Interpretability and Explainability of Logistic Regression Model for Breast Cancer Detection

Emina Tahirović[1] and Senka Krivić[2]
[1]*Faculty of Engineering and Natural Sciences, International Burch University, Sarajevo 71000, Bosnia and Herzegovina*
[2]*Faculty of Electrical Engineering, University of Sarajevo, Sarajevo 71000, Bosnia and Herzegovina*

Keywords: Logistic Regression, Explainable AI, Transparency, Healthcare.

Abstract: Artificial Intelligence techniques are widely used for medical purposes nowadays. One of the crucial applications is cancer detection. Due to the sensitivity of such applications, medical workers and patients interacting with the system must get a reliable, transparent, and explainable output. Therefore, this paper examines the interpretability and explainability of the Logistic Regression Model (LRM) for breast cancer detection. We analyze the accuracy and transparency of the LRM model. Additionally, we propose an NLP-based interface with a model interpretability summary and a contrastive explanation for users. Together with textual explanations, we provide a visual aid for medical practitioners to understand the decision-making process better.

1 INTRODUCTION

An accurate cancer diagnosis is essential for planning the best action and establishing a treatment plan. Over a hundred risk factors can simultaneously be involved in estimating a single post-test probability, making manual prediction incredibly difficult. Machine learning models can be beneficial for processing large numbers of variables and thereby bridging the gap between risk factors and risk estimation. However, legal and ethical accountability issues make fully independent AI medical systems unlikely. An alternative is the possibility of developing explainable AI systems, which would aid humans in decision-making while keeping the simplicity of its background processes.

Among the numerous computer models used for predicting clinical outcomes can be distinguished two main subcategories: models built by the statistics community and models built by the machine learning community. Logistic Regression (LR) is a statistical fitting model widely used to model medical problems, such as estimating disease risk in coronary heart disease, breast cancer, prostate cancer, postoperative complications, and stroke. Several studies (Ayer et al., 2010; Aviv et al., 2009) have shown that LR is a valuable tool in medical diagnosis since the methodology is well established.

As ML models penetrate critical and sensitive areas such as medicine, what becomes increasingly challenging is the inability of humans to understand these models (Lipton, 2018). It is of utter importance that ML models used in the medical domain can be trusted. Even if a model achieves high accuracy, it is still desirable that medical practitioners can decide if the diagnosis makes sense and that the output is interpretable even to a patient. In a hypothetical scenario, an LR model would output a cancer diagnosis and explain why this sample was classified as a benign or malignant tumor. The medical practitioner could conclude if the patient needs more invasive tests to be conducted or if the diagnosis is clear enough as is. A semantic explanation paired with the visualized aid that clarifies the deciding factors and features for a specific case can be used for this purpose, which we propose in this paper.

We propose an Explainable AI (XAI) system for breast cancer detection consisting of a classification model and semantic and visual models providing the decision of an LR model together with interpretations and explanations for a medical practitioner interacting with the system. The classification segment is based on a binary LR model, trained on breast cancer characteristics data. The semantic model is an NLP-based user interface using automated question-answering models, where the user is prompted to ask questions about the classification of the tumor. This way, we produce an output tailored to the user's needs. We combine the inherent interpretability of LR with a contrastive explanations approach. Figure 1 gives a diagram explaining this system.

Tahirović, E. and Krivić, S.
Interpretability and Explainability of Logistic Regression Model for Breast Cancer Detection.
DOI: 10.5220/0011627600003393
In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence (ICAART 2023) - Volume 3*, pages 161-168
ISBN: 978-989-758-623-1; ISSN: 2184-433X

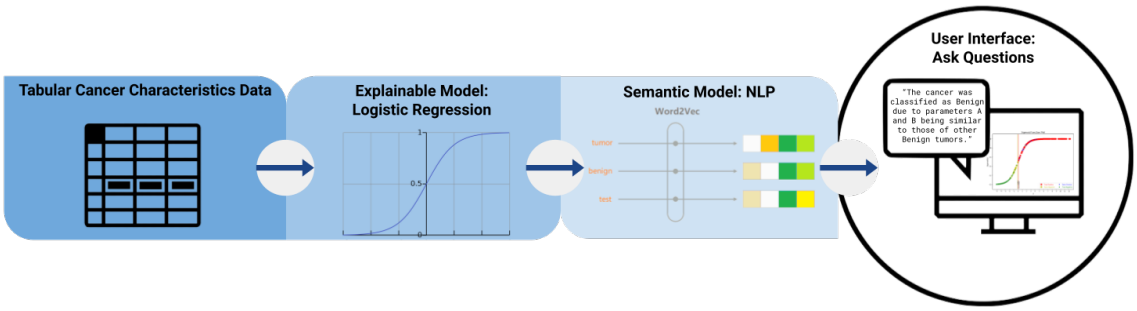

Figure 1: Diagram of the proposed XAI system for Breast Cancer Detection.

