Applying Machine-Learning to Predict Breast Cancer Classification from Adipokine Levels

Tyler Tish, Akhil Ramesh, Suraj Menon

Abstract

Breast cancer is a common form of cancer that can often be fatal, especially if not detected early. There are factors that can indicate the possible presence of breast cancer through the screening that is done at routine check-ups and simple blood work. Comparing data for obesity-related factors possibly linked to breast cancer between patients with breast cancer and a healthy control group using different artificial intelligence methods, allows us to discern a possible correlation between some of these factors and the presence of breast cancer. The artificial intelligence methods in order to analyze the relevance to predicting the presence of breast cancer were unsupervised learning and supervised learning models. Unsupervised learning algorithms were applied in order to determine how predictor data groups together and see if there are any patterns in the data. Supervised learning algorithms were applied in order to find which features are most relevant in order to determine breast cancer classification. From the supervised learning model, logistic regression was performed which shows that the BMI, resistin levels, and glucose levels were significant in predicting patient breast cancer classification. From this data it can be concluded that monitoring the BMI, resistin levels, and glucose levels can help determine the presence of breast cancer early on.

Introduction

Cancer is a disease in which cells in the human body grow at an unregulated rate, causing many health complications. Breast cancer is one the most common types of cancer found in women across the world. With 42,000 women and 500 men dying from breast cancer annually in the United States alone, predictive measures to catch the disease in its early stages are crucial to reducing the mortality rate of this condition. There are various factors that can be detected early on and indicate the presence of breast cancer. For example, obesity is a condition that has been closely correlated to increased cancer rates and increased severity of the disease as well as higher fat content and insulin levels can influence the progression of breast cancer (Simon et al., 2018). The objective for this project is to examine data in the dataset (size:116x10) provided to us by Coimbra University and use predictive modeling to find the obesity factors most closely correlated with the presence of breast cancer. The data was sampled from 116 Portuguese pre- and postmenopausal overweight and obese women. The original objective of the study that created this dataset was to illuminate the relationship between adipokines (leptin, adiponectin, etc.) and breast cancer as opposed to just looking at body weight as an obesity factor. Adipokines are cell-signaling molecules that play a major role in the energy and metabolic state of the body, inflammation, and obesity as well (Mahmood et al., 2020). The biological rationale behind examining BMI, glucose levels, and insulin resistance is that these are all the most common obesity factors. The study aims to determine if adipokines also play a role in the development of breast cancer and thus the original study measures leptin, adiponectin, resistin, and monocyte chemotactic protein-1 levels (Crisóstomo et al., 2016). This is done for all the features in the dataset. These features consist of age, BMI, glucose levels, insulin resistance, HOMA, leptin levels, adiponectin levels, resistin levels, and MCP1 levels. Included in the dataset is a classification of whether the subject is a healthy patient or a patient with breast cancer. A “1” indicates that the subject is a healthy control and a “2” signals that the subject is a breast cancer patient. The dataset consists of 64 breast cancer patients and 52 healthy controls. Implementing AI can help solve this problem by analyzing various factors from breast cancer patients and comparing them to the same factors recorded from the healthy control group. This analysis will show the

relevance of these factors and their correlation to the possible presence of breast cancer. Artificial intelligence allows very different tests to be run and conclusions can be drawn from each of these tests in order to compare results and most accurately determine early indicators of breast cancer.

Methods

The data for this project was obtained from Coimbra University as a Microsoft Excel spreadsheet and the data was then imported into MATLAB using the “readtable” function. When processing this data, the classification of the data was then converted from a string into an integer so that the data could be easily divided into a healthy dataset and a patient dataset for comparing the two. This data did not require much normalization as there were no missing values in the dataset.

A variety of data exploration techniques that were used for determining the relevance of these factors to the presence of breast cancer. Histograms were developed across adipokines to analyze differences in distribution of values for healthy patients vs diseased patients. Scatter plots and pie charts were not used as they do not display the information well and the data is too discrete to show differences between values for healthy and diseased patients. A ttest was applied in order to determine if there was any significant difference between the breast cancer patients and the healthy control for each of the factors recorded. Principal component analysis (PCA) was performed on z-score normalized adipokine data in order to reduce the dimensionality of data to increase interpretability. This was done by developing a matrix of principal component (PC) coefficients and transformed data, as well as a vector containinging variances explained by each PC. ANOVA tests were performed comparing healthy patients to diseased patients for each adipokine to find differences in means across groups. ANOVA p-values were analyzed at α=0.05 to determine statistical significance. K Means-clustering was performed on all adipokine data to find optimal clustering of patient data, and this optimal clustering was found by finding the mean of all euclidean distances from each observation to their respective clusters. The optimal cluster value corresponded to the highest of these means, assuming at least one mean value was greater than 0.5. Hierarchical clustering was performed using Euclidean distance as the basis for similarity and standardized patient data to analyze similarities in obesity predictor data across diseased patients, as well as identifying differing health profiles between diseased and healthy patients.

