy, robustness, and interpretability of the models. Here are some key performance factors that we will consider when predicting breast cancer:

F1-Score: The F1-score is the harmonic mean of precision and recall (sensitivity). It provides a balance between precision and recall, which is especially useful when dealing with imbalanced datasets.

Confusion Matrix: A confusion matrix provides a detailed breakdown of the model's predictions, including true positives, true negatives, false positives, and false negatives. It is valuable for understanding where the model is making errors.

- **False negatives** - Also referred to as "Type II errors" in the context of binary classification and hypothesis testing. These errors occur when a test or model incorrectly fails to identify a true positive or when it wrongly indicates the absence of an event or condition when it is present. It's very important to minimize false negatives, since missing a positive case could have serious consequences.
- **False positives** - Also referred to as "Type I errors" in the context of binary classification and hypothesis testing. These errors occur when a test or model incorrectly identifies an event or condition as present when it is absent. It's important to minimize false positives, since these can lead to unnecessary treatment or intervention, a lot of mental traumas for patients and loss of credibility and reliability of diagnostic tests.

CLASSIFICATION IMPLEMENTATION

The following methodology was used to implement the different classification methods. The code for this can be found in Github at [9] –

Load the data using Python and Google Collaboratory. We used the Wisconsin Breast Cancer Diagnostic dataset [5]. It has 33 columns and 569 rows.

Sanitize the data –

I. The data has an extra column "Unnamed:32" with null values. Drop that column.
II. All columns except diagnosis are float64. The "diagnosis" column is of type object. Since this contains only two values - "Malignant" or "Benign", change the values to -

- Malignant - 1
- Benign – 0

Separate input and result variable - remove result "diagnosis" into its own series and remove from the input variables.

Study the input variables –

We see that the following features are measured in the dataset as mean, standard error (se) or worst (mean of 3 largest values) –

- radius
- texture
- perimeter
- area
- smoothness
- compactness
- concavity

- symmetry
- fractal_dimension

Correlation and dropping of strongly correlated variables -

Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate) [8]. If two variables are very strongly correlated (values close to either 1 or -1), they do not convey any extra information and should be removed from the dataset. We can find the correlation between different variables by creating a correlation matrix with all variables.

From Fig 1, we see that most "*_worst" columns are closely correlated to the mean columns. Similarly, we will also eliminate "*_se" columns since they are measuring the same metrics.

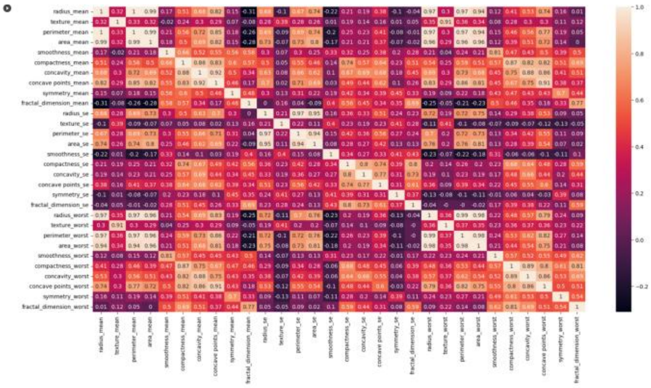

Figure 1

We also see that radius_mean is closely correlated to perimeter_mean and area_mean. That also makes sense since perimeter = 2*pi*radius and area equal pi*radius*radius. So, we will eliminate perimeter and area from our dataset. The final remaining columns in our dataset are –

- radius_mean
- texture_mean
- smoothness_mean
- compactness_mean
- concavity_mean
- concave points_mean
- symmetry_mean
- Fractal_dimension_mean

We build the correlation matrix again to ensure that variables are not highly correlated. It is shown in Fig 2.

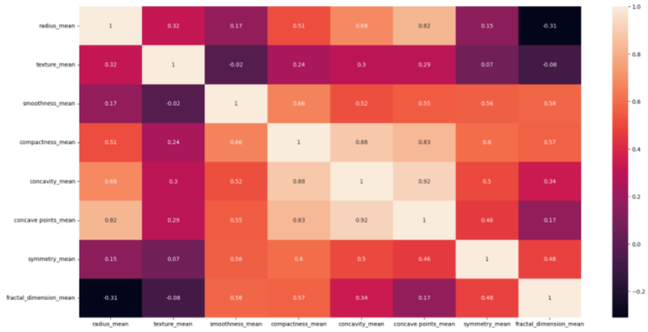

Figure 2

Normalize the data -
Feature scaling is the process of normalizing the range of features in a dataset. Real-world datasets often contain features that are varying in degrees of magnitude, range, and units. Therefore, for machine learning models to interpret these features on the same scale, we need to perform feature scaling. This was achieved by using StandardScaler from sklearn library.

Split the dataset into train and test -

- Test size - 25%
- Total rows in train = 426
- Total rows in test = 143

Run the different Machine Learning algorithms –
Train each model on the training dataset, make a prediction with the test data and then compare the results with the test dataset. Output F1 score and confusion matrix in each case. The results are summarized in the Results section.

