Explaining Highly Uninterpretable Breast Cancer Detection Models with Explainable AI

Taylor Bostick and Katherine Graham

***Abstract*—Breast cancer is the one of the most common cancers to affect middle-aged and late-aged women. Early detection of cancer has been proven to have a positive effect on remission outcomes. To further improve these outcomes, numerous machine learning techniques have since been adopted for cancer detection. However, some of the best performing models include less interpretable or black box models, which can be difficult for many clinicians and businesses to trust. Though these models are highly accurate, they notoriously fail to provide insight into the reasons behind their predictions. The fast-growing field of Explainable AI effectively threatens to unveil the decisions made by these models. In this paper, we extend the work of numerous authors before us and develop six neural network architectures to predict cancer for numerical data retrieved from digitized fine needle aspirate (FNA) images. We test the hypothesis that a smaller neural network with forced feature extraction performs better than baseline and multi-layered networks. We incorporate several explainable methods to provide insight into each model's decision.**

I. Introduction

Breast cancer is the second most common cancer among women after skin cancer and the second leading cause of cancer death behind lung cancer [1]. Typically occurring in middle-aged and older women, it makes up about 1 in 3, or 30%, of all new cancers affecting American women each year [2]. According to estimates provided by the American Cancer Society, about 287,850 new cases of invasive breast cancer will be diagnosed in women this year alone [2]. Additionally, it is estimated that 43,250 women in the U.S will die from some form of breast cancer by the end of 2022.

Accurate detection goes hand-in-hand with early detection. By increasing the accuracy of detection models, treatment can begin sooner. This may lead to earlier remission and positively impact remission outcomes. However, current methods of detection, including lab and imaging tests, biopsies, bone scans, magnetic resonance imaging (MRI), and x-rays, require an educated human eye to interpret results. This is often very time-intensive and has the effect of delaying treatment for patients. Acknowledging the need for faster, more efficient means of detecting cancer, data scientists and clinicians in the last decade have begun to incorporate machine learning into the detection process.

A. Novelty

Improving cancer detection has quickly become a prominent and well-studied area in the field of machine learning and the broader data science community. Some of the best performing models developed for detection include less interpretable or black box models. In this work, we define a black box as a high-performing model that makes decisions ambiguously. Rephrased, this would refer to any model that produces useful information but reveals little to no insight into how it delivers predictions. To add novelty to this report, explainable methods were used to provide more clarity in the decisions made by all tested models.

Expainable AI (XAI) is an emerging but fast-progressing field. Techniques developed in this field focus on improving interpretability and assessing systematic bias in less interpretable machine learning models. Though their performances may be unmatched, businesses often carry a natural distrust for these models since there is no way to explain their decisions.

In this work, six neural network architectures ranging in complexity were developed to predict cancer diagnoses. These include a baseline model, a smaller model with forced feature extraction, and four larger models with five, seven, ten, and fifteen hidden layers, respectively. All models were trained using the Wisconsin Diagnostic Breast Cancer dataset and explained at both the model-level and prediction-level using the following explainable methods:

- Receiver Operating Characteristic (ROC) Curve
- Permutational Variable Importance
- Accumulated Local Effects (ALE)
- Shapley Values
- Local Interpretable Model-agnostic Explanations (LIME)

We hypothesized that the best performing model would be the smaller model with forced feature extraction. Furthermore, we believed that all applied explainable methods would support this hypothesis and provide insight into model decisions.

II. Related Works

Breast cancer detection is a well-studied topic in the data science community, with many approaches relying heavily on machine learning techniques. Numerous papers have been published utilizing the Wisconsin Diagnostic Breast Cancer dataset in particular. These papers often approach detection as a binary classification problem and seek to quantify competitiveness between several or more models. One such paper published by authors S. Sharma et al. evaluates the performance between random forest (RF), k-nearest neighbor (kNN), and naive bayes models. Though it was found that all algorithms performed extremely well, each having an accuracy greater than 94%, the kNN model was found to be the most effective [5].

In another paper, authors D.A. Omondiagbe et al. exploited computer aided detection (CAD) to predict cancer diagnoses. CAD is a technology that utilizes machine learning to decrease observational oversights of physicians charged with translating medical images [8]. CAD has been proven to be effective particularly in early cancer detection. In the published report, performance was comparatively analyzed for support vector machine (SVM), artificial neural network (ANN), and naive bayes models. Each model's performance was further examined after applying several dimensionality reduction methods: (i) correlation-based feature selection (CFS), (ii) recursive feature elimination (RFS), and (iii) linear discriminant analysis (LDA). It was reported that each model utilizing LDA had the best overall performance with an accuracy of 99.94% [6].

A final notable work includes an evaluation performed by authors Yash Amethiya et al. Their research presents a comparative analysis of early breast cancer detection machine learning methods and biosensors [7]. Among the algorithms explored were the extreme learning machine – radial basis function (ELM-RBF), support vector regression (SVR), back-propagation neural network (BPNN), and ANN. Results were compared using accuracy and the confusion matrix performance metrics, precision and recall. Ultimately, the study found that biosensors and machine learning both demonstrated great potential in the field of early breast cancer detection [7].

III. Wisconsin Diagnostic Breast Cancer Dataset

In this report, we utilized the Wisconsin Diagnostic Breast Cancer dataset found on *Kaggle* [9]. 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* column that uniquely identifies the FNA image and 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.

IV. Exploratory Analysis

Our initial exploratory analysis of the Wisconsin dataset alerted us to a number of distinct dataset characteristics prior to model fitting. These findings can be outlined as follows:

- Several or more features critically skewed
- High correlation between features
- Imbalanced target class

The following subsections further detail each of these findings.

A. Data Transformations

We began our analysis of the Wisconsin dataset by first examining the distribution of its features. Upon testing for skewness, the majority of features were found to have skewness scores greater than one. More than several of these identified features were found to have critically high skewness scores greater than two. To treat skewness in the dataset, we applied square root transformations to each predictor.

