Final Report
CSC 859 SP 2021

Dr. Dragutin Petkovic, Ph.D.

18 May 2021

Team One
Members:
Lauren Luke
Nianzi Yi
Lynn Mari

I. Executive Summary

Abstract

Background: Given the data from the Patrício, et al. study, we applied ML methods to derive the accuracy of this prediction model that was designed to be used as a biomarker for breast cancer.

Method: Our method consisted of using the Random Forest Classifier to train the data set for 116 patients. The tools we used were Jupyter Notebook and Scikit Learn. Once the ML was applied we were able to extract the features that contributed the most importance to the study. Our study concluded by performing an ethical audit and evaluation of our results.

Results: The top 4 features of highest importance included Glucose, Resistin, Age and BMI with an overall 81% accuracy of the model. We can conclude that with these results this test may prove some help when detecting early signs of breast cancer but should not be used as the only factor for detection.

Conclusions: We concluded this prediction model to be a sufficient addition to use for the prediction and detection of breast cancer but should not be relied on solely. The ease of accessibility and non-invasive structure of the test are some of the benefits that come with this test.

II. Motivation, Problem Description and Case Study Goals

Motivation

Breast cancer is an ongoing battle throughout society today that without proper and accurate detection could be detrimental to someone's life. Improving the accessibility and reliability of testing for women's health is a problem that faces many different factors. New ideas and methods are being designed to create a new potential biomarker to detect the presence of breast cancer that is affordable and easily obtainable through routine blood analysis.

Problem Description

The accuracy of the diagnosis could be improved, only 81% accuracy. There is also the problem of whether or not the right features are being tested. Designing and creating a potential biomarker to detect early signs of breast cancer with the accuracy that was calculated shows that it should be not used in place of a mammogram. The results themselves with the accuracy rate the test calculated also could potentially misdiagnose a patient. One also has to keep in mind the fact that this test may be missing certain features that contribute to breast cancer and are not being tested for.

Case Study Goals

Once a group of participants was solidified, the goal was then to assess hyperresistinemia and metabolic dysregulation in breast cancer. Routine blood work would take place with every patient undergoing identical processes to extract the data for the 9 features being tested. Once all features were tested the top features with the highest importance were identified as being the most accurate biomarkers for detecting breast cancer for this specific study. This test in turn is not to be used solely to detect breast cancer or to be used in place of a mammogram but to be used as a predicting model.

III. ML Study

Description of The Application

The data set we used for this project is “Breast Cancer Coimbra Data Set”, and it is from UCI Machine Learning Repository. It contains 10 predictors, 9 quantitative clinical features and 1 binary dependent variable, to indicate the absence or the presence of breast cancer. The 9 features were observed for 64 patients with breast cancer and 52 healthy controls, and they can be gathered in routine blood analysis. The label predictor contains binary values 1 for healthy controls and 2 for patients. In this project, to train our model more conveniently, we changed label value to 0 for healthy controls and 1 for patients. The prediction model we trained is based on these 10 predictions.

Description of Training Data

The training database only has numerical features(columns), no missing data, and class label (0 or 1) in the last column. In the “breast.csv” file, there are 117 rows, out of which row 1 is titles of columns and row 2 to 117 (total of 116 rows) are samples, and 10 columns with 9 features because the last column called “Classification” is the class label column with value 0 indicating - class sample, and value 1 indicating + class samples. According to Excel, there are 52 negative samples (healthy samples) and 64 positive samples (patient samples) in the original dataset. The number of positive samples and negative samples are balanced.

ML Applied and Evaluation Methods

In this project, we chose the Random Forest Classifier model to classify training samples. As a result, the software tools we used are Jupyter Notebook and ScKit Learn Python ML tool. After we got the trained RF model, at the evaluation stage, we chose the RFEX Model explainer that contained some core components like Cumulative F1 score and Cohen distance, to help us explain the trained RF model in an ethical way.

ML experiment

First of all, to start the ML experiment, we read the file "breast.csv" and created a data frame as our training database.

```
In [2]: #create the dataframe by reading 'breast.csv' file
        df = pd.read_csv('breast.csv')
```

(Code of processing data file)

After that, we printed out the data frame to do a sanity check. We can see that the data frame is 116 rows x 10 columns, which means there are 116 samples, 9 features , and the 10th column is the class label. The number of features and samples match the original dataset. To do spot check, we picked the first 5 samples and the last 5 samples (as shown below), and compared feature values and labels of these 10 samples with the data in the original dataset. The data of the imported data frame also matches the original data.

```
In [3]: #display the dataframe
        df
```

Out[3]:

	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Classification
0	48	23.500000	70	2.707	0.467409	8.8071	9.702400	7.99585	417.114	0
1	83	20.690495	92	3.115	0.706897	8.8438	5.429285	4.06405	468.786	0
2	82	23.124670	91	4.498	1.009651	17.9393	22.432040	9.27715	554.697	0
3	68	21.367521	77	3.226	0.612725	9.8827	7.169560	12.76600	928.220	0
4	86	21.111111	92	3.549	0.805386	6.6994	4.819240	10.57635	773.920	0
...	...	...	...	...	...	...	...	...	...	...
111	45	26.850000	92	3.330	0.755688	54.6800	12.100000	10.96000	268.230	1
112	62	26.840000	100	4.530	1.117400	12.4500	21.420000	7.32000	330.160	1
113	65	32.050000	97	5.730	1.370998	61.4800	22.540000	10.33000	314.050	1
114	72	25.590000	82	2.820	0.570392	24.9600	33.750000	3.27000	392.460	1
115	86	27.180000	138	19.910	6.777364	90.2800	14.110000	4.35000	90.090	1