This paper is structured as follows. Section 2 describes related work. In section 3, we explain LR and its interpretability and explainability. In section 4, we start with a brief introduction to the dataset, move on to explaining our LR model, and finally present the NLP semantic user interface. Section 5 presents results and section 6 concludes the paper.

2 RELATED WORK

The development of machine learning methods made many breakthroughs in challenging clinical tasks such as assisted diagnosis, automatic image analysis, and prediction (Zhao et al., 2020). Logistic Regression is successfully used for several medical prediction tasks (Naji et al., 2021; Anisha et al., 2021). The LR model for binary data is probably the most widely used in medical research (Hastie et al., 2009). Ayer et al. (Ayer et al., 2010) compared the performance of LR and Artificial Neural Networks (ANNs) on the Breast Cancer Wisconsin (BCW) dataset, which we also use in this research. They note that in terms of clinical interpretation, LR models have better clinical references than ANNs. Sultana and Jilani (Sultana and Jilani, 2018) used Simple Logistic Regression. They trained the classifier on the BCW dataset. In their research, this accuracy was higher than any other classifier, some of which were Nearest Neighbor Classifier, Random Forest, and Decision Tree.

Limitations on the use of deep and ensemble learning models in the medical domain are reflected in their lack of interpretability (Chakrobartty and El-Gayar, 2021). Gunning and Aha (Gunning and Aha, 2019) define XAI as "AI systems that can explain their rationale to a human user, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future".

The measures and models involved in decision-making and solutions to explain them explicitly had been inspected by Yang et al. (Yang et al., 2022). They demonstrated the research trends toward trustable AI and showcased promising XAI results for the two most widely investigated classification and segmentation problems in medical image analysis.

The fundamental conceptual differences of Explainable Artificial Intelligence (XAI) for regression and classification tasks were analyzed by Letzgus et al. (Letzgus et al., 2021). They focus on 'post-hoc' explanation where they try to attribute the prediction for each data sample to its input features in a meaningful manner. They cite contextual utility and feature importance as the bases of early approaches toward understanding the decision processes of ML models. They concluded that explanation methods are favorable in the regression scenario and that XAI cannot be transferred between different types of ML problems without adaptation.

One point of view that has the potential to broaden the scope of XAI explanations is *contrastive questions*. Cashmore et al. (Cashmore et al., 2019) give an example of this type of question: *"Why A rather than B?"*. When answering a contrastive explanation, one must consider a situation where scenario B might be better suited than scenario A. In other words, one must consider why scenario B would be more appropriate by arguing why scenario A would be less so. This type of explanation is called a *contrastive explanation* Krarup et al. (Krarup et al., 2021) argue that users of explainable user interfaces must be able to ask explicit contrastive questions.

Natural Language Processing (NLP) may offer a helping hand in a quest to make medical domain regression tasks more explainable. Krieger (Krieger, 2016) identifies the medical NLP statements used by medical professionals as *gradation words*. E.g. *"X-ray technician A suspects that patient B suffers from breast cancer"*. Modeling such language in an NLP semantic output could lead to more trust in medical domain regression tasks. Therefore, we use such language to leave an approachable and professional impression.

3 METHODS

This section describes the proposed approach of an XAI system consisting of an LR classifier, interpretability, and explainability modules.

3.1 Logistic Regression

Logistic Regression determines the relationship between a binary outcome (dependent variable) and predictors (independent variables). It estimates the probability of an event's occurrence. It is widely used in biostatistical applications where binary responses (two classes) occur quite frequently (Hastie et al., 2009). In the medical problem, we address the two classes which make the binary response 1 for malignant tumors and 0 for benign. Here, the diagnosis is the dependent variable, and the parameters which describe the cancer cells are the independent variables.