Table III and **Table IV** in the **Appendix** show the before and after effects of this transformation, respectively. Post-transformation, nearly all features show skewness scores below 2. The most critically skewed feature, **area_se**, that had an original skewness score of 5.447 has a skewness score of 2.142 post-transformation.

After treating predictor skewness, min-max normalization was applied to guarantee that all features would have the same scale.

B. Feature Relationships

After treating predictor skewness and scaling the dataset, the relationship between features was examined through a heatmap. **Figure 1** illustrates these findings. It appears that many of the variables have high correlation coefficients, r. More precisely put, nearly a third of the features have $r > 0.7$.

C. Data Augmentation

A final noteworthy discovery from the Wisconsin dataset was the class imbalance observed in the target class, *Diagnosis*. Referring to **Figure 2**, it appears that the number of benign cases (64.57%) is nearly double that of the number of malignant cases (35.43%).

Failing to treat this imbalance prior to fitting a model could lead to untrustworthy results. For that reason, we chose to treat this issue by applying the synthetic minority oversampling technique (SMOTE) to our training set.

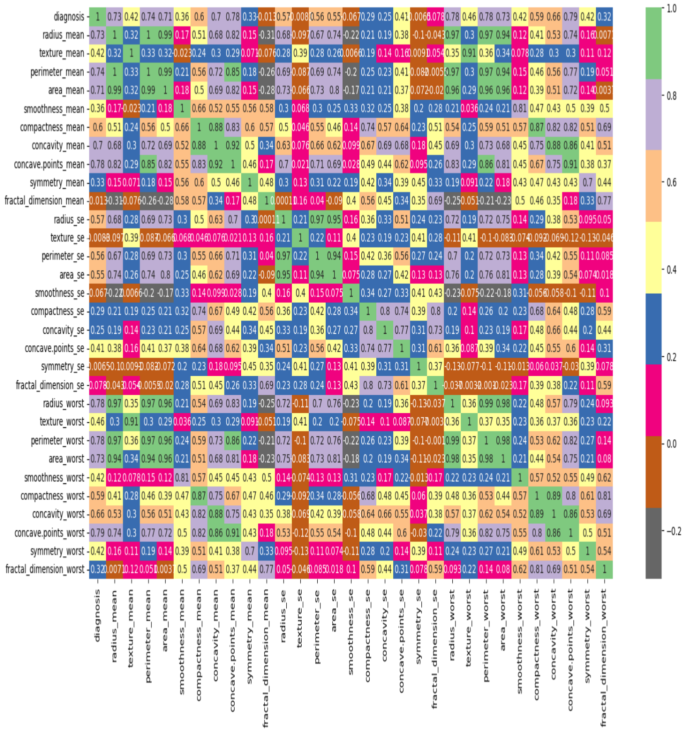


Fig. 1. Heat map of the Wisconsin Diagnostic Breast Cancer dataset.

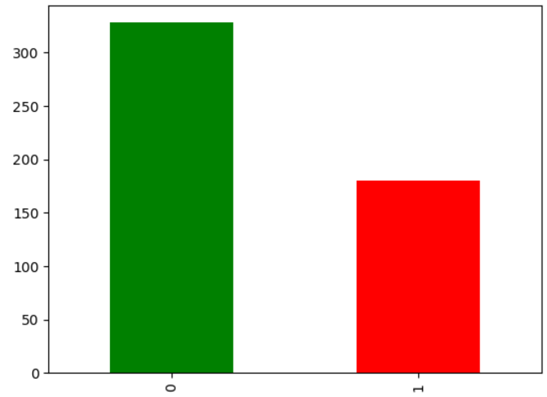


Fig. 2. Count of benign (0) and malignant (1) cases in the original Wisconsin dataset.

The Wisconsin dataset was split into training and test sets using a standard 70:30 split. SMOTE was then applied to the training set with the parameters **perc.over** = 400, **perc.under** = 120. These parameters generated 400 additional cases of the minority class (1) and 120 additional cases of the majority class (0).

Figure 3 shows the class distribution post-transformation. It can be observed that the number of malignant cases is now approximately equal to the number of benign cases. The final training set contains a total of 1,205 observations.

V. Model Architectures and Evaluations

We developed six neural network architectures for model training and evaluations. The six architectures consisted of *Baseline* (single hidden layer), *Small* (forced feature extraction), and four larger models. These larger models include *BigNet5*, *BigNet7*, *BigNet10*, and *BigNet15*, which consist of five, seven, ten, and fifteen hidden layers, respectively.

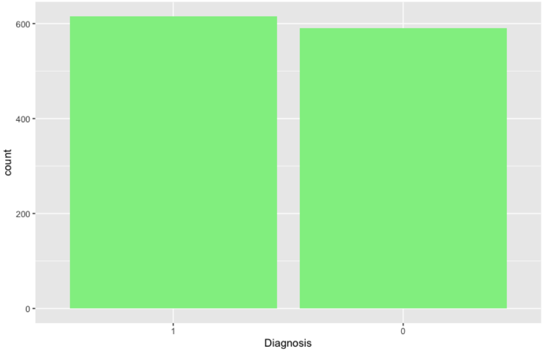


Fig. 3. Count of benign (0) and malignant (1) cases in the training set after applying SMOTE.

All models used rectified linear unit (ReLU) activation for the input and hidden layers, while sigmoid activation, the standard for binary classifiers, was specified in each output layer. To prevent overfitting and better treat highly correlated features, weight constraints were set within each hidden layer, and dropout layers were placed after every hidden layer. For evaluations, Adam was used for optimization, and binary cross entropy was used to quantify loss.

A. Baseline Model

For our baseline model, we built a simple network with a single hidden and dropout layer. A **MaxNorm** constraint of three was specified in the hidden layer, and a dropout rate of 0.2 was specified in the dropout layer. The constraint was set to prevent overfitting, and works by constraining the values of the weights in a network. The model took in all thirty predictors as input. It trained for 100 epochs in batches of 32.

Table I shows that *Baseline* performs well, with a testing accuracy of around 99% and a loss of 0.04811.

