15 July 2022

Final Report

1st Cycle Integrated Project

IMPLEMENTATION AND APLICATION OF THE SHAP AND LIME ALGORITHMS IN HEALTHCARE

João Monteiro 96610 | Maria Pinho 95952

Under: Nuno Silva | José Moreira | Raquel Santos | Adalgisa Guerra

Instituto Superior Técnico

Hospital da Luz Learning Health

Explainable Artificial Intelligence provides a promising solution to the lack of trust and use of AI models by medical professionals. In this internship, we performed an initial review of existing literature, in order to obtain a list of available types of explanations, and a definition with clear criteria of what makes a good explanation. Afterwards, we developed a survey to medical professionals in which we presented the distinct types of explanations and asked for an evaluation according to the previously defined criteria. Finally, we used SHAP and LIME methods to develop an online application which would provide a good explanation for a random forest predictive model of surgical margin in prostate cancer patients that had undergone surgery. The machine learning model has good performance (90% accuracy, 83% recall, 96% specificity, 95% precision).

Keywords: Explainable Artificial Intelligence; XAI; SHAP; LIME; Survey

1. Introduction

Prostate cancer is the fourth most common cancer, with the highest incidence in Europe, 33.5%, followed by Asia, 26.2%. It is most prevalent in these same continents, with 5-year prevalence being 37.8% in Europe and 23.7% in Asia (WHO, 2020). Although its exact causes are unknown, a few risk factors have been identified, such as an advanced age, endogenous hormones, and dietary fat. Progression of prostate cancer in patients is highly variable, as it is most often composed by a large number of small tumours which are asymptomatic. If detected at an early stage, prostate cancer is treated by radical prostatectomy, a surgical removal of the tumour (Quinn, 2002).

In the medical field, radiomics extracts a large number of features from radiographic images to objectively and quantitatively describe tumour phenotypes (Yip & Aerts, 2016) (Lambin, et al., 2012). It is comprised of four sequential processes: Image acquisition, Image segmentation, Feature extraction and Analysis, which may lead to the creation of clinical tools that can be useful to determine patient prognosis (Aerts, et al., 2014).

In certain areas of medicine, for example diagnosis, artificial intelligence models tend to have better results than their human counterpart (Lundberg, Nair, Vavilala, & al., 2018). However, due to the complexity of the problems, which leads to a growing complexity of the models, there is a lack of trust in this technology by medical professionals (Antoniadi, et al., 2021). It is in this context that explainable artificial intelligence (XAI) was created, providing an explanation for the results, so that they may be understood not only by the programmers who built the model, but also by any individual knowledgeable in its applied field. A model is considered explainable if it can be intrinsically interpretable – for example, a linear regression – or if it contains an adequate explanation (Markus, Kors, & Rijnbeek, 2021).

Machine learning is a branch of artificial intelligence that uses algorithms that leverage data to improve performance on a set of tasks (IBM, 2020). This way, it is possible to create programs trained around a sample dataset, known as training data, in order to make predictions with new data without the data scientists' interference. A black box machine learning model is an algorithm which can only be inspected in terms of its inputs and outputs, without any knowledge of its inner workings (Molnar, 2022). The lack of transparency of this type of algorithms tends to make it difficult to understand how exactly the model works and how it produced specific decisions, which imposes the need for XAI methods such as LIME and SHAP.

Lime or Local Interpretable Model-Agnostic explanations is an algorithm that was created by (Ribeiro, Singh, & Guestrin, 2016) which focuses on training interpretable models to explain individual predictions of black box machine learning models. By testing what happens to the model predictions when it is given variations of the sample data into the machine learning model, LIME generates a new dataset consisting of altered entries and the corresponding predictions of the black box model. Then an interpretable model is trained on this new dataset, weighting by the proximity of the sampled instances to the instance of interest. Even though this new model represents a good local approximation, it is not possible classify it has a good global approximation of the original machine learning model due to its higher complexity (Molnar, 2022). One of the pros of applying LIME to health-related problems is that as it is a post-hoc method, it allows health professionals to understand which variables are responsible for the prediction made by the model and compare them with their previous knowledge.

