Which Explanation Makes Sense? A Critical Evaluation of Local Explanations for Assessing Cervical Cancer Risk Factors

Celia Wafa Ayad WAFA.AYAD@POLYTECHNIQUE.EDU
Ecole polytechnique

Thomas Bonnier THOMAS.BONNIER@SOCGEN.COM
Société générale

Benjamin Bosch BENJAMIN.BOSCH@SOCGEN.COM
Société générale

Jesse Read JESSE.READ@POLYTECHNIQUE.EDU
Ecole polytechnique

Sonali Parbhoo S.PARBHOO@IMPERIAL.AC.UK
Imperial College London

Abstract

Cervical cancer is a life-threatening disease and one of the most prevalent types of cancer affecting women worldwide. Being able to adequately identify and assess factors that elevate risk of cervical cancer is crucial for early detection and treatment. Advances in machine learning have produced new methods for predicting cervical cancer risk, however their complex black-box behaviour remains a key barrier to their adoption in clinical practice. Recently, there has been substantial rise in the development of local explainability techniques aimed at breaking down a model's predictions for particular instances in terms of, for example, meaningful concepts, important features, decision tree or rule-based logic, among others. While these techniques can help users better understand key factors driving a model's decisions in some situations, they may not always be consistent or provide faithful predictions, particularly in applications with heterogeneous outcomes. In this paper, we present a critical analysis of several existing local interpretability methods for explaining risk factors associated with cervical cancer. Our goal is to help clinicians who use AI to better understand which types of explanations to use in particular contexts. We present a framework for studying the quality of different explanations for cervical cancer risk and contextualise how different explanations might be appropriate for different patient scenarios through an empirical analysis. Finally, we provide practical advice for practitioners as to how to use different types of explanations for assessing and determining key factors driving cervical cancer risk.

1. Introduction

Cervical cancer is a dangerous cancer of the uterus affecting women's health worldwide. It is the fourth most prevalent type of cancer in women, with an estimated 604 000 new cases and 342 000 deaths in 2020 alone (Sung et al., 2021). Left undetected and untreated, cervical cancer can result in damage to the tissue of the cervix and can gradually reach other areas of the human body, such as the lungs, liver, and vagina. A few risk factors such as prior exposure to Human Papillovirus (HPV), smoking, weakened immunity and starting sexual activity at a young age, are known to increase the likelihood of developing cervical cancer (Kashyap et al., 2019). Yet there may be many other unknown driving factors that increase a patient's chances of developing cervical cancer. Being able to accurately identify these factors is crucial for early detection and treatment.

Recent advances in AI have contributed to a growing body of research aimed at using machine learning (ML) algorithms for early assessment of cervical cancer risk. Among these, Ratul et al. (2022); Kruczkowski et al. (2022) compare the performances of random forests, deep learning and Naive Bayes for predicting cervical cancer risk, while Mehmood et al. (2021) present an algorithm known as *CervDetect* for feature selection and subsequently use a deep neural network to determine those variables, such as history of STDs and age, most correlated with elevated risk of cervical cancer. These methods, though predictive, exhibit high variance, particularly when applied to heterogeneous patients. Prior research has also proposed several methods (Chadaga et al., 2022; Lee et al., 2021; Curia, 2021; Conceição et al., 2019; Chekin et al., 2022) using ensembles of decision trees and neural networks applied to tabular and image data to predict cervical cancer risk. Though performant, these methods are not interpretable making them difficult to use in practice.

To overcome these issues, Mohanty and Mishra (2022) survey explainability methods to better understand the risk factors that are responsible for the development of certain types of cancer. Unfortunately however, these methods often provide competing explanations and there is little agreement as to which explanation makes sense for a particular context, nor are these accompanied by any uncertainty metric or objectively compared to one another (Krishna et al., 2022; Attanasio et al., 2022; Camburu et al., 2019; Bodria et al., 2021; Neely et al., 2021).

In our work, we review and synthesize properties, desiderata and definitions in the interpretable machine learning literature relevant for assessing cervical cancer risk. We provide a framework for assessing the quality of different explanations for cervical cancer risk and compute different metrics for determining which explanation makes the most sense for cervical cancer risk assessment. In our experiments, we provide, to the best of our knowledge, the first empirical study analysing the performances of different methods for explaining cervical cancer risk factors. For each method, we contextualise how different formulations of these explanations might be appropriate for different patient contexts and when an explainability technique may not be suitable for use. Finally, we provide advice for practitioners as to how to use different types of explanations in practice for assessing and determining key factors driving cervical cancer risk.

Generalizable Insights about Machine Learning in the Context of Healthcare

Explanations and our ability to interpret these explanations is highly dependent on their downstream application task (Doshi-Velez and Kim, 2017; Lage et al., 2018). However while our core focus in this paper is assessing cervical cancer risk using various interpretability methods, some of our contributions will certainly be valuable in other contexts. Specifically:

- Algorithm: We introduce an algorithm for comparing the performances of various interpretability tools, which could be used to assess how explanations may differ across patients, diseases or applications.
- Metrics: We provide an empirical comparison of how explanations may differ in terms of their accuracy, compactness, consistency and robustness using different metrics for each. These metrics may be applicable in other contexts, beyond our cervical cancer risk assessment task.
- Demonstration: We show that our method for analyzing and assessing the performances of various explanations for cervical cancer risk assessment provides a generalizable set of guidelines for clinicians to use to more easily audit a machine learning model's predictions in high-stake settings, while maintaining predictive accuracy.

2. Related Work

A number of ML approaches for assessing cervical cancer risk have been developed. These works are either not interpretable or do not always produce consistent or faithful predictions, particularly in applications with heterogeneous outcomes. Several works on interpretable and explainable ML have also focused on reviewing and characterizing what makes a good explanation in terms of properties and evaluation metrics. However, to the best of our knowledge, there has not been prior work that critically evaluates the quality of these explanations for assessing cervical cancer risk.

Cervical Cancer Risk Assessment Methods Prior research has proposed several approaches (Chadaga et al., 2022; Lee et al., 2021; Curia, 2021; Conceição et al., 2019; Chekin et al., 2022) to train highly performing models in order to accurately predict the cervical cancer disease, using different categories of models and types of data. Unfortunately, not all of these methods are interpretable. Other propositions include simple models such as decision trees and complex black-box models like neural networks and ensemble methods (Wang et al., 2021) which are applied on tabular and image datasets to predict cervical cancer risk. Similarly, Mohanty and Mishra (2022) used explainability methods to better understand the risk factors responsible for development of certain types of cancer. Unfortunately however, these methods often provide competing explanations and there is little agreement as to which explanation makes sense for a particular context. Unlike these, our work provides an empirical analysis comparing different types of local explanations for assessing cervical cancer risk. We provide guidance to clinicians who use AI to better understand which types of explanations are more suitable for cervical cancer risk assessment.

Local Explanations Local explanations provide explanation for *a specific input.* Ribeiro et al. (2016) show that using the weights of a sparse linear model, one can explain the decisions of a black box model in a small area near a fixed data point. Similarly, Singh et al. (2016) and Koh and Liang (2017) output a simple program or an influence function, respectively. Other approaches have used input gradients to characterize local logic (Maaten and Hinton, 2008; Selvaraju et al., 2016). However, such local explanations often do not match with human notions of contexts (Miller, 2018): a user may have difficulty knowing if and when explanations generated locally for input x translate to new inputs x' and research on which local explanations to use in different contexts remains limited. In our work, we empirically assess the properties of local explanations for use when applied to the task of assessing cervical cancer risk, and provide guidance as to which of these explanations may be suitable in different contexts.

Reviews of Explanation Types and Metrics Several review papers e.g. Agarwal et al. (2022); Chen et al. (2022); Zhou et al. (2021) have identified and described important properties and desiderata for explanations. Among these, Zhou et al. (2021) provide an overview of various metrics for evaluating explanation types. Liao et al. (2022) conduct a user survey to build a taxonomy of desired properties of explanations; Vitali (2022) provide a review of evaluation metrics based on how compliant they are with existing laws. Some of these papers focus on characterizing different *types of explanations* (Marcinkevičs and Vogt, 2020). Others such as Chen et al. (2022) provide a survey of *explanation quality* in terms of properties defined in interpretable machine learning papers, synthesizes them based on what they measure, and describe the theoretical trade-offs between different formulations of these properties. Our work is complementary to these works and provides an *empirical evaluation of these explanations*, specifically applying some of the metrics described in Chen et al. (2022) in the context of cervical cancer risk assessment. Unlike Agarwal et al. (2022), we do not use the faithfulness metric to compare explanations produced to ground truth feature importances, since in reality it is implausible to have access to these and we choose not to compute the unfairness metric, since there is increasing evidence that fairness metrics can in fact preserve or even perpetuate bias (e.g. Wachter et al. (2021)) which we want to avoid. Instead, we compute the ROAR metric to measure the impact of removing top features on the model performance.

3. Method: A Local Feature Contribution Assessment Framework

Our goal in this work is to provide a framework for assessing the quality of different explanations for cervical cancer risk and compute different metrics for determining which explanation makes the most sense for cervical cancer risk assessment. We propose a systematic approach for interpreting the predictions of a black-box model using multiple interpretability methods, and compare the explanations based on desired criteria. Our approach consists of three key phases: a) first we train a series of models for cervical cancer risk assessment and choose the best among these models; b) next, we interpret the models from a) using a series of local explainability techniques; c) we compute a series of metrics to assess the plausibility and coherency of each of the explainability methods considered. An overview of this framework is provided in Figure 1. Overall, our framework provides domain experts with a means of understanding not only which factors contribute to patient risk of cervical cancer, but also contextualises when certain types of explanations may be preferrable to others for cervical cancer risk assessment.

3.1. Problem Setup

Given a cohort $X \in \mathbb{R}^{N \times D}$ of N patients with D features, we propose a multistage analysis pipeline. First, we test the performances of supervised learning models $f : \mathbb{R}^{N \times D} \to \{0, 1\}$ for predicting cervical cancer risk. From these models, we find the model f^* that best

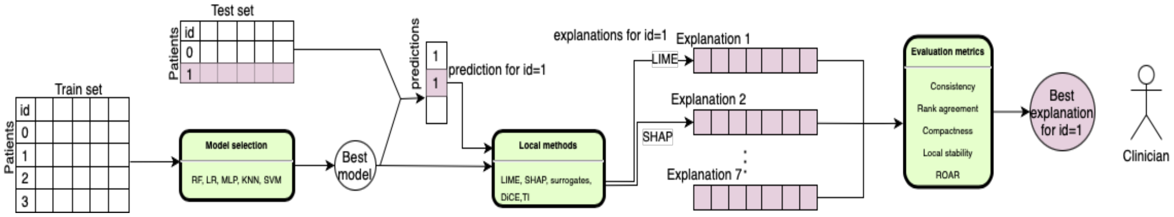

Figure 1: Ilustration of the two stages pipeline. (1) choosing the best model and (2) selecting the best explanation for individual patients.

predicts a patient's risk of cervical cancer based on their data. Assume $\mathcal{L}$ denotes the loss function used to train f, x represent an instance of interest and ψ represent a regularization term added to $\mathcal{L}$ to prevent overfitting.

$$\mathcal{L}(y, f(x)) := -\frac{1}{N}\sum_{i=1}^{N} y_i \log f(x_i) + (1-y_i)\log(1-f(x_i)) + \psi(x_i), \tag{1}$$

where y_i is the actual label (0 or 1) for the ith data point, and $f(x_i)$ is the predicted probability of the positive class for the ith data point. In our work, we consider five different model architectures for f that widely been used in prior literature for cervical cancer risk assessment, namely Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), k-Nearest Neighbors (KNN), and Multilayer Perceptron (MLP). The model that best predicts a patient's risk of cervical cancer is given by:

$$f^*(x_i) = \min_f \mathcal{L}(y_i, f(x_i)). \tag{2}$$

3.2. Generating Local Explanations

Next, we use the predictions of the chosen model to generate explanations for each patient using existing local explainability techniques. Our objective is to assess the quality of explanations for the predictions made by f^*. Let g represent an interpretable model used to produce local explanations of the predictions of f^*. The types of local explanations we consider in this paper are detailed in the subsections below. We focus on these methods as they are most widely used across several healthcare applications.

Local Interpretable Model-Agnostic Explanations. LIME is a model-agnostic technique (Ribeiro et al., 2016) for explaining the predictions of machine learning models on a local, instance-specific basis. LIME approximates the predictions of $f^*(x)$ by with a simpler $g(x)$ for an instance of interest x. The approximation is done using weighted linear regression, where the weights are determined based on the proximity of training samples to x, by minimizing:

$$\arg\min_g \mathcal{L}_{\pi_x}(f^*, g) + \Omega(g) \tag{3}$$

where $\mathcal{L}_{\pi_x}(f^*, g)$ measures the discrepancy between the predictions of $f^*(x)$ and the interpretable model $g(x)$ on a local neighborhood of x, π determines the weights for the training samples, and $\Omega(g)$ promotes simpler and more interpretable models. The explanation consists of the feature importance values derived from the model coefficients of g. In our work g is a linear regression.

SHAP. SHAP (Lundberg and Lee, 2017) values explain the output of any machine learning model by attributing a value to each feature for a prediction based on their contributions towards that prediction. This is done by considering all possible feature combinations. The general formula for computing SHAP values is based on the Shapley value concept and can be expressed as:

$$\phi_i \;=\; \sum_{S \subseteq D \setminus \{i\}} \frac{|S|!\;(|D|-|S|-1)!}{|D|!} f^*_{x \in \mathbb{R}^D}(x) \;-\; f^*_{x \in \mathbb{R}^S}(x)$$

where S is a subset of features D excluding feature d, and $f^*_{x \in \mathbb{R}^S}$ and $f^*_{x \in \mathbb{R}^D}$ are the predictions of the black-box model for subsets S and S with feature d included, respectively. Variants of SHAP include kernel SHAP, tree and deep SHAP, depending on the model f^* being explained.

Diverse Counterfactual Explanations. DICE (Mothilal et al., 2019) is a model-agnostic method for generating diverse and interpretable counterfactual explanations for individual predictions. DICE finds instances similar to original instance x, but with different predicted outcomes. Optimization requires minimizing a distance metric between the counterfactuals and x, subject to constraints that ensure dissimilarity among generated counterfactuals. Counterfactuals are generated by perturbing the features of x while staying within the feasible range of feature values. The optimization problem can be formulated as:

$$\min_{x_{c_i} \in \mathcal{X}} \delta(x_{c_i}, x_i) \in C(x_{c_i}) = c_i,\; x_{c_i} \neq x_i$$

where x_{c_i} is a counterfactual instance, $\mathcal{X}$ is the feasible range of feature values, $\delta(x_{c_i}, x_i)$ is a distance metric between the counterfactual and the original instance, $C(x_{c_i})$ is a constraint function that enforces the counterfactual to have a desired predicted outcome c_i, and $x_{c_i} \neq x_i$ ensures that the counterfactual is different from the original instance.

Tree Interpreter. Tree Interpreter (Li et al., 2019) is a model-specific method for interpreting predictions of tree-based models, such as random forests. It provides a way to attribute feature importance values for predictions made by tree-based models, by tracing the decision path of an instance through the tree and measuring the contribution of each feature towards the prediction. This is done by summing the changes in prediction associated with each decision node along the path, weighted by the proportion of instances that pass through each decision node. The contribution of feature i for a specific instance x can be computed as:

$$Contribution_i = \sum_{d=1}^{M} w_d \cdot \mathbb{I}(x \in R_d) \cdot \Delta p_d,$$

where R_d corresponds to the region associated with decision node d, Δp_d denotes the change in predicted probability associated with d, M is the number of decision nodes in the tree or ensemble, w_d denotes the weight associated with decision node d in the tree or ensemble and $\mathbb{I}$ represents the indicator function.