Let y denote the presence of the disease, where $y = 0$ or $y = 1$, and let $\mathbf{x}$ denote the vector of predictors (features). Given that p denotes the probability of breast cancer (the probability that $y = 1$), we can define Logistic Regression with

$$P(y = 0,1|\mathbf{x},\mathbf{w}) = \frac{1}{1+e^{-(\mathbf{w}^T\mathbf{x}+b)}} \quad (1)$$

Here, $(\mathbf{w},b)$ are weights, with b being a constant, $\mathbf{w}$ the regression coefficients to their respective predictor variables $\mathbf{x}$, and y the class label. These coefficients are estimated from the available data.

If the training instances are from the dataset $\mathcal{M} = (\mathbf{x}_1,y_1),...,(\mathbf{x}_n,y_n)$ with the labels $y_i \in \{0,1\}$, one estimates $(\mathbf{w},b)$ by minimizing the negative log-likelihood: $\min_{\mathbf{w},b} \sum_{i=1}^{n} log(1+e^{-(\mathbf{w}^T\mathbf{x}_i+b)})$ (2).

3.2 Interpretability of Logistic Regression

An oft-made claim is that linear models are preferable to deep neural networks because of their interpretability. The high accuracy of complex models comes at the expense of interpretability. Hence, even the contributions of individual features to the predictions of such a model become challenging to understand (Lou et al., 2012). Interpretability can be seen as a reflection of several different ideas than a monolithic concept (Lipton, 2018). Another view of interpretability is that it represents the degree to which a human can understand the cause of a decision (Miller, 2019).

While the output of the LR model is binary, the precedent of this output is a probability output in the range from 0 to 1. The probability output is transformed into the binary output by using a cutoff value of 0.5. A positive classification ($= 1$) is the result of a probability of ≥ 0.5, and a negative classification ($= 0$) is the result of a probability of < 0.5. We use this probability of each classification as an interpretability approach, essentially utilizing an aspect of LR's inherent interpretability. By analyzing Eq.(1) we get

$$\frac{P(y = 1|\mathbf{x},\mathbf{w})}{1-P(y = 1|\mathbf{x},\mathbf{w})} = \frac{P(y = 1|\mathbf{x},\mathbf{w})}{P(y = 0|\mathbf{x},\mathbf{w})} = e^{\mathbf{w}^T\mathbf{x}+b} \quad (2)$$

We can now examine how the feature weights and values affect the probability. Let us say that x is a vector of $x_1,...,x_n \in \mathbb{N}$ variables, and feature values of x_i changes from $x_k = a$ to $x_k = c$ where $c-a = 1$. Now we can examine how the change of the feature from a to c impacted the model's outcome by observing the ratio the probability gets scaled by a factor of e^{w_k} as we can deduce from:

$$\frac{e^{(b+w_1x_1+...+cw_k+...+w_nx_n)}}{e^{(b+w_1x_1+...+aw_k+...+w_nx_n)}} = ... = e^{(a+1)w_k-aw_k} = e^{w_k} \quad (3)$$

To highlight this impact for a user and provide the interpretability of the decision, we use the following plots:

1. Sigmoid plot, to clarify to the user where the currently observed cancer sample lies on the plot,
2. value counts plots to clarify the dominance of feature values for each classification, and
3. classification shift at the classification border with changing values of x_k.

3.3 Explainability of Logistic Regression

If we look at LR as a classification model, there is implicit knowledge associated with the class. This can serve as a piece of a potential explanation of LR (Letzgus et al., 2021). The explainability of a model depends on the level of known information about how the parameters of a model impact the decision process. In this respect, the coefficients of an LR model make it directly explainable. Hence, as valuable information for the user, we select visualization of feature importance.

Another important aspect that needs to be kept in mind when developing XAI models is the recipient of the information. The information must be tailored to the end user. As our interface is meant to be used by medical personnel, we tailor our explanations in a way that will help relate their pre-existing knowledge on the topic, while helping them understand how our model makes predictions. The explainability of our model relies on three different methods:

- identification of a relevant subset of features,
- interpretation of feature importance scores,
- explainability by contrastive example.

We present feature importance plots, parallel plots, classification shift plots, box plots, and heat maps to help the end user understand what kind of data leads to a certain prediction.

Contrastive Explanation: is a specific type of explanation which answers a question of type: *"Why A rather than B ?"*. It includes a contrastive example with the goal of presenting the difference in the decision-making to the user. It is commonly used in various AI fields (Krarup et al., 2021). In our medical XAI system case, we propose presenting two different feature vectors $\mathbf{x}_{ce}$ and $\mathbf{x}_{cf}$ as contrastive examples of feature vectors to the user. For the current input vector $\mathbf{x}_i$ and the decision of an LR classifier y_i, we define these contrastive vectors as follows. $\mathbf{x}_{ce}$ is the value of a feature vector with minimal Euclidean distance from the input vector $\mathbf{x}_i$ which results in the opposite class result $\neg y_i$. $\mathbf{x}_{cf}$ is a vector in which variation in a single feature value results in the prediction class change and becomes $\neg y_i$.

