# Breast Cancer Detection: Defining a "Black Box" Model using Explainable AI (XAI)

Taylor Bostick and Katherine Graham

***Abstract*—The purpose of this report is to detail some of the background work, motivation, implications, and exploratory analyses, as well as updates on current project progress, related to our final project on cancer detection and explainable AI. The project can be outlined succinctly into the following two areas:**

**(1) To first develop a complex and difficult-to-interpret, or "black-box," model to detect the presence of breast cancer in computed features from digitized breast mass fine needle aspirate (FNA) images.**

**(2) To define the quality of this final, "black box" model by incorporating explainable AI methods.**

**The Wisconsin Diagnostic Breast Cancer (WDBC) dataset will be used for the purposes of this project. From the initial exploratory analyis, it was discovered that the dataset was high dimensional with few observations (less than 1000). Transformations were applied to handle feature skewness and data normalization; the synthetic minority oversampling technique, or SMOTE, was used to treat target class imbalance and augment the total number of cases in the dataset. Feature selection was performed using recursive feature elimination, or RFE.**

**The preferred "black box" model that this dataset will be fed through will be a single-layer artificial neural network, since these models often result in incredibly high accuracy but lack significantly in interpretibility. Finally, explainable AI methods Shapley-Flow and Break-Down will be used to define and examine the quality of our artificial network.**

## I. Introduction

THE purpose of this project is to create a data science solution to a real-world problem. Our project is going to focus on using explainable AI (XAI) to define a "black box" model for breast cancer detection. A "black box" model is defined as a model that "can get highly accurate predictions from... after we train them with large datasets, but we have little hope of understanding the internal features and representations of the data that a model uses to classify a particular image into a category" [1]. We will first develop a deep learning model to classify input as "having cancer present" or "not having cancer present". Then, we will use XAI to evaluate and define the model.

XAI is still an emerging field, but it has the ability to improve interpretability and assess and handle systematic bias in models. Making sure that we come up with a solution that is accurate and easily interpreted is an important part of the project. Our team wants to use data science techniques to effectively detect breast cancer.

### A. Who Will This Benefit?

Breast cancer is the second most common breast cancer among women after skin cancer and also the second leading cause of cancer death behind lung cancer [2]. Accurate detection means that treatment can begin sooner, which means it will have a higher success rate. As stated previously, we are focusing on using data science methods to detect breast cancer, which will benefit women by getting them into treatment, and hopefully, remission faster.

Additionally, this is an issue that women all over the world encounter. Our project could have the ability to benefit both women in the United States as well as across the world. An accurate model could change the lives of women everywhere.

## II. Background Work

This is a topic that has been studied by a lot of different groups in the data science community. These approaches all focus heavily on machine learning to detect breast cancer. One article noted using three different machine learning algorithms, Random Forest, kNN (k-Nearest-Neighbor), and Naive Bayes, to accurately detect breast cancer in patients [3].

Another group used computer aided detection (CAD) to make diagnoses. CAD uses machine learning to effectively detect cancer earlier on. Early detection allows patients to get into treatment and hopefully remission faster. This paper used Support Vector Machine (SVM), Artificial Neural Networks, and Naive Bayes to get the most accurate approach. Linear discriminant analysis (LDA) was used to reduce the high dimensionality of features. They used the Wisconsin Diagnostic Breast Cancer Dataset and were able to achieve an accuracy of 98.82% [4].

While the topic has been studied deeply, our group plans to use other methods such as XAI to define these "black box" models.

## III. Novelty

As established previously, there has been plenty of work completed on breast cancer detection in the data science community. Our team is going to use Explainable AI to go one step beyond other people's work.

Machine learning models have often been defined as "black box". This means that the model produces useful information, but does not include any insights into how it

works. Explainable AI can be used to define these types of models so that we can understand how and why they are successful.

There are multiple different XAI methods that our group could use for this project. The two that we are focusing on are Shapley-Flow and Break-Down. These methods will be detailed further in subsequent sections.