Local Surrogates. Local Surrogates (Molnar, 2022) are model-specific methods for interpreting predictions of machine learning models that aim to provide insights into the decision-making process of the black-box model for a specific prediction, by fitting a simpler model, such as linear regression or decision tree, using the training data in the local neighborhood of the instance of interest. Local surrogates generate explanations in the form of interpretable models or feature importance values, depending on the specific method used.

3.3. Evaluation Metrics

Our objective is to assess the quality of each of the explanation techniques described earlier for the task of cervical cancer risk prediction. The quality of an explanation is determined by several desiderata quantified using the following metrics, based on those in Chen et al. (2022). Note that although these metrics are not necessarily specific to cervical cancer, there are several works from medicine which demonstrate these properties of explanations may be effective for AI in healthcare (see for instance, Jung et al. (2023); Amann et al. (2020)).

The Consistency Metric compares the explanation attributed to the same instance or set of instances by different feature importance-based explainability methods. This is done through computing an L2 distance between pairs of explanations (Slack et al., 2021). If two different methods attribute similar feature importance to the same instance or set of instances, the user's confidence and trust in the model's predictions increases.

The Stability Metric compares the explanation given to an instance in its neighborhood, meaning that if two instances have similar feature values and predictions, they should be attributed similar explanations (Petsiuk et al., 2018). Two instances having similar feature spaces and different predictions (those are in the boundary of the decision) will not necessary have the same explanation, especially that most of the explainability methods compute feature explanations from the model's predictions. In order to perform the stability analysis, we use the local Lipschitz metric for explanation stability (Alvarez-Melis and Jaakkola, 2018).

$$\hat{L}(x_i) = \underset{x_j \in B_\epsilon(x_i)}{\operatorname{argmax}} \frac{\|\phi(x_i) - \phi(x_j)\|_2}{\|x_i - x_j\|_2} \tag{4}$$

where x_i refers to an instance, $B_\epsilon(x_i)$ is the ϵ-sphere centered at x_i, and $\phi(x_i)$ and $\phi(x_j)$ are the explanation parameters for x_i and x_j. Lower values indicate more stable explanations.

The Compactness metric measures how many features are needed to reach a certain percentage of the prediction. This can be obtained by fixing the percentage we want to reach and compute the number of features needed to explain that fixed percentage of the prediction.

The Faithfulness Metric: RemOve And Retrain (ROAR) (Hooker et al., 2018) is a machine learning interpretability metric that involves iteratively removing a subset of features from a dataset, retraining the model on the reduced dataset, and then evaluating the changes in model accuracy or feature importance. Agarwal et al. (2022) employ a similar faithfulness metric but assume access to the ground truth feature importances. In practice, it may be difficult to have access to these. Instead, we consider faithfulness in terms of feature and rank agreement to measure the similarity between each pair of methods in terms of rank and sign. Other measures of faithfulness from Agarwal et al. (2022) can be viewed as variants of these.

A number of works deem stability, consistency, compactness and faithfulness as important facets of interpretability for healthcare domains overall. These include Amann et al. (2020) and Jung et al. (2023). Note that these quality metrics and explanations are not meant as a replacement for clinical expertise. We believe combining domain expertise and these explanations and performing external validation on another dataset is necessary to deduce context-specific explanations that are grounded in clinical utility, which is the focus of future work.

3.4. Algorithm for Assessing Local Feature Contribution

We summarize our approach for assessing the quality of different explanations in Algorithm 1. We start by cleaning and balancing the dataset, then we test existing supervised machine learning models and select Random Forest as it is the most performing one in terms of AUC. Next, we use the selected explainability methods in order to explain the predictions attributed to each patient in the dataset. Finally, we choose the method that satisfies the desired properties that are fixed by the clinician. Code is available at [1].

Algorithm 1 A Local Feature Contribution Assessment Framework

```
Data: $\mathbb{D} = \{(X, Y)\}, X \in \mathbb{R}^{NxD}, Y \in \{0, 1\}, f \in F, g \in G, m \in M, w \in W$
Result: $\theta^*$
$\theta \leftarrow \emptyset$
$\mathbb{D} \leftarrow ADASYN(\mathbb{D})$                                  ▷ Balance the dataset.
$X_{train}, Y_{train}, X_{test}, Y_{test} \leftarrow split(\mathbb{D})$
while $f$ in $F$ do
    $f(x) = \min \mathcal{L}(y, f(x))$                          ▷ Learn the best model.
end
$f^* \leftarrow \min_f \mathcal{L}(y, f(x))$                    ▷ Generate plausible local explanations.
while $g \in G$ do
    $\theta \leftarrow +feature\_importance(g, X_{test})$
end
while $m \in M$ do
    $scores \leftarrow evaluate(\theta, m)$     ▷ Evaluate each explainability method with each metric.
end
$\theta^* \leftarrow \text{argmax}(\sum_{i=1}^{M} score_i * w_i)$        ▷ $W$ are defined by the clinician for the desired
    explanation properties.
```

[1] https://github.com/cwayad/Local-Explanations-for-Cervical-Cancer

4. Cervical Cancer Risk Assessment

The following section provides details of our cohort selection and data processing for predicting and assessing the quality of explanations for cervical cancer risk.

Cohort Selection. We use Cervical cancer risk factors dataset X from the UCI repository (Fernandes et al., 2017). This data contains 858 female patients characterized by $D = 35$ features including demographic information such as age and number of pregnancies, clinical tests such as Hinselmann, Schiller and Citology, many Sexually Transmitted Diseases such as HPV and AIDS, and diagnosis taken by the patients such as HPV and CIN. A summary of cohort characteristics and demographics can be found in Table 1.

Age	No. of sexual partners	First sexual intercourse	No. of Pregnancies	Smokes	Cancer Diagnosis	Contraceptives	Total STDs	No. of Tests
20's	2.54	17.01	2.10	0.14	0.03	0.72	0.14	0.24
30's	2.67	18.09	2.83	0.17	0.04	0.75	0.15	0.26
40's	2.50	18.59	3.29	0.07	0.05	0.68	0.23	0.27
50's	2.80	16.40	4.80	0.40	0.20	0.40	0.00	1.00
70+	2.50	19.75	7.25	0.50	0.00	0.00	0.00	0.25
Teen	2.25	15.08	1.40	0.11	0.01	0.56	0.18	0.20

Table 1: A summary of cohort characteristics and demographics based on age.

We also contextualise the results we obtain by examining different patient instances from the cohort. The summary statistics of these patients relative to the population mean are provided in Table 3 in Appendix B.

Data Processing. To conduct our experiments, we impute missing values and generate synthetic additional samples for class 1 (presence of cancer) using the ADASYN (Adaptive Synthetic Sampling) technique in order to balance balanced the dataset by oversampling the minority class. After preprocessing with ADASYN, the new dataset contains 1677 patients. We split the balanced dataset into 80% for training and 20% for testing, ensuring an unbiased evaluation of our models.

Model selection for f^*. f is trained to predict risk of cervical cancer y from X. For our experiments, we trained five different models for f namely LR, RF, SVM, KNN and MLP. Specifically, each of the models was trained using 10-fold cross validation and the parameters for each were selected by conducting a grid search over parameters and choosing those values that produce the best accuracy. f^* was subsequently obtained by selecting the model minimizing the loss in Eqn 1. For our experiments this was a RF model. These results are consistent with prior studies that showed RFs as one of the top-performing ML models used for predicting cervical cancer risk.

Local Explanation Generation. We generated feature importance explanations using LIME, three variants of SHAP (Tree-SHAP (TSHAP), Kernel-SHAP (KSHAP) and Sampling-SHAP (SSHAP)), Tree Interpreter, DICE and Local Surrogates. These provide local explanations for individual instances, highlighting the contribution of each feature towards the model's prediction. We apply these methods to the test set.

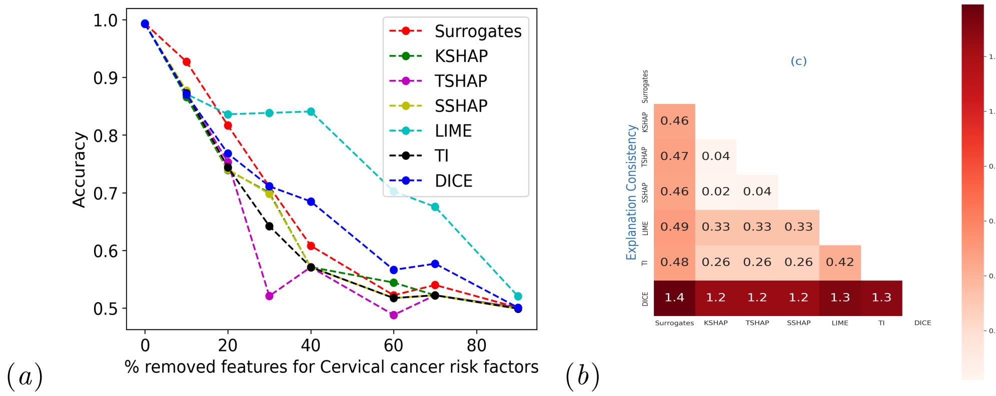

Figure 2: (a) The decrease in model's accuracy after removing a % from the top features and model retraining on the new feature set using ROAR. The feature rankings are taken from the mean feature contributions that have been computed locally. Removing the top 30% most important features given by Tree SHAP decreases the accuracy of the RF by 50%. On the other hand, removing between 30% and 50% of the feature ranking provided by LIME doesn't affect the model's accuracy, making it the model with the most faithful explanations across the cohort.(b) For each pair of methods, Consistency calculates the distance between the contributions for all instances using L_2 norm. Tree SHAP and Sampling SHAP is the most consistent pair, while Local Surrogate and DiCE is the least consistent pair.

5. Results

We first compared the quality of each of the local explanation techniques when applied to f^* and examined the top features produced by each of these techniques. This gives us those features that best explain the optimal model f^*. Next, we measure the decrease in accuracy of f^* when we successively removed a fraction of the top features for each explanation. A summary of these results across the test set can be found in Figure 2. Overall, we see that LIME is the most robust to removal of features, while the model accuracy of all other explanations drops significantly after removing the top features.

The top 30% important feature given by TreeSHAP have the most impact on model learning. We observe very different accuracy drop after removing 30% of the features and model retraining. Indeed, training the model without the top 30% of features given by TreeSHAP drops the model's accuracy to 53%, meaning that those features have the most significant impact on the model learning. Similar accuracy drops can be seen for other variants of SHAP. This is because models trained with SHAP predominantly rely on one feature for predicting cervical cancer risk namely prior HPV infection. Dropping the same percentage of top features given by LIME will only decrease the accuracy by 15%, making LIME the model that produces the most faithful explanations. Unlike SHAP methods, DICE and Surrogates, LIME makes use of multiple features to produce a model explanation. Here, dropping the top features do not drastically decrease the model's accuracy as other features may still be predictive. Table 2 further describes how many features each explanation uses to produce a model with 90% accuracy, as well as the mean stability and coherency of these explanations.

Methods	# Features for 90% Accuracy	Accuracy with 5 features(%)	Mean Stability	Mean Consistency
SSHAP	**1**	12	0.87	**0.34**
TSHAP	**1**	12	0.36	**0.34**
KSHAP	**1**	12	1.44	**0.34**
LIME	**1**	45	0.66	0.46
TI	**1**	19	0.59	0.42
DiCE	9	**100**	1.66	1.11
Local Surrogates	3	**100**	**0.22**	0.54

Table 2: Compactness, stability and consistency of local explainability methods for predicting cervical cancer risk. Some methods predominantly require only one feature to achieve 90% prediction accuracy. Local surrogates have the highest mean stability, while SHAP variants have the highest mean consistency.

All explainers agree on HPV being the most important risk factor for cervical cancer. Next, we examined the feature contributions produced by each explanation technique for each of the patients from Table 3 in Appendix B. The results for Patient 0 and 1 are shown in Figure 3. We see that for Patient 1, all explainers identify prior incidence of HPV as a predominant determinant of cervical cancer. In contrast, all explainers identify use of hormonal contraception and IUD as the third driving risk factors in determining cervical cancer. Local surrogates on the other hand, uses smoking, and age as key risk factors.

These explanations, though different in terms of the risk factors used, are all plausible explanations for cervical cancer assessment and are consistent with existing literature. We compared these results to those of Patient 0 from Table 3 in Appendix B, who is diagnosed as not having cancer. These results are shown in the right sub-figures in Figure 3. Notably the first two driving factors in both cases are similar. Interestingly, except for Local surrogates, all explainers show that starting sex intercourse after 18-years old may help prevent from being diagnosed with cervical cancer. Finally, SHAP explainers indentify contraceptives (hormonal and IUD) may lead cervical cancer. In contrast, LIME and Tree Interpreter are unsure of its positive or negative impact on cervical cancer.

SHAP explainers have exactly the same top 10 most important features for Patient 1. Next we examined the explanations produced by each technique in terms of feature and rank agreements for the same patients: (a) Patient 0 and (b) Patient 1. These results are shown in Figure 4. We observe similar top features given by SHAP explainers for Patient 1, and comparable feature ranking between TSHAP, KSHAP and Tree Interpreter for Patient 0. On the other hand, DiCE is have the most distinct feature and rankings among the explainers, which can be justified by the unsigned nature of feature importance identified by DiCE.

The most unstable explainers are those that depend on creating local neighborhoods. We also compared each explanation in terms of its local stability and compactness.

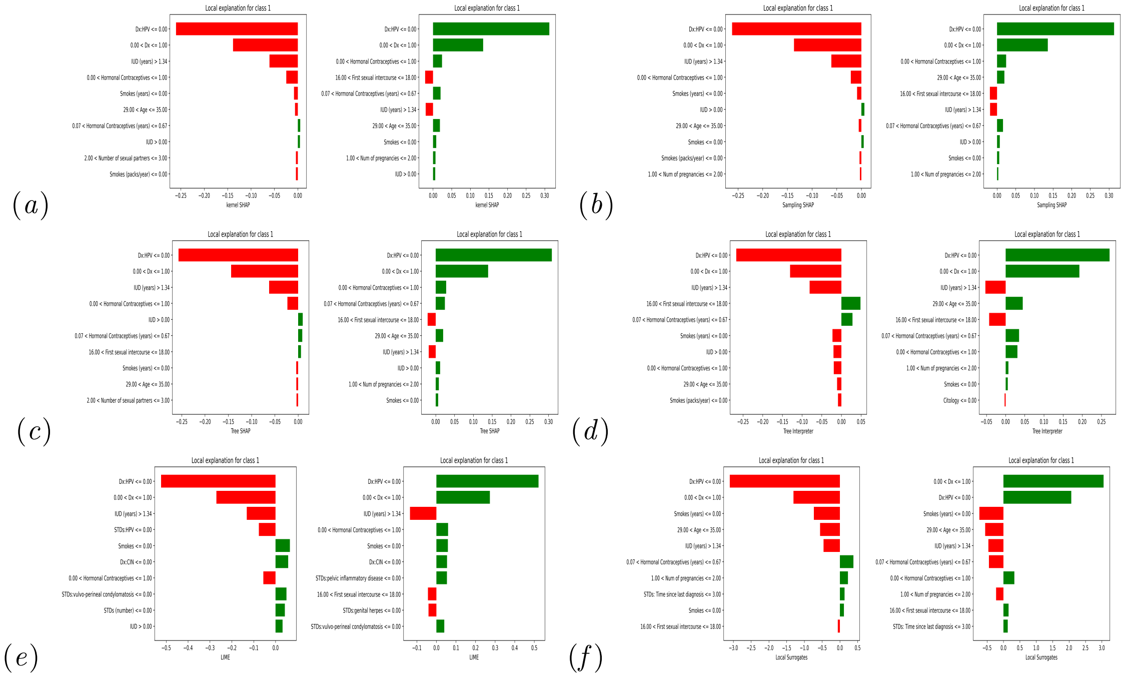

Figure 3: Comparison of feature importance for two similar patients (Patient 0 and Patient 1). On the left, Patient 0 diagnosed as not having cancer (Dx:Cancer=0) and on the right, Patient 1 diagnosed with cancer (Dx:Cancer=1). The key driving factors for this patient are similar to those of a patient with cancer.