116 rows × 10 columns

(visualize data frame)

Next, we used isna() to check if there is missing data in the data frame. According to our chart below, there are no rows containing missing data. Also we found the numbers of positive samples and negative samples are 64 and 52, both of them match the original dataset.

```
In [31]: #check if there is no missing data, this will tell us how many missing data we have
         df[df.isna().sum(axis=1) > 0]
```

Out[31]:

Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Classification

```
In [32]: #find number of negative(-) class
         num_of_negative = len(df[df.Classification == 0])
         num_of_negative
```

Out[32]: 52

```
In [33]: #find number of positve(+) class
         num_of_positive = len(df[df.Classification == 1])
         num_of_positive
```

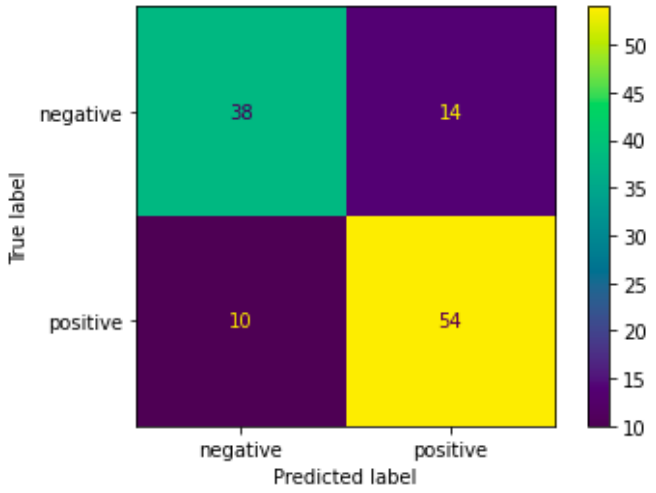
Out[33]: 64

(sanity check)

To train our RF model, we chose 500 and 1000 for NTREE, which is the number of trees in RF. For MTRY, we tried threes numbers which are 0.5*SQRT(number of features), SQRT(number of features), and 2*SQRT(number of features). Since the number of features is 9, the three values of MTRY would be 2, 3, 6. We use these NTREE and MTRY numbers for our grid search to find the best pair of NTREE and MTRY, which can generate the most accurate RF model. Also, because we use the RandomForestClassifier in SciKit, we keep the default value of CUTOFF which is (0.5,0.5) for majority vote. After we are done with grid search, the best pair of NTREE and MTRY we got is (ntree = 500, mtry = 6). We use this pair of values to train our final RF model.

Experiment Result

Here is the misclassification matrix of the best trained RF model. X axis is the predicted classification result during the training phase, and Y is the true classification label in the original dataset. The first green set is True Negative(TN), which means the best trained RF model classifies 38 negative samples correctly. The purple cell on the right of TN is False Positive (FP). The value 14 means the model classifies 14 negative samples into the positive class. The other purple cell is False Negative (FN), and the value 10 means the model classifies 10 positive samples into the negative class. The last yellow cell is True Positive (TP), and the model classifies 54 positive samples correctly. In the original dataset, the number of negative samples is 52, and the number of positive samples is 64. According to this misclassification matrix, the number of negative samples is 38+14=52, and the number of positive samples is 10+54=64. Both of them equal the total numbers of positive samples and negative samples in the original dataset.



(misclassification matrix for base model)

Then, we use "oob_score_" and "f1_score()" to find the F1 score and the OOB score of the model. From the tool, the output of F1 score is 0.81818, and the output of OOB score is 0.7931.

F1 Score

```
In [52]: #prediction is based on training set
         #get F1 score based on training phase
         print('F1 Score: ', f1_score(train_label, prediction))

         F1 Score:  0.8181818181818182
```

OOB Score

```
In [59]: #get OOB score based on training phase
         print('OOB Score: ', model.oob_score_)

         OOB Score:  0.7931034482758621
```

(Code of calculating F1 score and OOB score)

We check the result of F1 score and OOB score manually according to the misclassification matrix the tool gives to us. The F1 score and OOB score which are calculated based on misclassification matrix are the same as the scores from the tool's output.

1. **Recall** = TP / (TP+FN) = 54 / (54+10) = 0.84375
   Precision = TP / (TP+FP) = 54 / (54+14) = 0.7941176
   F1 score = 2*(recall * precision) / (recall + precision) = 2*(0.84375*0.7941176) / (0.84375+0.7941176) = 0.81818
2. **OOB score** = 1 – oob_error = 1 – (total number of error / total number of samples) = 1 - (24/116) = 0.7931

From the above part, we calculate F1 score manually using recall based on positive class. The F1 score we get is 0.81818, which is not high for a medical predictor. As a result, this classification accuracy shows that our model is not a very good RF classification model if it is applied to predicting breast cancer.

