A Lightweight Logistic Regression Model for Breast Cancer Diagnosis (IMSE 514 HW3)

Samuel Bellaire
ECE Department
University of Michigan-Dearborn
Dearborn, MI, USA
srbellai@umich.edu

***Abstract*—Breast cancer is a highly prevalent disease amongst women worldwide. However, with an early diagnosis, it is also among the most survivable forms of cancer. We propose a high-accuracy logistic regression model trained on 5 different features from the Wisconsin Breast Cancer (Diagnostic) dataset to assist medical professionals in correctly diagnosing tumours in breast masses as either malignant or benign.**

***Index Terms*—Breast Cancer, Machine Learning, Logistic Regression**

I. Introduction

Breast cancer is one of the most common types of cancer in women worldwide [1], with over 12% of women in the United States being diagnosed with breast cancer at some point during their lifetime [2]. Fortunately, with modern medical care, the long-term survival rate of breast cancer is very high (over 80%) [1], [3], though the tumour must be detected as early as possible to maximize the probability of a good clinical outcome.

II. Related Work

While many researchers have modeled the development of breast cancer cells [4], [5] in the pursuit of more effective treatment strategies, artificial intelligence researchers have turned their attention towards developing automated prognostic and diagnostic tools to assist medical professionals in making a diagnosis.

Perhaps one of the most famous studies conducted is the formation of the Wisconsin Breast Cancer (Diagnostic) dataset [6]. Even though it was created in 1993, it is still used today by both machine learning beginners and academic researchers. A number of recent works in the literature have used the WBCD dataset [7]–[9] to classify whether a fine needle aspirate (FNA) sample originated from a benign or malignant tumour.

Machine learning is also used in a number of other medical applications. The authors of [10] developed an electronic nose system that analyzes the human breath for biomarkers indicative of COPD. In 2020, [11] identified 8 biomarkers that can be used to diagnose patients with COVID-19.

In this work, we propose a lightweight logistic regression framework that uses a low-dimensional feature space (5 features) to correctly classify breast mass tumours with a high degree of accuracy.

III. Methodology

A. Data Analysis

The WBCD dataset includes 569 records with 10 different metrics for each observation. Each of these metrics has a mean, standard deviation, and maximum value, which extrapolates to a total of 30 features in the dataset.

We can begin with the assumption of each sample being independent. Since each sample was collected from different test subjects with (presumably) no relationship to each other (e.g. genetic), it is reasonable to make this assumption of independence.

It should also be noted that there is a significant disparity in the scale between different features. The mean area exceeds 1000 in many samples, but the mean smoothness is less than 0.2 in every instance of the dataset. To achieve normalization, the range of each column of the input matrix is restricted to $[0, 1]$ by dividing each column by the maximum value of the corresponding feature. While not strictly necessary, this normalization process vastly increases convergence rate in the presence of highly disparate predictor values.

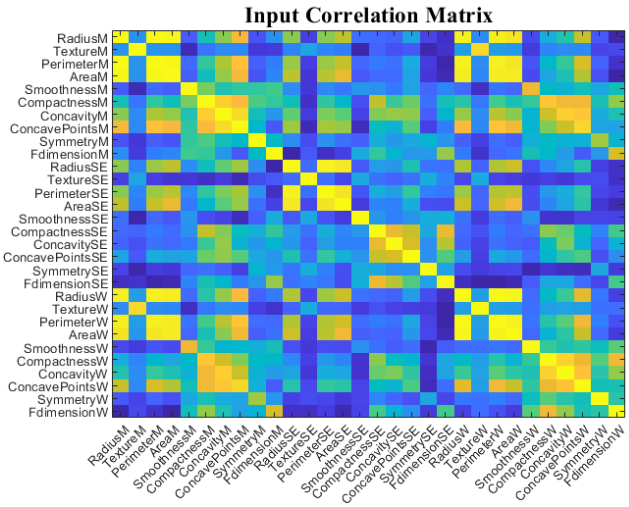

Fig. 1. Input Correlation Matrix (Yellow = 1, Blue = 0)

Next, the data must be analyzed for confounding effects by plotting the correlation matrix, which can be seen in Fig. 1.

We observe several regions with a high degree of correlation; most notably between radius, perimeter, and area. These three parameters exhibit a correlation of nearly 1. This occurs because these three parameters are mathematically related, even for objects that are not quite circular: perimeter is given by $P = 2\pi r$, and area by $A = \pi r^2$. Prior to training the model, some of these features were removed.

We also observe that compactness, concavity, and number of concave points are related, though the correlation is not as high. The confounding effects of these parameters can be reduced using L2 regularization while training the model. Interestingly, the compactness feature does not exhibit strong correlation with perimeter or area. The authors of the dataset note that compactness is computed by eq. 1.

$$\text{compactness} = \frac{P^2}{A - 1} \tag{1}$$

Clearly, there is a nonlinear relationship between the perimeter and area with the compactness feature, though it is not recognized by the linear correlation computation. This must be taken into consideration when building the model.