By including XAI in our project, we are able to establish novelty from other people's work on this topic and provide an explanation of "black box" models.

## IV. Dataset

We will be using a dataset that we found on the website Kaggle. The dataset is titled Breast Cancer Wisconsin (Diagnostic) Data Set. The features in this dataset are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. This is a type of biopsy that collects samples of cells from a lump or mass in the breast. The features describe characteristics of the cell nuclei present in the image from the FNA.

There are 30 different features in the dataset in addition to an id that uniquely identifies the FNA image and the diagnosis, either malignant(M) or benign(B). The 30 features boil down to 10 features taken from the image. Those 10 features are:

- Radius
- Texture
- Perimeter
- Area
- Smoothness
- Compactness
- Concavity
- Concave Points
- Symmetry
- Fractal Dimension

For each of those 10 features, the mean, standard error, and largest(mean of the three largest values) was taken. This provides us with the 30 features in the dataset. The dataset consists of 569 rows, of which 357 are benign and 212 are malignant.

This differs from the initial dataset that we had planned on using because instead of it being histology images that we would have to extract characteristics from, this dataset has already completed that step. Additionally, this dataset uses FNA, which is a technique used in cytology, a branch of pathology. The dataset that we had been looking into previously used histology, another branch of pathology. This switch in datasets allowed us to progress forward with the project without switching too much of our solution pipeline.

## V. Wisconsin Diagnostic Breast Cancer (WDBC) Exploratory Data Analysis

### A. Notable Findings

The exploratory analysis provides a comprehensive first glimpse into the distinct characteristics of the WDBC feature set. Some notable findings of the analysis include:

- Small sample size, high dimensionality
- Numerous features are highly correlated
- Seven features have skewness value greater than 2
- Target class, *Diagnosis*, suffers from moderate class imbalance
  - 64.57% of cases are *Benign*
  - 35.43% of cases are *Malignant*
- Identified 21 significant features

The following subsections include details on each of these findings.

### B. Small n, Large p

The first step of our exploratory analysis involved getting a total count of records and fields present within the dataset. The original dataset was found to contain 569 rows and 31 columns.

Though this is an alarmingly small dataset in terms of observations, dimensionality is also high. This combination typically would envoke concerns over the "curse of dimensionality", where error increases directly with the number of features. However, shallow neural networks have been observed to handle high dimensionality datasets fairly well [5]. Hence, we express more concern over the small sample size.

Treatment consists of applying the Synthetic Minority Oversampling Technique, or SMOTE, to the training set, which also works to treat the class imbalance issue observed in the target class. Subsequent sections further detail this treatment.

### C. Data Transformations

Evaluating the skewness of each field in the WDBC dataset reveals that the majority of features have a skewness score greater than 1, with seven of those having a score greater than 2. Though there is some grayness over how greatly the accuracy of neural networks are affected by skewed data fields, it is known that they do handle skewed data significantly better than other classification methods [6]. Still, we opted to treat fields by means of a square root transformation.

Viewing **Figures 4** and **5** in the Appendix, we can see that this transformation has both successfully and significantly decreased the severity of the skewness across all features.

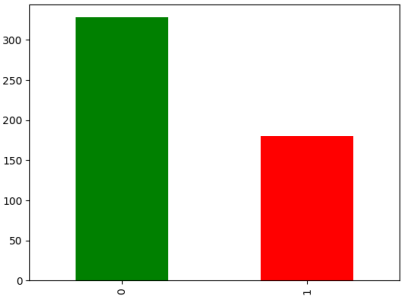
Directly after applying this square root transformation, we applied Min-Max Normalization to the dataset. Furthermore,

we address and remove outliers by means of calculated z-scores. After applying these transformations, the newly transformed dataset contains a total of 508 observations.

### D. Data Augmentation

A third notable finding of the WDBC dataset is the moderate class imbalance observed in the target class, *Diagnosis*. Looking to **Figure 1** below, it is clear that the number of benign cases is nearly double that of the number of malignant cases. More precisely, approximately 64.57% of all records are observed to be benign, and 35.43% of records are observed to be malignant.