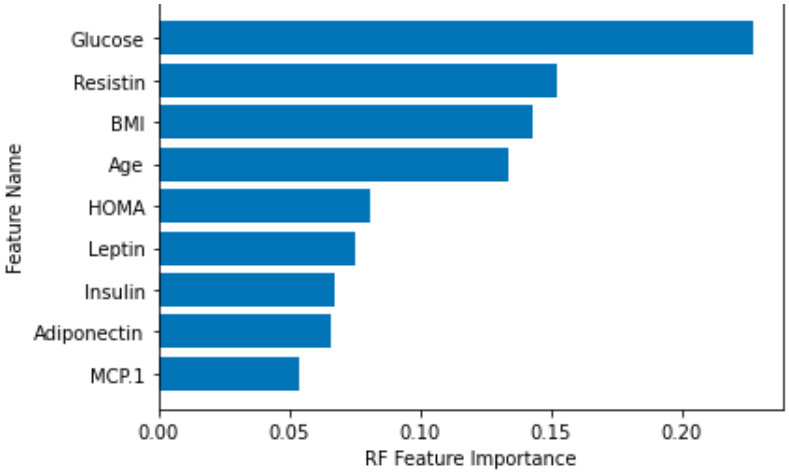
Explainability Report

In this section, we use the RFEX model explainer to analyze the prediction results of our model ethically.

MDA Feature Ranking

First, to build the RFEX Model Summary table, we used MDA to rank features. Since there are only 9 features in this dataset, we picked all of them. The right side is a list of top 9 ranked features. Each element of this list contains the name of the feature and its importance. The order of the list is from the most important feature to the least important feature among top 9 features. The right side is a bar chart that shows the result of top 10 ranked features. The X axis is the importance of features, and the Y axis is the top 9 ranked features' names. From top to bottom, these 10 features are ordered from the most important to the least important. We can see that the most important feature is 'Glucose', which takes 0.2276 importance, and the least important feature is 'MCP.1', which takes 0.0541 importance. The result in the bar chart matches the result in the list.

```
[('Glucose', 0.22763148138007147),
 ('Resistin', 0.15216146844477346),
 ('BMI', 0.14295925304554677),
 ('Age', 0.13365917453644227),
 ('HOMA', 0.0808891523147242),
 ('Leptin', 0.0750579365989072),
 ('Insulin', 0.06763327073708585),
 ('Adiponectin', 0.06588999639426381),
 ('MCP.1', 0.054118266548184996)]
```

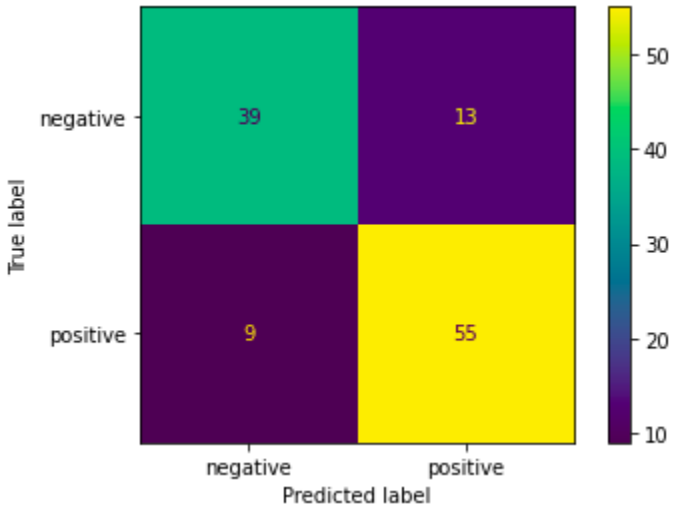


(list and chart of feature ranking based on MDA)

Cumulative F1 score

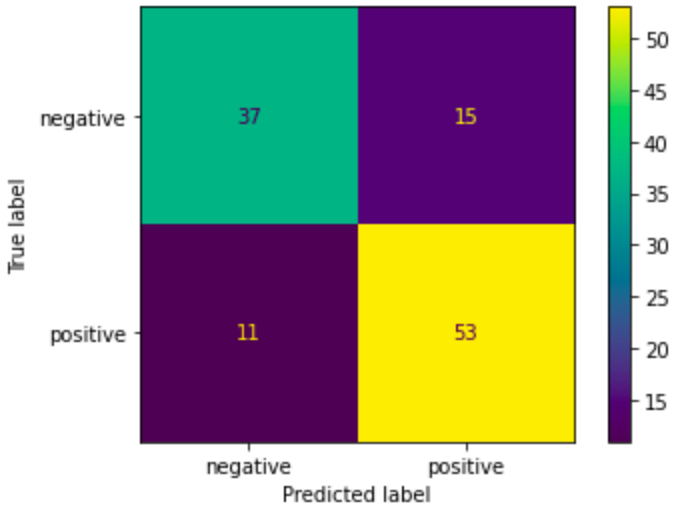
Next, we calculate cumulative F1 scores based on top 4 ranked features and top 7 ranked features.

- **Cumulative F1 Score for Top 4 Ranked Features:**
  We chose 1, 2, 3 for MTRY and 500 and 1000 for NTREE to do grid search. The best pair of MTRY and NTREE we found is (MTRY=2, NTREE=500). The cutoff is (0.5,0.5), which is the default value in ScKit Learn. The cumulative F1 score based on top 4 ranked features is 0.83333, which is a bit higher than the base F1 score (0.81818). Here is the corresponding misclassification matrix:



(misclassification matrix based on top 4 features)

- **Cumulative F1 Score for Top 7 Ranked Features:**
  We chose 2, 3, 4 for MTRY and 500 and 1000 for NTREE to do grid search. The best pair of MTRY and NTREE we found is (MTRY=4, NTREE=500). The cutoff is (0.5,0.5), which is the default value in ScKit Learn. The cumulative F1 score based on top 7 ranked features is 0.80303, which is a bit lower than the base F1 score (0.81818). Here is the corresponding misclassification matrix:



(misclassification matrix based on top 7 features)

RFEX Model Summary

In this section, we will provide an RFEX Model Summary table that is used to explain our model. Features in this table are ranked by their importance (their MDA value). The 5th and 6th columns contain basic stats of positive class and negative class for each feature. AV is the average value of feature values. SD is the standard deviation of feature values. MIN and MAX are the minimum and maximum values in each feature column. The last column in this table is the cohen distance between + and - class samples for each feature. This is used to measure feature separation for user confidence. Cohen Distance = (ABS(AV+ - AV-) / SDMax), where SDMax is the larger standard deviation of positive class's SD and negative class's SD, and ABS is absolute value.

