A Comparison of Two Different Linear Classification Techniques on Various Datasets

Group 101 - Jonas Lehnert, Daniel Borisov, and Jeffrey Hyacinthe

Abstract—Two commonly used benchmark datasets in machine learning are the wine quality dataset, and the breast cancer classification dataset. The red wine dataset provides various features such as acidity and sugar content, and associates a subjective taste quality with it. Similarly, the breast cancer dataset incorporates multiple features in order to determine a classification of benign or malignant with it. Two linear classification models, logistic regression and linear discriminant analysis, were tested on these datasets to determine their classification accuracy and to contrast the methods. It was found that both models achieved comparable accuracies for classifying both the breast cancer data and the wine quality data. Both had high classification rates for classifying breast cancer data, and mediocre classification rates for wine quality. We found that the logistic regression model was more computationally efficient, having the quickest runtime. Different logistic regression parameters were tested, and it was found that a decaying learning rate without regularization, and the implementation of a couple of interaction terms, had the best performance with the wine dataset classification. Furthermore, a principle component analysis showed that only the first three principle components of the wine dataset were required to achieve a comparable accuracy to the full feature set. Further work would include attempting multi-class classification of the wine and a different model, such as a quadratic discriminant analysis. As well, we hope to test other datasets with these models to compare subjective and objective classifications.

1 Introduction

One of the central problems in Machine Learning concerns itself with the performance of various algorithms. In a paper by google, it was shown that very different algorithms can perform equally well under the condition that data is not sparse [1]. When data sets are very limited, this however often does not hold, as briefly discussed in [2]. In cases of small data sets, certain algorithms may perform better than others.

For this assignment we made use of two small data-sets, the wine data-set was a subset of the wine Quality data-set (Paulo Cortez 2009) containing only the red vinho verde samples and a Cancer data-set from Breast Cancer Wisconsin (Diagnostic) data-set (William H. Wolberg) . The wine-data set incorporates several measurements of wine properties. Neural network models and support vector machines have been tested on this data-set [3]. Here, a neural network was able to achieve an accuracy of 84.3%, while support vector machines achieved 86.8%. The cancer data-set deals with different medical measurements in patients with benign or malignant tumors. Several groups have already built models on this data set. In [4], artificial neural networks multivariate adaptive regression splines (MARS) were used that to achieve an average correct classification rate of 98.25%. The authors compare their method to simpler LDA classifiers that achieve only 95.91%. Other research groups applied binary logistic regression on the data set, and achieved an accuracy of 98.9% [5]. Two simple classification algorithms in machine learning are *logistic regression* (LR) and *Linear Discriminant Analysis* (LDA). While LR is discriminative and attempts to learn $P(y|\mathbf{x})$, LDA makes assumptions about $P(\mathbf{x}|y)$ and uses Bayes rule to estimate $P(y|\mathbf{x})$. In this report, we will compare the two methods by applying them on the wine and cancer data-sets via k-fold cross-validation. We report differences in performance of the two methods and implore regularization to investigate it's effects on performance.

2 Datasets

The wine data-set contained 1599 samples, multiple (11) features and the quality was measured on a scale of 10. It did not contain any missing values and all the features were real values but used different metrics with varying scales. We normalized the data via z-score averaging in order to counteract the impact that varying numerical magnitudes could

Jonas Lehnert and Jeffrey Hyacinthe are with the Department of Quantitative Life Sciences
Daniel Borisov is with the Department of Physiology
Project submitted September 28, 2019