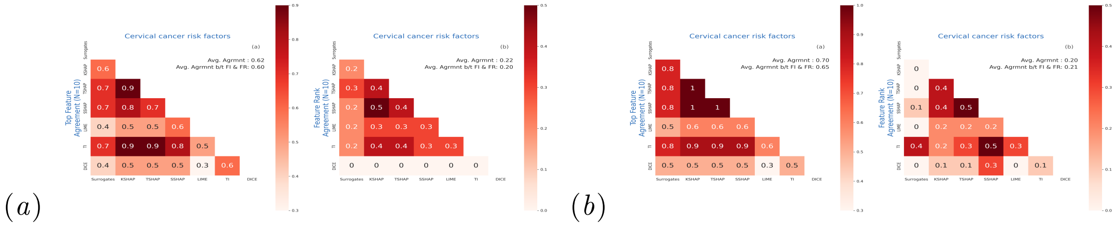

Figure 4: Feature and rank agreements for the Patients 0 and 1. For each patient in subfigures, (a) Feature agreement measures the fraction of common features between the sets of top-10 features of each pair of the rankings, and (b) Rank agreement checks that the feature order is comparable between each pair of the rankings. Tree SHAP and Kernel SHAP have the highest feature and rank agreements for the first patient, while DiCE and Local surrogates have the least feature and rank agreements.

These results are shown in Figure 7 in Appendix A and Figure 5 respectively. Figure 7 in Appendix A. shows the stability in the neighborhood of five features, namely: age, smokes, HPV, first sexual intercourse, and IUD. We observe that LIME, KSHAP and Local surro-

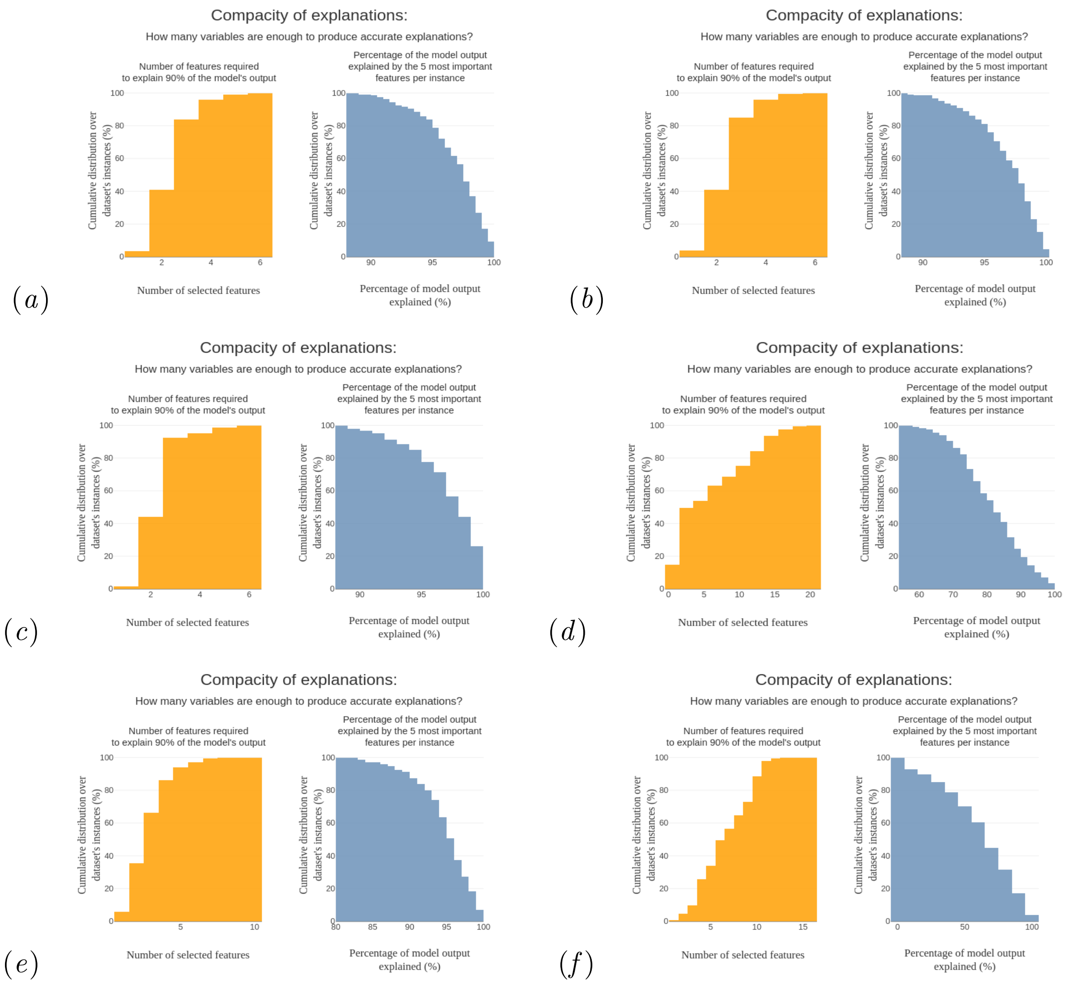

Figure 5: Compactness of the explanations generated by (a) Kernel SHAP, (b) Sampling SHAP, (c) Tree SHAP, (d) LIME,(e) Tree Interpreter, (f) Local surrogates.

gates are the least stable explainers as their feature importance for the given features vary between negative and positive values. LIME is the less varying compared to the other two surrogates.

Local surrogates and DiCE approximate 100% model's output with 5 features. Figure 5 shows the complexity of the selected local methods in terms of how many features are needed in order to explain 90% of the accuracy of the model and how much accuracy achieved with fewer features, here 5. Local surrogates and DiCE need respectively 3 and 9 features in order to achieve 90% of model accuracy, while the other explainers can approximate it with only 1 feature on average. While Local surrogates and DiCE are the only two features that can approximate full model accuracy with 5 features, Tree Interpreter achieves only 12 % of the model accuracy with 5 features.

No single explanation performs optimally across patients and metrics. Local explanation methods perform differently across different patients. Eg, for high risk patients (Cohen et al., 2019; international collaboration of Epidemiological Studies of Cervical Cancer, 2007). Counterfactuals and Local Surrogates give more compact explanations compared to other methods. LIME is the most stable and SHAP the most consistent in terms of feature and rank agreements. Notably, local explanations are not meant to replace insights from aggregation. Aggregation enables showing the tendencies and average importance of the features for a global understanding of cervical cancer, but for personalized treatment, using a local approach is preferable to find those risk factors most likely to affect patients individually. E.g. some global risk factors for cervical cancer include exposure to herpes and immune system deficiency (international collaboration of Epidemiological Studies of Cervical Cancer, 2007) . Yet for Patient 1, we see contraception may play a role, which is not always the case for other patients in the cohort (see appendix for details)

6. Discussion

In this paper, we presented a framework to compare local feature attribution methods in order to identify key risk factors causing cervical cancer for individual patients and demonstrated on two patients having very similar characteristic but different predictions (cancer and no cancer). Overall we observe that though the explainers may agree on the importance of some features, there is no single explanation or technique that performs optimally across all metrics of consistency, compactness, stability, faithfulness and accuracy. Rather, a clinician may choose an explanation method based on the context or choose to compute a weighted sum of these metrics. The best explanation would then correspond to the method producing the highest cumulative score overall, but the weights of each metric in the combination should be left to clinical experts to choose. Local explanations are also not meant to replace insights from aggregation, but may be preferable to determine those risk factors most likely to affect patients individually.

Regardless of the nature of the SHAP used, one obtains very similar feature importance. However, SHAP values are all largely determined by one predominant driving risk factor namely prior HPV infection and explanation quality significantly drops if this feature is not available. Methods such as Local surrogates, LIME and Kernel SHAP learn local interpretable models in the neighborhood of the instance we desire to explain, however can be sensitive to the choice of neighbourhood for two instances with the same characteristics and predictions. Clinicians should exercise caution with these methods unless they have experience in identifying which groups a patient may be most similar to. If global stability is desirable, local surrogates may be the most suitable method to use.

Yet, if a clinician is treating an older patient at higher risk of developing cervical cancer, they may desire explanations that are more compact and stable to isolate factors chiefly responsible for risk to develop a more focused treatment plan. If however, a clinician is treating a patient with many other comorbidities, it might be more useful to have a less compact explanation to be able to view the impact all comorbidities may have on overall risk. Counterfactual generation may be suitable when a patient has a genetic predisposition to a cervical cancer and wants to reason about possibilities under which they may be at higher risk of getting the disease by isolating the most important features from an initial set, or reason about alternative ways to reduce this risk.

Limitations Firstly, the manner in which local explanations are aggregated to draw conclusions on the decrease in model accuracy after removing top features needs careful consideration. Currently, using the mean of local explanations may inadvertently assign higher importance to irrelevant features, resulting in issues with feature ranking during each step of feature removal. To mitigate this, a new aggregation method can be developed that does not adversely affect the aggregation of explanations from each method used in the evaluation of explanations. Additionally, the use of generated additional patients to balance the dataset raises concerns about whether these generated patients deviate from the original distribution, which could impact the correctness of the explanations. Future research can explore the use of additional data or take precautions when generating new instances to ensure they align with the original distribution. Moreover, the inclusion of only seven local methods that are compatible with random forest may limit the comprehensiveness of the approach. Future work can expand the set of methods used to include as many relevant methods as possible, thereby enhancing the robustness and applicability of the framework.

7. Conclusion

While cervical cancer remains a devastating disease for women's health worldwide, machine learning shows promise as an effective tool for early detection and treatment. This is especially crucial as clinicians strive to accurately identify the root causes of the disease and prevent its onset, rather than simply treating it after it has occurred. In this article, we presented a framework and demonstrated its application on a cervical cancer dataset. Our approach allows selecting the suitable explanations to reason about each patient's risk of developing cervical cancer, while satisfying several desired explanatory properties. There are many potential avenues for future research, such as extending our framework to other healthcare applications and areas where explanation is needed. This could open up new possibilities for leveraging machine learning in general and explainability methods in particular to understand patient outcomes and address critical healthcare challenges. Finally, future work could perform a user study with clinicians to assess how different weighted sums may lead to context-specific explanations.

References

Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. Openxai: Towards a transparent evaluation of model explanations. *Advances in Neural Information Processing Systems*, 35:15784–15799, 2022.

David Alvarez-Melis and Tommi S. Jaakkola. On the Robustness of Interpretability Methods, June 2018.

Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, and Vince I Madai. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making*, 20(1):1–9, 2020.

Giuseppe Attanasio, Eliana Pastor, Chiara Di Bonaventura, and Debora Nozza. Ferret: A Framework for Benchmarking Explainers on Transformers, August 2022.

Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. Benchmarking and Survey of Explanation Methods for Black Box Models, February 2021.

Oana-Maria Camburu, Eleonora Giunchiglia, Jakob Foerster, Thomas Lukasiewicz, and Phil Blunsom. Can I Trust the Explainer? Verifying Post-hoc Explanatory Methods, December 2019.

Krishnaraj Chadaga, Srikanth Prabhu, Niranjana Sampathila, Rajagopala Chadaga, Swathi KS, and Saptarshi Sengupta. Predicting cervical cancer biopsy results using demographic and epidemiological parameters: a custom stacked ensemble machine learning approach. *Cogent Engineering*, 9(1):2143040, 2022.

Nasrin Chekin, Haleh Ayatollahi, and Mojgan Karimi Zarchi. A Clinical Decision Support System for Assessing the Risk of Cervical Cancer: Development and Evaluation Study. *JMIR Medical Informatics*, 10(6):e34753, June 2022. ISSN 2291-9694.

Zixi Chen, Varshini Subhash, Marton Havasi, Weiwei Pan, and Finale Doshi-Velez. Does the explanation satisfy your needs?: A unified view of properties of explanations. *arXiv preprint arXiv:2211.05667*, 2022.

Paul A Cohen, Anjua Jhingran, Ana Oaknin, and Lynette Denny. Cervical cancer. *The Lancet*, 393(10167):169–182, 2019.

Teresa Conceição, Cristiana Braga, Luís Rosado, and Maria João M. Vasconcelos. A Review of Computational Methods for Cervical Cells Segmentation and Abnormality Classification. *International Journal of Molecular Sciences*, 20(20):5114, October 2019. ISSN 1422-0067.

Francesco Curia. Cervical cancer risk prediction with robust ensemble and explainable black boxes method. *Health and Technology*, 11(4):875–885, July 2021. ISSN 2190-7196.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Kelwin Fernandes, Jaime S Cardoso, and Jessica Fernandes. Transfer learning with partial observability applied to cervical cancer screening. In *Pattern Recognition and Image Analysis: 8th Iberian Conference, IbPRIA 2017, Faro, Portugal, June 20-23, 2017, Proceedings 8*, pages 243–250. Springer, 2017.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A Benchmark for Interpretability Methods in Deep Neural Networks, June 2018.

international collaboration of Epidemiological Studies of Cervical Cancer. Comparison of risk factors for invasive squamous cell carcinoma and adenocarcinoma of the cervix: collaborative reanalysis of individual data on 8,097 women with squamous cell carcinoma and 1,374 women with adenocarcinoma from 12 epidemiological studies. *International journal of cancer*, 120(4):885–891, 2007.

Jinsun Jung, Hyungbok Lee, Hyunggu Jung, and Hyeoneui Kim. Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. *Heliyon*, 2023.

Nainakshi Kashyap, Nadiya Krishnan, Sukhpal Kaur, and Sandhya Ghai. Risk factors of cervical cancer: a case-control study. *Asia-Pacific journal of oncology nursing*, 6(3): 308–314, 2019.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. *arXiv preprint arXiv:1703.04730*, 2017.

Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective, February 2022.

Michał Kruczkowski, Anna Drabik-Kruczkowska, Anna Marciniak, Martyna Tarczewska, Monika Kosowska, and Małgorzata Szczerska. Predictions of cervical cancer identification by photonic method combined with machine learning. *Scientific Reports*, 12(1):3762, 2022.

Isaac Lage, Andrew Ross, Samuel J Gershman, Been Kim, and Finale Doshi-Velez. Human-in-the-loop interpretability prior. *Advances in neural information processing systems*, 31, 2018.

Eunsaem Lee, Se Young Jung, Hyung Ju Hwang, and Jaewoo Jung. Patient-Level Cancer Prediction Models From a Nationwide Patient Cohort: Model Development and Validation. *JMIR Medical Informatics*, 9(8):e29807, August 2021.

Xiao Li, Yu Wang, Sumanta Basu, Karl Kumbier, and Bin Yu. A Debiased MDI Feature Importance Measure for Random Forests. *arXiv:1906.10845 [cs, stat]*, jun 2019. URL http://arxiv.org/abs/1906.10845. arXiv: 1906.10845.

Q Vera Liao, Yunfeng Zhang, Ronny Luss, Finale Doshi-Velez, and Amit Dhurandhar. Connecting algorithmic research and usage contexts: A perspective of contextualized evaluation for explainable ai. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 147–159, 2022.

Scott Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. *arXiv:1705.07874 [cs, stat]*, November 2017.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Ričards Marcinkevičs and Julia E Vogt. Interpretability and explainability: A machine learning zoo mini-tour. *arXiv preprint arXiv:2012.01805*, 2020.

Mavra Mehmood, Muhammad Rizwan, Michal Gregus ml, and Sidra Abbas. Machine learning assisted cervical cancer detection. *Frontiers in public health*, 9:788376, 2021.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 2018.