(Base RF accuracy is F1=0.81818, for ntree=500, mtry=6, and cutoff (0.5,0.5))

Feature index	Feature name	MDA value	Cumulative F1 score	AV/SD + class [MIN, MAX]	AV/SD - class [MIN, MAX]	Cohen Distance
1	Glucose	22.8	N/A	88.2/10.2 [60,118]	105.6/26.6 [70,201]	0.65

2	Resistin	15.2	N/A	11.6/11.4 [3.3,82.1]	17.3/12.6 [3.2,55.2]	0.43
3	BMI	14.3	N/A	28.3/5.4 [18.7,38.6]	27/4.6 [18.4,37.1]	0.24
4	Age	13.4	0.83333	58.1/19 [24,89]	56.7/13.5 [34,86]	0.07
5	HOMA	8	N/A	1.6/1.2 [0.47,7.1]	3.6/4.6 [0.5,25.1]	0.46
6	Leptin	7.5	N/A	26.6/19.3 [4.3,83.5]	26.6/19.2 [6.3,90.3]	0
7	Insulin	6.8	0.80303	6.9/4.9 [2.7,26.2]	12.5/12.3 [6.3,58.5]	0.46
8	Adiponectin	6.6	N/A	10.3/7.6 [2.2,38]	10/6.2 [1.7,33.8]	0.04
9	MCP.1	5.4	N/A	499.7/292.2 [45.8,1256.1]	563/384 [90.1,1698.4]	0.16

The value of Cohen Distance less than 0.2 is small separation, between 0.2 and 0.5 is medium separation, and between 0.5 and 0.8 is large separation. Higher Cohen Distance value means the feature can show better separation between + and - classes. There are 5 features out of 9 features in our dataset denote medium separation, and the left 4 features denote small separation. According to the summary table above, for most features, highly ranked features like “Glucose” and “Resistin” have high separation, and the separation is low for those not important features like “MCP.1” and “Adiponectin”. This also means highly ranked features are more powerful than lowly ranked features when this model is classifying healthy controls and patients. However, there are two exceptions --- “HOMA'' and “Insulin”. For these two features, their Cohen distance is very high, but their feature importance values are low. This shows that “HOMA” and “Insulin” can classify healthy controls and patients well, but they did not play important and powerful roles during the classification process since their importance value is not high.

By looking at the cumulative F1 score based on the top 4 features, which are “Glucose”, “Resistin”, “BMI”, “Age”, the value 0.83333 is higher than the base F1 score which is 0.81818, but the difference is not big. This is not an ideal result. “Glucose”, “Resistin” and “BMI” are the top three ranked features with large Cohen distance values. They can separate positive and negative samples well. However, after adding the 4th feature “Age”, because its Cohen distance value is very small at the same time when its feature importance is large, the “Age” feature pulls down the classification ability of the top three features. As a result, the cumulative F1 score based on the top 4 features is only a bit higher than the base F1 score. The cumulative F1 score based on top 7 features is 0.80303, which is a bit lower than the base one. We think the reason that might cause this situation is that the F1 score of top 7 features misses classification contribution of the last two features (“Adiponectin” and “MCP.1”), as a result, the F1 score of top 7 features would be a bit lower than the base F1 score.

IV. AI Ethics Audit and Evaluation of Results

Method Overview

Our team (Team 1) used the Random Forest Classifier to train the data collected in the Patrício, et al. study. In our data-training process, we found that the overall accuracy rate of the model was only 81%. We also found that of the nine features Patrício et al. collected in their dataset, it was Glucose and Resistin that functioned as the two most significant and powerful features. The significance was determined by the MDA value, while the strength of the features' abilities to distinctly separate positive and negative classes was determined by the Cohen Distance. Our model yielded 84% Recall and 79% Precision, which means that our model is better able to classify patients (positive class) than it is able to classify the healthy control group (negative class.)

Glucose was the most important feature with the highest MDA value, followed by Resistin as the second most important feature with the second highest MDA value. Furthermore, in both features, the Cohen Distance was considerably higher than in the remaining seven features: Glucose classifies the positive (patient) class and the negative (control/healthy) class most accurately. Resistin classifies the positive class and the negative class second most accurately. The positive group (i.e., the patient group) had both lower glucose and lower Resistin levels than the negative (i.e., control/healthy) group. BMI was not as significant when we factored in MDA and it was not as powerful when we factored in the Cohen Distance, and neither was *Age*.

Application of Ethics Audit the ML Study: The Ethical Implications

To address the ethical considerations for the *Breast Cancer Coimbra Data Set* collected by Patrício et al., we relied upon the *Ethics Guidelines for Trustworthy AI: High-Level Expert Group on Artificial Intelligence* which was produced by the European Commission.[1] In their study Patrício, et al.[2] find that for a certain combination of patients' age, BMI, Glucose levels, and Resistin, serve as good

[1] High-Level Expert Group on AI (2019). *Ethics guidelines for trustworthy AI* (Report). European Commission.