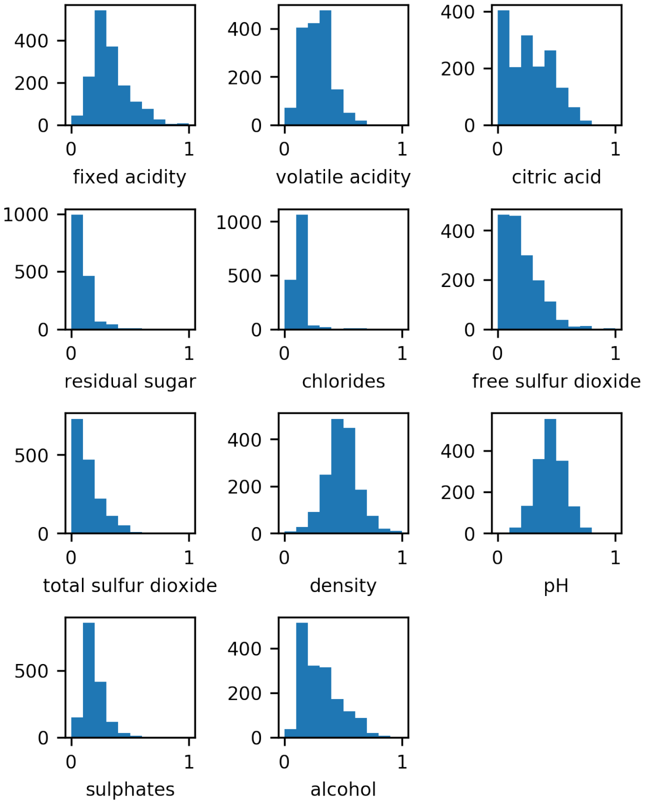

Fig. 1. Distribution of the various features of the wine data-set.

have on learning. In order to do our classification, we transformed the quality scale into a binary variable of 1 for qualities over 5 and 0 otherwise. We explored the data with multiple plots to see how the values for each features were distributed and observed that they were skewed towards 0 and often bi-modal(**Fig.1**, **Fig.2**). Before our normalization, all features except free and total sulfur dioxide had means close to 0 and a lot of outliers (**Fig.3**).

The cancer data-set had 699 samples with 11 features. The data-set featured some missing values with the value '?', we discarded all patient samples that had any missing values. The first feature, being the sample ID, was removed as a patient ID should not predict tumor malignancy. In total, 16 samples were removed. All features were real integers on a scale from 0 to 10. The class feature, the one we aimed to predict, was either 2 or 4 and we transformed it into a simpler 0, 1 binary scale (2=0, 4=1). Most of the features have a mean of 0, with a first quantile reaching about 5 and there was enough variance for the error bars to cover the full scale (0 to 10). However, some features with a bit less variance had some outliers such as marginal adhesion and uniformity of cell shape(**Fig.3**). We normalized this set via z-score averaging as well.

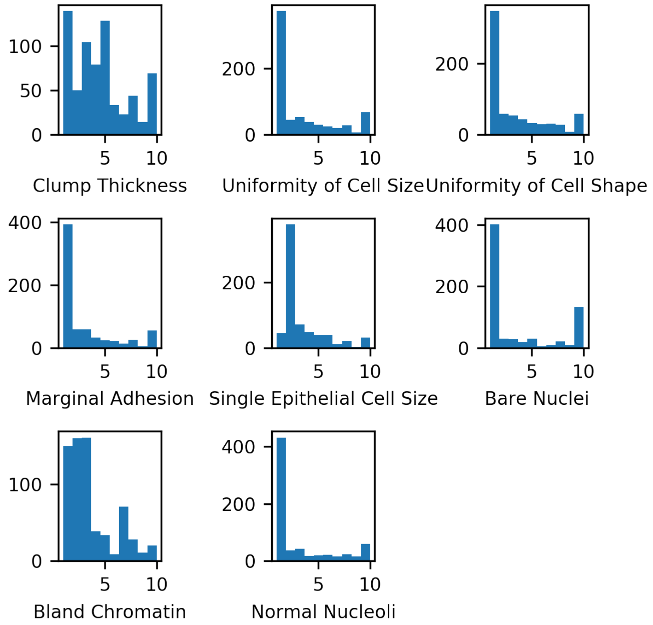

Fig. 2. Distribution of the various features of the cancer data-set.

Before the analysis, we shuffled the data-sets for k-fold cross validation to remove any potential bias from the order of samples.

The use of classification on the wine data-set could lead to ethical issues, if the classification was used to determine the perfect features for a high-quality wine rising up to in a worst case scenario some addictive wines. While trying to understand what features make a product of quality is certainly an understandable thing to do, we must be careful not to let it lead us too much and homogenize commercial products.

The cancer data-set leads to much more obvious ethical concerns. First, we should always be careful with the identification of health data, it is of utmost importance to protect it in order to prevent that information from reaching individuals that are not authorised. With the classification itself, there is clear good that can come from the ability to predict the class or any feature of cancer growth. However, we must be careful on how we use that data such that it is not used to diagnose people without their knowledge and without an expert to corroborate it.