4 IMPLEMENTATION DETAILS XAI MEDICAL SYSTEM

This section explains the dataset used for this research and the implementation process. In the first subsection, we go through the nature of the dataset and its features. In the following subsections, we explain the implementation of the LR model and the NLP semantic user interface.

4.1 Dataset Description

The dataset used for this research is the well-known Wisconsin Breast Cancer Dataset obtained from the UC Irvine Machine Learning Repository (Mangasarian et al., 1995). The nine predictor variables analyzed for this research are visually assessed cytological characteristics of a Fine Needle Aspiration (FNA) sample. These variables take integer values between 1 and 10. Wolberg and Mangasarian (Wolberg and Mangasarian, 1990) chose the nine features based on statistical analysis, which showed a significant difference in values for benign and malignant samples. Therefore, they are prominent candidates for predictor variables for our LR model. These predictor variables are uniformity of cell shape, uniformity of cell size, clump thickness, marginal adhesion, single epithelial cell, bare nuclei, bland chromatin, normal nucleoli, and mitosis. The feature *class* is the predicted or the dependent variable in this dataset.

4.2 Logistic Regression Model

Our LR model is implemented from scratch, in Python, on the Jupyter Notebook platform. It is a binary LR model with an added bias term. The motivation behind the from-scratch implementation is the interpretability of each step of LR, where one can choose to print out different values and interpret them separately, as opposed to a black-box approach to LR with a library-provided model.

The original model is trained on all 9 features and is evaluated using stratified 5-fold cross-validation. These preliminary scores are *97.10%* for accuracy, and *95.65%* for F_1-score. Models trained on different subsets of features are cross-validated individually. Each subset consists of n most important features, as determined by feature importance scores (see subsection 5.2). The results are recorded and the best-performing model is selected. The optimal subset size was determined to be 7 features, which are, in order of importance: *bare nuclei*, *clump thickness*, *mitosis*, *uniformity of cell shape*, *marginal adhesion*, *bland chromatin*, and *normal nucleoli*. This model is our finished LR model.

4.3 NLP Models and User Interface

The output of our framework is based on an NLP explainable user interface, which produces a combination of textual and pictorial explanations to provide aid for medical practitioners to better understand the decision-making process of the LRM. In the following subsections, we explain the models used to implement this user interface.

Word2Vec: is a word embedding model. It is an unsupervised learning algorithm that generates low-dimensional vector representations of words. The variation used in this research is Skip-gram, which predicts the context, given the word.

GloVe: or global vectors is an unsupervised learning algorithm that obtains low-dimensional vector representations for words and performs dimensionality reduction on the word context matrix. It thereby keeps track of the frequencies of word co-occurrences in a text, with an additional weight parameter dependent on the distance between the words.

Question Database: consists of twenty-two questions and their respective answers. Some of these answers are static, while others are assembled with the help of the information obtained from the model. The database is shown in table 1. Answers to questions

13-22 are explained in section 5.2.

Automated Question Answering: The next level of interpretability is reached through an NLP automated question-answering model based on Word2Vec or GloVe embeddings. We use cosine similarity to estimate how similar the user-asked question is to the questions in the database. The process is illustrated from start to finish in figure 2.

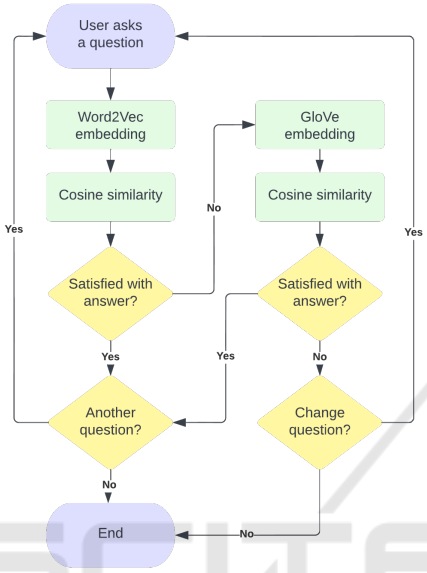

Figure 2: User interface state diagram.