[2] Patrício M, Pereira J, Crisóstomo J, Matafome P, Gomes M, Seiça R, Caramelo F. Using Resistin, glucose, age and BMI to predict the presence of breast cancer. BMC Cancer. 2018 Jan 4;18(1):29. doi: 10.1186/s12885-017-3877-1. PMID: 29301500; PMCID: PMC5755302.

indicators and predictors of the presence of breast cancer. In our data-training process we were able to affirm their assertions, and to narrow in on glucose and Resistin as the two most important features.

The *High-Level Expert Group on Artificial Intelligence* states that a trustworthy AI system has three components that should always be met throughout the system's duration. The AI system should be **lawful** therefore in compliance with all applicable laws and regulations, it should be **ethical** by adhering to ethical principles and value, and it should be **technically and socially robust** enough to withstand and mitigate unforeseeable and unintended harms.[3] Each of these components, by itself, is necessary but not sufficient for the realization of trustworthy AI, and when tension between the components arise, society should set out to align the lawfulness, ethicality, and robustness of an AI system.[4] For these conditions to be met, a system would need to meet certain criteria. We will consider the criteria below vis-à-vis the data and study:

1. **Human Agency and Oversight:** An AI system should constantly work towards ensuring fundamental human rights, human agency, and human oversight.[5]

We find that in their study, Patrício et al., are motivated by a realization of a fundamental human right, viz., the right to preemptive and mitigative care. What is great about the study is that a simple test, like Resistin, might serve to function as an indicator of the presence of cancer. Resistin, as the study indicates, is a simple €20 test that can be administered when conducting routine annual blood panels for patients.[6] Since age and BMI are easily empirically observable, and glucose levels get tested on any given random blood test, we found that the one major indicator would be Resistin.

As the study clearly indicates in the report, the goal of this is to make breast cancer detectable at an early stage, but not to replace mammography as a way of detection.[7] In their study, they also indicate

[3] European Commission, p. 2.
[4] Ibid.
[5] Id., 14.
[6] Patricio, p. 6.
[7] Ibid.

that this will make breast cancer screenings accessible to women who are not of the testing-age —i.e., if they are younger than 40 or older than 75. Another possible advantage to this study, which the scientists do not mention, is that it would make breast cancer screenings accessible to any human with mammary glands be they men or women (because men, as well, get breast cancer.) The study does not seem to over-inflate its results, and seems to present them in a realistic manner.

2. **Technical Robustness and Safety:** AI systems should work to minimize unintentional and unexpected harm, as well as preventing unacceptable harm, they should constantly implement changes to both technical processes and human processes that work against the upholding of the ethical moral code of trustworthiness. A system should be accurate and have a fallback plan to ensure safety. It should also be reliable and reproducible.

The *Breast Cancer Coimbra Data Set* collected by Patrício et al., seems to have been gathered through ethical means: The women recruited for the study were newly diagnosed and had never had breast cancer treatment up to that point and samples collected were naïve samples (i.e., collected before treatment).[8]

The goal of their study, as they state, is to measure *hyperresistinemia* and metabolic dysregulation in women with breast cancer, and this is done by gauging Resistin levels in the patients.[9] First, to ensure an even baseline, it was ascertained that all selected volunteers (the 52 healthy control group and the 64 patients) had no underlying conditions, comorbidities, or acute infectious diseases. Women with a BMI over $40 kg/m^2$, were excluded from the study.[10]

Furthermore, Patrício, et al., ensured that the conditions under which the samples were collected were identical for all 116 participants. The data collected was the participants' age, weight, height, and menopausal status. All participants were fasting for the blood drawing and all blood

[8] Id., 2.
[9] Ibid.
[10] Id., 3.

samples were collected at the same time.[11] To ensure that the system is reliable and reproducible, the Gini coefficient was used to measure the total decrease in node impurities associated with splitting the variable in RF algorithms averaged over all trees.[12] The previous section of this paper delves deeper into this.

However, some possible data-gathering ethical issues arose for us:

- **The Accuracy Rate** (F1 Score) is only 81%. While 81% may appear high, it is pretty low where human life is concerned.

- **The Sample Size** of 116 patients seems a bit too small as it barely covers a fraction of a fraction of what would stand to represent a class of a population.

- **The Diversity in the Sample Size is Never Mentioned**: it therefore does not indicate whether the participants were of the same region —e.g., are they Mediterranean women, Portuguese, or are they from one certain geographical section of Portugal? Has the study been applied to Northern European women, African women, Asian women, etc.?

- **Sample History and Background:** The study does not mention the deeper medical backgrounds of their patients (or participants on the whole). While they ensured that no comorbidities or acute infectious diseases were present across the board, the study never mentions whether or not the patients were BRCA1/BRCA2 positive or negative. Moreover, they do not provide any information on their Oncotype DX status. They never mention the medical histories of hormonally induced cancers in the families of the patients. More specifically, they never mention a family history of breast cancer.

- **Causation vs. Correlation:** It seems that the doctors conducting the study were able to consistently predict, with an 82-88% accuracy, cancer in menopausal or post-menopausal women. And in our data-training

[11] Ibid.
[12] Ibid.

process of the *Breast Cancer Coimbra Data Set* collected by Patrício et al., we were able to arrive at 81% accuracy.

