Building Machine Learning Models to Predict the Malignancy of Breast Cancer Tumors

1st Aaron Fryzel
School of Biomedical Engineering
Drexel University,
Philadelphia, United States
af3285@drexel.edu

***Abstract*—Breast cancer remains one of the most prevalent type of cancer worldwide, necessitating effective diagnostic methods. This study explores the development and improvement of machine learning models to predict the malignancy of breast cancer tumors using the Wisconsin Breast Cancer Dataset. I implemented and compared the performance of three different supervised learning models: Logistic Regression, Naïve Bayes, and Support Vector Machine (SVM). Each model was trained and tested on pre-processed data, and their performance was evaluated based on accuracy, precision, recall, and F-measure metrics. The results indicated that the Naïve Bayes model slightly outperformed the Logistic Regression and SVM models, achieving the highest accuracy of 97.3percent. My findings suggest that machine learning techniques, can provide robust and reliable diagnostic tools for identifying malignant breast cancer tumors, potentially enhancing the clinical decision-making processes.**

***Index Terms*—Machine Learning, SVMs, Breast Cancer, Logistic Regression, Naive Bayes**

I. Introduction

According to the NIH, in 2024 roughly two million people are expected to be diagnosed with cancer while around 600,000 will die from it. Of the many different types of cancer, breast cancer is the most common and one of the most well studied. Due to breast cancers commonality and complexity processing the large amount of data in order to make accurate decisions in a clinical setting provides a deep issue, particularly in the characterization of tumors. Because of this problem many healthcare providers struggle in prescribing proper treatments to there patients due to the high-risk and invasive nature of most commonly used cancer treatments. As such the goal of this study is to attempt to create supervised machine learning models capable of diagnosing breast cancer tumors as being either malignant or benign using the Diagnostic Wisconsin Breast Cancer Database for potential usage in a clinical setting.

II. Relevant Works

Both the dataset I am using and the topic I am studying have been quite popular over the years since their respective inceptions and as such I will not be able to talk about all of the work that has been done in a reasonable time frame, instead I will be focusing on two papers I believe to be of particular interest that will give a good overview of this field of inquiry The first work that I would like to speak of is titled "Nuclear feature extraction for breast tumor diagnosis" this paper was originally published in 1993 and was a part of the early push for machine learning in clinical settings, it was also the paper from which the dataset I am using was originally sourced. This paper proposed a linear programming model known as a Multi-Surface Method Tree(MSM-T) based approach to this dataset. The second paper I would like to focus on is titled "Machine learning in medicine: a practical introduction" which was published in 2019, in this paper the authors give an overview of how machine learning methodologies can be applied in a clinical setting, as an example they use the dataset Diagnostic Wisconsin Breast Cancer Dataset to build a variety of models which they report performance metrics for. The main models the researchers use in this paper are a single-layer ANN, a Generalized Linear Regression model, and a SVM Both of these papers are of interest because they represent how machine learning has evolved, their usage of the same dataset allows for me to draw direct comparisons between the efficacy of my own models when compared to older ones and more modern implementations.

III. Methods

I began my analysis by loading the Diagnostic Wisconsin Breast Cancer Dataset using the pre-built sklearn function after which I separated class information from the rest of the data. After loading the data, I then began the pre-processing portion of the analysis in which I applied z-score normalization to all non-class data, after which I then added a bias feature to my data. Following this, I then separated my data into training (80 percent) and testing (20 percent) sets. Dividing my data in this manner allows for me to perform a technique called hold-out validation which will allow for the proper testing of machine learning models. Going forward from this point all of my models were trained using the training data and then tested using the testing data.