B. Small Model

The idea behind the creation of a smaller model was that, by condensing the feature representation space in the hidden layer, we could force a type of feature extraction on the network. This model was defined with the same number of hidden and dropout layers, constraints, and dropout rate as *Baseline*, but took fifteen input neurons rather than thirty. Training was also set for 100 epochs and completed in batches of 32.

Table I shows that the performance of *Small* is on par with *Baseline*, though loss has slightly increased.

C. Larger Models

Our motivation for developing larger models was to observe how interpretations and performances changed when model complexity increased. The only details changed for these larger models were the number of hidden and dropout layers. Similarly to the idea motivating the smaller model, the first hidden layer in each of the larger models was set to take in all thirty input neurons while remaining hidden layers were set to take only fifteen.

Table I outlines the performance of these larger models. While training metrics for the models have some fluctuations, testing metrics show a continued downhill performance across all models except for *BigNet10*, which has as accuracy of 97%. The worst performing model in this set is *BigNet15*, which has a testing accuracy of 94%.

D. Model Evaluations Summary

Below, **Table I** depicts a concise view of the performance between models.

Model	Train Loss	Train Acc	Test Loss	Test Acc
Baseline	0.0466	0.9842	0.04811	0.9883
Small	0.0623	0.9849	0.0519	0.9883
BigNet5	0.0223	0.9932	0.1386	0.9532
BigNet7	0.0615	0.9815	0.103	0.9474
BigNet10	0.0393	0.9884	0.09286	0.9708
BigNet15	0.0786	0.9733	0.1062	0.9415

TABLE I
Model Evaluations

The top three performing models intitally identified are *Baseline*, *Small*, and *BigNet10*. *Baseline* and *Small* tie for accuracy at nearly 99%, though the former has a slightly lower loss.

VI. Explainable Interpretations

Previous approaches using the Wisconsin Breast Cancer Diagnostic dataset to incorporate explainable methods are few in number and brief in their investigations. These works also have a tendency to draw explanations for single-layer neural networks and random forest models. To further set our own research apart, we performed a full-scale, multi-level exploration of all six architectures discussed and evaluated in the previous section. Explanations were drawn at both the model-level and the prediction-level.

A. Model-level Explanations

Our exploration begins at the global level, where we seek to determine (i) overall fit, (ii) variable importance, and (ii) variable effects on the average prediction for each of the six neural network architectures.

Receiver Operating Characteristic (ROC) curves were used to illustrate the diagnostic ability of each classifier and evaluate overall fits. Curves were further supplemented with confusion matrix statistics. Permutational variable importance plots were used to determine variable significance within each model. Finally, accumulated local effects (ALE) plots were implemented to visually demonstrate the influences of model features on the average prediction.

1) ROC Curve Interpretations: **Figures 4-9** in the **Appendix** depict the ROC curves for each model. Though it is difficult to see the difference in performance between models, it can be seen that each model appears to be high-performing. Performances are significantly better than that of a random classifier.

Below, **Table II** quantifies the results illustrated in each curve.

Model	Sensitivity (TPR)	Specificity (TNR)	Precision	F1-Score
Baseline	0.984	0.9907	0.984	0.984
Small	0.984	0.9907	0.984	0.984
BigNet5	0.9259	1.0	0.8873	0.9403
BigNet7	0.9683	0.9352	0.8971	0.9313
BigNet10	0.9841	0.9630	0.9394	0.9612
BigNet15	1.0	0.9074	0.863	0.9265

TABLE II
Confusion Matrix Statistics

Table results confirm that each model is indeed high-performing. The worst performing model, *BigNet15*, has a balanced accuracy score of 92.65%. The best performing models are *Baseline* and *Small*. Both predict positive diagnoses and negative diagnoses correctly 98.4% and 99.07% of the time, respectively. They have a balanced accuracy score of 98.4%. Overall, table results closely mirror those from initial model evaluations.

2) Permutational Variable Importance Plot Interpretations: Permutation-based variable importance is one model-agnostic method used to provide insights about the significance of the features within a model. Similar to the leave-one-covariate-out (LOCO) approach, permutation-based variable importance seeks to assess a feature's importance by comparining the initial model with the model on which the effect of that feature is removed [14]. However, the latter removes the effect of a feature through a random reshuffling of the data rather than retraining without the feature.

Figures 10-15 in the **Appendix** depict the permutation variable importance plots for each model. Model dropout loss values for the original dataset are shown on the x-axis of each plot; important variables are depicted on the y-axes. The length of each bar corresponds to the drop-out loss post dataset permutations. The most significant observation of note is that all models mark predictors 6, 7, 13, and 10 as highly important. These numbers correspond respectively to **concavity_mean**, **concave.points_mean**, **area_se**, and **radius_se**.

3) ALE Plot Interpretations: Similar to partial dependence (PD) plots, ALE plots are used to evaluate how individual features influence model predictions on average. Though the latter can be more difficult to interpret than the former, interpretations of strongly correlated features can be trusted. PD plots for features that are strongly correlated with other features can lead to averaging predictions for instances of artificial data that are very unlikely to occur [15]. This induces a type of bias into plots, which renders interpretations unusable. ALE plots effectively block the effects of correlated features by calculating the differences rather than the averages between predictions.

Figures 16-19 in the **Appendix** depict ALE plots for the features identified in each model's permutational variable importance plot: **concavity_mean**, **concave.points_mean**, **area_se**, and **radius_se**. ALEs of diagnosis are situated on the y-axis and represent units of the prediction variable; feature values are depicted on the x-axis.

Referencing **Figure 16**, it can be seen that for diagnoses with an average concave.points_mean $\geq$ 0.5, each model predicts an up-lift toward a positive diagnosis with respect to the average prediction. Additionally, for diagnoses with an average concave.points_mean < 0.5, the feature effect on the prediction becomes negative. Expressly, the probability of having a positive diagnosis falls.

Figures 17-19 have similar interpretations. For **concavity_mean** values ≥ 0.4 and < 0.4, **radius_se** values ≥ 0.19 and < 0.19, and **area_se** values ≥ 0.18 and < 0.18.