This of course raises an important point: hormonal cancers tend to show up during hormonal changes like during and around menopause. Resistin is a hormone that was named as a portmanteau of *Resistance to Insulin* and it is hypothesized that it functions to indicate how fat cells trigger insulin resistance. [13]

The hormonal imbalances women experience which are associated with being close to or around menopause, seem to disrupt insulin resistance in bodies, which is precisely why most women experience some form of weight gain in varying degrees during, and around, menopause.

Interestingly enough, the mean age of the women in the study was 58.1 for the healthy control group and 56.7 for the patient group —i.e., women around their menopause years.[14] The women, for the most part, were perimenopausal, menopausal, or postmenopausal with 59% of the patient group, and 63% of the healthy control group were found to be postmenopausal.[15]

There would, therefore, seem to be a menopause-related reason for a disruption in glucose levels and BMI, and possibly Resistin (a.k.a., insulin resistance)—and of course, the age bit, is just part of it all.

The study gets a bit murky when we consider that the conclusions of this study might be neglecting to take into account that certain changes in 'biomarkers' might *themselves* be caused by the hormonal changes brought about by the onset, or nearing, of menopause. Therefore, it is worth considering, that these biomarkers *themselves* might actually be co-existing merely correlatively and contingently but not causally. This would, therefore, bring the absoluteness of Resistin's putative legitimacy into question —especially when taking into account the 81% accuracy rate.

[13] Berger, Abi. "Resistin: a new hormone that links obesity with type 2 diabetes." *BMJ : British Medical Journal* vol. 322,7280 (2001): 193.
[14] Patricio, p. 4.
[15] Ibid.

3. **Privacy and Data Governance:** is a crucial element in harm-prevention. AI system developers should work on producing adequate data governance that cover both the quality and integrity of the data collected and used.[16] This is achieved by making sure of the following:

a. **AI systems ought to guarantee privacy and data protection:** The Patrício, et al. study does not mention the privacy of the data collected and catalogued. They do however mention that the volunteers were told exactly what the study entailed and that written consent was given.

b. **A system ought to acquire data that is of high quality and of integrity:** while the Patrício, et al. study does collect clean data samples, and while they do seem to thoroughly test and document at every step of the process from planning to training, to testing, to deployment, we find other issues in their data acquisition viz., those that pertain to the sample size, background, and diversity, which we cover in the sections above. Therefore, while the training data might have a *prima facie* disposition to appear free of biases, inaccuracies, and errors, our concerns mentioned above raise questions about the integrity of the scope of the putative results adduced.

c. **Access to data** is another step to be met in the privacy and data governance tenet. It mainly has to do with strongly establishing from the beginning who can access users' data and under what circumstances and that only duly qualified and competent personnel should be allowed to access the data on a strictly need-to-access basis.[17] The Patrício, et al. study does not mention anything about that, but that would not seem to be a possible issue for them.

4. **Transparency:** is closely linked to the ethical *principle of explicability* and when it comes to ML systems this is achieved by ensuring

[16] Commission, p. 17.
[17] Ibid.

traceability, explainability, and communication:[18] Traceability of data is what ensures auditability and explainability, and as far as traceability goes, Patrício, et al. seem to be following a proper protocol of cataloguing and documenting their data-collection, as well as their chosen algorithms, which in turn helps us identify errors in the system chain and would therefore help us mitigate them.

As far as explainability goes, Patrício, et al. not only explain the technical processes of their ML system, but they also explain human decisions. Their technical explainability seems to be understood and traced by humans. We also imagine that the explainability Patrício, et al. are offering, is accessible by laypeople so that even those functioning at the legislative and organizational level would be able to know what this accuracy or explainability entails.

Patrício, et al. communicate their ML system's level of accuracy and disclose its limitations to both practitioners and end-users in a comprehensible manner.

5. **Diversity, Non-Discrimination, and Fairness:** A trustworthy AI must be inclusive and diverse throughout all its stages and this is achieved by ensuring that the AI system provides equal access to all stakeholders through inclusive design and equal treatment.[19] Patrício, et al. would need to diversify their sample, and they would need to clearly reiterate that this Resistin testing, as it stands, is not an adequate solution to early detection, but that it *might* be on its way to other discoveries.

6. **Societal and Environmental Wellbeing:** For an AI system to be fair and to prevent harm, the list of stakeholders of an AI system should also include the environment and other sentient beings.[20] AI systems should also understand that the legacy it leaves future generations is important and should consider future generations part of its stakeholder base. AI Trustworthiness, therefore, entails that AI system be:

[18] Id., 18.
[19] Id.,18.
[20] Id.,19.

- **Sustainable and Environmentally Friendly:** A system's development, deployment, and usage, should never come at the expense of the planet and the environment.[21] Nothing in the Patrício, et al. study suggests that it is in contention with the environment or other sentient beings.

- **Consider its Social Impact:** AI Systems should consider not only their positive effects on human life, but also its negative effects and these effects ought to be closely monitored and considered —e.g., the deterioration of the social fabric, the effect of the technology on mental and physical wellbeing, etc.[22]

  A possibly problematic angle of the Patrício, et al. study, that functions on the social and societal level, seems to be in its placing women's weight as something to be perpetually speculated upon. And it somehow seems to function, at least implicitly, to blame women for getting breast cancer especially at a time of severe uncomfortable hormonal disruptions which happen *to* women, and are not *caused by* women.