3 Results

Various experiments were carried out to assess the performance results of simple logistic regression and linear discriminant analysis algorithms for predicting wine quality ratings and breast cancer classification. Accuracy was determined using a 5-fold cross validation algorithm, with the average

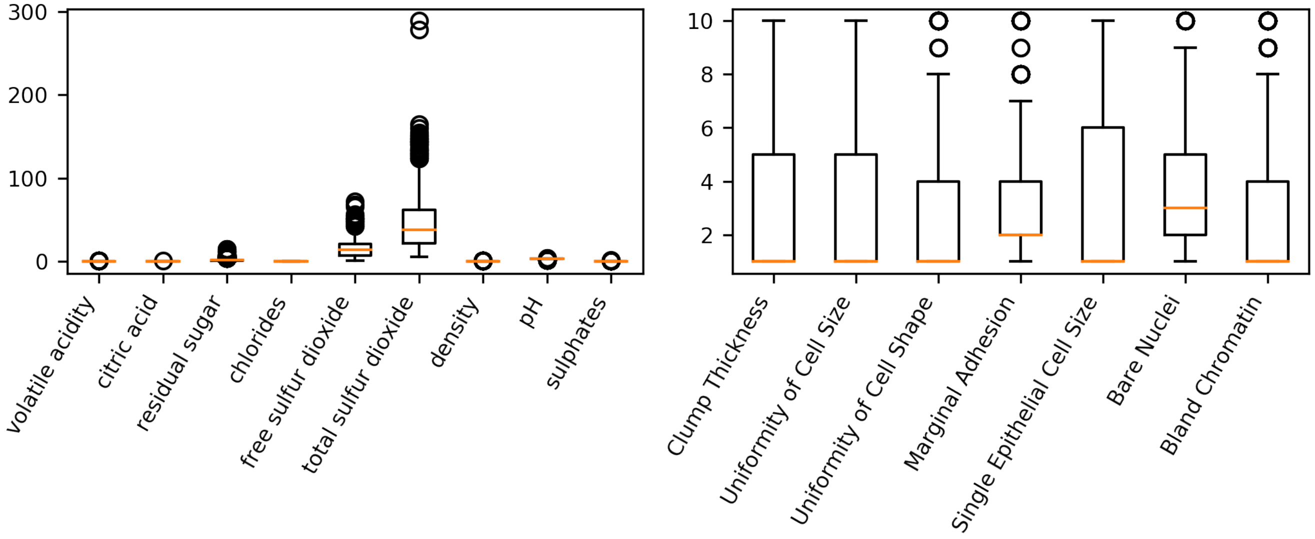

Fig. 3. Comparison of the mean and standard deviation of the different features via box-plot. In particular, observe the large differences in deviation and mean in the wine data-set.

accuracy used as the final performance metric. Furthermore, a variety of different logistic regression parameters were characterized in their influence on prediction accuracy on the two datasets. The first parameter analyzed was the iteration count, in which the learning rate was varied from 0.05 to 0.2 to assess its impact in predicting wine quality based on the provided data-set features. The results, shown in **Fig. 4**, **5**, illustrate that there exists a minimum learning rate required to allow for the model to converge in a reasonable amount of iterations, and that there also exists a point where an increase in learning rate begins to decrease the accuracy, indicating a failure to stabilize at a local minimum. Also shown, in red, are the results of a model implementing a schedule learning rate decay, utilising a step decay, in which the learning rate cuts by half every iteration. A decaying learning rate allows the model to stabilize at a level comparative to an ideal learning rate where it is large enough to allow convergence, while staying small enough to allow for stabilization.

The experiment also analyzed the impact of iteration rate on the accuracy of the model, utilising a stable learning rate as determined from the previous experiment, while also contrasting it with the impact of iterations on a model implementing a learning rate decay system. The results are shown in **Fig. 4**, **Fig. 5**. As shown, the learning rate decay model converges to a stable state quicker than the non-learning rate decay model. As a result, the learning rate decay model is more computational efficient, requiring less iterations to converge.