Aryan Mohanty and Sushruta Mishra. A Comprehensive Study of Explainable Artificial Intelligence in Healthcare. In Sushruta Mishra, Hrudaya Kumar Tripathy, Pradeep Mallick, and Khaled Shaalan, editors, *Augmented Intelligence in Healthcare: A Pragmatic and Integrated Analysis*, Studies in Computational Intelligence, pages 475–502. Springer Nature, Singapore, 2022. ISBN 978-981-19107-6-0.

Christoph Molnar. *Interpretable Machine Learning*. 2 edition, 2022. URL https://christophm.github.io/interpretable-ml-book.

Ramaravind Kommiya Mothilal, Amit Sharma, and Chenhao Tan. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations, December 2019.

Michael Neely, Stefan F. Schouten, Maurits J. R. Bleeker, and Ana Lucic. Order in the Court: Explainable AI Methods Prone to Disagreement, July 2021.

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized Input Sampling for Explanation of Black-box Models, September 2018.

Ishrak Jahan Ratul, Abdullah Al-Monsur, Bushra Tabassum, Abrar Mohammad Ar-Rafi, Mirza Muntasir Nishat, and Fahim Faisal. Early risk prediction of cervical cancer: A machine learning approach. In *2022 19th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pages 1–4. IEEE, 2022.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016.

Sameer Singh, Marco Tulio Ribeiro, and Carlos Guestrin. Programs as black-box explanations. *arXiv preprint arXiv:1611.07579*, 2016.

Dylan Slack, Sophie Hilgard, Sameer Singh, and Himabindu Lakkaraju. Reliable Post hoc Explanations: Modeling Uncertainty in Explainability, November 2021.

Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

Fabio et al. Vitali. A survey on methods and metrics for the assessment of explainability under the proposed ai act. In *Legal Knowledge and Information Systems: JURIX 2021:*

The Thirty-fourth Annual Conference, Vilnius, Lithuania, 8-10 December 2021, volume 346, page 235. IOS Press, 2022.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Bias Preservation in Machine Learning: The Legality of Fairness Metrics under EU Non-Discrimination Law. 123:735, 2021.

Ruiqi Wang, Mohammad Ali Armin, Simon Denman, Lars Petersson, and David Ahmedt-Aristizabal. Towards interpretable attention networks for cervical cancer analysis. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3613–3616, 2021. doi: 10.1109/EMBC46164.2021.9629604.

Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593, 2021.

Appendix A. Contribution of the Data Points to the Prediction Making.

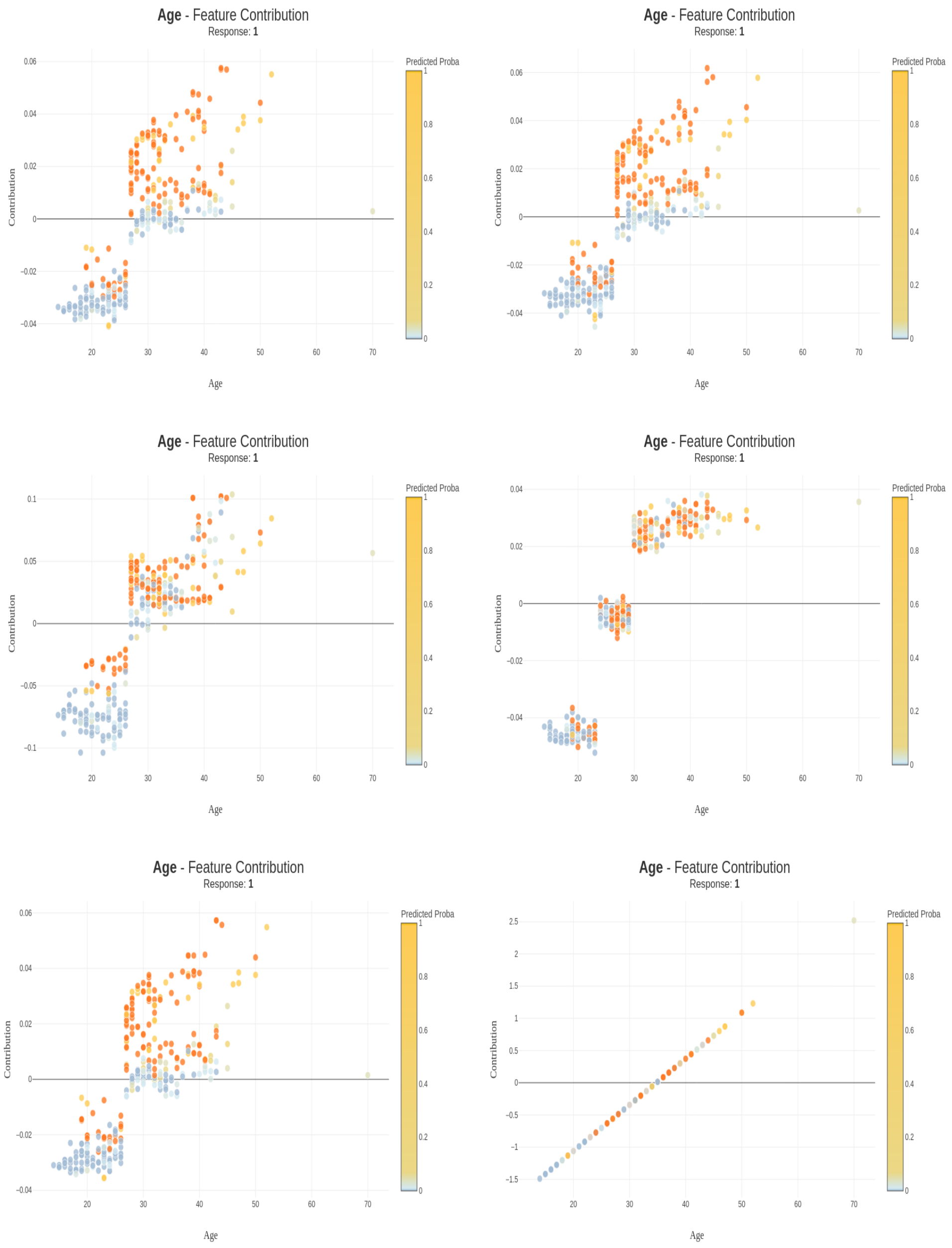

Figure 6: Contribution of the Age of all patients to the class 1 (diagnosed with Cancer).

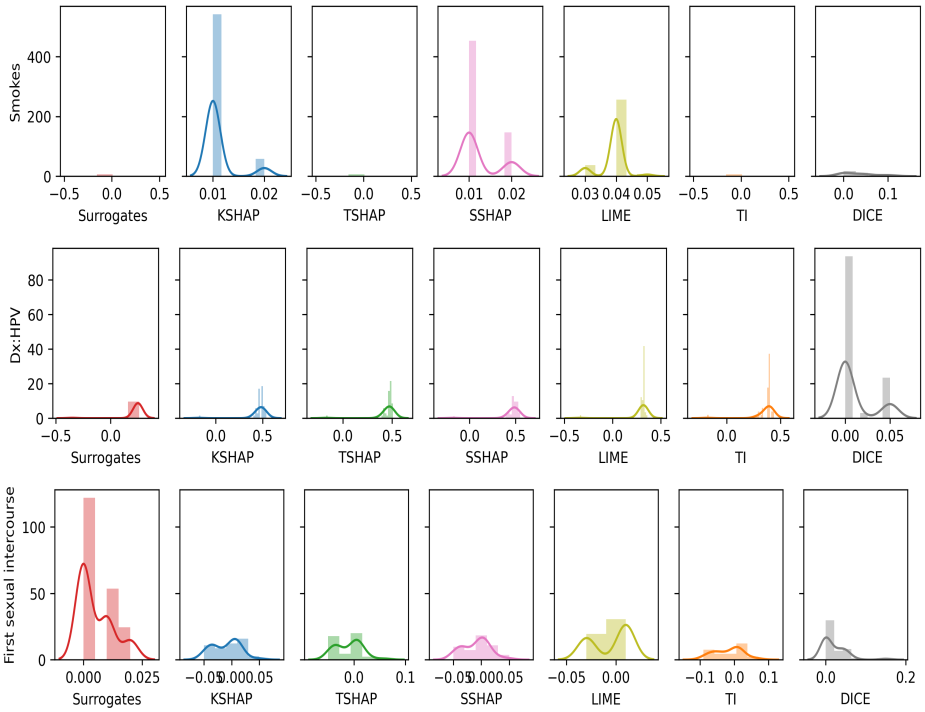

Figure 7: Local variability of the feature importance for Patient 1. We compute the feature importance of 30 nearest neighbors to this patient, among these we keep the patients having the same explanations and predictions and compute the distance between each data point and Patient 1. Small distances mean more stable explanations. For example, Local surrogates is the most stable method for feature Dx:HPV of the Patient 1.

Appendix B. Comparing Different Patient Explanations

Feature	Patient 0	Patient 1	Patient 2	Patient 3	Mean in population
Age	27.00	27.00	14.00	70.00	26.82
Number of sexual partners	2.00	2.00	2.00	1.00	2.51
First sexual intercourse	19.00	14.00	14.00	16.00	17.00
Num of pregnancies	2.00	3.00	1.00	10.00	2.26
Hormonal Contraceptives (years)	7.00	0.86	0.00	0.00	2.04
STDs: Time since first diagnosis	4.00	4.00	4.00	4.00	4.18
STDs: Time since last diagnosis	3.00	3.00	3.00	3.00	3.23
HPV	0.00	1.00	0.00	0.00	0.02
IUD	0.00	0.00	0.00	1.00	0.10

Table 3: Summary statistics of four different patients diagnosed with cervical cancer relative to the mean of the population.

B.1. Patient 2 with Age =14 and Dx:Cancer= 0

B.1.1. Feature importance attributions

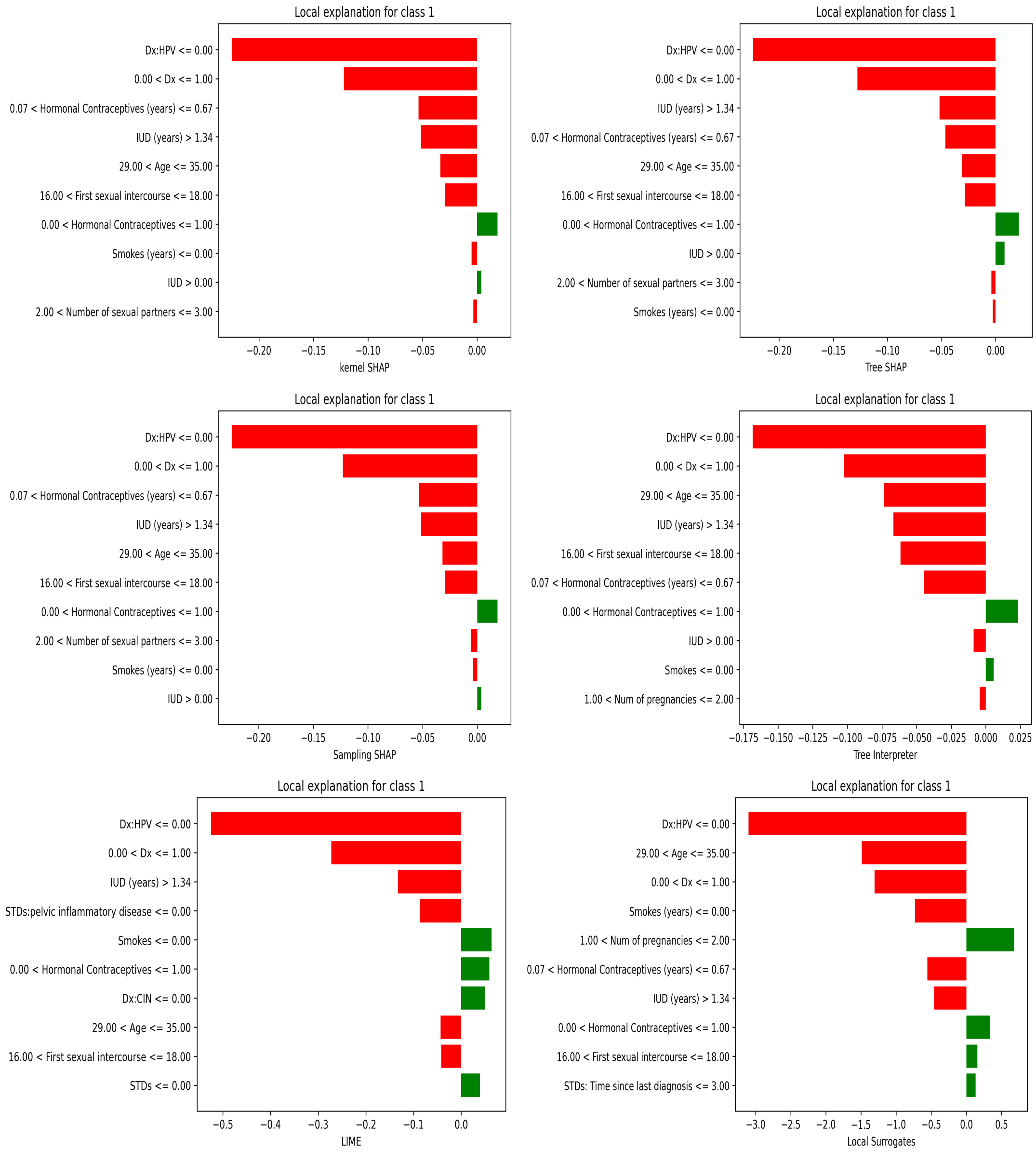

Figure 8: Feature importance attribution for Patient 2, with Age =14 and Dx:Cancer= 0.

B.1.2. Stability of the features in the neighborhood

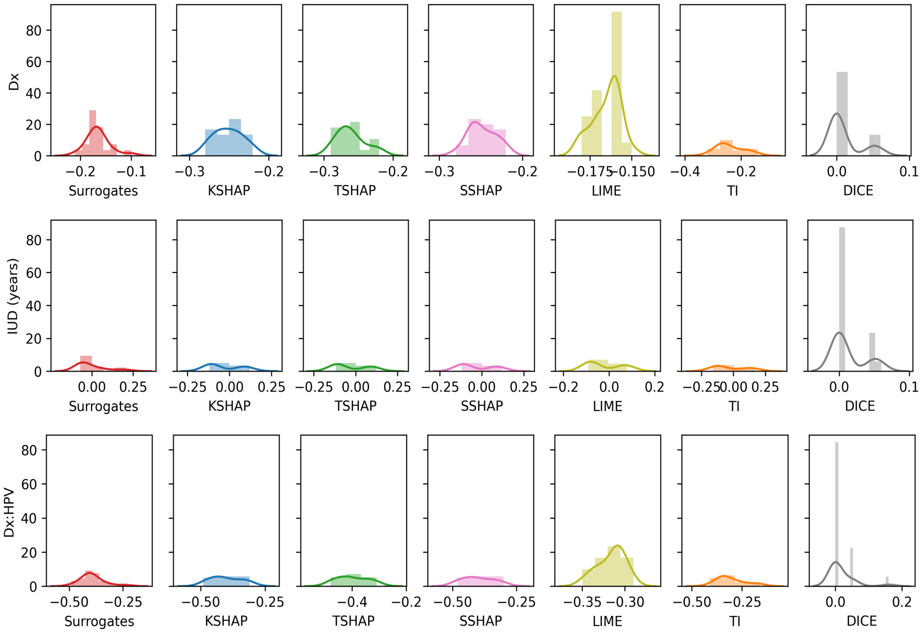

Figure 9: Local variability of the feature importance for Patient with id=2. We compute the feature importance of 30 nearest neighbors to this patient, among these we keep the patients having the same explanations and predictions and compute the distance between each data point and Patient with id=2. Small distances mean more stable explanations. For example, LIME is the most stable method for feature Dx of the Patient with id=2.