5 EXPERIMENTAL RESULTS

This section presents the prediction, interpretability, and explainability results of our models. We provide textual, as well as pictorial explanations of our results.

5.1 LR Prediction Evaluation

We present the prediction results of our model in Table 2 for the test subset and the original dataset. For the performance evaluation, we compare the CV score, accuracy, precision, recall, and F1 score. Table 3 shows the confusion matrices for the test set and complete dataset. These results indicate that our model performs well, as the rates of falsely classified data are low. The ratio of each portion (true positives, false positives, false negatives, and true negatives) is well-preserved in the test set, as evident from comparing the test and entire dataset confusion matrices.

5.2 Interpretability Results

When the user begins the interaction with our interface, they are presented with some basic information about how their tumor is seen by our model. They are told what their classification is (*benign* or *malignant*), and how probable it is that the tumor is malignant.

Figure 3 shows the Sigmoid plot for the entire dataset, where the colors of the dots are clarified by the legend. The two vertical bars represent the right-most negative and the left-most positive classification, along with their respective $\boldsymbol{Z}$-values. The currently observed cancer sample is represented by a cyan dot.

This figure is presented along with the answer to question no. 13 in the semantic output (see Table 1). Its purpose is to clarify to the user where the currently observed cancer sample lies on the plot and to get a visual sense of how similar this sample is to other such classifications.

In this way, we maintain a level of transparency as to how our model classifies data. This plot is presented in combination with a textual output stating the density of falsely classified samples in the 0.5-neighborhood of the current cancer sample (along the $\mathbf{Z}$-axis), the values of each of the 7 relevant features, and the space between the left-most positive and right-most negative classification. These statements aim to help the user interpret the classification.

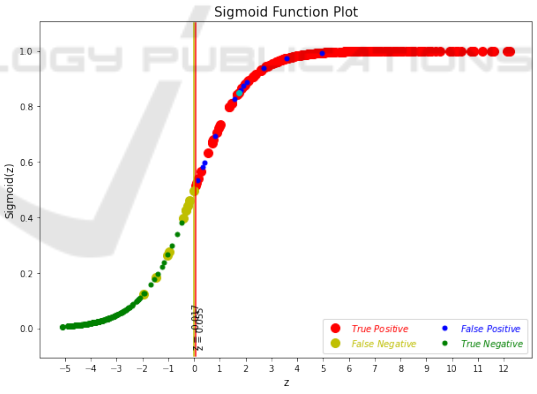

Figure 3: Sigmoid plot for the entire dataset.

Figure 4 shows the percentages of value counts for *bare nuclei*, for benign and malignant samples. Analogous plots are generated for all other features. These plots are used in questions 14 and 15 to clarify the differences in the values that the individual features typically take for different classifications. This way, the user can clearly see that for example *bare nuclei=1* is most often seen in benign samples.

Figure 5 shows the shift of the classification probability of the observed tumor sample with the change of a single feature, in this case, *bare nuclei*. This plot

Table 1: Questions database.

	Questions	Answers
1	Why is the tumor classified as malignant?	The parameters of this tumor produced a similar function value to the tumors which have previously been marked malignant by a human doctor. Your parameters are:
2	Why is the tumor classified as benign?	The parameters of this tumor produced a similar function value to the tumors which have previously been marked benign by a human doctor. Your parameters are:
3	What does 'malignant' mean?	Malignant tumors are cancerous (i.e., they invade other sites).
4	What does 'benign' mean?	Benign tumors are those that stay in their primary location without invading other sites of the body.
5	How similar is this tumor to tumors which have been classified as False Positive?	Your tumor falls in the range of tumors that are known to have been correctly classified as negative, i.e. were True Positive and had no similar False Positive points nearby.
6	How similar is this tumor to tumors which have been classified as False Negative?	Your tumor is not classified as benign, therefore it is not similar to other such data points.
7	Should I get additional tests done?	Additional tests are not considered necessary in cases similar to yours. However, consult your specialist about the best course of action.
8	What do 'benign' and 'malignant' mean?	Benign tumors are those that stay in their primary location without invading other sites of the body. Malignant tumors are cancerous (i.e., they invade other sites).
9	What is the difference between 'benign' and 'malignant'?	Benign tumors are those that stay in their primary location without invading other sites of the body. Malignant tumors are cancerous (i.e., they invade other sites).
10	What is the next course of action?	The next course of action for your tumor should be consulted with your specialist.
11	What are some similar samples and their likelihoods?	Point X was found to be in the 5 most similar points to your sample. It has a Y probability of being malignant.
12	What is the treatment plan for my tumor?	The treatment plan should be consulted with your specialist.
13	Can I see my tumor's data point visualized?	Here is your tumor data visualized:
14	Which feature values prevail for malignant tumors?	Here you can see approximate percentages of either classification per feature per value:
15	Which feature values prevail for benign tumors?	Here you can see approximate percentages of either classification per feature per value:
16-22	How does the classification shift with changing the value of feature X?	This is how your tumor classification changes with changing the value of feature X:

Table 2: Evaluation of LR model.