Fig. 1. Count of observations by target class.



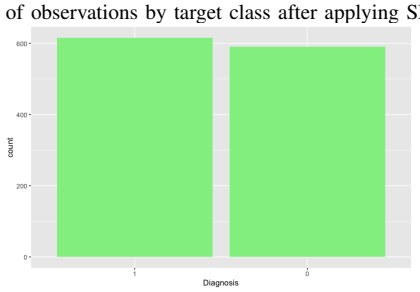

This imbalance is not incredibly severe; however, treating this imbalance would ensure overall better classification accuracy. Hence, we opted to treat this imbalance by applying SMOTE.

After separating the dataset into train and test sets by means of a 70:30 split, we applied the function call,

**SMOTE(diagnosis ∼ ., train, perc.over = 400, perc.under = 120)**

to the training set. The parameter **perc.over** specifies the number of extra cases generated for the minority class; in this case, we generated 400 additional cases. The parameter **perc.under** specifies the numbers of extra cases generated for the majority class; here, we specified 120. In **Figure 2** below, it can be observed that the number of malignant cases is approximately equal to the number of benign cases.

Fig. 2. Count of observations by target class after applying SMOTE.

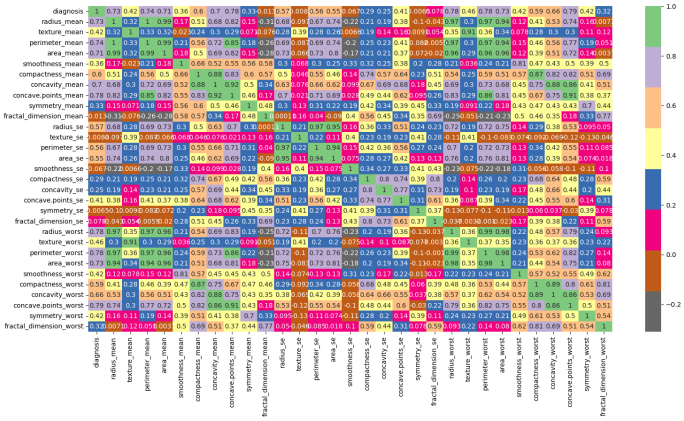



After making this function call and applying SMOTE to our training set, we addressed not only target class imbalance but also effectively augmented the WDBC dataset. After SMOTE, the final training set contains a total of 1205 observations.

### E. Feature Selection

The final notable finding concluding our exploratory analysis of the WDBC dataset involves the number of feature fields found to be directly correlated to each other. Looking to the heatmap in **Figure 3**, we observe numerous features with correlation scores greater than 0.6. It can also be noted that nine of the thirty potential predictors have correlation scores greater than 0.7 relative to *Diagnosis*.

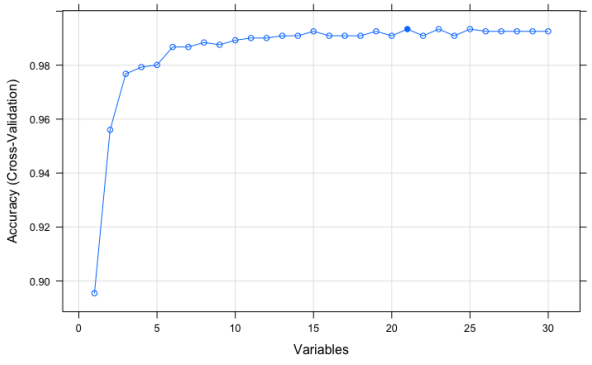
Fig. 3. Heat map of original WDBC dataset.




Following these observations, it was decided to further explore the most significant features by means of feature selection using recursive feature elimination, or RFE. Though artificial neural networks have the capability to determine which features are most significant in the model, applying feature selection methods prior to pushing data points through a network can save on computational time by reducing the number of redundant features, or noise, to be evaluated [8].

