Breast Cancer Prediction using Machine Learning

Melvin Oswald Sahaya Anbarasu
Robotics
Univeristy of Delaware
Newark, Delaware
oswald@udel.edu

Nii Otu Tackie-Otoo
Civil and Environmntal Engineering
University of Delaware
Newark, Delaware
niiotu@udel.edu

Annamalai Muthupalaniappan
Robotics
University of Delaware
Newark, Delaware
annamala@udel.edu

Atharva Vichare
Computer Science
University of Delaware
Newark, Delaware
atharvav@udel.edu

***Abstract*—Breast cancer is a serious worldwide health concern since it is one of its most prevalent and serious malignancies that affect both men and women. Given that it is one of the leading causes of death for women, early detection is crucial to improving patient outcomes. This project concentrates on the development and implementation of a machine learning-based model for the prediction of breast cancer. The dataset is sourced from the well-known website Kaggle and includes individual clinical and demographic data that is taken from the digital picture of a breast lump Fine Needle Aspirate (FNA). The project's goal is to create a machine learning model that can identify benign and malignant breast tumors deploying widely used machine learning methods, such as *Random Forest*, *Support Vector Machine*, *kNN(k-Nearest Neighbors)*, and *Logistic Regression*. In order to further understand the model, we tune a select group of hyperparameters of each model and study how each hyperparameter in a model affects its performance**

***Index Terms*—Breast Cancer, Breast Cancer Prediction, Logistic Regression, Support Vector Machine, Random Forest, kNN**

I. Introduction

In today's world, among the various cancers found, breast cancer is one of the deadliest cancers found among both women and men which has affected a massive number of people. Breast cancer alone captures a major chunk of nearly one-quarter of them which equals to an approximate estimation of 2.1 Million people in just last year alone. The rise in the number of cases every year has made it a growing concern in the medical community and made it a global health concern. Therefore detection of breast cancer in the budding/initial stages would be highly beneficial for the affected patient. Breast Cancer originates when the cells present in the breast tend to grow out of control. These outgrowing cells form the so-called tumors and lumps which are observed in X-rays. This project concentrates on the development and implementation of a machine learning-based model for the prediction of breast cancer. The key challenge against the detection of breast cancer is how to classify them into malignant and benign. Since this being a medical problem we require an authentic medical related dataset which is derived from the prominent website Kaggle [1].The dataset contains clinical and demographic information on individuals that was retrieved from a digitized picture of a fine Needle aspirate (FNA) [2] of a breast lump, with a binary classification task to discriminate between malignant and benign tumors.

The methodology involves *preprocessing* and standardizing the data, label encoding and employing machine learning models taught in the class and decided to explore models taught out of class, out of which we chose Random Forest [5], [6](ensemble models) for our classification task. The reason we went with fewer models on top of comparing the best models for our task is learning more about the models by tweaking the values of *hyperparameters* and learning how they influence the performance of the model.

The results obtained are inferred below in the report which depicts a comprehensive study of all the machine learning models involved. The analysis of those models shows their behavior with respect to the fine-tuning of the hyper parameters. The findings from the study indicate the potential of the machine learning models and their pros,cons and trade-offs of each model which can be integrated into existing diagnostic processes to enhance the accuracy by reducing the false negatives since this is a medical problem. The implications of this project can be extended to a broader field of medical diagnostics, displaying the use cases of machine learning in improving healthcare outcomes.

II. Related Works

In this section, we will review some of the work done by others on the related front. In paper [7], the authors attempted to predict breast cancer on the same dataset as ours. They used *kNN*, *Random Forest [5], [6]* and *Naive Bayes models* whereas we have used *Logistic Regression*, *kNN*, *SVM* and *Random Forest* [5], [6] for the same task. In their Setup, instead of Naive Bayes, we have tried a *Logistic Regression* and *Support vector machine* [4]. We were able to achieve similar results on the common models between our setups.

In Paper [8], The Authors compares the performance of many machine learning models on different datasets related to breast cancer particularly on the dataset used in our project.

Some of the models trained and evaluated include *SVM*, *Adaboost*, *Naive Bayes*, *ANN*. we were able to achieve similar results for the common models in our setup.

III. Data

A. Description

The data set contains *569* observations, 37% of which are from breast cancer patients. Each observation comprises variables obtained from a small part of a digitized fine needle aspirate slide. The forms of the boundaries of a group of nuclei are decided as a result of this. This enables accurate automated examination of *nuclear size*, *shape*, and *texture*.The target variable is *'Diagnosis'*. Ten characteristics are calculated for each nucleus. For each patient, the *mean value, maximum (or 'worst') value, and standard error* of each characteristic is obtained across a range of isolated cells. The nucleus has the following characteristics: *radius (the mean distance from the center to points on the perimeter), texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractional dimension.* [2].