B.1.3. Rank and feature agreements

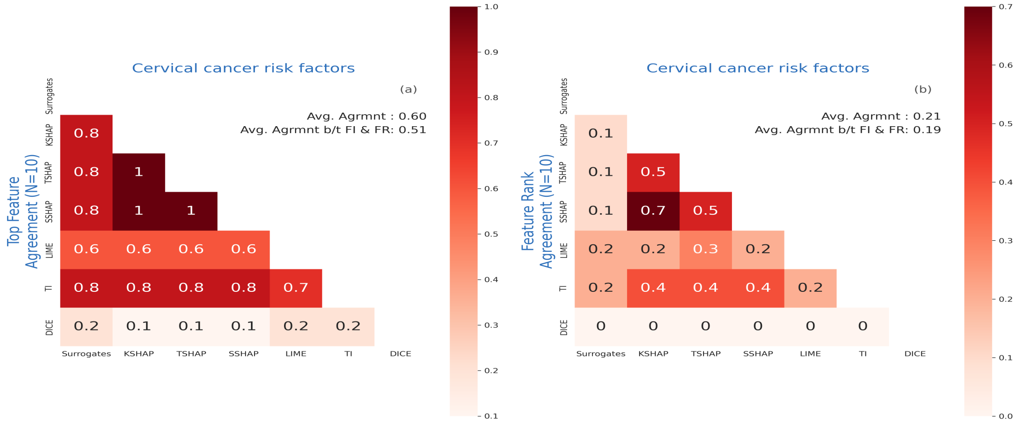

Figure 10: Feature agreement for patient 2 (Age=14).

B.2. Patient 3 with Age =70 and Dx:Cancer= 0

B.2.1. Feature importance attributions

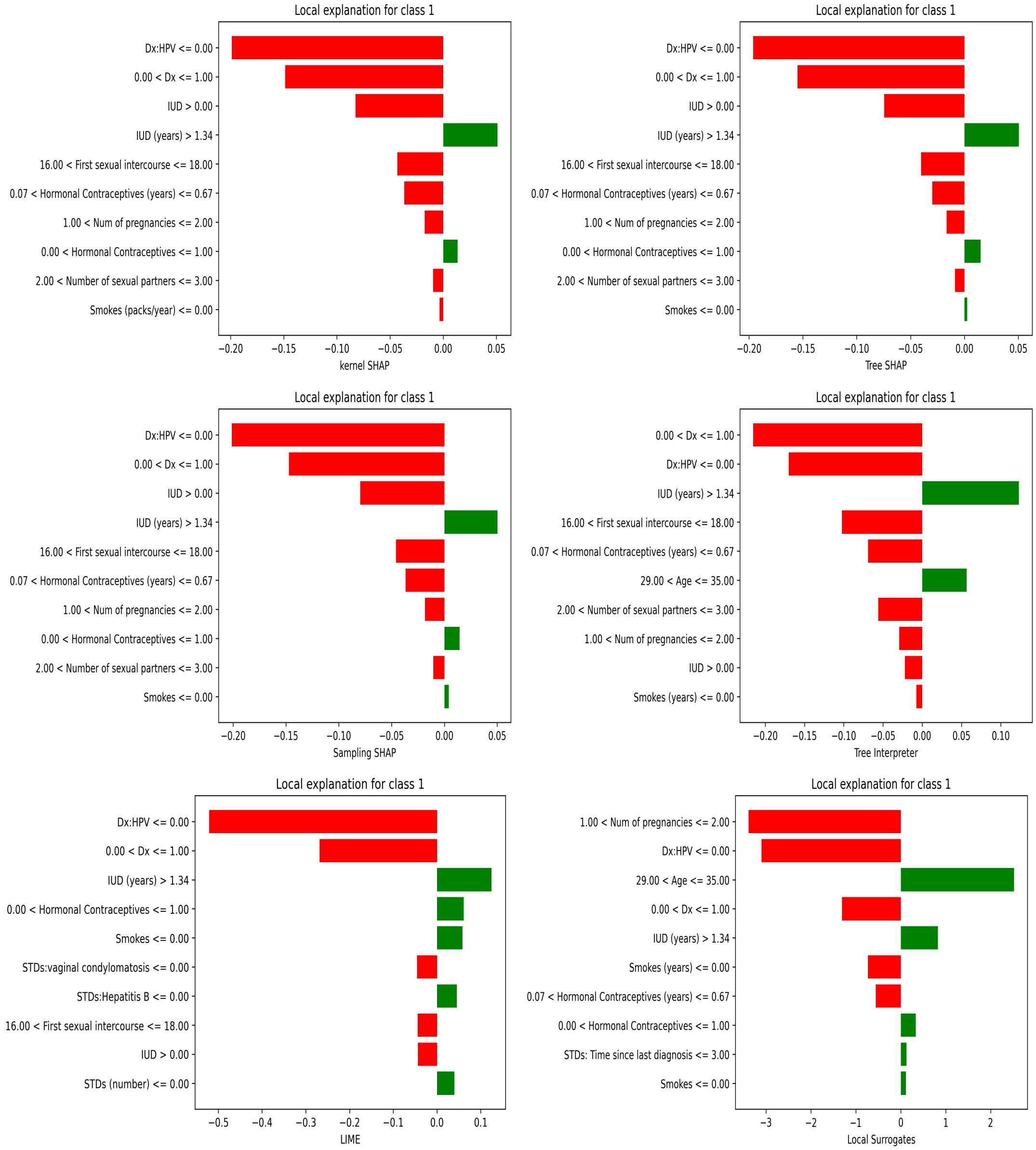

Figure 11: Feature importance attributions for Patient 3, with Age =70 and Dx:Cancer= 0.

B.2.2. Stability of the features in the neighborhood

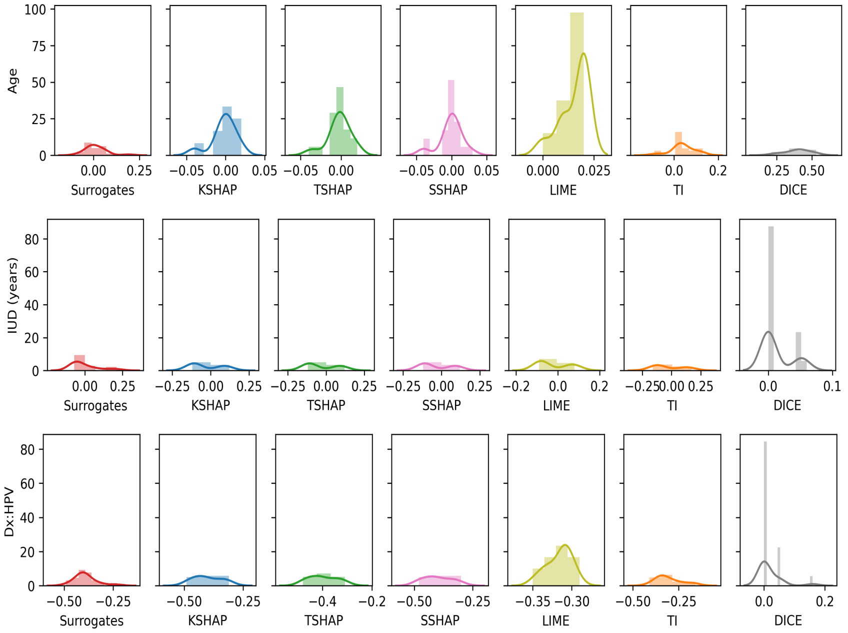

Figure 12: Local variability of the feature importance for Patient with id=3. We compute the feature importance of 30 nearest neighbors to this patient, among these we keep the patients having the same explanations and predictions and compute the distance between each data point and Patient with id=3. Small distances mean more stable explanations. For example, LIME is the most stable method for feature Dx:HPV of the Patient with id=3.

B.2.3. Rank and feature agreements

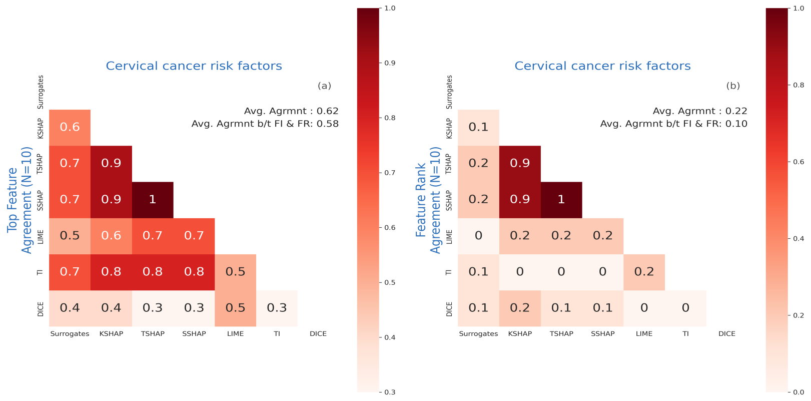

Figure 13: Feature agreement for patient 3 (Age=70).

Appendix C. Fooling Explanations with Random Variables

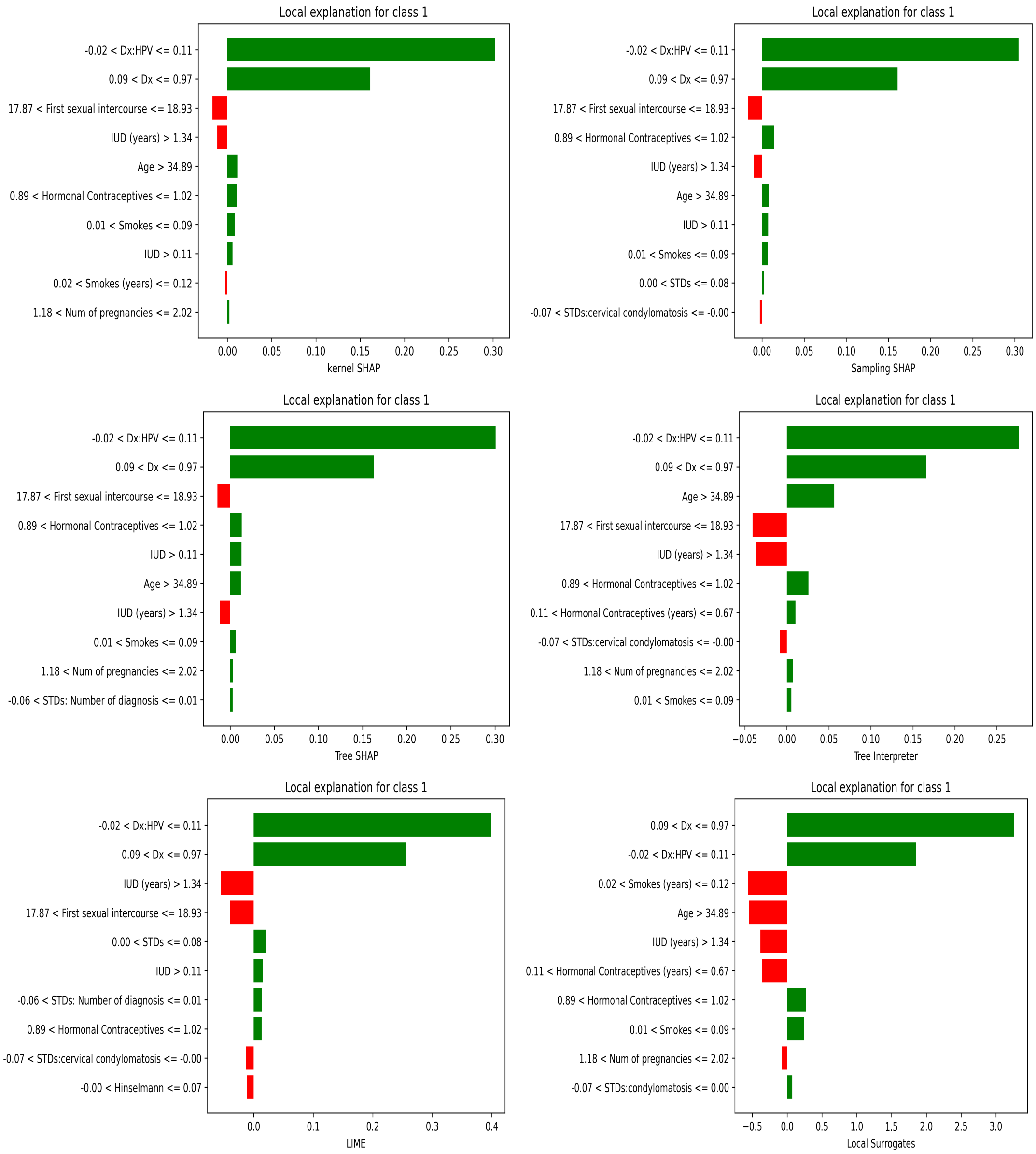

Figure 14: Adding a binary, a continuous random variables and noise to the features ($\epsilon \in \mathcal{N}(0, .1)$).

Appendix D. Effect of Changing a Feature Value on the Explanations for Patient 1.

D.1. Changing Age from 27 to 80

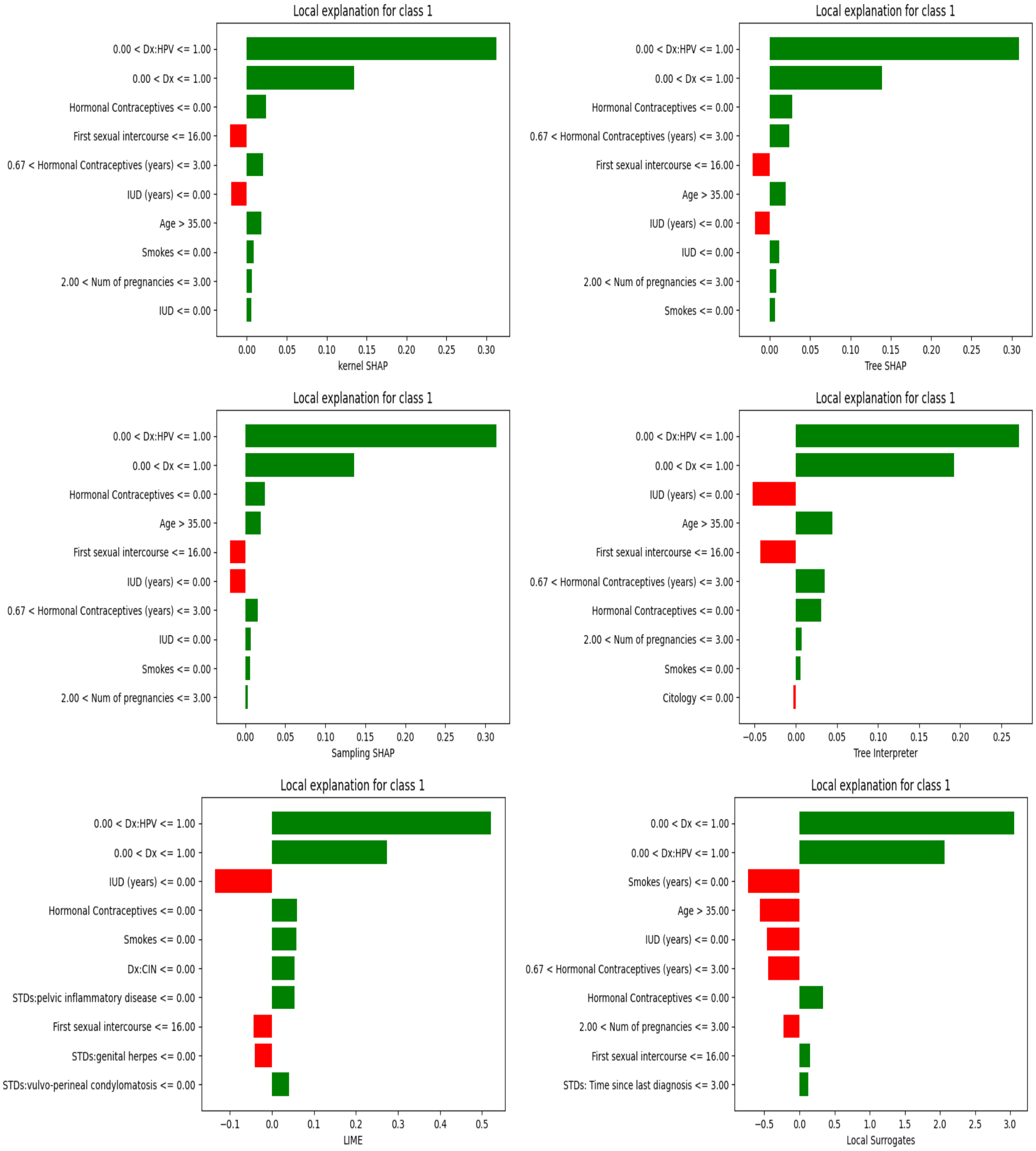

Figure 15: Feature importance attributions for Patient with id=291 and Age = 80.