In short, RFE is a popular feature selection method that applies a backward selection process to find the combination of features most relevant to predicting the target variable. It first builds a model with all features present, calculates the importance of each feature relative to the target, then ranks features by importance and recursively removes those with least importance.

Fig. 4. RFE plotted results identify twenty-one relevent predictors.




We opted to apply the random forest algorithm configured to evaluate all possible subsets on the SMOTE-balanced

training set for each iteration. 10-fold cross validation was used to evaluate each random forest model. Looking to **Figure 4**, it can be observed that only twenty-one of the thirty predictors were found to be relevant to predicting *Diagnosis*.

Referring to **Figure 8** of the Appendix, all 30 features and their importance scores can be observed. It can also be observed from the output that the top five most significant features include:

- texture_worst
- area_worst
- texture_mean
- perimeter_worst
- radius_worst

## VI. Next Steps

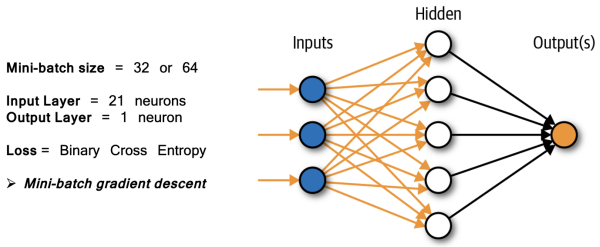
With preprocessing, data augmentation, and feature selection steps sufficiently completed, next steps involve model fitting and evaluation.

### A. Model Architecture

The WDBC dataset has been evaluated across numerous deep learning models. While some have developed more complex models with multiple hidden and drop out layers [9], others have constructed simpler network structures, relying on only a single hidden layer [10]. Overall model performance varies little despite variations in model architecture; across the model architectures we explored, overall accuracy varied around to 95%. For the purposes of our project, we are more interested in properly evaluating and defining the inner workings of a ”black box” model; hence, we opted for a simplified model architecture.

Pictured below in **Figure 5** are the current stipulations of our chosen model architecture.

Fig. 5. Structure of ANN.




Our model architecture consists of an input layer, a single hidden layer, and an output layer. Both input and hidden layers rely on the ReLu activation function; the output layer utilizes the sigmoid activation function. Furthermore, the input layer is set to take twenty-one neurons, and the output layer is due to take just one, the latter being standard for binary classification problems. Rather than iterating through the entire dataset or a single observation at a time, we will utilize mini-batch gradient descent to push batches of n = 32 or n = 64 through the network at a time.

### B. Overview of XAI Methods

Some brief investigating into the WDBC dataset reveals that certain XAI methods have been used previously to evaluate dataset performance on ”black box” models [12]. However, these methods are limited to more prominently used methods, such as SHAP, LIME, Morris sensitivity analysis, and partial dependence plots [12].

We have decided to constrict methods to Break-Down and Shapley-Flow.

*1) Overview: Break-Down:* The Break-Down method delivers local explanations and calculates each feature’s contribution in the prediction of a target variable as changes occur in prediction values ”along with a growing set of variables described by a specific order” [11]. In other words, diagnostics start with the mean expected model response, and continue to add more features as the model is conditioned. The ordering of features added to the model is significant; different orderings of the fields can lead to different contributions [11].

Whereas SHAP values average over all subsets of variable orderings, ultimately neglecting interactions between variables, the Break-Down method analyzes the feature orderings and visualizes interactions in the model [11]. Final attributions are then determined using a ”single ordering, which is chosen based on greedy heuristics” /citexai.

*2) Overview: Shapley-Flow:* Shapley-Flow is a graph-based, more recent method of SHAP and Asymmetric Shapley Values, or ASV. It allows for the use of variable dependency structure in the explanation process, where relationships are described using a causal graph [11]. Attributions are assigned to the relationships between features, or edges of the causal graph. These edge attributions have boundaries for each generated explanation that hold the standard Shapley values.