The second method applied is SHAP (SHapley Additive exPlanation), created by (Lundberg & Lee, 2017) which is based upon Shapley values. These are a feature of cooperative game theory, in which the value of an individual action is measured by the overall impact it has on the result given. In terms of predictive models, this translates to measuring the transition from the expected median value of the result, into an individual output by altering each variable at a time. If variables are not independent or the model is non-linear, their value is the average of all possible orderings. As consequence, SHAP can be applied to explain individual specific predictions as well as overall model behaviour. SHAP values can then be applied to a variety of plot-based explanations, from which we chose waterfall and force plots, due to their interpretability and ease of use.

The aim of this study is to apply XAI principles and evaluate the best approach to implement the SHAP and LIME methods in healthcare, particularly with the creation an online application with a predictive model for the classification of prostatic cancer surgical margin.

2. Methods

2.1 Literature Review

In our review of existing literature, our objectives were:

1. Defining explainable artificial intelligence.
2. Obtaining criteria for evaluating explanations.
3. Obtaining a list of available types of explanations.
4. Defining a good explanation.

We did not follow systematic review guidelines, such as PRISMA, because the field of XAI is quite recent, and there is no definite terminology used by all investigators. For example, some papers use 'interpretability' and 'explainability' interchangeably, while others consider 'interpretability' to be a requisite part of 'explainability'. As such, we used the search terms 'explainability' or 'explainable', 'interpretability', 'artificial intelligence', 'machine learning' and medicine' or 'medical' or 'clinical'. As sources we used PubMed, Google Scholar, arXiv and Web of Science.

After we had answers to all our objectives, we searched for existing SHAP, and LIME models applied in a clinical context. Using the same sources, we looked for 'SHAP' or 'LIME' and 'model'.

2.2 Initial Assessment - Survey

Our goal when building the survey was to answer the questions:

1. What do medical professionals know about XAI?
2. How trustworthy is XAI to medical professionals?
3. What explanation(s) do medical professionals prefer?

As such, we built a questionnaire, following the six principles of question design: reliability, one person should tend to answer the question the same way on different occasions; validity, the question should measure what is intended; discrimination, different answer options should be distinct and easily distinguishable; same meaning for all respondents; response rate, all questions should have a high response rate (if during testing one question is answered less frequently, then it should be altered); and relevance (de Vaus, 2013).

These principles require simple questions, with clear examples and answers in a valid scientifically approved format. We explored answer options, having created examples for each one, in a first version of our questionnaire. It was then shown to a small testing group, and feedback was used to choose and reformulate the questions and the answers' format in order to improve ease of response.

Upon obtaining a final version, we opted for online distribution, for which we use the platform https://freeonlinesurveys.com, which, despite having a time limit, as surveys can only be distributed in a period of two weeks, provided for a better user interface for respondents. The final questionnaire was then given to Hospital da Luz Learning Health for distribution.

2.3 Data Science

The machine learning model that the created web application was built around, was based in a 121-entry dataset with 16 variables (listed in the cohort table present in Appendix 1 and in our web application), of which five are numeric, one is symbolic, and the rest are binary, and two classes, given to us by our project advisors.

In order to be able to compare different alterations done to the original dataset, and verify if they increased the overall model performance, we created a simple neural network model for each modification and compared the correspondent performance metrics (accuracy, recall, precision, specificity).

To obtain the best model performance this process was divided in four steps: data profiling, data preparation, data classification and prediction.

2.4 App development

Our biggest goal in the development phase was creating the least complex and best organized application to present all the information necessary to understand the model created and that would allow health professionals to enter their patients' data and obtain the expected surgical margin explained by the methods SHAP and LIME.
In order to achieve the desired goal, we searched for a framework that could be easily implemented and hosted online for everyone to access. So, for the construction of this app, DASH was used, a python framework used for data visualization without requiring advanced knowledge of web development.