D.2. Changing Number of pregnancies from 3 to 0

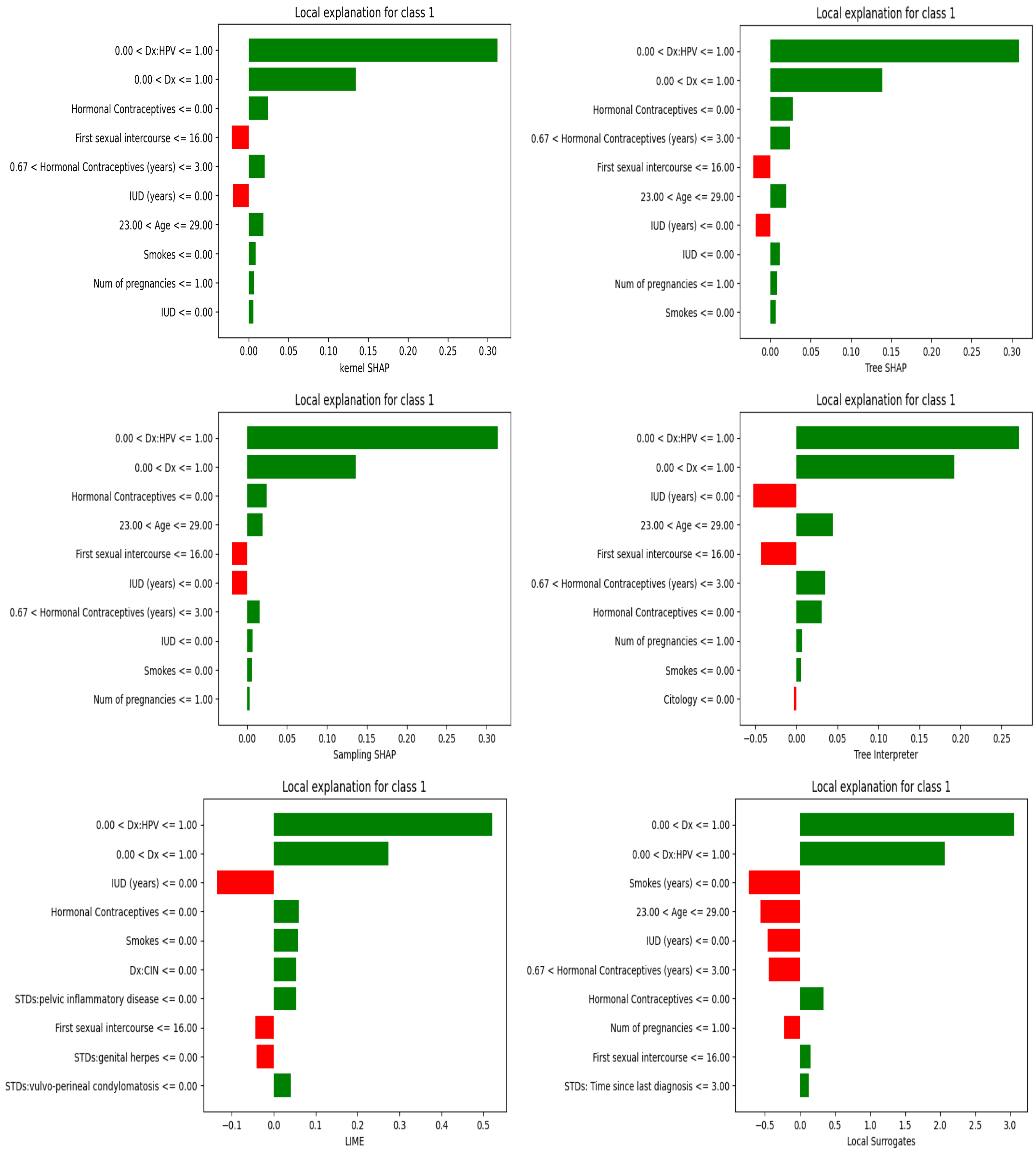

Figure 16: Feature importance attributions for Patient with id=291 and Number of pregnancies = 0.

D.3. Changing Smokes from 0 to 1

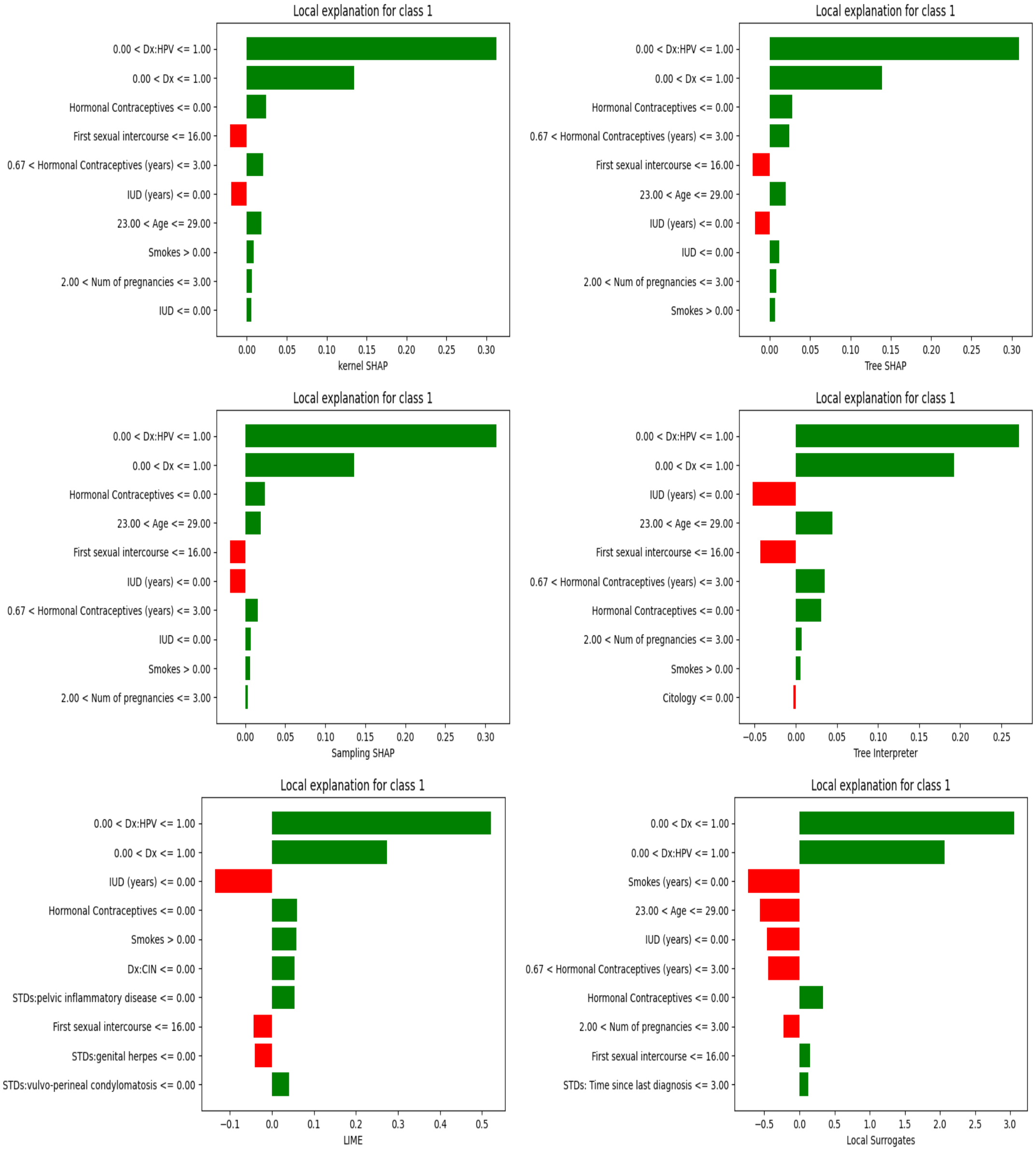

Figure 17: Feature importance attributions for Patient with id=291 and Smokes = 1.

D.4. Changing Number of sexual partners from 2 to 60

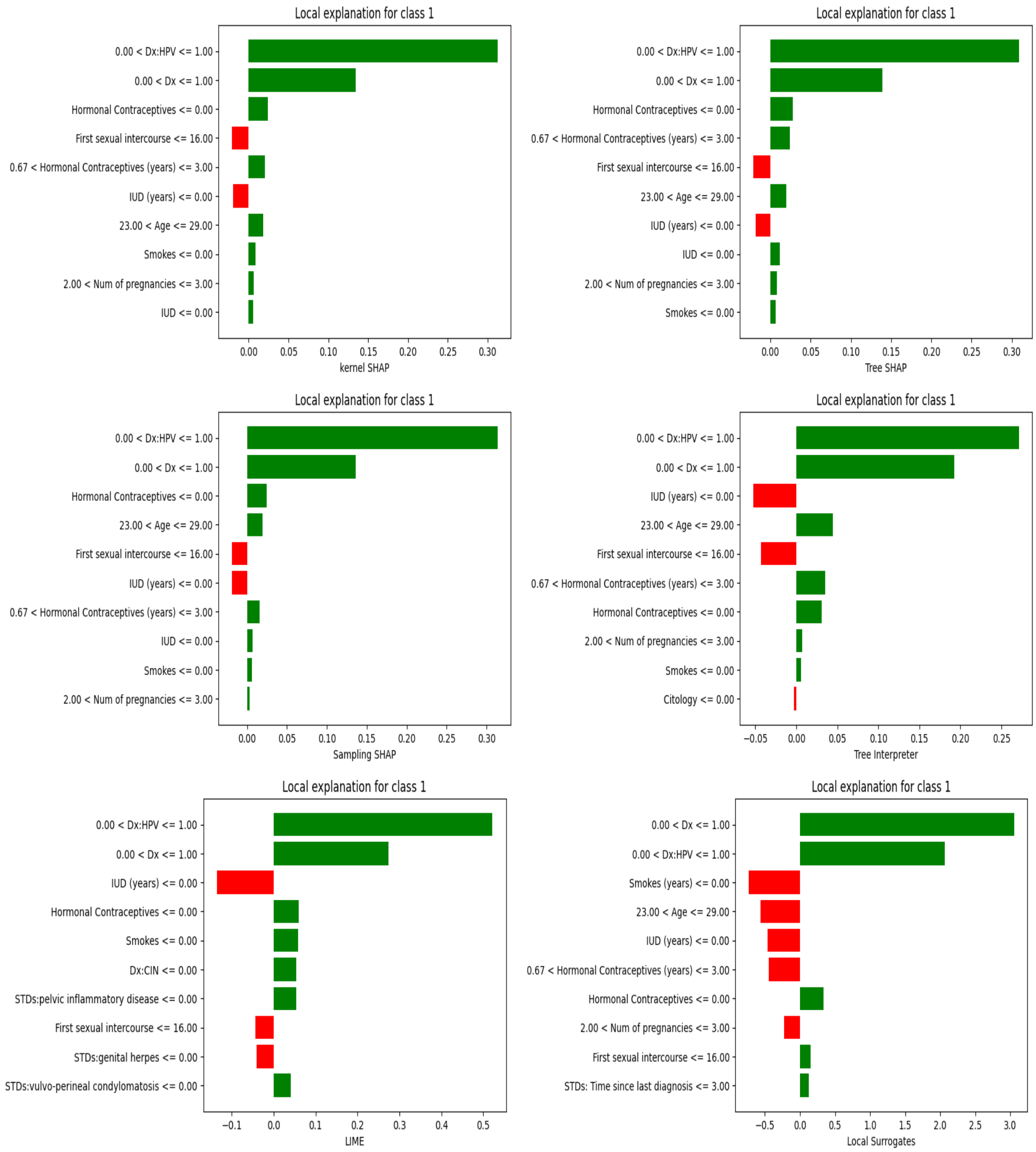

Figure 18: Feature importance attributions for Patient with id=291 and Number of sexual partners = 60.

D.5. Changing First sexual intercourse from 14 to 40

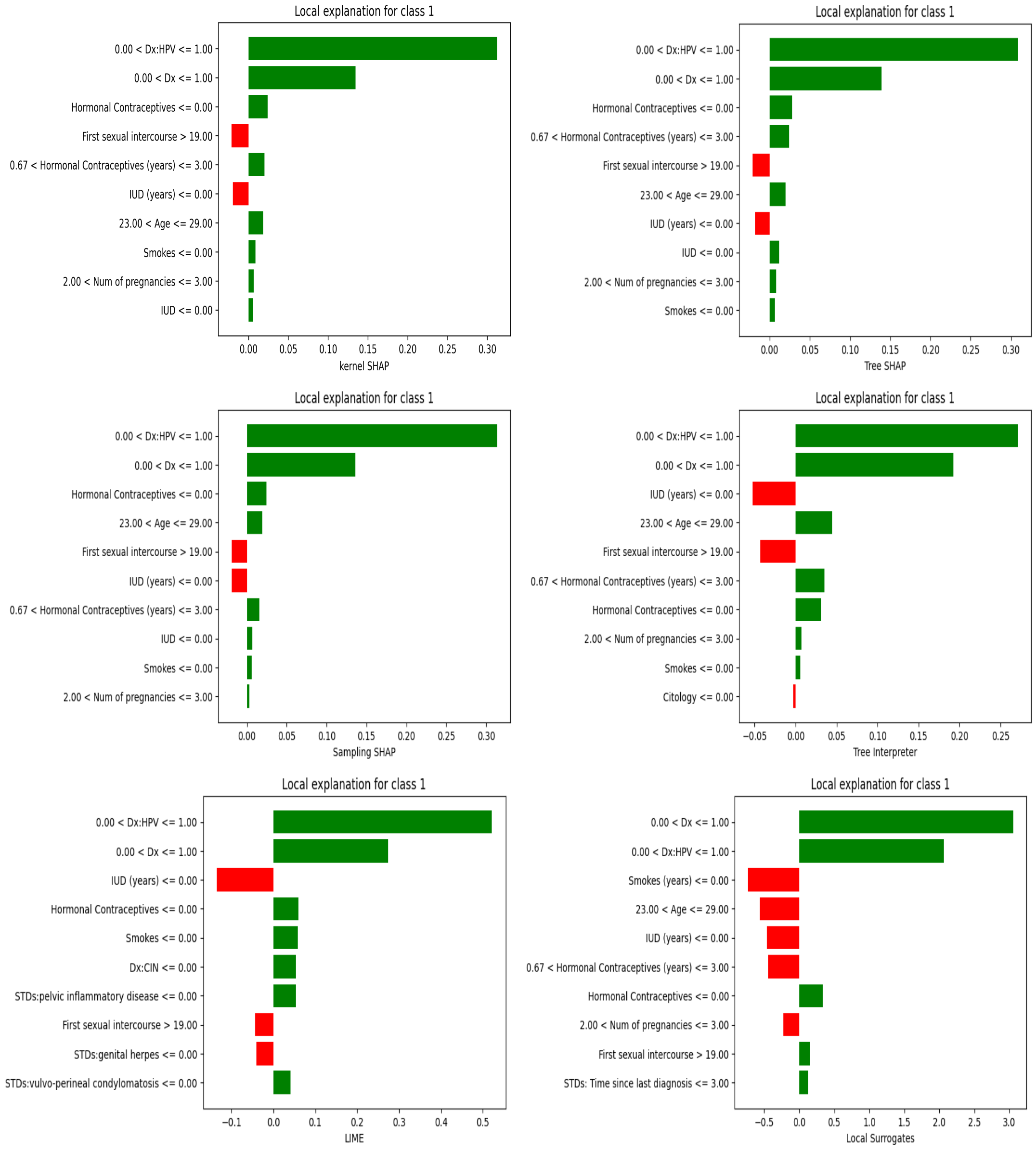

Figure 19: Feature importance attributions for Patient with id=291 and First sexual intercourse = 40.