Furthermore, this method is advantageous in the way that large amounts of information about the structure of variable relationships and explanation boundaries are retained [11]. However, a directed causal graph limited to a small number of variables (for readability) must be constructed prior to applying Shapley-Flow. The method also requires a reference observation, where choices of this observation may result in varied explanations.

# APPENDIX A

Fig. 6. Skewness scores for original WDBC fields.

```
area_se                  5.447186
concavity_se             5.110463
fractal_dimension_se     3.923969
perimeter_se             3.443615
radius_se                3.088612
smoothness_se            2.314450
symmetry_se              2.195133
compactness_se           1.902221
area_worst               1.859373
fractal_dimension_worst  1.662579
texture_se               1.646444
area_mean                1.645732
compactness_worst        1.473555
concave.points_se        1.444678
symmetry_worst           1.433928
concavity_mean           1.401180
fractal_dimension_mean   1.304489
compactness_mean         1.190123
concave.points_mean      1.171180
concavity_worst          1.150237
perimeter_worst          1.128164
radius_worst             1.103115
perimeter_mean           0.990650
radius_mean              0.942380
symmetry_mean            0.725609
texture_mean             0.650450
diagnosis                0.528461
texture_worst            0.498321
concave.points_worst     0.492616
smoothness_mean          0.456324
smoothness_worst         0.415426
dtype: float64
```

Fig. 7. Skewness scores for transformed WDBC fields.

```
area_se                  2.141981
fractal_dimension_se     1.765806
perimeter_se             1.635855
radius_se                1.477656
symmetry_se              1.343412
smoothness_se            1.206641
fractal_dimension_worst  1.180458
area_worst               1.107839
fractal_dimension_mean   1.068065
concavity_se             0.937892
area_mean                0.933839
compactness_se           0.903304
symmetry_worst           0.890223
perimeter_worst          0.786099
radius_worst             0.783175
texture_se               0.741415
perimeter_mean           0.653536
radius_mean              0.622870
compactness_worst        0.604870
compactness_mean         0.564793
diagnosis                0.528461
symmetry_mean            0.441233
concavity_mean           0.360016
texture_mean             0.309895
concave.points_mean      0.243789
smoothness_mean          0.190934
texture_worst            0.180188
smoothness_worst         0.135253
concavity_worst          0.027867
concave.points_se       -0.421049
concave.points_worst    -0.443414
dtype: float64
```

Fig. 8. RFE results.

```
Recursive feature selection

Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

 Variables Accuracy  Kappa AccuracySD KappaSD Selected
         1   0.8955 0.7914   0.032235 0.06412
         2   0.9560 0.9122   0.034998 0.06977
         3   0.9768 0.9536   0.011547 0.02311
         4   0.9793 0.9585   0.013648 0.02733
         5   0.9801 0.9602   0.005764 0.01156
         6   0.9867 0.9735   0.010460 0.02093
         7   0.9867 0.9735   0.008887 0.01779
         8   0.9884 0.9768   0.009703 0.01943
         9   0.9876 0.9751   0.009750 0.01952
        10   0.9892 0.9785   0.010351 0.02072
        11   0.9901 0.9801   0.010168 0.02036
        12   0.9901 0.9801   0.010168 0.02036
        13   0.9909 0.9818   0.009905 0.01983
        14   0.9909 0.9818   0.009905 0.01983
        15   0.9925 0.9851   0.008236 0.01649
        16   0.9909 0.9818   0.009905 0.01983
        17   0.9909 0.9818   0.009905 0.01983
        18   0.9909 0.9818   0.009905 0.01983
        19   0.9925 0.9851   0.008236 0.01649
        20   0.9909 0.9818   0.009905 0.01983
        21   0.9934 0.9867   0.008555 0.01713        *
        22   0.9909 0.9818   0.009905 0.01983
        23   0.9934 0.9867   0.008555 0.01713
        24   0.9909 0.9817   0.008246 0.01651
        25   0.9934 0.9867   0.008555 0.01713
        26   0.9925 0.9851   0.008236 0.01649
        27   0.9925 0.9851   0.008236 0.01649
        28   0.9925 0.9851   0.008236 0.01649
        29   0.9925 0.9851   0.008236 0.01649
        30   0.9925 0.9851   0.008236 0.01649

The top 5 variables (out of 21):
   texture_worst, area_worst, texture_mean, perimeter_worst, radius_worst
```