B. Preprocessing

1) Feature Scaling: Feature scaling is one of the important data processing techniques that should not be overlooked which is also termed as data normalization. It is a process of bringing all the independent variables or features under the same range of values. This step is crucial in preventing the features with higher values dominating the learning process, this step ensures that all the features contribute equally to the model's predictions. Two of our models namely *kNN* and *Support Vector Machine* [4] that are based on distance metrics usually converge faster, whereas running them on unscaled data takes forever to run as they depend on the distance metrics. We used *StandardScaler* [3] (from *sklearn*) which involves removal of the mean and then scaling the features to the variance of one.

2) Label Encoding: Label encoding is a technique that is used to convert categorical values to numerical values as it can be fitted in the machine learning model in a better way. In our dataset, the target variable column has categorical values (*namely Malignant (M) , Benign (B)*). It is smart to convert these values to Values of *0* and *1* as they are easy to interpret both for humans and machine learning models. We used Label encoding to map *Malignant* to *1* and Benign to *0* as *Malignant* being the cancerous tumors and *Benign* being the non-cancerous one.

IV. Models

We aim to build a machine learning model for detecting breast cancer using prominent machine learning methods such as *Logistic Regression*, *kNN*, *Support Vector Machine* [4], and *Random Forest* [5], [6]. First, we trained all the models in out-of-the-box or default configuration, we used accuracy as the evaluation metric. In order to further analyze the models in depth, we configured the values of the *hyperparameters* in each model compared them against the accuracy to get an overall perspective of how each hyperparameter influences the performance of the model.

We also used C*ross-Validation(GridSearchCV [3] from sklearn)* to get the best optimal *hyperparameters* for our model. *GridsearchCV* [3] is a function that performs an extensive search over the specified parameter grid to find the best *hyperparameter* combination that gives the best evaluation metric(accuracy in our case). The dataset is divided into 'k' folds so that the model may be trained and assessed 'k' times. Every time a fold is chosen for testing, the remaining folds are used for training.

V. Experiments

A. Experimental Setups

Our experimental setup comprises of training all of the models on the dataset, including *Logistic Regression*, *kNN*, *Support Vector Machine* [4], and *Random Forest* [5], [6]. The dataset split is in the ratio *80:20 (train and test sets)*. As mentioned before, the training data was scaled by standardizing features by the removal of the mean and then scaling the features to the variance of one. (*StandardScaler [3] from sklearn*).

As mentioned above, all models were trained on out-of the box or default configuration. *GridsearchCV* (cross-validation) [3] was performed on all models to acquire the best *hyperparameters* and then trained and tested on the data. We used accuracy as the evaluation metric for understanding the importance of selecting the optimal *hyperparameters* for the model.

For *Logistic Regression*, the chosen parameter grid was { ***Regularization Strength, C = [0.001, 0.01, 0.1, 1, 10, 100], Penalty = [l1, l2]*** }. Here different values of Regularization strength and two different penalties were chosen as the parameters in the grid. For kNN, the chosen parameter grid was { ***Number of Neighbors : [2, 3, 4, 5, 6, 7, 8, 9, 10, 11], Metric : Manhattan, Euclidean*** }. Number of neighbors and two different metrics were chosen as the parameters in the grid.

Kernel: Linear, polynomial, radial basis function (rbf), sigmoid, Regularization Parameter(C): [0.001, 0.01, 0.1, 1, 10] was the parameter grid used for *Support Vector Machine* [4]. Different kernels and two different Regularization parameter values were chosen as the parameters in the grid. For *Random Forest* [5], [6], the chosen parameters grid was ***'Number of Estimators': [10, 20, 30, 50, 100],'Max Depth': [1, 2, 3, 4, 5, 6, 7])***. Different values of estimators and level of depth were chosen as the parameters in the grid.

B. Experimental Results

We will go over the project's findings and a few observations in this section.

The accuracy of all models as shown in the Table I in their default configuration are compared with accuracy of the models trained with best hyperparameters returned by the GridSearchCV(cross-validation) [3] to emphasize the importance of choosing the optimal hyperparameter combination to attain its peak performance.