Lastly, we examine the distribution of the two classes: malignant and benign. The dataset contains 357 samples of benign tumours, and 212 of malignant tumours. Thus, in the null model, there is a slight bias towards the benign class (62.7%), though this is not extreme enough to require data augmentation or additional data collection.

B. Model Selection and Tuning

Since the aim of this work is to develop a binary diagnostic algorithm (benign or malignant), we implemented a logistic regression classifier. The parameters of the model were estimated using numerical methods, with the parameter gradients being given by eq. 2, where $\mathbf{X}$ is the design matrix, $\hat{\mathbf{y}}$ is given by the logistic function, and $\mathbf{y}$ is the observed values for each row.

$$\Delta\beta = \mathbf{X}^T(\hat{\mathbf{y}} - \mathbf{y}) \tag{2}$$

The system was trained and evaluated on a train-test split of 80-20 using stratified 5-fold cross-validation. Stochastic Gradient Descent with L2 regularization and a batch size of 16 was used as the optimizer, and the model was trained for 500 epochs. A few additional hyperparameters are listed below.

- $\eta = 1 \times 10^{-3}$ (Learning Rate)
- $\lambda = 1 \times 10^{-4}$ (L2 Regularization Coefficient)

The final model was trained using 5 of the 30 features in the dataset, which are also listed below. The process used to select these features is detailed in Sec. III-C.

- Max Concave Points
- Mean Radius
- Mean Texture
- Max Smoothness
- Standard Error of Concavity

C. Feature Selection

The method used in this work to perform feature selection is the bottom-up design approach. Typically, this approach is very time consuming and expensive compared to alternative feature selection strategies. However, the WBCD dataset is of small enough numerosity (569) and dimensionality (30) that a full bottom-up approach is not unreasonable.

Using the same hyperparameters that the final model will end up using (discussed previously in Sec. III-B), features are added one at a time and the model's accuracy with that additional parameter is re-evaluated 25 times. We average the accuracy over 25 independent training sessions to mitigate the probability of random chance resulting in a feature being selected as the best. The feature that results in the best model performance is then added to the final model, and this process is re-iterated until satisfactory performance is reached, or all features are added to the model.

After performing a bottom-up search, the 5 features identified in Sec. III-B were chosen, with the model having a preliminary accuracy metric of 95.8%. Interestingly, it appears that the Max Concave Points feature is highly indicative of whether a tumour is benign or not. A model using **only this feature** can achieve an accuracy of up to 90%. However, to increase the robustness of the model, we opt to use 5 features, which is still relatively lightweight compared to the initial set of 30 features.

IV. Results and Discussion

Table I shows the 5-fold cross-validation results for 5 separate training instances. In all 5 trials, it can be seen that the classification accuracy for each fold is high ($> 90\%$), which indicates that the model is robust and does not overfit on the training data. The model's overall accuracy is 96.46%.

TABLE I
5-Fold Cross-Validation Results

Trial	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
1	94.7%	94.7%	94.7%	92.9%	98.2%
2	94.7%	96.5%	95.6%	94.7%	92.9%
3	94.7%	96.5%	96.5%	94.7%	92.0%
4	97.3%	91.2%	96.5%	93.8%	96.5%
5	92.0%	91.2%	97.3%	96.5%	96.5%

Table II shows the confusion matrix for the logistic regression classifier, indicating that most testing instances are correctly classified. The model exhibits maximum precision at 100.0%, but the false-negative rate consequently suffers with a recall of 90.48%. The F1-score of the model is 95.0%.

TABLE II
Logistic Regression Confusion Matrix

	Pred. +	Pred. -
True +	38	4
True -	0	71

The ROC curve is shown in Fig. 2, with an AUC metric of 0.9956. By adjusting the classifier threshold, it can be seen that the false-positive rate of the classifier can be kept very low while maintaining a high level of sensitivity.

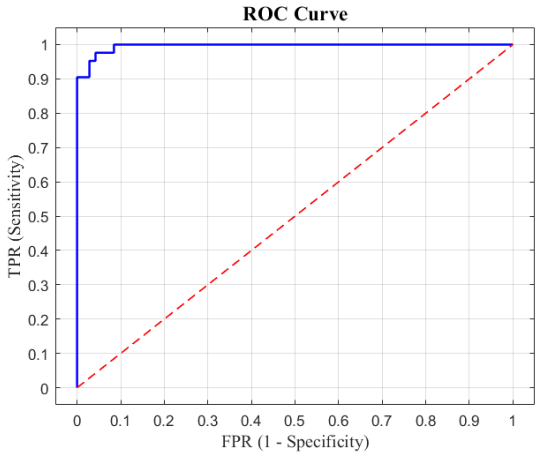

Fig. 2. ROC Curve

Arguably, for the automated prognostics domain, machine learning algorithms should aim to maximize the recall, even if the precision or other metrics of the model suffer. This can be accomplished by selecting an operating point on the ROC curve closer to the right edge of the graph. By doing so, the number of missed positive conditions is minimized, which considerably reduces the probability of a disease "flying under the radar."