3. Results

3.1 Literature Review

An artificial intelligence or machine learning model is considered explainable if it is intrinsically comprehensible, for example a linear regression, or if the model is not interpretable then it provides an explanation along with the result, which approximates the model's behaviour (Markus, Kors, & Rijnbeek, 2021). Explainability can then be considered an accentuation of the parts of the algorithm which most contribute to its predictions, implying correlation between variables, not causation. XAI can also be distinguished from and require an explanation interface, in which the user can see the explanation and the result (Holzinger, Langs, Denk, Zatloukal, & Müller, 2019).

An evaluation of XAI can, as such, be made from the model itself or from the interface and should quantify its effectiveness, satisfaction, and usability by the final user, in this case, medical professionals. Based on (Markus, Kors, & Rijnbeek, 2021) we decided upon four criteria to evaluate explanations:

a. Clarity: the explanation is unambiguous, providing similar explanations in similar cases.
b. Parsimony: the explanation is concise and comprehensible, inversely proportional to the complexity of the explanation.
c. Soundness: the explanation is truthful to the original model.
d. Completeness: the explanation provides sufficient information.

There are many types of explanations available, so categorization is usually based on scope and model dependency. By scope, methods are separated into global explanations, which explain model behaviour for all model predictions, and local explanations, which explain each model prediction independently. Methods can also be model-specific, i.e., dependent on previous knowledge of the model they are applied to, or model-agnostic, i.e., they can be applied to any AI model. Finally, explanations depend on the nature of the input data: tabular data XAI models are more common in studies, often resulting in explanation by graphs, followed by imaging data and least often, text-analysis models (Antoniadi, et al., 2021).

The SHAP and LIME methods used are classified as model-agnostic local explanations (Markus, Kors, & Rijnbeek, 2021). These types of explanations follow an already built model and approximate its behaviour – post-hoc explanations – so they have the potential to present plausible but misleading explanations. Depending on the model they are applied to, they can also satisfy different criteria, so its important to consider which is more important to the final users.

3.2 Initial Assessment - Survey

We proposed as survey a questionnaire in five sections. The first one focuses on ascertaining the medical professional's familiarity with the area of XAI and their preconceived trust in machine learning models. Afterwards we present three types of explanations – with images, graphs, and text - each with a dedicated section, where an example is presented of a question, answer and explanation and the respondent is asked to classify each explanation in the four selected criteria. Classification is done in a 1-5 horizontal scale, as seen in the following example.

Figure 1 - Questionnaire section with an example of an explanation by text, with the following evaluation by the four criteria - clarity, parsimony, completeness, and soundness

The final section asks for respondents' preference between the three explanations presented in the previous sections, as well as what is the minimum model performance considered acceptable for it to be considered reliable. The questionnaire was written in Portuguese for ease of understanding by the target respondents (Portuguese medical professionals from Hospital da Luz) and can be accessed in its totality in an annexed pdf "Survey XAI Questionário".

3.3 Data Science

The machine learning model performance evaluation was made empirically by comparing the variation of the performance metrics as a consequence of dataset modifications. In the data profiling step, there were no identified plausible outliers, but six missing values were found associated with the variable "regra". To impute missing values, we tried three different methods, deleting the correspondent rows, changing the input to 0

and changing the input to the most frequent value. By creating a neural network model around each of these methods, it was found that the most accurate model was the one that deleted the rows where missing values where present.

In the data preparation step we applied scaling, balancing and discretization techniques to the dataset in order to better the performance of our model. The best scaler obtained was the minmax scaler, the best balancing technique was an oversampling method called SMOTE and no discretization techniques were used since there were no performance improvements by using them.

In the prediction step there were developed the necessary functions to modify the input data so it can be correctly evaluated after the final classification model was chosen in the last step.

In the last step, data classification, there was chosen the best classifier from a list ranging from simple probability classifiers such as naive Bayes and tree-like model of decisions such as decision trees to more complex classifiers such like random forests and neural networks. From the tested classifiers the best performing one was random forests, as it can be seen in the figure below.