  When we place the blame on BMI, and glucose levels (especially with 81% accuracy on a sample of 116 participants), we seem to be telling women that it is because they are fat, they get cancer and that they deserve it. There is no indication whatsoever that fatter women are at higher risks of getting breast cancer, or any cancer —especially with missing genetic and genomic histories of patients. Resistin, as a hormone, is name that is a portmanteau of "resistance to insulin"[23] and functions to explain the link between obesity and diabetes.[24] However, when our team trained the data collected and catalogued by Patrício, et al., we found that Resistin and Glucose were consistently *lower* in the positive patient group than they were in the negative healthy control group. It is crucial to note that we are not asserting that these metrics are not important, but rather, that deploying these

[21] Ibid.
[22] Ibid.
[23] Berger, (2001).
[24] Ibid.

biomarkers for gauging breast, and other cancers, should be handled with a high level of care and accuracy.

- **Society and Democracy:** should also be considered when assessing the impact of a system during the development, deployment, and usage stages. For example, AI systems should never hinder the democratic process by disrupting elections.[25]

  This could be an issue for this study if the results get inflated and touted as absolutely conclusive. The results of this study would *prima facie* appear beneficial, because mammograms can be expensive, and because not all patients under oppressive medical coverage plans have access to them. However, the findings of the study have only been able to report an 80-82% accuracy, and this itself does not entail that this methodology is adequate enough for thorough testing, thus not suitable for meeting the requisite standard of care. And because medical systems seem to be motivated by profit, we worry that some medical insurance systems would use the findings as excuses for denying women over forty their routine annual mammograms. We therefore, would recommend not being overly optimistic when releasing their results, as money-hungry laypersons in power might take advantage of the 80-82% accuracy (81% in our data-training process) and use these results to get out of offering women the basic legislated levels of care.

7. **Accountability** which is achieved by:

   a. **Auditability:** entails the enablement, and the making knowledge available of the assessment of algorithms, data, and design processes whether by internal or external auditors,[26] and Patrício, et al. have been meeting that criterion every step of the way.

   b. **Minimization and Reporting of Negative impacts:** the ability to report on actions and decisions that yield a particular outcome, and

[25] Ibid.
[26] Ibid.

the ability to respond to such consequences should always be ensured.[27] For Patrício, et al. this seems to be *prima facie* met.

c. **Trade-offs:** if there arises a need to address one tenet at the expense of the others, the decision should be studied and measured carefully.[28] For Patrício, et al. this might include further testing to account for larger and more diverse sample sizes, or through gauging genetic and genomic backgrounds of patients and the healthy controls.

d. **Redress:** Whenever an unjust or adverse outcome arises, the system should be designed for redress to be possible because knowing that redress is possible ensures trust —especially when the perpetually marginalized are involved.[29] That is what we are hoping Patrício, et al. would achieve upon the expansion and diversification of their sample sizes.

[27] Id., 22.
[28] Ibid.
[29] Ibid.

V. Summary and Recommendation

We find that, as far as going forward with the study goes, we recommend increasing the sample size, increasing the group's diversity, and to at least make sure that the BRCA1/BRCA2 status of the participants is collected and documented. If Oncotype DX testing is feasible, that might help as well. Including BRCA1/BRCA2 classification in the data-training and data-testing phase, might help solidify Resistin and glucose as conclusive.

As far as our position on the results as they stand at the moment, we find that, if the objective of the study is to assert that testing for Resistin on routine checkups, in addition to already existing and established testing methods, could be conducted for early prediction and detection of breast cancer, then this certainly is a beneficial outcome that would function ethically by giving more women (and men, because they too get breast cancer) access to routine checkups. At €20, Resistin testing would be financially insignificant and non-invasive enough, to make it a desirable mode of prediction and detection. And it might also make regular testing accessible to those outside the permissible annual mammogram range.

However, because of the low accuracy rate, coupled with the other concerns raised above in this paper, we find that if the findings would be utilized to push for Resistin testing in place of mammograms, genetic testing, or MRIs (for high-risk groups) —and Patrício, et al. assure us that this is *not* the case—then this study would *clearly* be insufficient to function as an adequate replacement no matter how much money it could potentially save insurance companies. Human life is always more important than any financial gain, and expediency in diagnosis should never come at the expense of thoroughness and certainly never at the expense of human life and human health. We, therefore, find that any attempt to make it appear as a viable replacement to mammograms, ultrasounds, genetic-testing, or MRIs, would cause harm and be in direct violation of *The Harm Principle* and that would render it inherently unethical and immoral. Luckily for us, Patrício, et al., assure us that this is not the case.

VI. References

Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., & Caramelo, F. (2018). Using Resistin, glucose, age and BMI to predict the presence of breast cancer. BMC Cancer, 18(1). Retrieved May 12, 2021, from https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra

Petkovic, D., Alavi, A., Cai, D., Yang, J., & Barlaskar, S. RFEX: Simple Random Forest Model and Sample Explainer for non-Machine Learning experts. Retrieved May 12, 2021, from https://cs.sfsu.edu/sites/default/files/technical-reports/Petkovic%20RFEX%20SFSU%20TR_0.pdf

Petkovic, D. (April 17, 2021). Random Forest Model and Sample Explainer for Non-Experts in Machine Learning — Two Case Studies. Retrieved May 12, 2021, from https://ilearn.sfsu.edu/ay2021/pluginfile.php/1210033/mod_resource/content/3/RFEX%20Random%20Forest%20Explainability%20and%20two%20case%20studies%202021.pdf

High-Level Expert Group on AI (2019). *Ethics guidelines for trustworthy AI* (Report). European Commission.