Furthermore, it can be noted that the ALE for *BigNet7* is consistently the flattest curve across each plot. We conclude that these features have less effect on the average prediction in the 7-layer model. It is also noted that **concave.points_mean** and **concavity_mean** in *Baseline*, *Small*, and *BigNet5* have larger effects on average predictions than in the other models. Similarly, it appears that the features **radius_se** and **area_se** have larger effects on average predictions in *Baseline* and *Small* than in the other models.

B. Prediction-level Explanations

Our exploration concludes at the prediction-level, where we Shapley values and Local Interpretable Model-agnostic Explanations (LIME) plots are used to determine which variables contribute to the selected prediction.

1) Shapley Value Interpretations: The Shapley value is another model-agnostic method that delivers explanations at the local level. Based on cooperative game theory, it works by computing the average marginal contribution of a feature's value over all possible orderings [15]. Since computation time increases exponentially with the number of predictors, Shapley value calculations are generally slower for larger models.

Figures 20-25 in the **Appendix** portray the Shapley value results of each model. Red bars are indicative of negative contributions toward diagnosis, and green bars are indicative of the positive contributions. Expressly, features with red bars negatively impact the probability of a positive diagnosis, while features with green bars positively impact positive diagnosis probability. Prediction or contribution values are depicted on the x-axis; the y-axis portrays significant variables and their value for the observation.

The only features identified across all models as having significant negative impact on diagnosis are **concavity_mean**, **concave.points_mean**, **radius_se**, and **texture_worst**. The top negative contributor in *BigNet5*, *BigNet7*, and *BigNet10* is **texture_worst**; in *Baseline* and *Small*, the top negative contributors are **concave.points_mean** and **concavity_mean**; the top negative contributor in *BigNet15* is **area_worst**. One final point to note is that, in all models except for *BigNet10*, **compactness_se** has a significant positive impact on diagnosis.

2) LIME Interpretations: Local surrogate models are often used to explain single predictions of highly uninterpretable models. In the original paper, LIME was proposed as a concrete implementation of local surrogate models, which are trained to approximate the predictions of underlying black box models [15].

The algorithm for training a surrogate model can be outlined as follows:

1) Select an instance of interest to generate an explanation of its black box model prediction
2) Perturb the dataset to retrieve black box model predictions for new data points
3) Weight the new samples by their closeness to the instance of interest
4) Train the weighted surrogate on the perturbed dataset

[15]

Figures 26-31 in the **Appendix** portray the LIME plots for each of the six models. Interpretations are somewhat similar to those drawn using Shapley values. Predictors with blue bars represent negative contributions toward a positive diagnosis, while predictors with orange bars represent the positive contributions. One important feature to note is that all models appear to identify **concavity_mean** (predictor 6) as having the largest negative influence on a positive diagnosis. Other variables identified across each model as having a notable negative impact include **texture_worst** (predictor 21), **area_se** (predictor 13), and **radius_se** (predictor 10).

Additionally, all models collectively identify **compactness_se** (predictor 15) as having a significant positive influence. All models except for *BigNet10* also identify **texture_se** (predictor 11) as having a significant positive impact on a positive diagnosis.

VII. Conclusion

Our project has successfully implemented explainable AI into the topic of breast cancer detection. Through this, we have a deeper understanding of factors that have a large impact on the detection of breast cancer. Though we initially hypothesized that *Small* would be the best performing model, our results indicate that *Baseline* has the best overall performance and effectively minimizes loss. However, we believe that the smaller model remains competitive and that the explanations provided by the implemented explainable methods support this.

Concave.points_mean, **concavity_mean**, **radius_se**, and **area_se** were identified across all models as highly impactful on predictions. *Baseline* and *Small* further identified **concave.points_mean** and **concavity_mean** as having a noticeable negative impact on positive diagnosis outcomes.

A. Future Work and Societal Impact

In the future we are hopeful that this type of model can be adapted to detecting other types of cancer. Early detection is an important part of any cancer diagnosis and has the ability to save lives in the long run. By adapting our work to other cancer types, we could help a larger population of people live a longer, healthier life.

B. Work Breakdown

We worked for about 2-3 hours per week over the course of the 15-week semester. We attempted to split the writing and computational parts of the project up evenly, but Taylor did slightly more of the computing while Katherine took on more of the writing a research role. Katherine did the majority of the background research and exploratory data analysis, while Taylor did most of the XAI computations. We believe that the report was split up rather evenly and we worked on the different parts collaboratively. The Ignite and Final presentations were divided evenly as well, and we believe that the team did a good job working together for the entirety of the project.

APPENDIX A

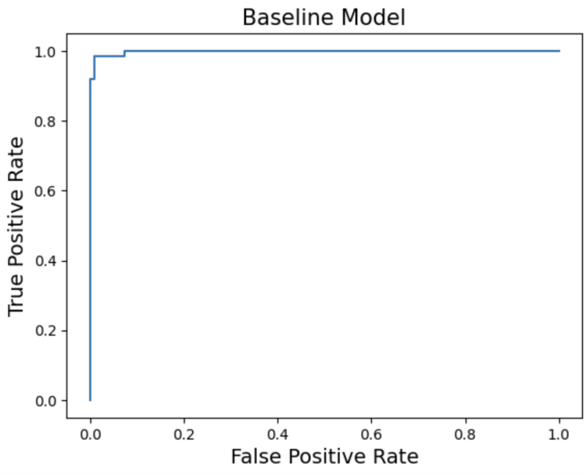


Fig. 4. ROC curve for baseline model.

Wisconsin dataset fields	Skewness scores
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

TABLE III
SKEWNESS SCORES FOR ORIGINAL WISCONSIN DATASET FIELDS.

WDBC fields	Skewness score
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

TABLE IV
SKEWNESS SCORES FOR TRANSFORMED WISCONSIN FIELDS.