We would like to add that the performance data that influence these decisions and the correspondent python code can be accessed in this GitHub page – https://github.com/GuberXZ/PIC

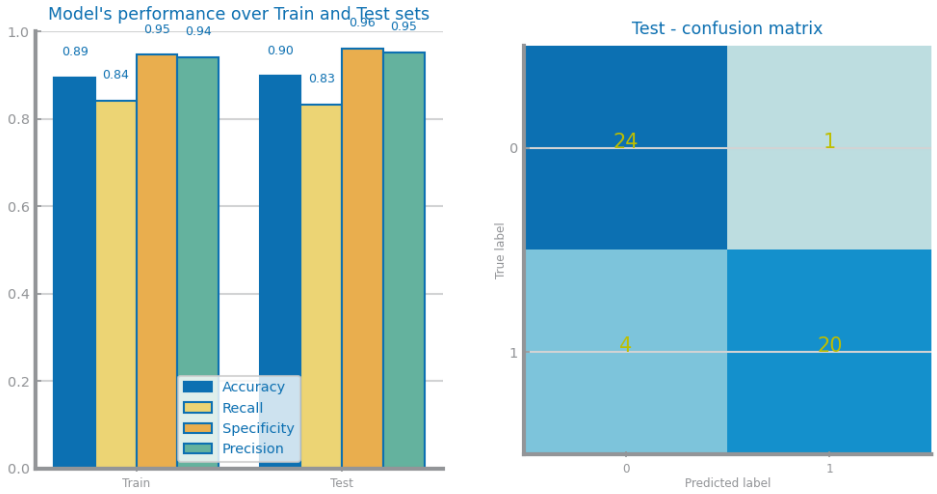


Figure 2 - Model performance for test and train sets and test dataset confusion matrix

3.4 App development

We created a web application based on the python framework DASH, named Prostatic Cancer Surgical Margin Prediction Tool, which can be accessed in limedash.herokuapp.com. The web page is divided in two main sections and one collapsible region.

The first section, on the left, agglomerates the patient information, which is characterized by all the inputs necessary to make a prediction. After the final input is defined, in the right-side section, it can be seen the SHAP or LIME explanation and the correspondent likelihood of the predicted surgical margin.

The SHAP graph indicates the impact (magnitude of increase or decrease in log-odds) of factors on the model prediction of the patient's surgical margin likelihood. Blue bars indicate a decrease and red bars indicate an increase in surgical margin likelihood. The final risk value at the top of the figure is equal to log(p/(1-p)) where p is the predicted likelihood.

The LIME graph indicates the feature importance of for each of the classified class of factors on the model prediction of the patient's surgical margin likelihood. The pie chart on the left represents the most important features for the classification of Surgical Margin as negative and the pie chart on the right for its classification as positive. Since the python LIME package is has a limited number of feature importance graphics, there was created a pie chart for each class that increases in size with its correspondent class (surgical margin 0 or surgical margin 1) predicted likelihood.

The bottom collapsible region, named Predicted Model Information has all the information that led to the creation of the machine learning model, such as the data source, cohort table with the distribution of each variable to both classes and model training information. In this section there are also present the graphs with the model performance for the train and test sets and the confusion matrix for the test set. The feature importance graph demonstrates the nine most important features for the positive surgical margin prediction of the global data, where the red values characterize negative influence and the green one's positive influence.

Three screen captures of the web application can be seen in Appendix 3: An example patient prediction with SHAP explanation, the same prediction with LIME explanation and the model information collapsible region.

4. Discussion

In our review of literature, we found that although explanations are clearly categorized in a generally acceptable manner, categorizations focus mainly on the strategy the methods use to explain the algorithm and are not concerned with the explanation interface, by which final users can perceive it. As such, there is no direct connection between the classification of any one method by scope or model dependency and how good the explanation is according to the four criteria, as although there are some general tendencies that can be observed (Markus, Kors, & Rijnbeek, 2021). For example, SHAP and LIME are classified as post-hoc, global, model-agnostic explanations, which usually satisfy soundness and parsimony, but not completeness and/or clarity (Markus, Kors, & Rijnbeek, 2021).