	Test set	Whole set
CV score	/	100.00%
Accuracy	98.55%	96.93%
Precision	95.83%	95.42%
Recall	100.00%	95.82%
F1-score	97.87%	95.62%

Table 3: Confusion matrix for the test set (left) and the whole dataset (right) for LR model trained on 7 features.

		Actual	
		Positive	Negative
Predicted	Positive	45	1
	Negative	0	23

		Actual	
		Positive	Negative
Predicted	Positive	433	11
	Negative	10	229

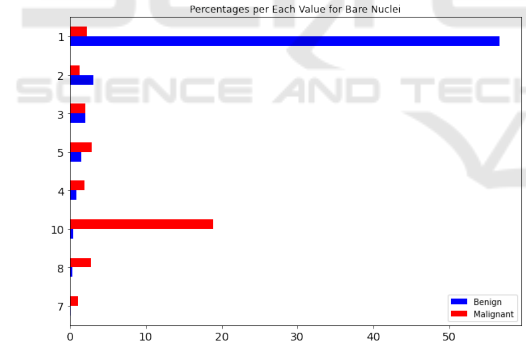

Figure 4: Class percentages plot for *bare nuclei*.

is shown along with the answer to questions no. 16-22 (one question per feature), for the purposes of visually clarifying the impact of the individual features on the classification shift of the currently observed cancer sample. In this way, we clearly show how changing the value of just one feature impacts the classification.

5.3 Explainability Results

Figure 6 shows the feature importance scores for our LR model, which are also the model weights. These scores explain which features had the most impact on the prediction. It can be noticed from the plot that

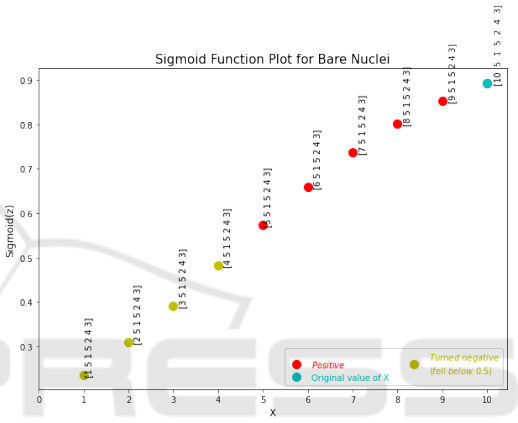

Figure 5: Changing the value of *bare nuclei* for the observed tumor sample.

the feature *bare nuclei* contributes most to the classification outcome, which means that it will have the greatest "pull" in deciding how the tumor is classified. This plot clearly shows how important each feature is for the prediction outcome.

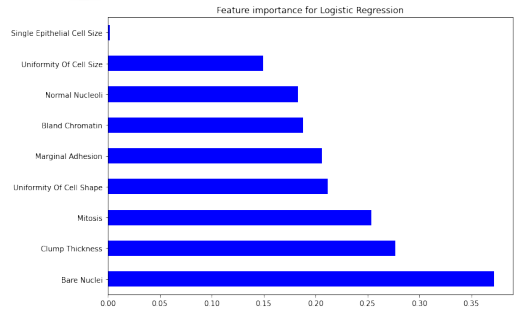

Figure 6: Feature importance scores for the LR model.

Figure 7 shows the heat maps for the malignant and benign portions of the data. Upon visual inspection, it becomes obvious that these two portions differ. For example, features *marginal adhesion* and *bland chromatin* have a correlation of *0.33* for the posi-

tive portion, and *-0.02* for the negative portion of the dataset. Many more similar examples can be found in the heat maps. This plot shows how features manifest different relationships between malignant and benign tumors.