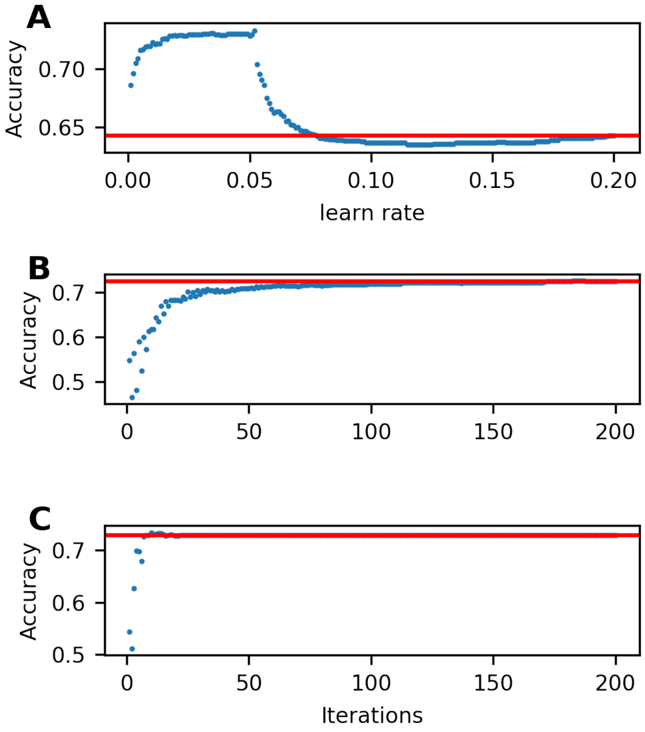

Fig. 4. Performance of Logistic regression on the wine-data set using: **A** Different learn rates. **B** Learning rate decay starting from a learning rate of $\lambda = 0.4$. **C** Learning rate decay starting from a learning rate of $\lambda = 5$. This underlines that learning rates seem to insignificantly effect how well logistic regression performs on the cancer data set.

Regularization was implemented and its influence on accuracy was assessed on both the breast cancer as well as the wine data-sets. Shown in **Fig. 6**, the model shows that while regularisation improves classification accuracy on a model that does not implement learning decay, the model fails to im-

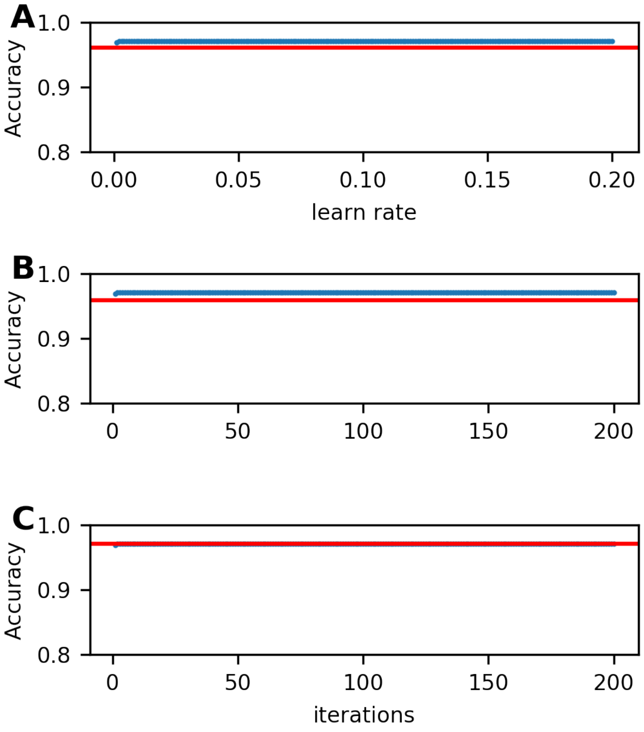

Fig. 5. Performance of Logistic regression on the cancer-data set using: **A** Different learn rates. **B** Learning rate decay starting from a learning rate of $\lambda = 0.4$. **C** Learning rate decay starting from a learning rate of $\lambda = 5$.

prove the accuracy beyond the learning rate decay model, and regularization implemented alongside a learning rate decay actually decreases classification performance as the λ term is increased.

Attempts at improving wine quality classification by increasing model complexity failed to significantly improve accuracy, in some cases decreasing it. As shown in **Fig. 7**, adding quadratic feature complexity reduced accuracy in both a learning rate decay and a non-decaying model. Looking at feature interaction, accuracy was slightly performed by calculating the free to total sulphur dioxide ratio to include as a feature, as well as including the ratio of volatile acidity to fixed acidity.