Model	Accuracy(Default)	Accuracy(CV)
Logistic Regression	0.974	0.9912
kNN	0.947	0.965
Support Vector Machine	0.9825	0.9825
Random Forest	0.956	0.965

TABLE I: Accuracy Comparison for Different Models

In Figure 1, We can interpret the maximum accuracy is achieved with a combination for *C= 0.1* and for the *regulariztion value of 12*. Lower value of C indicates stronger regularization, this prevents overfitting by penalizing the larger coefficients, also *l2 regularization(ridge regression)* which penalizes the squares of coefficient indicates it gives importance to a broad set of features without completely ignoring any rather by preventing the model from having very large weight for any particular feature.

In Figure 2, for *kNN*, The optimal distance metric is euclidean which means it uses the similarity between data points for determining the nearest neighbors. The combination of *n=5* and *euclidean distance* shows that the model takes a balanced approach between capturing local patterns(smaller *n_neighbor* value) and uses standardized distance metric for finding the similarity between neighbors.

GridSearchCV [3] returns the optimal parameter combinations from the given grid of values for different parameters.The optimal parameters chosen using 5-fold cross-validation for each of the models on the scaled data are shown in TableII

Model	Optimal Hyperparameters (GridSearchCV)
Logistic Regression	Regularization Strength (C) = 0.1
	Penalty = 'l2'
kNN	Number of Neighbors = 5
	Metric = 'Euclidean'
Support Vector Machine	Kernel = 'linear'
	Regularization Parameter (C) = 0.1
Random Forest	Number of Estimators = 50
	Max Depth = 6

TABLE II: Optimal Hyperparameters

In Figure 5, for *SVM* [4], the choice of *linear kernel* states that the model performs well considering the linear relationship between features and target variable. It indicates the relationship is almost linear. Also C value of 0.1 indicates a moderately regularized model in this case that aims to balance fit the training data while avoiding *overfitting*.

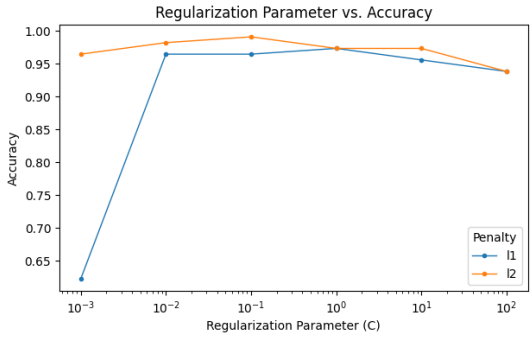


Fig. 1: Logistic Regression

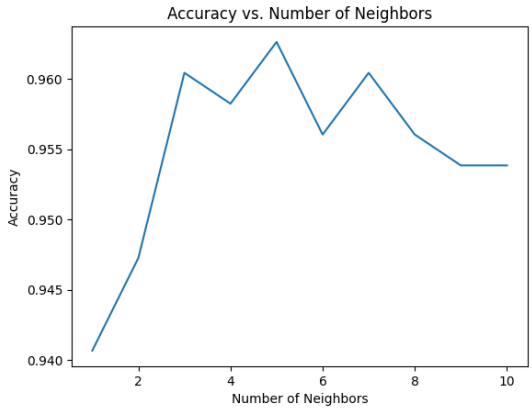


Fig. 2: kNN

In Figure 3, for *Random Forest*. The combination of *50 estimators* and a *maximum depth of 6* indicates the model balances between complexity in capturing patterns and generalization by avoiding overfitting. Trees with more max depth can capture more intricate patterns but are subjected to *overfit* noise in the training.

Taking a look at the confusion matrix (Figure 4) for three of the models with the highest accuracies (*Logistic Regression, SVM [4] and Random Forest* [5], [6]), there is one key thing we have to point out. Although all models used in this analysis had extremely high accuracies, this problem is a medical diagnosis one. In the medical industry, much emphasis is not placed on how accurate the model is at predicting either benign or malignant (accuracy). Neither are fully focused on the proportion of the model's right prediction for malignant cases (precision). What is of greater concern is the recall, which is the proportion of total malignant cases the model was able to predict correctly. In this instance, a higher recall implies that we are identifying close to all the malignant cases, hence those patients can begin treatment as soon as possible. In our case, the recall scores for *Random Forest* [5], [6], *kNN*, *Support Vector Machine* [4], and *Logistic Regression* are *0.93, 0.93, 0.95, and 0.98*, respectively. With that being said, it is of a much bigger problem if the model predicts a malignant case as benign compared to predicting a benign case as malignant.