Depending on the model they are applied to, and the interface built using these explanations, evaluation according to the criteria can vary greatly. The target userbase for the interface is also relevant, as within "medical professionals", there can be different needs. For example, an Emergency Department clinician might consider a longer explanation to be unnecessary complex (not satisfying parsimony), while another in the Intensive Care unit might require a longer more detailed explanation in order to satisfy completeness (Tonekaboni, Joshi, McCradden, & Goldenberg, 2019). For example, in the web application developed, the SHAP waterfall graph chosen was altered to only include three decimal places in the SHAP values of each variable, as upon receiving feedback, more decimal places lead to a decrease in parsimony.

We have also found that most studies which create or explore XAI integration in medicine, fail to adequately explain the feedback received, focussing mainly on model performance and not acceptance and use by the medical professionals its intended for (Antoniadi, et al., 2021).

In the short term, we have approached this lack with the creation of the proposed questionnaire, to be distributed to the medical professionals from Hospital da Luz, in different specialties, so as to properly identify their distinct needs and preferences. In order to prevent confusion while answering and promote full answers (each user answering all questions), we have limited it to the three most common types of explanations, by graphs, images, and text. This purposefully ignores the possibility of combining and presenting them simultaneously, in order to create easily distinguishable options. Following distribution would be a period of answer analysis, from which conclusions could be taken to improve our web application.

At the moment, the final web application created demonstrates the culmination of the procedural process of machine learning model creation followed by the explanation methods' implementation and its integration in the web.

By analysing the model performance for the test and train sets that can be seen in Figure 2 we can conclude that it was possible to avoid overfitting, a process that is extremely common in machine learning models associated with a small sample data, which happens when a model is able to make perfect prediction based on training data, but fails to generalize its predictive power to other sets of data (Nelson, 2020).

The final product presents a responsive software that corresponds to our stipulated goals, since we were able to develop a data input section were the medical staff can clearly insert the patients' information and an explanation section with clear, concise, and comprehensible graphs that demonstrate the influence of each feature in the predicted surgical margin likelihood. Following the same evaluation method applied in the survey, the web application is considered:

Soundness – very good: it gives similar importance to the same variables as the model itself.

Clarity – good: requires further verification with a medical professional. We have verified it provides similar explanations in similar results; however, the opinion of a medical professional is needed to verify.

Completeness – neutral: requires further verification with a medical professional. We consider it provides all information needed to understand the result given, however the opinion of a medical professional is needed to verify.

Parsimony – neutral: requires further verification with a medical professional. We consider it provides concise and adequate information needed to understand the result given, however the opinion of a medical professional is needed to verify.

Even though we think that the web application is the best that could be done within the time frame, there are some limitations associated with its functioning. One of those restraints is the fact that the dataset that the final model was based on is relatively small, which led to need to create synthetic data during the balancing process. Thus, influencing the predicted surgical margin likelihood negatively, when comparing the obtained class with the surgical margin related to real cases of prostate cancer. This limitation can be eliminated in future works by increasing the initial sample size, by not only having a dataset with more than the 121 entries used in this project, but also having extra data samples that could help with the final model verification.

In the short term, we would like to propose a meeting with a specialist in radiomics, in order to present our web application and receive feedback. We would ask questions regarding firstly, an evaluation of the model itself, i.e., whether results are plausible and correct; and secondly an evaluation of the explanation interface, according to the four criteria, but also regarding ease of use, appearance, and other subjective factors.

By analysing the mean SHAP value graph that can be found in Appendix 2 it can be seen that the five most important variables of our global model are "smooth capsular bulging", "capsular contact length", "regra" and "irregular contour". With this in mind we would also like to question an expert in the area if these are the features that should have more impact on the prediction of the surgical margin of a prostatic cancer patient.

In the long term, we propose the introduction of our web application into use as a prototype, and a study to be made of model behaviour. Being as it gives a prediction of whether there will be tumoral growth in the surgical margin of prostate cancer patients, medical professionals could introduce their patients' data and verify results directly. Another study proposed would be a study of hours of use, which could quantify the usefulness and useability of the web application and certify the validity of the criteria used to evaluate it a priori.