Furthermore, a principle component analysis was performed to assess whether the model could be improved through dimensionality reduction, as well as to assess how many principle components are required to reach the performance levels seen with the full feature set. As shown in **Fig. 8**, only the first three principle components are required to create a model with comparable performance to a model utilising the full feature set. The analysis also shows that utilising the first 10 of the 11 principle components produces more accurate results than all 11 components. Dimensionality reduction can be useful, as it makes model fitting computationally less expensive, and can potentially decrease our chance of over-fitting.

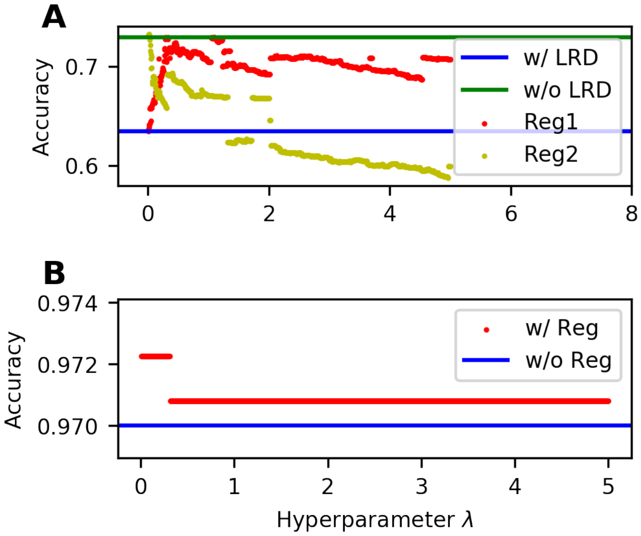

Fig. 6. Testing of regularization parameterization on accuracy of the models. **A** Wine data classification showing regularization without a learning rate decay, shown in red, compared to a baseline without regularization, shown in blue. Similarly, regularization was tested with a learning rate decay, shown in yellow, compared with the learning rate decay without regularization, shown in green. **B** Regularization of the breast cancer dataset, showing the regularization accuracy of a model implementing a learning rate decay, shown in red, when compared to baseline without regularization, shown in blue.

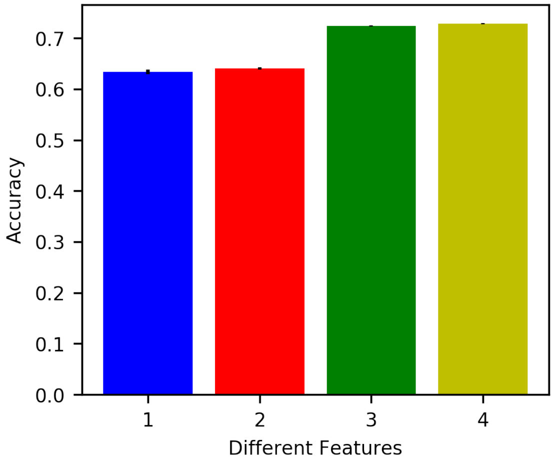

Fig. 7. Graph showing results of the inclusion of squared feature values, increasing the model complexity. Quadratic feature inclusion with no learning rate decay is shown in blue, with no inclusion in red. Similarly, the inclusion of quadratic features with a learning rate decay model is shown in green, and without the inclusion of features in yellow. As seen, the inclusion of quadratic features slightly decreases accuracy of the model in a 5-fold cross validation test.

Finally, we compared the results of the logistic regression with the results of a linear discriminant analysis model (LDA), shown in **Fig.9**. The results show that the LDA model has comparable results

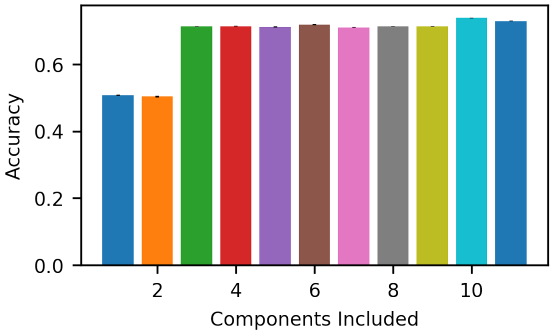

Fig. 8. Graph showing wine classification performance based on the inclusion of different principle components. As shown, the first three principle components can allow for comparable accuracy to the performance of the model that utilises the full feature set.