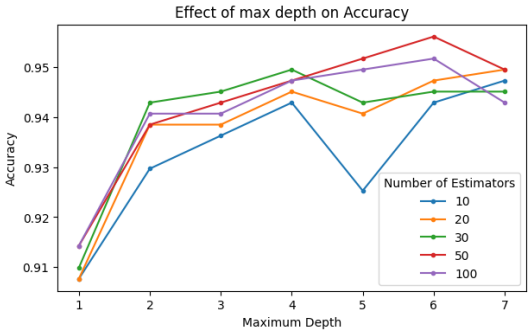


Fig. 3: Random Forest

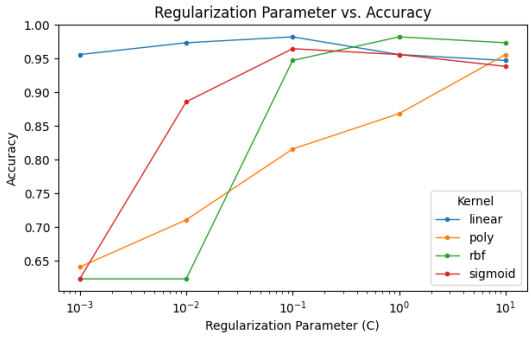
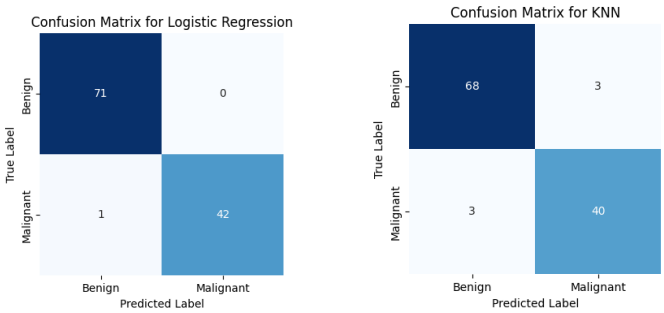
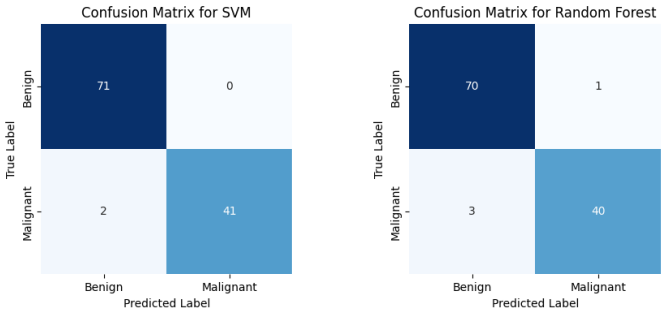


Fig. 5: Support Vector Machine



(a) Logistic Regression (b) kNN



(c) SVM (d) Random Forest

Fig. 4: Confusion Matrices

VI. Conclusion

Our project focused on breast cancer prediction that made use of machine learning models like *Random Forest [5], [6]*, *kNN*, *Support Vector Machine*, and *Logistic Regression*. Cross validation was utilized to determine the *ideal hyperparameters* for achieving the greatest model performance. The most effective model was proven out to be logistic regression demonstrating high accuracy of 0.9912. Since, this is a medical problem, our concern is recall rather than a model with high accuracy. In our case, Logistic Regression has a high recall score of 0.98 which is also the one with the highest accuracy for our dataset. This might not always be the case when considering different datasets. We tuned different values of hyperparameters plotted against accuracy to determine the relationship between them and also understand how each hyperparameter influences model's performance. The study provides a structure that can be adapted to other datasets and models, making it a basis for machine learning projects in the future.

References

[1] Kaggle, https://www.kaggle.com/.

[2] Bazila Banu, Ponniah Thirumalaikolundusubramanian, "Comparison of Bayes Classifiers for Breast Cancer Classification," Asian Pac J Cancer Prev, vol. 19 (10), pp. 2917-2920.

[3] sklearn, https://scikit-learn.org/stable/

[4] Corinna Cortes, Vladimir Vapnik, "Support Vector Networks,", Machine Learning, vol. no. 20, pp. no. 273–297, 1995.

[5] Breiman, Leo., "Bagging Predictors,", vol. 24, no. 2, pp. 123–140, August 1996.

[6] Breiman, Leo, "Random Forests," Statistics Department, University of California, Berkeley, January 2001.

[7] S. Sharma, A. Aggarwal and T. Choudhury, "Breast Cancer Detection Using Machine Learning Algorithms," 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), Belgaum, India, 2018, pp. 114-118, doi:10.1109/CTEMS.2018.8769187.

[8] Tahmooresi, M., Afshar, A., Bashari Rad, B., Nowshath, K. B., Bamiah, M. A. (2018). Early Detection of Breast Cancer Using Machine Learning Techniques. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 10(3-2), 21–27.