5. Conclusion

In conclusion, this project aimed to review XAI methods available and evaluate the best approach to implement the SHAP and LIME methods in healthcare, particularly with the creation an online application with a predictive model for the classification of prostatic cancer surgical margin. As such, after an initial review of literature we chose four criteria – clarity, soundness, parsimony, and completeness – to evaluate explanations, we created a survey for medical professionals to evaluate three different options: explanations by text, graphs, and images. Finally, we developed a classification model with good performance (90% accuracy, 83% recall, 96% specificity, 95% precision) from a database of 121-entries and implemented it with SHAP and LIME explanations into a web application.

Acknowledgments

We would like to thank our mentors Nuno André da Silva, José Moreira, Raquel Santos and Adalgisa Guerra for their insights, support, and guidance during this project. Additionally, we would like to thank Hospital da Luz Learning Health and Instituto Superior Técnico for providing the opportunity to work with them in this integrative project.

References

Aerts, H. J., Velazquez, E. R., Parmar, C., Grossmann, P., Carvalho, S., Lambin, P., & al, e. (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications 5.1*, 1-9. doi:https://doi.org/10.1038/ncomms5006

Antoniadi, A., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B., & Mooney, C. (2021). Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. *Appl. Sci.*, 11. doi:https://doi.org/10.3390/app11115088

de Vaus, D. (2013). *Surveys in Social Research* (6th ed.). Routledge.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery, 9*(4), e1312. doi:https://doi.org/10.1002/widm.1312

IBM, C. E. (2020). *IBM*. Retrieved 07 2022, from https://www.ibm.com/cloud/learn/machine-learning

Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., Van Stiphout, R. G., & al, e. (2012). Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer 48.4*, 441-446. doi:https://doi.org/10.1016/j.ejca.2011.11.036

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems, 30*.

Lundberg, S., Nair, B., Vavilala, M., & al., e. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng, 2*, 749-760. doi:https://doi.org/10.1038/s41551-018-0304-0

Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 103655. doi:https://doi.org/10.1016/j.jbi.2020.103655

Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (2 ed.). Retrieved from https://christophm.github.io/interpretable-ml-book

Nelson, D. (2020). *What is overfitting?* Retrieved 07 14, 2022, from https://www.unite.ai/what-is-overfitting/

Quinn, M. a. (2002). Patterns and trends in prostate cancer incidence, survival, prevalence and mortality. Part I: international comparisons. *BJU international 90.2*, 162-173. doi:https://doi.org/10.1046/j.1464-410X.2002.2822.x

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should {I} Trust You?": Explaining the Predictions of Any Classifier. *CoRR, abs/1602.04938*. doi:https://doi.org/10.48550/arXiv.1602.04938

Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: contextualizing explainable machine learning for clinical end use. In *Proceedings of the 4th Machine Learning for Healthcare Conference* (pp. 359--380). PMLR. Retrieved from https://proceedings.mlr.press/v106/tonekaboni19a.html

WHO, W. H. (2020). *Prostrate Cancer Fact Sheet.* Retrieved from https://gco.iarc.fr/today/data/factsheets/cancers/27-Prostate-fact-sheet.pdf