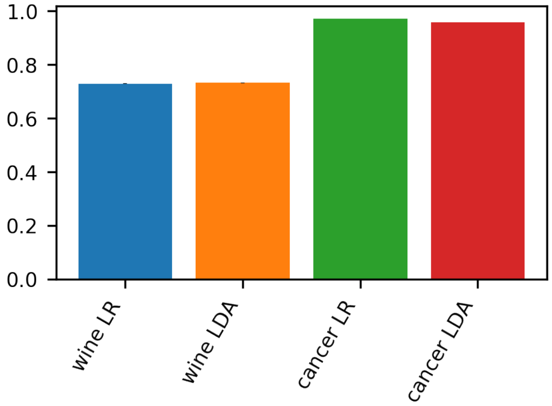

Fig. 9. Graph contrasting the performance of the logistic regression model and the linear discriminant analysis model in classifying the wine data and the breast cancer data. As shown, both models had comparable accuracies for both data-sets.

to the logistic regression model. Furthermore, we analysed the model efficiency to determine whether the LDA or logistic regression model had a shorter run-time for training and classifying the results in the 5 fold cross validation. We found that the Logistic regression model was more computationally efficient, with the LDA model having a run-time of 1.23 seconds (wine data)/ 0.73 seconds (cancer data), compared to the 0.025 seconds (wine data)/ 0.011 seconds (cancer data) of the logistic regression model for the breast cancer data-set.

4 Discussion

We were able to demonstrate that both the logistic regression model and the linear discriminant analysis model are both valid and comparable models for classifying the wine quality data-set and the breast cancer class data-set. Both have very high accuracy classifying the breast cancer data-set, exceeding 95% classification accuracy. However, both the logistic regression model and the linear discriminant analysis model are only able to achieve accuracies of around 70-75 %. It was demonstrated that the logistic regression model was able to improve classification by utilising some interaction terms, namely ratios for acidity and sulfur content, as well as by utilising a learning rate decay system.

One interpretation of the results could say that while both the LDA and logistic regression model are accurate models of objective classifications, they both are unable to classify subjective measures, such as wine quality. To test this hypothesis, future work might try and utilise the same model implementation on different data-sets that include both objective classification results, as well as subjective measures as determined by human satisfaction scales, and comparing the accuracy of the results. Another interpretation of the results could be that the breast cancer data-set measured the best features required for cancer type classification, while the features used to attempt to describe the wines tested are uncorrelated with the quality of the wine, and that unknown features that were not accounted for might better classify the quality.

Future work would also include testing a quadratic linear analysis model for wine classification, as well as utilising a multi-class model to determine whether a model could perform better by estimating the quality measure directly, compared to trying to determine whether the wine quality is simply good or bad. Apart from these findings, we investigated the effects of regularization onto logistic regression performance. To our surprise, regularization did not improve model performance when the learning rate decayed, but it did increase performance when learning rates stayed constant. We also showed that the wine data can be reduced to 3 dimensions via principle component analysis which is useful for computational efficacy.

5 Statement of Contributions

The workload of the project was equally distributed across all three group members. Jonas Lehnert was responsible for the implementation of the logistic regression class, Jeffrey Hyacinthe was responsible for the implementation of the linear discriminant analysis class, and Daniel Borisov was responsible for utilising these classes to run the experiments. Similarly, write-up breakdown was equally distributed across the Introduction, Dataset analysis, Results/Discussion to Jonas, Jeffrey and Daniel, respectively. The abstract was collaboratively written by the group.

REFERENCES

[1] M. Banko and E. Brill, "Scaling to very very large corpora for natural language disambiguation," in *Proceedings of the 39th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2001, pp. 26–33.

[2] A. Géron, *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*. " O'Reilly Media, Inc.", 2017.

[3] P. Cortez, J. Teixeira, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Using data mining for wine quality assessment," in *International Conference on Discovery Science*. Springer, 2009, pp. 66–79.

[4] S.-M. Chou, T.-S. Lee, Y. E. Shao, and I.-F. Chen, "Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines," *Expert systems with applications*, vol. 27, no. 1, pp. 133–142, 2004.

[5] A. F. Seddik and D. M. Shawky, "Logistic regression model for breast cancer automatic diagnosis," in *2015 SAI Intelligent Systems Conference (IntelliSys)*. IEEE, 2015, pp. 150–154.