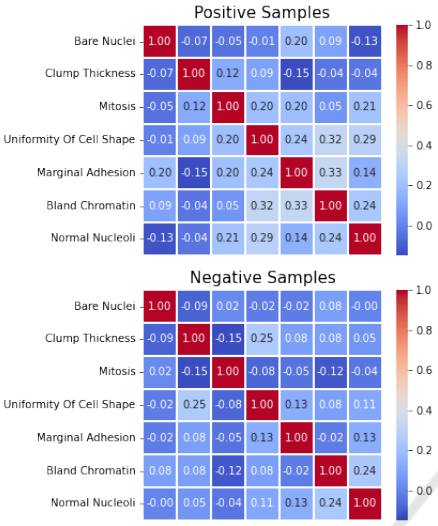

Figure 7: Heat maps for the LR model for positive (malignant) and negative (benign) portions of the data.

Figure 8 shows the parallel coordinates plot for the malignant and benign portions of the data. One can infer simply by looking at this graph that the features tend to take lower values for the benign portion of the dataset, while the malignant portion tends to take higher values. This can also be inferred from figure 4.

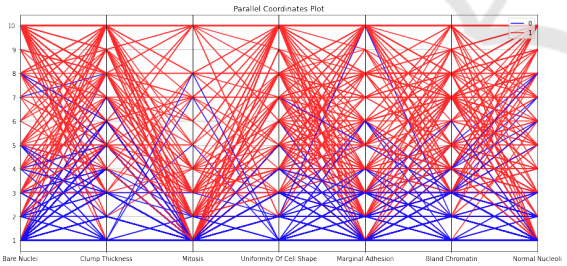

Figure 8: Parallel coordinates plot for positive (1=malignant) and negative (0=benign) portions of the dataset.

Figure 9 shows the box plots for positively and negatively classified samples. Again, it is clear that the positive and negative samples behave differently. For instance, the median values and ranges are different for all features. Feature values for 5 out of 7 features in the negative portion are scattered across the plot. These plots confirm that the malignant and benign portions of the data have different statistical properties.

To better explain the effect that the values of the features have on the classification of the tumor, we

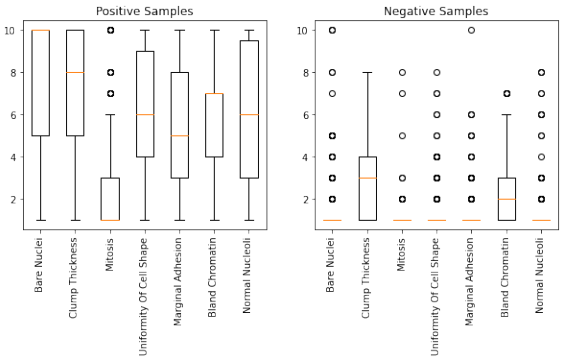

Figure 9: Box plots for the whole dataset.

find the space of possible solutions for $Sigmoid(\mathbf{Z}) - 0.5 < \varepsilon$, $\varepsilon = 1 \times 10^{-6}$. This way, we show how moving the values along their respective ranges affects the shift in classification at the classification border.

Figure 10 shows the classification shift at the classification border while moving the values of the feature *bare nuclei* along its range. Analogous figures are generated for all other features. These figures help explain how changing even a single feature's value affects the classification.

If we try to imagine the vector of 7 features as a vector in 7-dimensional Euclidean space, where each feature represents a single dimension, then we can also imagine that the illustrated changes are step-wise changes per dimension. This way, especially illustrative are the figures where the classification had changed, but the step-wise shift per dimension is only one step away from the original vector. This can, for example, be observed for feature *bare nuclei*, where changing the value of *bare nuclei* from *3* to *2*, while the values of the other features remain constant, changes the output of *Sigmoid(**Z**)* from close to *0.5* but positive, to significantly less than *0.5*.

Another interesting connection can be made between the weights of the features and the classification shifts plot. Namely, it can be observed that the $Sigmoid(\mathbf{Z})$ values always increase as X_i increases, where i is the currently observed feature. This is because the weights of the features are all positive.

6 CONCLUSION

In this paper, we presented a framework for breast cancer prediction and output interpretation based on Logistic Regression and NLP methods. The inherent interpretability of Logistic Regression was used as the foundation of the quest toward interpretability, while the pursuit of explainability focused on explaining the preferred data of each class. While Logistic Regression is commonly used in clinical and health services

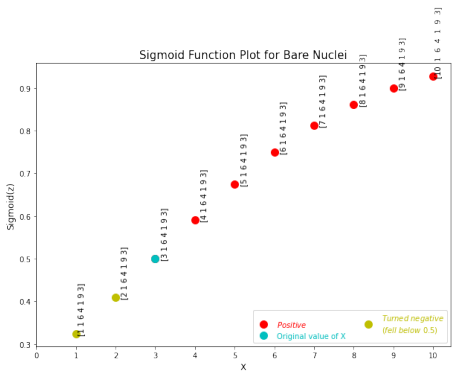

Figure 10: Classification shift at classification border for *bare nuclei*.

as a powerful tool, it is important not to forget its limitations, such as the assumption of the absence of high intercorrelations among the predictors. Therefore, in future work, we will examine the XAI setup for medical applications with more powerful methods such as Deep Learning.