Since these algorithms are intended to be used as an aid to trained healthcare professionals rather than replacing them, the decrease in precision as a result of maximizing recall is easily handled. If a false-positive result is given, the medical professional can manually review the diagnosis and classify it as a false-positive.

Fig. 3 shows the sigmoid transfer characteristic of the classifier from a linear domain into a probability space. Although the classifier's performance is high, the transfer characteristic is actually quite poor, with most training instances lying in the transition region. Thus, the model is very sensitive to adjustments in the classification threshold. The addition of more features improves the distribution of points on the transfer characteristic curve, though it comes at the cost of increased model complexity.

V. Conclusion

Our work has presented a lightweight logistic regression framework using just 5 of the 30 features in the WBCD dataset. The classifier was able to achieve a high accuracy, with a F1-Score of 95.0%, though the recall (arguably more important than precision in automatic diagnostics and prognostics) is only 90.48%. In future work, we plan to further tune this model to achieve a higher recall at the expense of the model's precision.

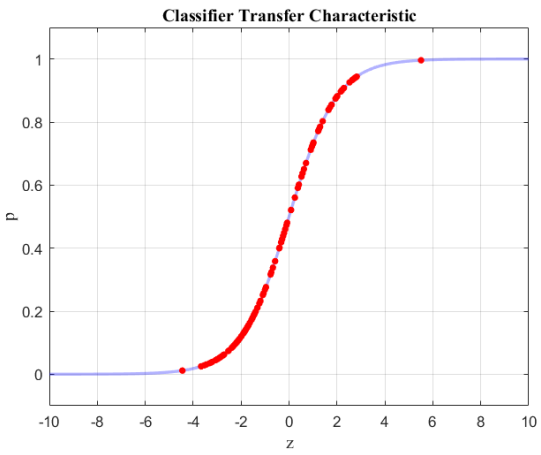

Fig. 3. Logistic Regression Transfer Characteristic

References

[1] N. Harbeck, F. Penault-Llorca, J. Cortes, M. Gnant, N. Houssami, P. Poortmans, K. Ruddy, J. Tsang, and F. Cardoso, "Breast cancer," *Nature reviews Disease primers*, vol. 5, no. 1, pp. 1–31, 2019.

[2] A. G. Waks and E. P. Winer, "Breast cancer treatment: a review," *Jama*, vol. 321, no. 3, pp. 288–300, 2019.

[3] Y.-S. Sun, Z. Zhao, Z.-N. Yang, F. Xu, H.-J. Lu, Z.-Y. Zhu, W. Shi, J. Jiang, P.-P. Yao, and H.-P. Zhu, "Risk factors and preventions of breast cancer," *International journal of biological sciences*, vol. 13, no. 11, p. 1387, 2017.

[4] I. Holen, V. Speirs, B. Morrissey, and K. Blyth, "In vivo models in breast cancer research: progress, challenges and future directions," *Disease models & mechanisms*, vol. 10, no. 4, pp. 359–371, 2017.

[5] D. L. Holliday and V. Speirs, "Choosing the right cell line for breast cancer research," *Breast cancer research*, vol. 13, pp. 1–7, 2011.

[6] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," in *Biomedical image processing and biomedical visualization*, vol. 1905. SPIE, 1993, pp. 861–870.

[7] A. Marcano-Cedeño, J. Quintanilla-Domínguez, and D. Andina, "Wbcd breast cancer database classification applying artificial metaplasticity neural network," *Expert Systems with Applications*, vol. 38, no. 8, pp. 9573–9579, 2011.

[8] M. H. Alshayeji, H. Ellethy, R. Gupta *et al.*, "Computer-aided detection of breast cancer on the wisconsin dataset: An artificial neural networks approach," *Biomedical Signal Processing and Control*, vol. 71, p. 103141, 2022.

[9] M. I. H. Showrov, M. T. Islam, M. D. Hossain, and M. S. Ahmed, "Performance comparison of three classifiers for the classification of breast cancer dataset," in *2019 4th International conference on electrical information and communication technology (EICT)*. IEEE, 2019, pp. 1–5.

[10] M. Rodríguez-Aguilar, L. D. de León-Martínez, P. Gorocica-Rosete, R. P. Padilla, I. Thirión-Romero, O. Ornelas-Rebolledo, and R. Flores-Ramírez, "Identification of breath-prints for the copd detection associated with smoking and household air pollution by electronic nose," *Respiratory Medicine*, vol. 163, p. 105901, 2020.

[11] P. Pan, Y. Li, Y. Xiao, B. Han, L. Su, M. Su, Y. Li, S. Zhang, D. Jiang, X. Chen *et al.*, "Prognostic assessment of covid-19 in the intensive care unit by machine learning methods: model development and validation," *Journal of medical Internet research*, vol. 22, no. 11, p. e23128, 2020.