Three supervising machine learning algorithms were used. Logistic Regression was used to identify significant relationships between multiple factors, all adipokines, and patient classification. To do this, a logistic regression model was fitted to the data to generate model coefficients, of which only coefficients with a corresponding p-value greater than 0.05 were used in the model. A decision tree was also used, which consists of a set of decisions based on training data, and uses binary/ternary rules to calculate a target value, in this case, patient classification. To do this, a decision tree model was fitted to the data to find features most important to predicting patient classification. In the decision tree, hyper-parameters were optimized to develop the most efficient model possible. Random forests consist of multiple decision trees to develop more accurate predictions. The process that was used to fit data to the decision tree was also used in building the random forest model with optimized hyperparameters.

Training and test sets were generated during hold-out validation, where 80% of the data was partitioned into the training set and the remaining into the test set. All supervised learning models were evaluated using n-fold cross-validation, where the data was partitioned into n subsets, where one set served as a test set and each model was trained on the remaining data and tested using this test set. Seeing as predicted data was categorical, accuracy from this cross-validation was used to evaluate these models. Parameters such as mean absolute error, correlation, or precision would not yield meaningful results when interpreting categorical data.Random guesses were based on generating random numbers within three standard deviations of the mean(99.7% of all data) for each predictor in the X training data and

fitting them to a supervised learning model, of which resulting predictions were compared to accuracy scores for the models containing real data.

Results

Feature	Median (Healthy)	Range (Healthy)	Median (Patient)	Range (Patient)
Age (years)	65	65	53	52
BMI (kg/m^2)	27.694	19.909	27.408	18.739
Glucose (mg/dL)	87	58	98.5	131
Insulin (µU/mL)	5.4835	23.504	7.58	56.028
HOMA	1.1397	6.64449	.50794	24.54206
Leptin (ng/mL)	21.495	79.171	18.878	83.9461
Adiponectin (µg/mL)	8.1278	35.8457	8.4464	32.094
Resistin (ng/mL)	8.9292	52.005	14.372	78.8083
MCP-1(pg/dL)	471.32	1210.257	465.37	1608.31

Figure 1: Summary Statistics for Healthy and Patient Groups in Coimbra Dataset

PC1 Coefficient	Feature
0.49285	HOMA
0.44398	Insulin
0.43902	Glucose
0.33149	Leptin
0.28174	Resistin
0.26043	BMI
0.25463	MCP_1
0.12457	Age
-0.17261	Adiponectin