Appendix E. Difference in the Explanations for Different Age Categories

E.1. Patients with Age 14 and 19

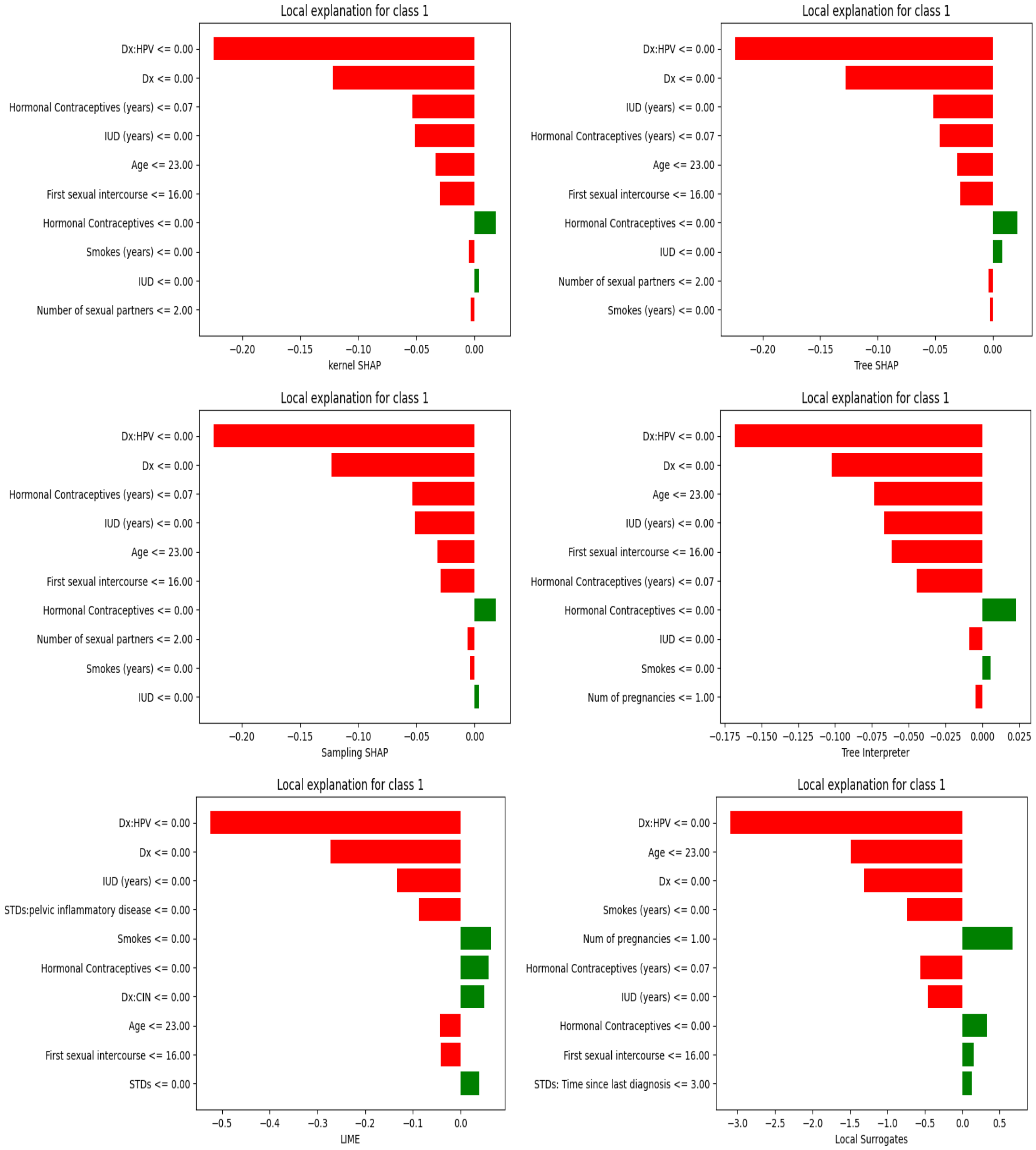

Figure 20: Feature importance attributions for Patient with id=29.

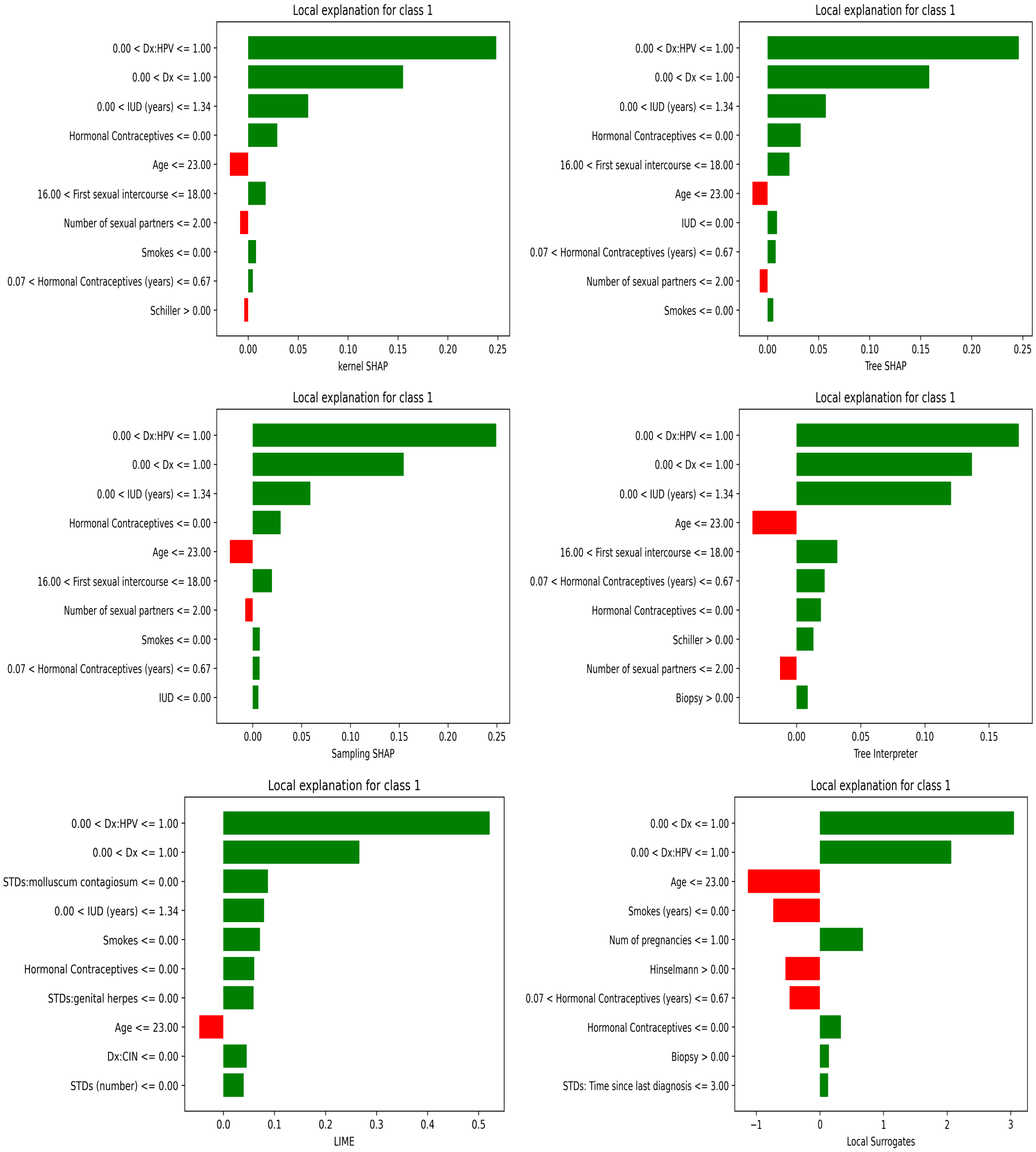

Figure 21: Feature importance attributions for Patient with id=19.

E.2. Patients with Age=24

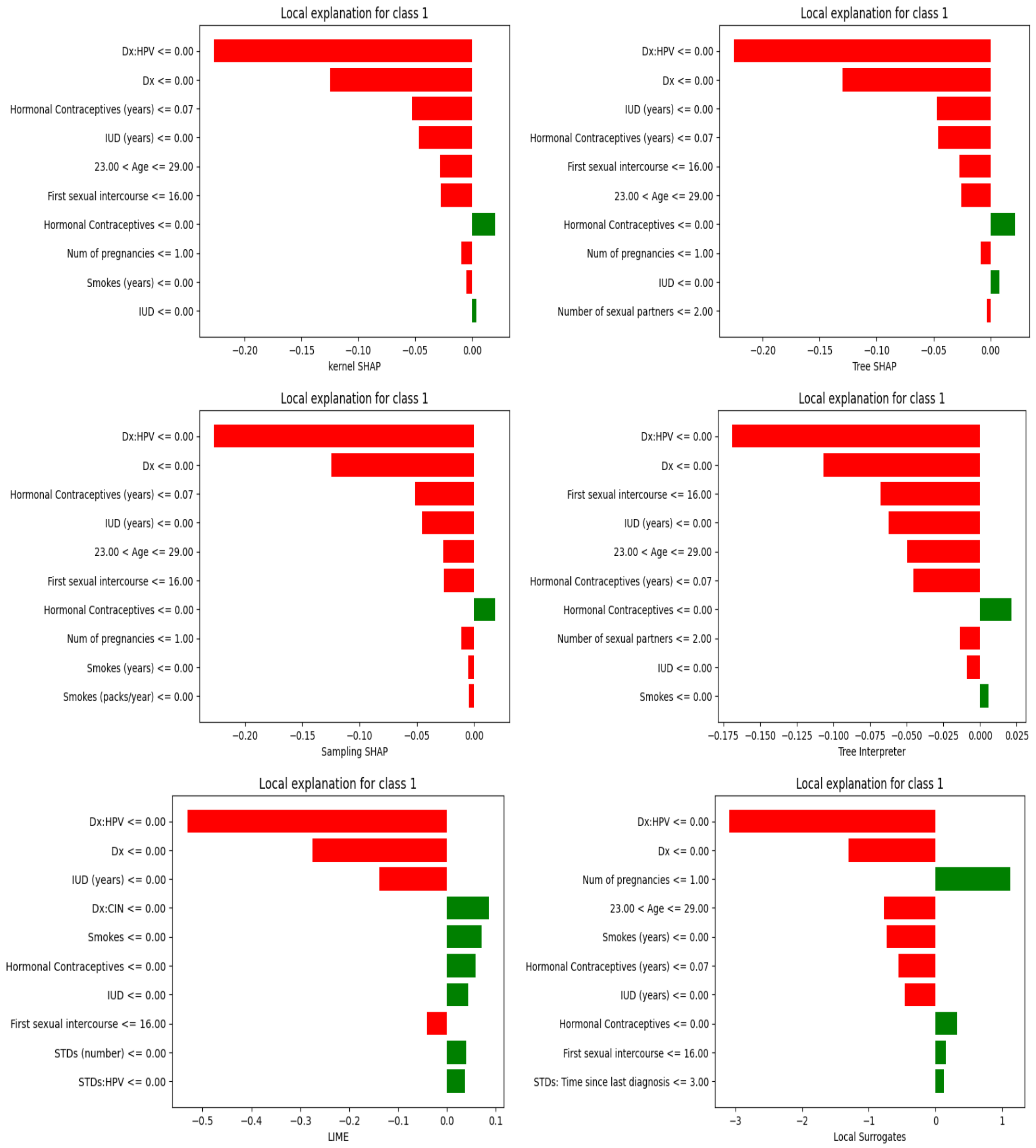

Figure 22: Feature importance attributions for Patient with id=11.

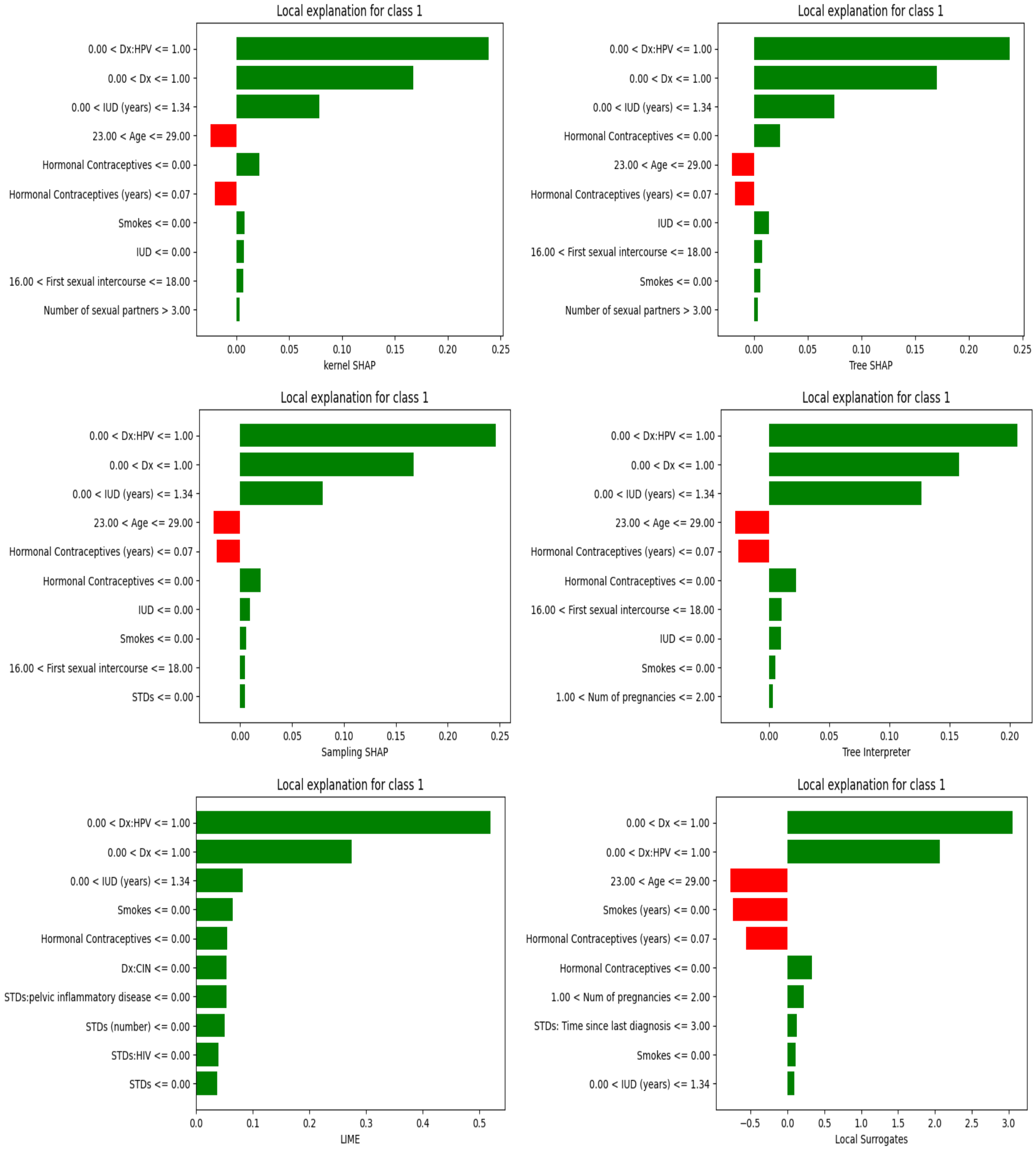

Figure 23: Feature importance attributions for Patient with id=136.