Berger, Abi. "Resistin: a new hormone that links obesity with type 2 diabetes." *BMJ : British Medical Journal* vol. 322,7280 (2001): 193.

VII. Appendix

1. Base Case

Use Grid Search to Find Best Mtry and Ntree

```
In [42]: #get the number of features
         num_feature = len(train_feature.columns)

In [43]: #find three values of Mtry
         mtry_1 = 0.5 * math.sqrt(num_feature)
         mtry_2 = math.sqrt(num_feature)
         mtry_3 = 2 * math.sqrt(num_feature)

In [44]: #take the value of mtry_1 as 2 to let the input of max_feature be int
         mtry_1
Out[44]: 1.5

In [45]: #take the value of mtry_2 as 3 to let the input of max_feature be int
         mtry_2
Out[45]: 3.0

In [46]: #take the value of mtry_3 as 6 to let the input of max_feature be int
         mtry_3
Out[46]: 6.0

In [47]: #parameters list for tuning
         mtry = [2,3,6]
         ntree = [500,1000]
```

```
In [48]: #do grid search
         for i in range (2):
             for j in range (3):
                 nTree = ntree[i]
                 mTry = mtry[j]
                 #create a RF model
                 model = RandomForestClassifier(n_estimators=nTree, max_features=mTry, oob_score= True, random_state = 0)
                 #train the RF model
                 model.fit(train_feature, train_label)
                 #use decision funciton estimated on training set to find prediction based on training set
                 prediction = np.argmax(model.oob_decision_function_, axis = 1)
                 #print out ntree value and mtry value with their corresponding f1 score
                 print("nTree = ", nTree, "     mTry = " , mTry)
                 print("F1: ", f1_score(train_label, prediction))
```

(grid search that is used to find the best pair of NTREE and MTRY)

Use the Best Mtry and Ntree to Create the Best RF Model

```
In [49]: #use the best split to create best RF model
         model = RandomForestClassifier(random_state = 0, max_features = 6, n_estimators = 500,oob_score = True)
         #apply the training data to the model
         model.fit(train_feature, train_label)
```

(RF Model training algorithm)

2. Top 4 features' cumulative F1 score

Cumulative F1 score

top 4 feature

```
In [65]: train_feature_top_4 = train_feature[['Glucose','Resistin','BMI','Age']]

In [66]: #get the number of features
         num_feature = len(train_feature_top_4.columns)

In [67]: #find three values of Mtry
         mtry_1 = 0.5 * math.sqrt(num_feature)
         mtry_2 = math.sqrt(num_feature)
         mtry_3 = 2 * math.sqrt(num_feature)

In [68]: #parameters list for tuning
         mtry_4 = [1,2,3]
         ntree_4 = [500,1000]

In [69]: #do grid search
         for i in range (2):
             for j in range (3):
                 nTree = ntree_4[i]
                 mTry = mtry_4[j]
                 #create a RF model
                 model = RandomForestClassifier(n_estimators=nTree, max_features=mTry, oob_score= True, random_state = 0)
                 #train the RF model
                 model.fit(train_feature_top_4, train_label)
                 #use decision funciton estimated on training set to find prediction based on training set
                 prediction = np.argmax(model.oob_decision_function_, axis = 1)
                 #print out ntree value and mtry value with their corresponding f1 score
                 print("nTree = ", nTree, "     mTry = " , mTry)
                 print("F1: ", f1_score(train_label, prediction))
```

(Retrain the model with top 4 ranked features)

```
In [70]: #use the best split to create best RF model
         model_4 = RandomForestClassifier(random_state = 0, max_features = 2, n_estimators = 500,oob_score = True)
         #apply the training data to the model
         model_4.fit(train_feature_top_4, train_label)
```

(RF Model training algorithm based on top 4 ranked features)

3. Top 7 features' cumulative F1 score

top 7 feature

In [74]:
```
train_feature_top_7 = train_feature[['Glucose','Resistin','BMI','Age','HOMA','Leptin','Insulin']]
```

In [75]:
```
#get the number of features
num_feature = len(train_feature_top_7.columns)
```

In [76]:
```
#find three values of Mtry
mtry_1 = 0.5 * math.sqrt(num_feature)
mtry_2 = math.sqrt(num_feature)
mtry_3 = 2 * math.sqrt(num_feature)
```

In [77]:
```
#parameters list for tuning
mtry_7 = [2,3,4]
ntree_7 = [500,1000]
```

In [78]:
```
#do grid search
for i in range (2):
    for j in range (3):
        nTree = ntree_7[i]
        mTry = mtry_7[j]
        #create a RF model
        model = RandomForestClassifier(n_estimators=nTree, max_features=mTry, oob_score= True, random_state = 0)
        #train the RF model
        model.fit(train_feature_top_7, train_label)
        #use decision funciton estimated on training set to find prediction based on training set
        prediction = np.argmax(model.oob_decision_function_, axis = 1)
        #print out ntree value and mtry value with their corresponding f1 score
        print("nTree = ", nTree, "     mTry = " , mTry)
        print("F1: ", f1_score(train_label, prediction))
```

(Retrain the model with top 7 ranked features)

In [44]:
```
#use the best split to create best RF model
model_7 = RandomForestClassifier(random_state = 0, max_features = 4, n_estimators = 500,oob_score = True)
#apply the training data to the model
model_7.fit(train_feature_top_7, train_label)
```

(RF Model training algorithm based on top 7 ranked features)