Our contribution is a completely transparent, interpretable, and explainable model, developed with the purpose of aiding medical personnel in the decision-making process. It contributes towards extending XAI to regression models, by adapting an NLP method as a way to access desired explanations. In future work, we plan to perform user experiments in order to rate the helpfulness of our model.

REFERENCES

Anisha, P., Reddy C, K., Apoorva, K., and Mangipudi, C. (2021). Early diagnosis of breast cancer prediction using random forest classifier. *IOP Conference Series: Materials Science and Engineering*, 1116:012187.

Aviv, R. I., d'Esterre, C. D., Murphy, B. D., Hopyan, J. J., Buck, B., Mallia, G., Li, V., Zhang, L., Symons, S. P., and Lee, T.-Y. (2009). Hemorrhagic transformation of ischemic stroke: Prediction with ct perfusion. *Radiology*, 250(3):867–877.

Ayer, T., Chhatwal, J., Alagoz, O., Kahn, C. E., Woods, R. W., and Burnside, E. S. (2010). Comparison of logistic regression and artificial neural network models in breast cancer risk estimation. *RadioGraphics*, 30(1):13–22.

Cashmore, M., Collins, A., Krarup, B., Krivic, S., Magazzeni, D., and Smith, D. E. (2019). Towards explainable AI planning as a service. *CoRR*, abs/1908.05059.

Chakrobartty, S. and El-Gayar, O. F. (2021). Explainable artificial intelligence in the medical domain: A systematic review. In Chan, Y. E., Boudreau, M., Aubert, B., Paré, G., and Chin, W., editors, *27th Americas Conference on Information Systems*. Association for Information Systems.

Gunning, D. and Aha, D. W. (2019). Darpa's explainable artificial intelligence (XAI) program. *AI Mag.*, 40(2):44–58.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition*. Springer Series in Statistics. Springer.

Krarup, B., Krivic, S., Magazzeni, D., Long, D., Cashmore, M., and Smith, D. E. (2021). Contrastive explanations of plans through model restrictions. *J. Artif. Intell. Res.*, 72:533–612.

Krieger, H. (2016). Capturing graded knowledge and uncertainty in a modalized fragment of OWL. In van den Herik, H. J. and Filipe, J., editors, *Proceedings of the 8th International Conference on Agents and Artificial Intelligence, Vol.2*, pages 19–30. SciTePress.

Letzgus, S., Wagner, P., Lederer, J., Samek, W., Müller, K., and Montavon, G. (2021). Toward explainable AI for regression models. *CoRR*, abs/2112.11407.

Lipton, Z. C. (2018). The mythos of model interpretability. *Commun. ACM*, 61(10):36–43.

Lou, Y., Caruana, R., and Gehrke, J. (2012). Intelligible models for classification and regression. In *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, 2012*, pages 150–158. ACM.

Mangasarian, O. L., Street, W. N., and Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. *Oper. Res.*, 43(4):570–577.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267:1–38.

Naji, M. A., Filali, S. E., Aarika, K., Benlahmar, E. H., Abdelouhahid, R. A., and Debauche, O. (2021). Machine learning algorithms for breast cancer prediction and diagnosis. *Procedia Computer Science*, 191:487–492.

Sultana, J. and Jilani, A. K. (2018). Predicting breast cancer using logistic regression and multi-class classifiers. *International Journal of Engineering & Technology*, 7(4.20).

Wolberg, W. H. and Mangasarian, O. L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the National Academy of Sciences*, 87(23):9193–9196.

Yang, G., Ye, Q., and Xia, J. (2022). Unbox the black-box for the medical explainable ai via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Information Fusion*, 77:29–52.

Zhao, Y., Yang, F., Fang, Y., Liu, H., Zhou, N., Zhang, J., Sun, J., Yang, S., Menze, B. H., Fan, X., and Yao, J. (2020). Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 4836–4845. Computer Vision Foundation / IEEE.