Figure 2: Ranked Table of PC1 Coefficients

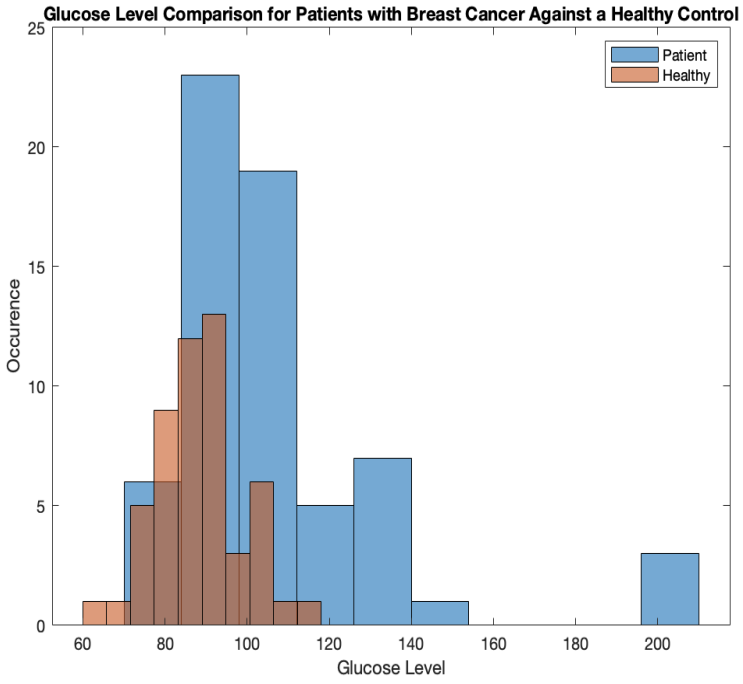

Figure 3: Glucose Level Histogram

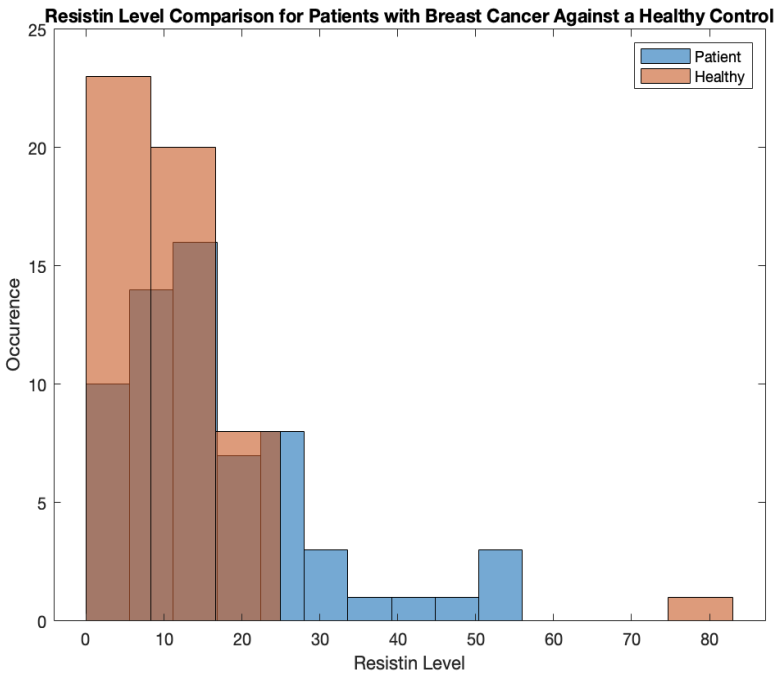

Figure 4: Resistin Level Histogram

Figure 5: Correlation Heat Map for Coimbra Breast Cancer Dataset Features

Feature Name	P Value
Age	0.6425
BMI	0.1560
Glucose	0.0000
Insulin Resistance	0.0026
HOMA	0.0020
Leptin	0.9908
Adiponectin	0.8355
Resistin	0.0141
MCP-1	0.3293

Figure 6: Two Sample T Test between Patient and Healthy Groups

Applying a two-sample t test to the dataset revealed the factors that had a statistically significant difference in means between the patient and control groups. As demonstrated by Figure 6, these factors were glucose levels, insulin resistance, HOMA, and resistin levels.

Using ANOVA, it was determined that there were differences in the means between diseased and healthy patients for Glucose,Insulin,HOMA, and Resistin levels. As figures 3 and 4 show, Glucose and Resistin showed the greatest variances in distribution of levels when comparing healthy to diseased patients.

As shown by Figure 2, HOMA, Insulin, and Glucose have the largest coefficient values associated with PC1. Breast cancer patient values cluster more closely to PC1, while healthy patient values cluster around PC2. For the clustergram, Patients seem to be closest in Euclidean distance when looking at HOMA and Insulin values. Assuming k values 1-10, k-means clustering shows that patient data is best split into 3 groups, as this was the k value corresponding to the most optimal silhouette score (0.7617).

Performing logistic regression, it was found that only 3 factors had model coefficients that were statistically significant , with those factors being BMI(p=0.0261), Glucose(p=.0024), and Resistin(p=.0498). Using the corresponding coefficient values, the logistic regression model is $\ln(\frac{\pi healthy}{\pi breastcancer})=5.6512-0.1501X_1+0.1056X_2+0.0586X_3$, where X_1, X_2, and X_3 are the predictor variables BMI, Glucose, and Resistin, respectively. The model reports an average accuracy of 0.1279 when performing 5-fold cross validation.