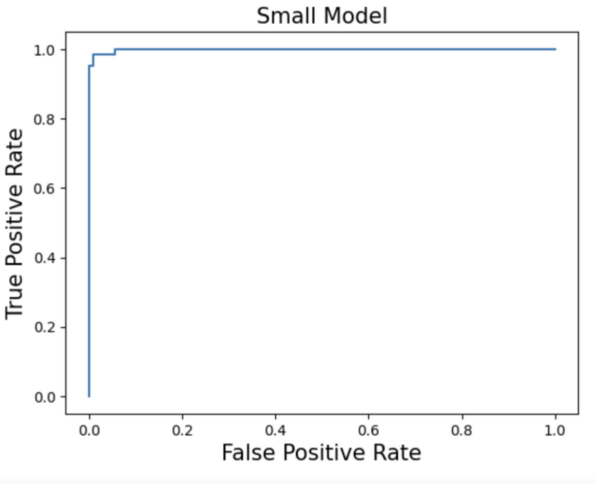


Fig. 5. ROC curve for smaller model.

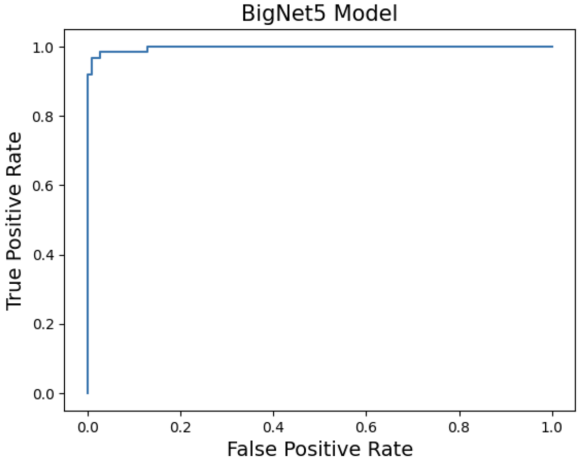


Fig. 6. ROC curve for 5-layer model.

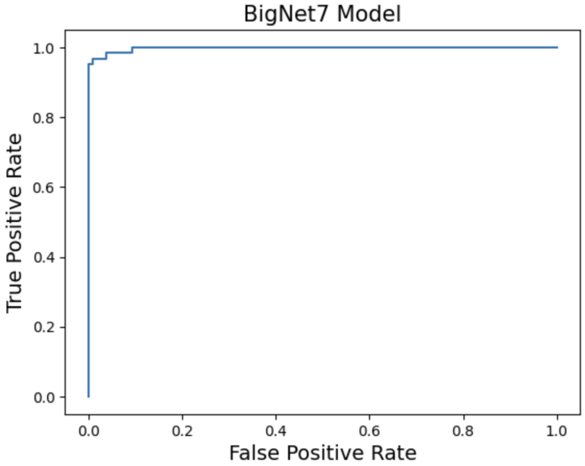


Fig. 7. ROC curve for 7-layer model.

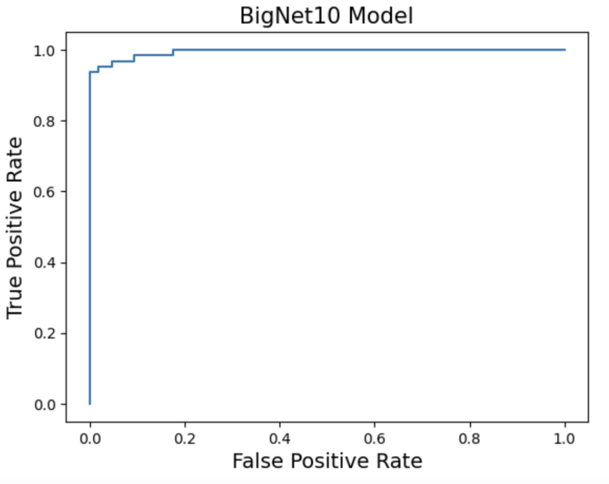


Fig. 8. ROC curve for 10-layer model.

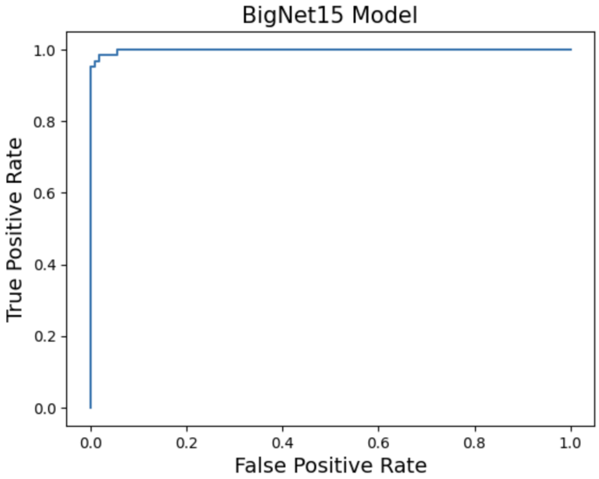


Fig. 9. ROC curve for 15-layer model.

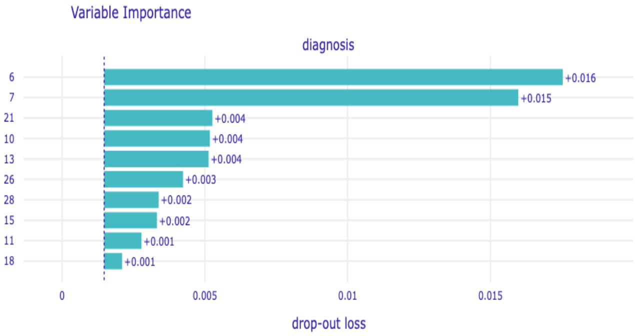


Fig. 10. Permutational variable importance plot for baseline model.

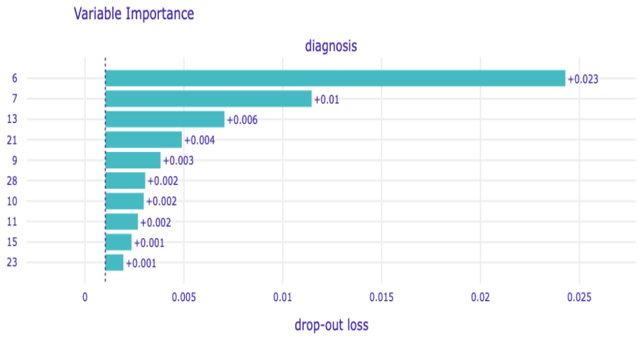


Fig. 11. Permutational variable importance plot for smaller model.