$$z = \frac{x - \mu}{\sigma}$$

Moving on from the pre-processing stage to the actual model building part of the project I created three models which I nested within separate callable functions that took my pre-processed and split data and returned the accuracy, precision, f measure, and recall of the given model. The first model that I chose to implement was a Logistic Regression

model that was optimized using gradient descent. To do this I used the derivative equations listed below which allowed me to find theta values corresponding to each feature within the training data, these values can then be multiplied on a per-sample basis after which there products can be added together to give the probability of a given sample being malignant or not. After finding theta and generating a set of prediction probabilities which I converted to binary class data using a sigmoid function I was then able to calculate loss within my gradient descent while loop which updated my theta values using a learning rate of 0.01 it would then be terminated after it either hit a loss value of 2^{-23} or hit a max iteration threshold of 3500 this would then produce the final optimized Logistic Regression model. I chose to use a Logistic Regression model in this instance due to its simplicity which I believed would be advantageous when working with the relatively low dimensional data I had access to. When calling this model, I chose to repeat the function ten times within a for loop and take the average of each of the performance metrics in order to somewhat account for the random chance present within the model's architecture.

$$\theta = (X^T X)^{-1} X^T y \tag{1}$$

$$\text{Predictions} = x_0 * \theta_0 + x_1 * \theta_1 ...$$

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \left(y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})) \right)$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_j^{(i)}$$

$$\theta := \theta - \alpha \frac{\partial J(\theta)}{\partial \theta}$$

The second model I chose to use was a Naïve Bayes model, this model was not commonly used in previous studies using this data, this is likely due to the independence assumption which is fundamental to the model's architecture which is violated in most cases of biological data. However, I believe this to be an oversight as due to the low feature count contained within this data taking an approach where we consider each feature to be independent may not be as harmful in this case as it would be in higher dimensional data. I began my implementation of Naïve Bayes by calculating the respective means for both benign and malignant class-splits after which I also found the class priors for the two classifications. From there I then found the log-likelihood corresponding with each class which when multiplied with its respective class prior converted into log-space allows for us to predict the probability of each samples class which can then be converted into binary class prediction data.

$$\log P(y|x) \propto \log P(y) + \sum_{i=1}^{n} \log P(x_i|y)$$

The final model that I examined in this study was a Support Vector Machine (SVM) model which I chose due to its common usage in biomedical contexts up until its recent supplantation by neural networks. As mentioned in my relevant work section many of the studies that made use of the Wisconsin dataset in the past chose to use this model. Unfortunately, due to my lack of familiarity with this models architecture I was not able to build a bespoke model within the timeframe given for this analysis and as such I opted to use the SVM functions available within the sklearn libraries for a SVM model with a linear kernel. I chose to still include this model in my study in-spite of this because it provides us with some interesting results within the greater context of the study.

As mentioned previously each of the functions I built for the individual models returned accuracy, recall, precision, and f-measure as performance metrics. These metrics were chosen because they give a simplistic but relatively complete overview of each models performance while also being easily comparable to other studies using this dataset

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F-measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

IV. Results and dataset information

Before diving into the results I would like to first look at the dataset that was used for this project. The Diagnostic Wisconsin Breast Cancer Database contains 30 features and 569 samples, as mentioned previously it was first published in 1993 as part of a paper titled "Nuclear feature extraction for breast tumor diagnosis." [1] The continuous value features that make up this dataset were calculated from digitized images of fine needle aspirate (FNA) of breast cancer tumors. The dataset also includes binary classification data with 1 representing that the tumor is benign and 0 representing that it is malicious.

As can be seen in figure 1, the Naive Bayes model slightly outperformed the other two models across almost all of the used performance metrics, with the optimized Logistic Regression model on average outscoring it in recall. This is an interesting result as in most biological use cases a Naive Bayes tend to under-perform due to their independence assumption. Because of this the results here indicate that there is some amount of independence in existence amongst the datasets features.