# REFERENCES


[1] Sarkar, Tirthajyoti. "Google's new 'Explainable AI' (xAI) service." Medium, Towards Data Science, 25 Nov. 2019, https://towardsdatascience.com/googles-new-explainable-ai-xai-service-83a7bc823773

[2] CDC. "Breast Cancer Statistics." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 6 June 2022, https://www.cdc.gov/cancer/breast/statistics/index.htm.

[3] S. Sharma, A. Aggarwal, and T. Choudhury, "Breast cancer detection using machine learning algorithms," in 2018 International conference on computational techniques, electronics and mechanical systems (CTEMS), 2018, pp. 114–118.

[4] D. A. Omondiagbe, S. Veeramani, and A. S. Sidhu, "Machine learning classification techniques for breast cancer diagnosis," in IOP Conference Series: Materials Science and Engineering, 2019, vol. 495, no. 1, p. 012033.

[5] Wojtowytsch, S., amp; E, W. (2020). (tech.). Can Shallow Neural Networks Beat the Curse of Dimensionality? A mean field training perspective. Princeton University. Retrieved October 23, 2022, from https://arxiv.org/pdf/2005.10815.pdf.

[6] https://iopscience.iop.org/article/10.1088/1757-899X/523/1/012070/pdf A Larasati et al 2019 IOP Conf. Ser.: Mater. Sci. Eng. 523 012070

[7] SubbaNarasimha, P.N. Arinze, B. Anandarajan, Murugan. (2000). Predictive accuracy of artificial neural networks and multiple regression in the case of skewed data: Exploration of some issues. Expert Systems with Applications. 19. 117-123. 10.1016/S0957-4174(00)00026-9.

[8] Wu, Z. (n.d.). (rep.). Feature Selection in Convolutional Neural Network with MNIST Handwritten Digits. Australian National University. Retrieved October 23, 2022, from http://users.cecs.anu.edu.au/ Tom.Gedeon/conf/ABCs2018/paper/ABCs2018_paper_156.p

[9] Zhang, Y. (2019, December 21). Deep Learning in Wisconsin Breast Cancer Diagnosis. Medium. Retrieved October 23, 2022, from https://towardsdatascience.com/deep-learning-in-winonsin-breast-cancer-diagnosis-6bab13838abd

[10] Elsafty, M. K. (2022, October 19). Breast cancer prediction using ANN. Kaggle. Retrieved October 23, 2022, from https://www.kaggle.com/code/mohamedkhaledelsafty/breast-cancer-prediction-using-ann

[11] Holzinger, A., Saranti, A., Molnar, C., Biecek, P., Samek, W. (2022). Explainable AI Methods - A Brief Overview. In: Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, KR., Samek, W. (eds) xxAI - Beyond Explainable AI. xxAI 2020. Lecture Notes in Computer Science(), vol 13200. Springer, Cham. https://doi.org/10.1007/978-3-031-04083-2_2

[12] Gupta, V. (2022, April 5). Explainable AI — Wisconsin Breast Cancer Dataset. Explainable AI — Wisconsin Breast Cancer Dataset — Kaggle. Retrieved October 23, 2022, from https://www.kaggle.com/code/vibhu10616/explainable-ai-wisconsin-breast-cancer-dataset

[13] Wang, J., Wiens, J., amp; Lundberg, S. (2020). (rep.). Shapley Flow: A Graph-based Approach to Interpreting Model Predictions. University of Michigan. Retrieved October 23, 2022, from http://proceedings.mlr.press/v130/wang21b/wang21b.pdf.