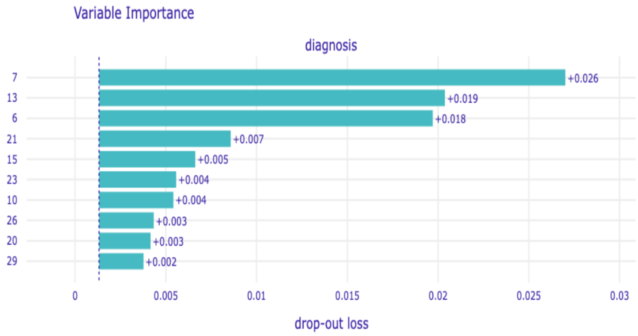


Fig. 12. Permutational variable importance plot for 5-layer model.

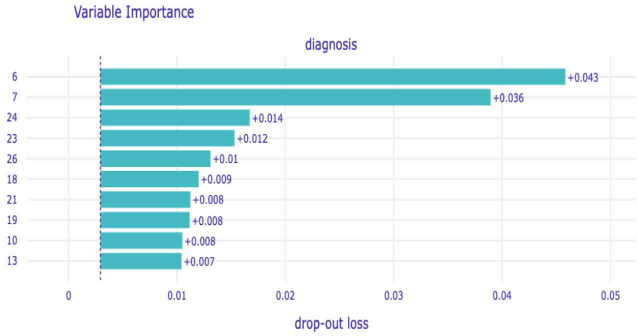


Fig. 13. Permutational variable importance plot for 7-layer model.

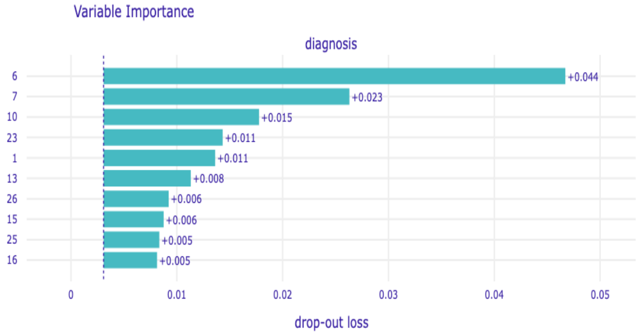


Fig. 14. Permutational variable importance plot for 10-layer model.

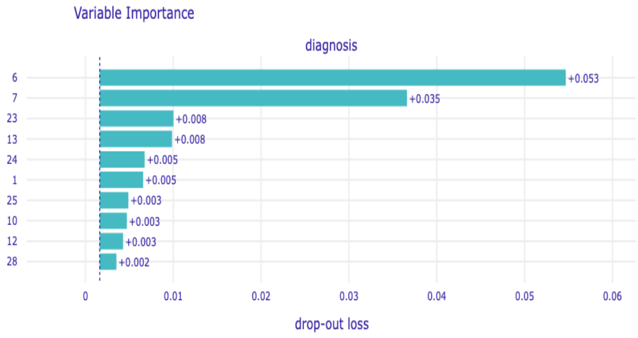


Fig. 15. Permutational variable importance plot for 15-layer model.

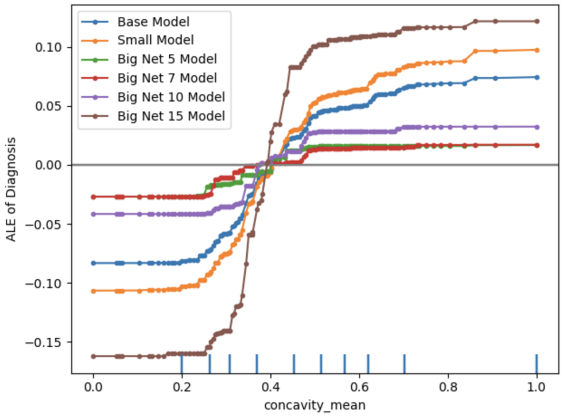


Fig. 16. ALE plot of **concavity_mean**.

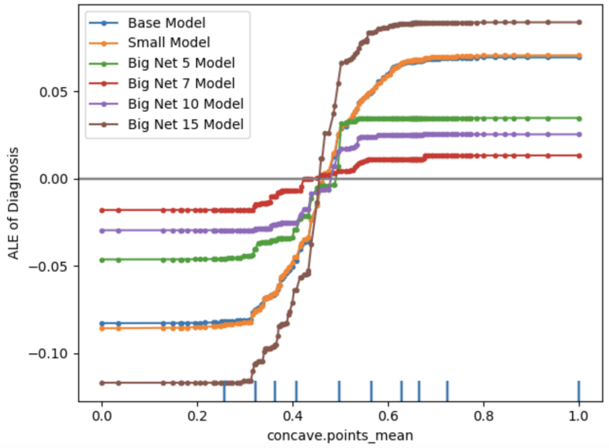


Fig. 17. ALE plot of **concave.points_mean**.

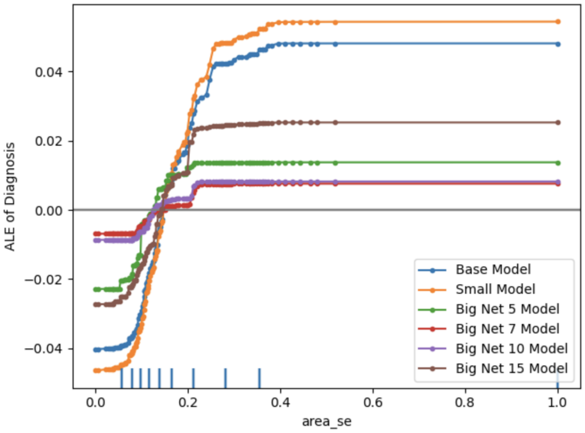


Fig. 18. ALE plot of **area_se**.