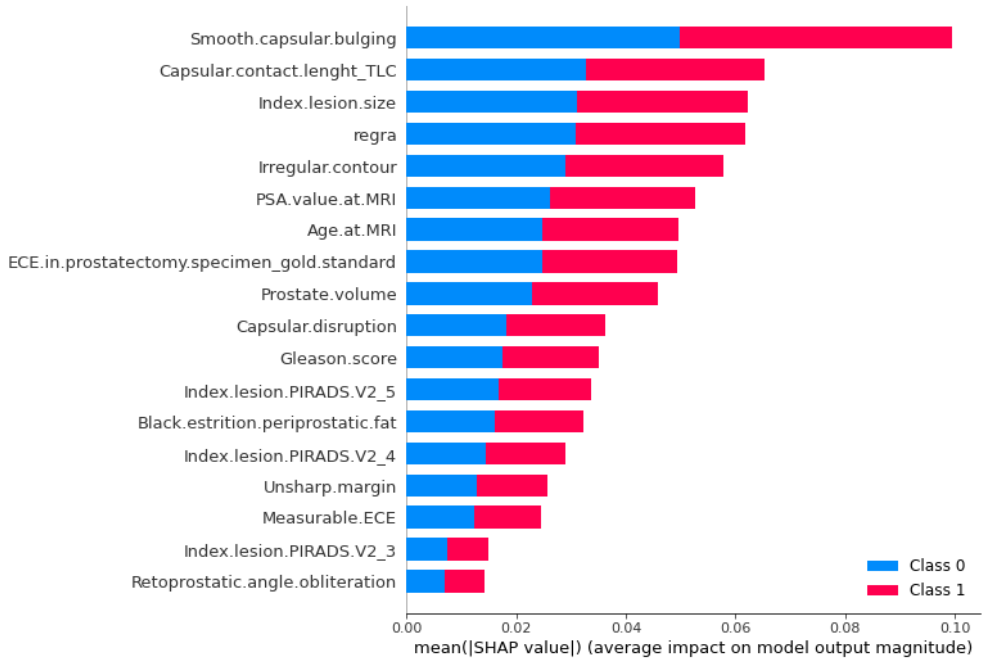
Yip, S. S., & Aerts, H. J. (2016). Applications and limitations of radiomics. *Physics in Medicine & Biology, 61*, 13. doi:https://doi.org/10.1088/0031-9155/61/13/R150

Appendix

Appendix 1 – Cohort Table

		Missing	surgycal margin-0	sugycal margin-1
n			83	38
Age.at.MRI, median [min,max]		0	61.2 [45.4,75.8]	62.1 [45.5,73.8]
Index.lesion.PIRADS.V2, n (%)	3	0	7 (8.4)	2 (5.2)
	4		41 (49.4)	20 (52.6)
	5		35 (42.2)	16 (42.2)
Prostate.volume, median [min,max]		0	43.5 [18.0,148.0]	42.9 [20.0,122.0]
PSA.value.at.MRI, median [min,max]		0	6.8 [2.63,21.2]	7.7 [2.2,20.0]
Index.lesion.size, median [min,max]		0	13.9 [5.0,32.0]	14.8 [7.0,30.0]
Capsular.contact.lenght_TLC, median [min,max]		0	11.9 [0.0,35.0]	14.6 [0.0,40.0]
Smooth.capsular.bulging, n (%)	no	0	38 (45.8)	10 (26.3)
	yes		45 (54.2)	28 (73.7)
Capsular.disruption, n (%)	no	0	45 (54.2)	15 (39.5)
	yes		38 (45.8)	23 (60.5)
Unsharp.margin, n (%)	no	0	39 (47.0)	16 (42.1)
	yes		44 (53.0)	22 (57.9)
Irregular.contour, n (%)	no	0	48 (57.8)	19 (50.0)
	yes		35 (42.2)	19 (50.0)
Black.estrition.periprostatic.fat, n (%)	no	0	65 (79.5)	28 (73.7)
	yes		17 (20.5)	10 (26.3)
Retoprostatic.angle.obliteration, n (%)	no		78 (94.0)	36 (94.7)
	yes		5 (6.0)	2 (5.3)
Measurable.ECE, n (%)	no	0	71 (85.5)	32 (84.2)
	yes		12 (14.5)	6 (15.8)
ECE.in.prostatectomy.specimen_gold.standard, n (%)	no	0	65 (78.3)	23 (60.5)
	yes		18 (21.7)	15 (39.5)
Gleason.score, n (%)	no	0	58 (69.9)	26 (68.4)
	yes		25 (30.1)	12 (31.6)
regra, n (%)	no	6	20 (24.1)	10 (26.3)
	yes		63 (75.9)	22 (57.9)

Appendix 2 – Average feature impact on model output for each class (surgical margin 0 and 1)



Appendix 3 - Web Application with a) SHAP explanation; b) LIME explanation; c) Model Information