Furthermore these results are of compounded interest due to the Naive Bayes being the most mathematically simplistic model of the three investigated model architectures. However, given the amount of random chance that is present within Logistic Regression models as, the small gap in performance, and the potential for further tuning within my implementation it is difficult to directly proclaim the Naive Bayes model as being superior in this use case. What can be determined however is that in this use case an SVM model is inferior to

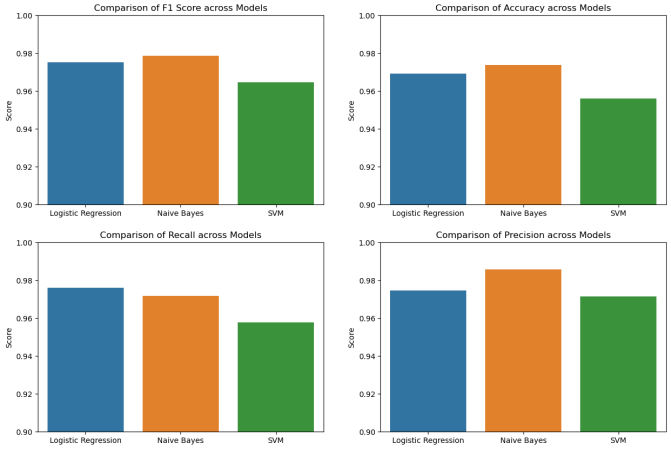

Fig. 1. Performance of each model across different metrics

the other two models used in this study due to its stability and lower performance. This may be due to a variety of causes such as the independence of the feature data or even the dimensionality of the data that is being worked with.

The results shown here slightly outperform the original model that was proposed in 1993 in the original paper which had an accuracy of 97 percent while my Naive Bayes model had an accuracy of 97.3 percent,[1] my models also slightly outperform the models created in the paper I mentioned previously from 2019 as there metrics only managed to get up to a 94 percent , however, after arranging there models into a voting ensemble the researchers from 2019 were able to achieve a similar accuracy of 97 percent.[2]

Conclusion

Machine learning has the potential to completely alter the way biomedical data is processed and analyzed, the performance metrics shown by the algorithms mentioned in this paper provide a strong case for the usage of machine learning in clinical settings. Furthermore, given the robustness found in my Naïve Bayes model for predicting malignancies in Breast Cancer I believe a strong case can be made for the usage of machine learning in this use case. Interestingly enough papers like the ones I mentioned previously have already catalyzed the use of machine learning in clinical settings. Of particular note is the usage of Convolutional Neural Networks (CNNs) to analyze numerous types of data such as; gene expression information, mammography data, and biopsy data analysis.[3]

While the usage of more complex models and techniques is certainly exciting and has been shown to be effective in many use cases I believe that something important can be taken away from the under-performance of my SVM model as well. I believe that these results show a strong potential use case for more simplistic models such as Naive Bayes in the clinical setting.

Future Work

Due to explosion in availability of biological data over the past few years there are many directions a continuation of this project could take. One approach could involve the usage of gene expression information in addition to past tumor characteristics to predict the recursion of cancer. Another route that could be taken is to attempt to create similar models to predict malignancy in other types of cancer. This study could also be further fine-tuned by using a feature selection methodology as well as other optimization techniques.

However, of all the future approaches I believe that a Deep Learning based approach that has the ability to parse both image-based data and genomic data has the potential to bear the most fruit due to Deep Learning's superior ability to account for dependencies within data which is absolutely crucial in a biomedical setting. This would most likely have to done in some form of ensemble-based approach but it has the potential to provide interesting insights into the links between different gene expression levels and tumor characteristics in the future.

References

[1] Street, W.N., Wolberg, W.H., and Mangasarian, O.L. (1993). Nuclear feature extraction for breast tumor diagnosis. Electronic imaging.

[2] Sidey-Gibbons, J.A., and Sidey-Gibbons, C.J. (2019). Machine learning in medicine: a practical introduction. BMC Medical Research Methodology, 19.

[3] N. Fatima, L. Liu, S. Hong and H. Ahmed, "Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis," in IEEE Access, vol. 8, pp. 150360-150376, 2020,