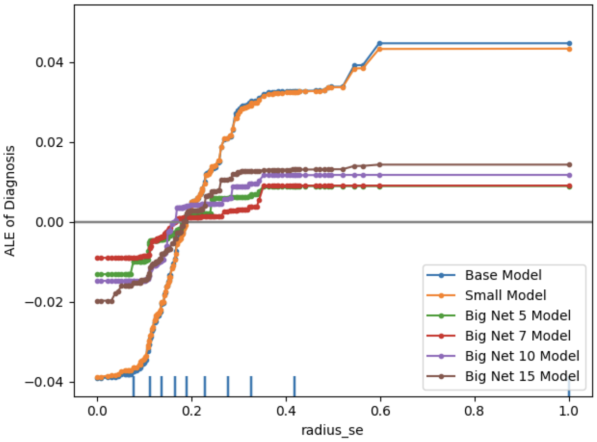


Fig. 19. ALE plot of **radius_se**.

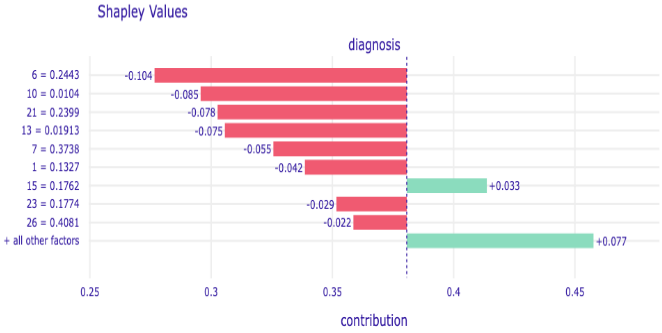


Fig. 20. Shapley values plot for baseline model.

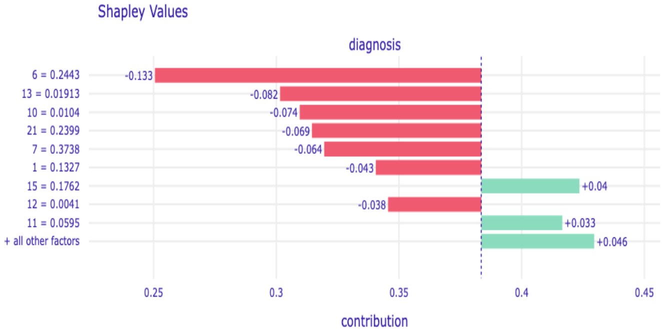


Fig. 21. Shapley values plot for smaller model.

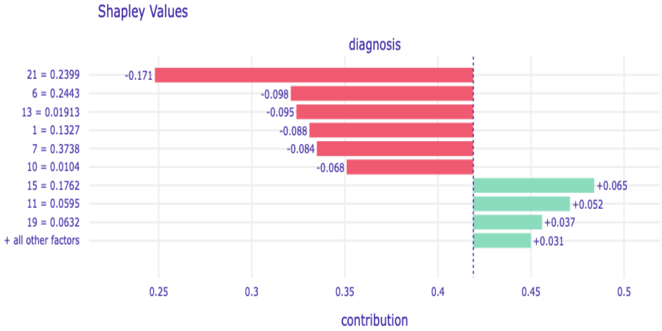


Fig. 22. Shapley values plot for 5-layer model.

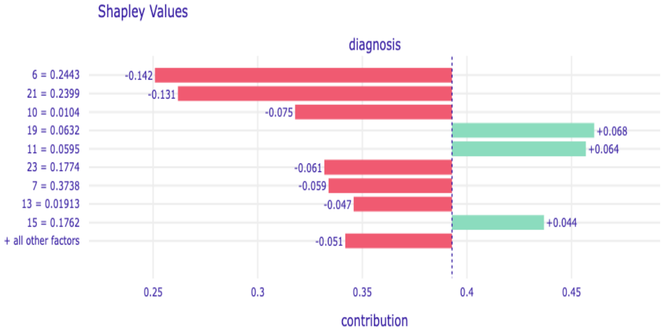


Fig. 23. Shapley values plot for 7-layer model.

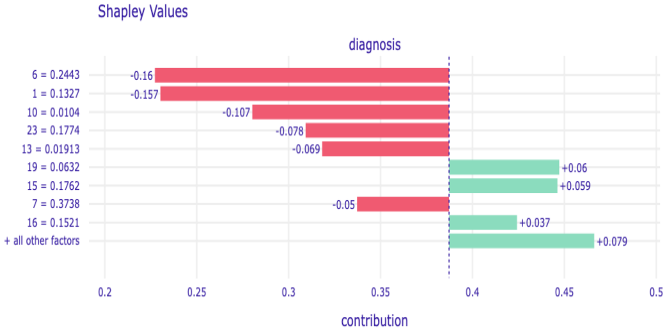


Fig. 24. Shapley values plot for 10-layer model.

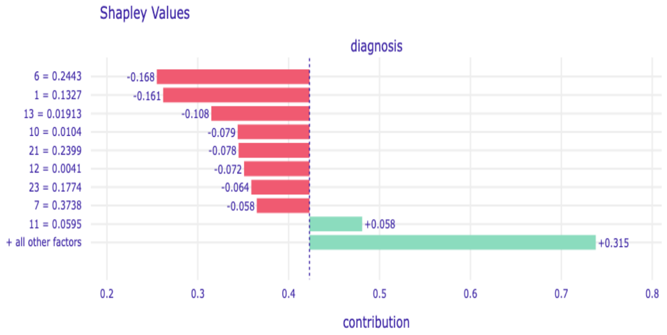
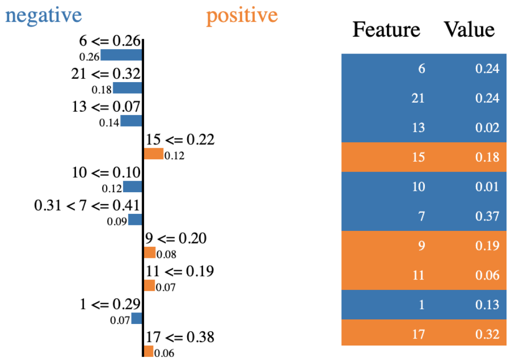
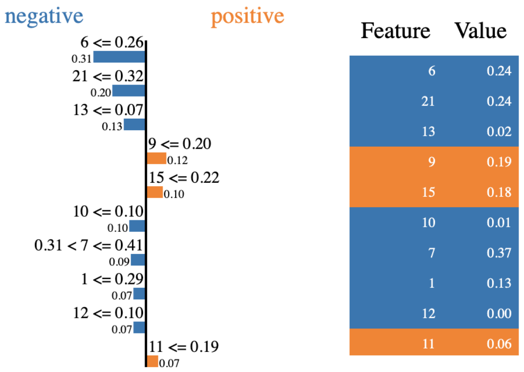


Fig. 25. Shapley values plot for 15-layer model.