The decision tree reported a minimum leaf size of 2, the maximum number of branch splits as 92, and the minimum number of samples in each branch node as 10. When evaluated using 3-fold cross validation, this model reported an average accuracy of 0.6817. This accuracy is higher than that of a random guess, which reported an accuracy of 0.6522.

When evaluated using 3-fold cross validation, the random forest model reported 0.6730 accuracy, lower than when compared to a random guess, 0.7826.

Discussion and Conclusion

For the clustergram, not many decisive conclusions can be drawn. Breast cancer patients do seem to somewhat cluster together; however, this clustering is not very strong and there are plenty deviations from this. K-means clustering shows the data is well-clustered when split into 3 groups. PCA shows that HOMA, Insulin, and Glucose are the 3 features most responsible for patterns seen in this data, as they have the largest coefficient values associated with PC1, which explains the most variance in the data. The t test shows that there is a significant difference between the healthy control and the patients with cancer for glucose, insulin, HOMA, and resistin levels.

Using these results, it is hypothesized that HOMA, BMI, Insulin, and Glucose will be the features with coefficients that are most significant to a supervised ML model.

The stated hypothesis successfully predicted the importance of BMI and Glucose in the logistic regression model; however, Resistin was not predicted to be significant. A negative coefficient value in the logistic regression model represents increases in BMI levels indicating a greater likelihood that a patient has breast cancer, while negative coefficients represent increases in Glucose and Resistin indicating a greater likelihood that a patient is healthy. These results align with finds in current literature (Tamaki et al., 2014) (Barba et al., 2012) (Georgiou et al., 2016). However, the model reports a very low accuracy score, and it is hypothesized that this may be because only three of 10 model coefficients were statistically significant, meaning the remaining coefficients were included in the model while testing even though they were not

significant to the model. To improve this model accuracy, more patient samples may be needed to have more statistically significant model coefficients. This model performed similarly across all subsets of data, but fits closest to BMI, Glucose, and Resistin data.

With 92 branch splits, the decision tree model has very high depth, which indicates high complexity and likely that the model overfits the data. No outlier data values and accuracy scores were found for either the decision tree or random forest model. It is hypothesized that accuracies for these models could be improved by standardizing training data to minimize the effect of outliers and altering hyperparameter values in the decision tree to find which ones have the greatest impact on model accuracy. For both models, it was found that Age, Glucose, and Resistin were the most important features in predicting the model. Out of these features, Glucose was the only feature predicted to be significant in the earlier hypothesis.This feature grouping makes sense, as scientific literature has shown strong correlations between age and resistin (Acquarone et al., 2019), glucose levels and age (Basu et al., 2003), and resistin and glucose levels (Li et al., 2009). These models perform similarly across all subsets of data.

The limitations of this study are the data used in this project was collected from Portuguese women and may not be indicative of women across the globe as these people all live in the same region. While the factors found to be relevant may be consistent for women across the globe, data that has a wider regional variation in samples would be required in order to determine whether these factors are consistently significant in different countries. Another limitation of this study is that this data was collected from only women despite men also suffering from breast cancer, although it is much less common for men to have breast cancer than it is for women. This means that the conclusions drawn from this study do not necessarily apply to men and more data and testing would be required to determine if the significance of these factors is consistent across genders.

Future directions of this project could involve doing this testing on men and to collect data from women from other regions of the world, as mentioned above. This testing would help solidify the significance of these findings and would help motivate hospitals to regularly test these factors and monitor their levels to help determine early signs of cancer and treat this cancer before it progresses.

Bonus

A binary Gaussian kernel classification model was also fitted to the dataset, in which a support-vector machine(SVM) analyzed the data for classification, based on a Gaussian kernel distribution, to find correlations between obesity predictors and patient classification. This was evaluated using 3-fold cross validation, and reported an average accuracy of 0.5430 across 3-fold cross validation. This relatively low accuracy score indicates that while a Gaussian kernel distribution fits the data relatively well, it may not be optimally suited for the distribution of data. Accuracy values could be improved by standardizing the data to have it fit more to a Gaussian kernel distribution.

Additionally, a neural network was used to make predictions on the dataset. Neural networks take input data and feed it into a hidden layer, which is a modular stack of equations used to combine the data non-linearly, to return classification values, in this case, patient classification. Neural networks with 1 and 3 hidden layers were created. The model with one hidden layer reported an accuracy of 0.8183 after validation, while the 3-hidden layer model reported a score of 0.5936 after validation. It is hypothesized that this difference in accuracy stems from some over-fitting that may occur when the additional layers are added to the network, which could cause some noise in the prediction data. Accuracy values could be improved by standardizing the data set more to make predictor data more linear, so that the neural network can combine it more efficiently.