KEY OBSERVATIONS

Here are some of the key observations about the techniques used in analyzing the breast cancer dataset:

1. **Logistic regression** works well when linearity is sufficient for classification. But if the relationship between the independent variables and the dependent variable is non-linear, as was found here, more complex models like decision trees or neural networks may be more appropriate.
2. When the dataset gets large and/or continuous retraining of the model may be desirable, **stochastic gradient descent with logistic regression** provides better results but may come at the cost of more hyperparameter tuning. In this case, since we have a small dataset, Logistic Regression performed better.
3. When non-linearity is assumed in a dataset, as results below indicate, **support vector machines** with proper selection of a kernel function can yield significantly better results for a dataset with high dimensionality.
4. **Decision trees** are a simple, non-parametric way to model a dataset, but are prone to overfitting and typically fall short on accuracy score yardsticks.
5. **Gradient boosted decision trees** capture complex relationships between input features and the target variable by combining multiple decision trees. The ensemble nature of GBDT allows it to reduce both bias and variance, resulting in robust and accurate predictions. They reduce overfitting and can come in very handy if there's an appetite to deploy more computational resources.
6. **Random forest** combines multiple decision trees, each trained on a random subset of the data and features, which helps reduce overfitting and produces robust and accurate predictions. The ensemble approach mitigates the biases and variances associated with individual decision trees and provides excellent generalization and robustness to overfitting at the expense of explain ability. It can yield great results with minimal tuning.
7. **Neural networks** can model complex, highly non-linear

relationships in a high-dimensional & large dataset, capturing complex patterns and interactions in the data but need a lot of tuning & computational power. With a much larger dataset, (VinDr-Mammo [10]), this can yield good results with adequate tuning.

Results

Seven different machine learning algorithms were trained on the exact same dataset to predict a classification problem - Malignant or Benign for breast cancer. The table below compares the results of the different algorithms.

Table I

Algorithm	F1 Score	False Negatives	False Positives
Logistic Regression	0.895	2	13
Logistic Regression with stochastic gradient descent	0.825	2	23
Support Vector Machines	0.930	1	9
Decision Tree	0.909	4	9
Gradient Boosted Decision Tree	0.923	2	9
Random Forest	0.923	2	9
Neural Network	0.937	1	9

Conclusion

In this study, the Neural Network showed highest accuracy and low false negatives. Support Vector Machines are a close second, with slightly lower accuracy. But there are several more factors to consider when choosing the algorithm -

- These tests were performed on a very small dataset. It is very hard to remove bias from such a small dataset. The results might change when we use very large datasets.
- A very basic three-layer Neural Network was used. With sufficiently large dataset and a large set of parameters, a finely tuned Deep Neural Network could yield even better results.

FUTURE WORK

- A global dataset could be created that has adequate data inputs from each of the mega-regions in the world where the data collection techniques are similar.
- A high-accuracy model, indexed by mega-region, could be created & used for prediction. The model could be accessed over the cloud by all the data labs in the world that help collect & provide imagery of the FNA data.
- The labs could use the model to predict the probability of cancer the moment they have the image of the FNA data, before sending it over to their diagnostic arms for further investigation, should it be deemed necessary by the model prediction layer.
- All the labs continue to enrich the model with more labeled data, which helps with periodic retraining of the model, to continue to make accuracy improvements.
- With enough advancements in FNA data collection exercises, this could turn into a self-serve detection exercise available to the masses in the palm of their hands, where one could collect their own data, get it imaged through the high-fidelity camera on their phone, and through an LLM supporting vision, turn into the characteristics of data, that the model on the cloud could use for predicting the nature of breast cancer. Prevention could be personalized for the patient, on the app, leading to another exemplary use case for technology, in saving millions of lives across the globe, and creating an equality in the level of health care across the developing & developed nations.

References

[1] Key facts on breast cancer by World Health Organization https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[2] Survival rates for breast cancer by the American Cancer Society https://www.cancer.org/cancer/types/breast-cancer/understanding-a-breast-cancer-diagnosis/breast-cancer-survival-rates.html

[3] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021 May;71(3):209-249. doi: 10.3322/caac.21660. Epub 2021 Feb 4. PMID: 33538338.

[4] Arnold M, Morgan E, Rumgay H, Mafra A, Singh D, Laversanne M, Vignat J, Gralow JR, Cardoso F, Siesling S, Soerjomataram I. Current and future burden of breast cancer: Global statistics for 2020 and 2040. Breast. 2022 Dec;66:15-23. doi: 10.1016/j.breast.2022.08.010. Epub 2022 Sep 2. PMID: 36084384; PMCID: PMC9465273.

[5] Wolberg,William, Mangasarian,Olvi, Street,Nick, and Street,W.. (1995). Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B.

[6] Fine Needle Aspiration (FNA) of the Breast https://www.cancer.org/cancer/types/breast-cancer/screening-tests-and-early-detection/breast-biopsy/fine-needle-aspiration-biopsy-of-the-breast.html

[7] Gareth James, Daniela Witten, Trevor Hastie, Rob Thibshrani: *An Introduction to Statistical Learning*

[8] What is Correlation? https://www.jmp.com/en_us/statistics-knowledge-portal/what-is-correlation.html

[9] Agarwal, A. (2023). Classification Techniques in Machine Learning for Breast Cancer Prediction (Version 1.0.0) [Computer software]. https://github.com/aashiagarwal2006/classificationPredictionUsingML

[10] Nguyen, H.T., Nguyen, H.Q., Pham, H.H. et al. VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. Sci Data 10, 277 (2023). https://doi.org/10.1038/s41597-023-02100-7

Aashi Agarwal is a senior at Mission San Jose High School, in Fremont California. She is passionate about applications of Data Science in Public Health.