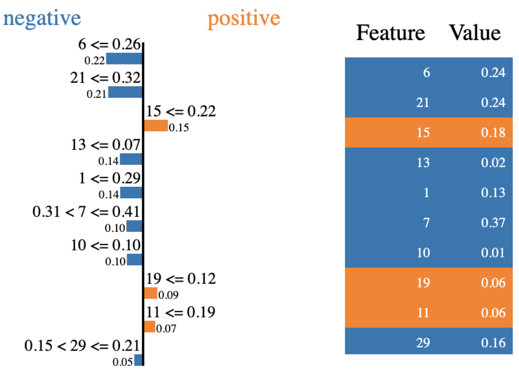
Feature	Value
6	0.24
21	0.24
13	0.02
15	0.18
10	0.01
7	0.37
9	0.19
11	0.06
1	0.13
17	0.32

Fig. 26. LIME plot for baseline model.



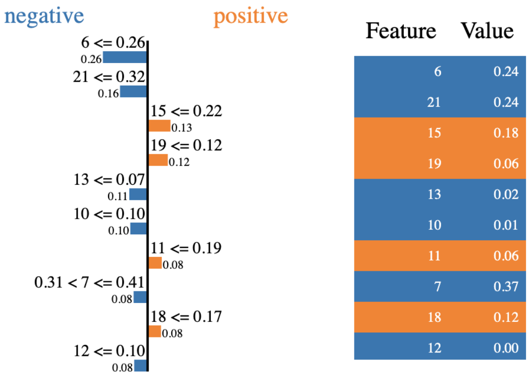
Feature	Value
6	0.24
21	0.24
13	0.02
9	0.19
15	0.18
10	0.01
7	0.37
1	0.13
12	0.00
11	0.06

Fig. 27. LIME plot for smaller model.



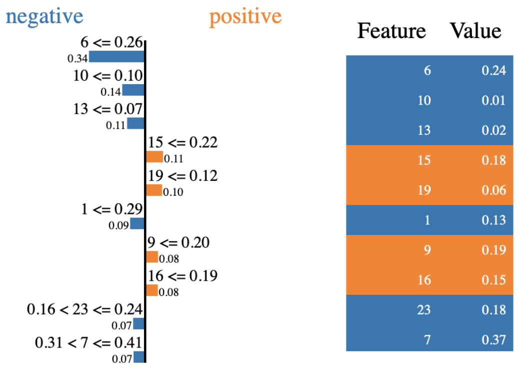
Feature	Value
6	0.24
21	0.24
15	0.18
13	0.02
1	0.13
7	0.37
10	0.01
19	0.06
11	0.06
29	0.16

Fig. 28. LIME plot for 5-layer model.



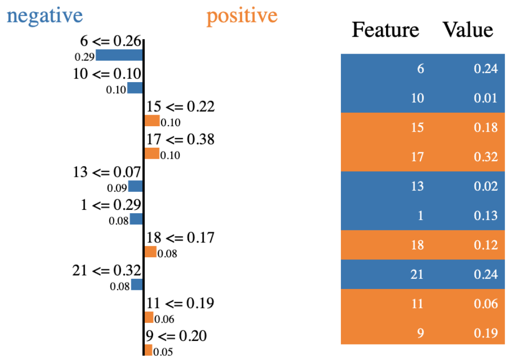
Feature	Value
6	0.24
21	0.24
15	0.18
19	0.06
13	0.02
10	0.01
11	0.06
7	0.37
18	0.12
12	0.00

Fig. 29. LIME plot for 7-layer model.



Feature	Value
6	0.24
10	0.01
13	0.02
15	0.18
19	0.06
1	0.13
9	0.19
16	0.15
23	0.18
7	0.37

Fig. 30. LIME plot for 10-layer model.



Feature	Value
6	0.24
10	0.01
15	0.18
17	0.32
13	0.02
1	0.13
18	0.12
21	0.24
11	0.06
9	0.19

Fig. 31. LIME plot for 15-layer model.

References

[1] 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.

[2] https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html

[3] 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

[4] https://www.darpa.mil/program/explainable-artificial-intelligence

[5] 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.

[6] 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.

[7] Yash Amethiya, Prince Pipariya, Shlok Patel, Manan Shah, ”Comparative analysis of breast cancer detection using machine learning and biosensors,” Intelligent Medicine, Volume 2, Issue 2, 2022, Pages 69-81, ISSN 2667-1026, https://doi.org/10.1016/j.imed.2021.08.004.

[8] Castellino, Ronald A. “Computer aided detection (CAD): an overview.” Cancer imaging : the official publication of the International Cancer Imaging Society vol. 5,1 17-9. 23 Aug. 2005, doi:10.1102/1470-7330.2005.0018

[9] UCI Machine Learning (2016). Breast Cancer Wisconsin (Diagnostic) Data Set. Kaggle. https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data?select=data.csv

[10] J. Brownlee, “Smote for imbalanced classification with python,” MachineLearningMastery, 16-Mar-2021. [Online]. Available: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/. [Accessed: 10-Dec-2022].

[11] 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.

[12] 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.

[13] 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

[14] A. Kozak, “Basic xai with DALEX-part 2: Permutation-based variable importance,” Medium, (01-Nov-2020). Available: https://medium.com/responsibleml/basic-xai-with-dalex-part-2-permutation-based-variable-importance-1516c2924a14.

[15] Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.). christophm.github.io/interpretable-ml-book/