References

Acquarone, E., Monacelli, F., Borghi, R., Nencioni, A., & Odetti, P. (2019). Resistin: A reappraisal. *Mechanisms of Ageing and Development*, *178*, 46–63. https://doi.org/10.1016/j.mad.2019.01.004

Barba, M., Sperati, F., Stranges, S., Carlomagno, C., Nasti, G., Iaffaioli, V., Caolo, G., Mottolese, M., Botti, G., Terrenato, I., Vici, P., Serpico, D., Giordano, A., D'Aiuto, G., Crispo, A., Montella, M., Capurso, G., Fave, G. D., Fuhrman, B., & Botti, C. (2012). Fasting glucose and treatment outcome in breast and colorectal cancer patients treated with targeted agents: results from a historic cohort. *Annals of Oncology*, *23*(7), 1838–1845. https://doi.org/10.1093/annonc/mdr540

"Basic Information about Breast Cancer." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 20 Sept. 2021, www.cdc.gov/cancer/breast/basic_info/index.htm#:~:text=Each%20year%20in%20the%20United, breast%20cancer%20than%20White%20women.

Basu, R., Breda, E., Oberg, A. L., Powell, C. C., Dalla Man, C., Basu, A., Vittone, J. L., Klee, G. G., Arora, P., Jensen, M. D., Toffolo, G., Cobelli, C., & Rizza, R. A. (2003). Mechanisms of the age-associated deterioration in glucose tolerance: contribution of alterations in insulin secretion, action, and clearance. *Diabetes*, *52*(7), 1738–1748. https://doi.org/10.2337/diabetes.52.7.1738

Christodoulatos, Gerasimos Socrates et al. "The Role of Adipokines in Breast Cancer: Current Evidence and Perspectives." Current obesity reports vol. 8,4 (2019): 413-433. doi:10.1007/s13679-019-00364-y

Crisóstomo, Joana, et al. "Hyperresistinemia and Metabolic Dysregulation: A Risky Crosstalk in Obese Breast Cancer." Endocrine, vol. 53, no. 2, 2016, pp. 433–442., doi:10.1007/s12020-016-0893-x.

Georgiou, G. P., Provatopoulou, X., Kalogera, E., Siasos, G., Menenakos, E., Zografos, G. C., & Gounaris, A. (2016). Serum resistin is inversely related to breast cancer risk in premenopausal women. *Breast (Edinburgh, Scotland)*, *29*, 163–169. https://doi.org/10.1016/j.breast.2016.07.025

Li, F.-P., He, J., Li, Z.-Z., Luo, Z.-F., Yan, L., & Li, Y. (2009). Effects of resistin expression on glucose metabolism and hepatic insulin resistance. *Endocrine*, *35*(2), 243–251. https://doi.org/10.1007/s12020-009-9148-4

Mahmood, T., Arulkumaran, S., & Chervenak, F. (2020). *Obesity and Gynecology* (2nd ed.). Elsevier.

Patricio, M., Pereira, J., Cristomo, J., Matafome, P., Seica, R., & Caramelo, F. (2018). UCI Machine Learning Repository: Breast Cancer Coimbra Data Set. Retrieved March 18, 2022, from archive.ics.uci.edu website: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra

Simon, Stacy. “How Your Weight May Affect Your Risk of Breast Cancer.” American Cancer Society, American Cancer Society, 4 Oct. 2018, www.cancer.org/latest-news/how-your-weight-affects-your-risk-of-breast-cancer.html#:~:text=Having%20more%20fat%20tissue%20can,some%20cancers%2C%20including%20breast%20cancer.

Tamaki, K., Tamaki, N., Terukina, S., Kamada, Y., Uehara, K., Arakaki, M., Miyashita, M., Ishida, T., McNamara, K. M., Ohuchi, N., & Sasano, H. (2014). The correlation between body mass index and breast cancer risk or estrogen receptor status in Okinawan women. *The Tohoku Journal of Experimental Medicine*, *234*(3), 169–174. https://doi.org/10.1620/tjem.234.169