E.3. Patients with Age=34

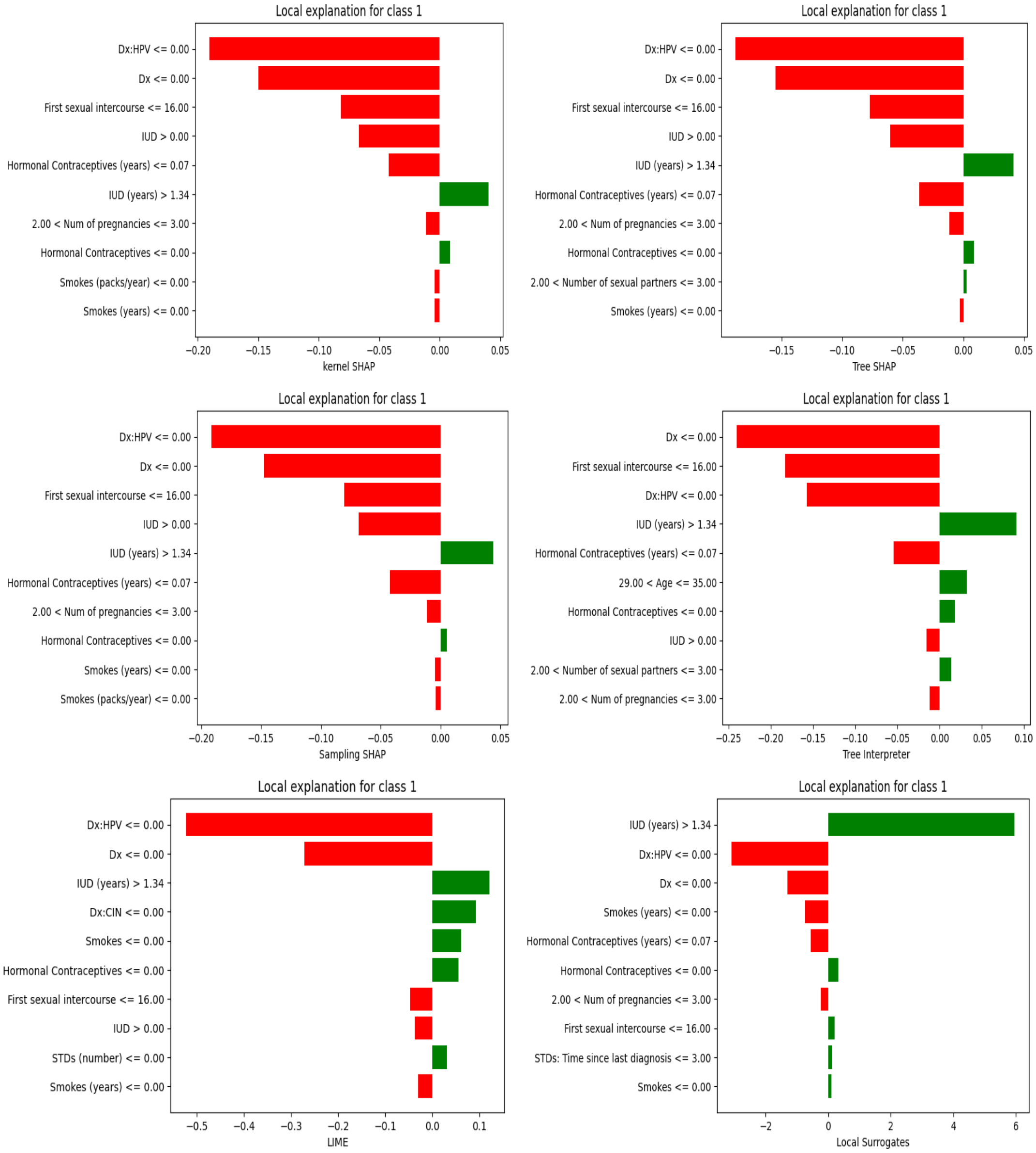

Figure 24: Feature importance attributions for Patient with id=6.

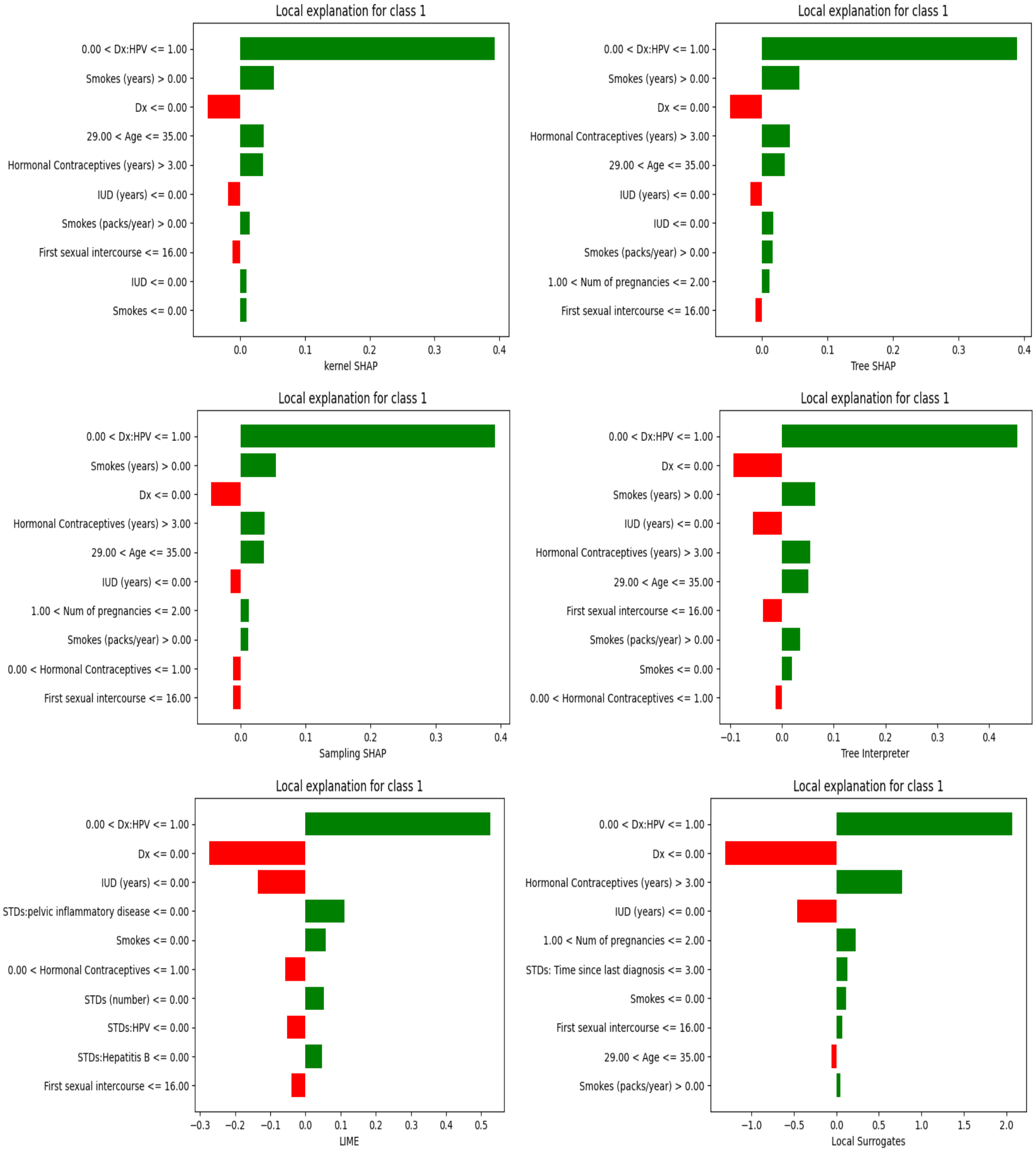

Figure 25: Feature importance attributions for Patient with id=307.

E.4. Patients with Age=44 and 45

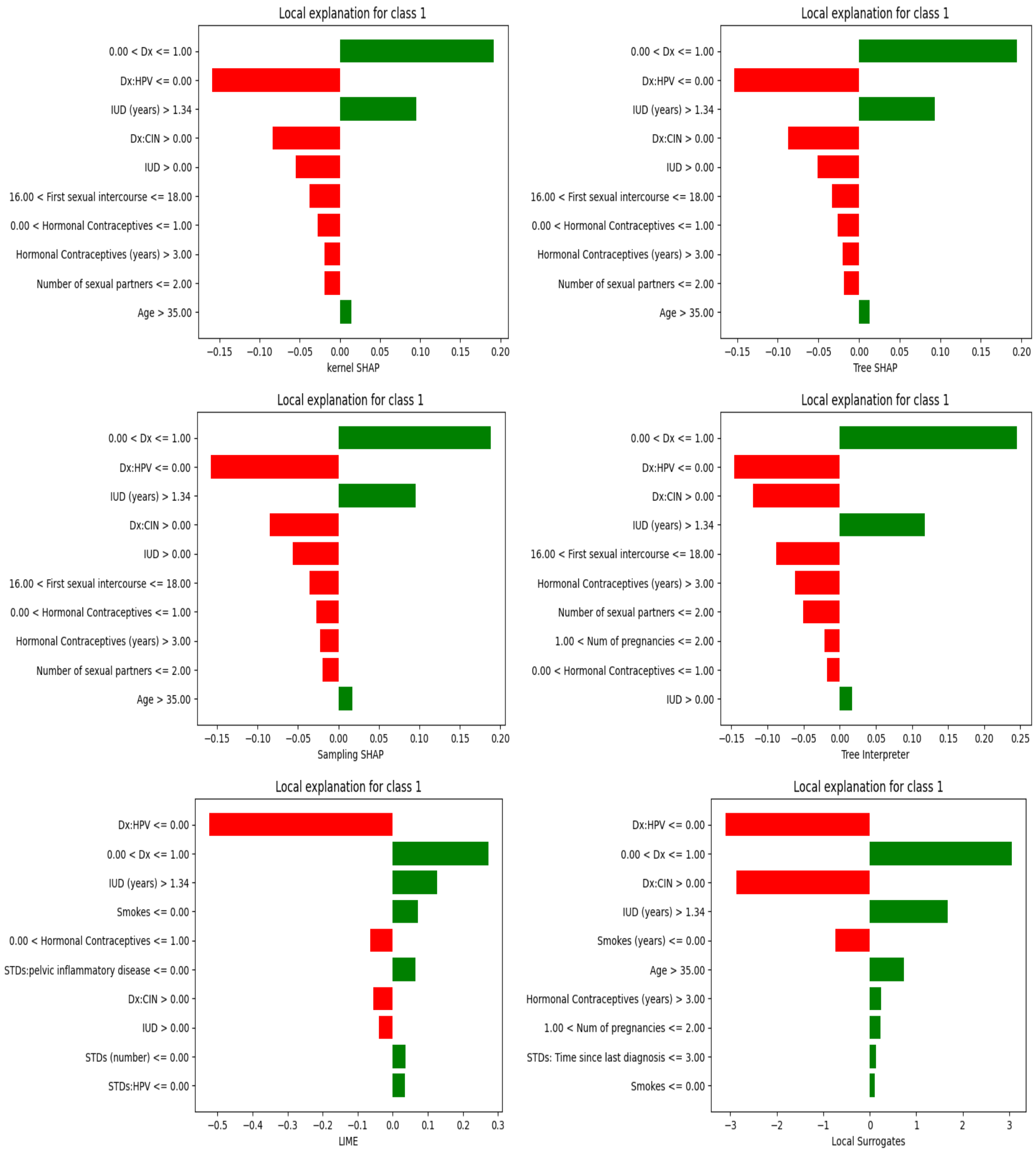

Figure 26: Feature importance attributions for Patient with id=45.

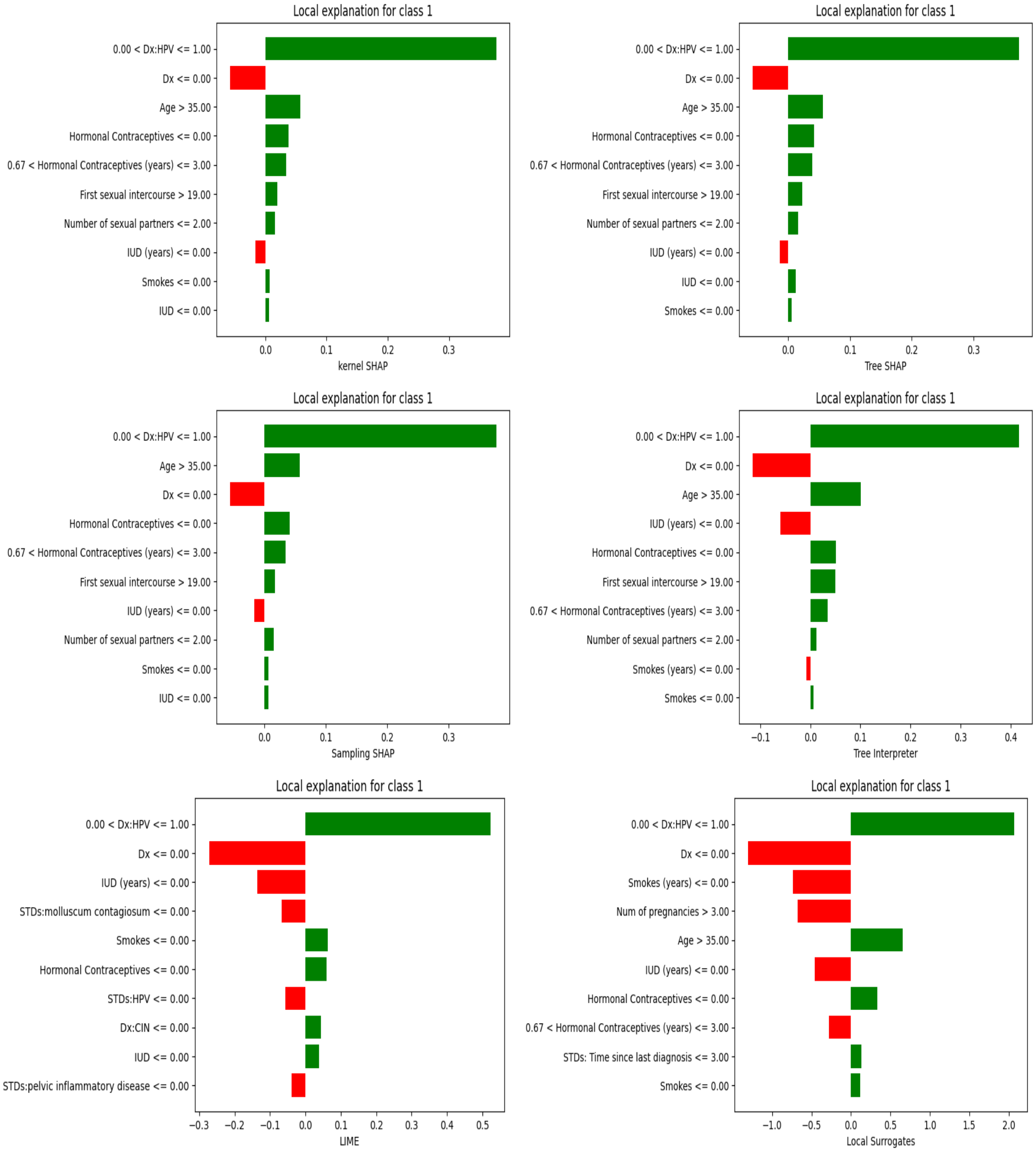

Figure 27: Feature importance attributions for Patient with id=158.

E.5. Patients with Age=54

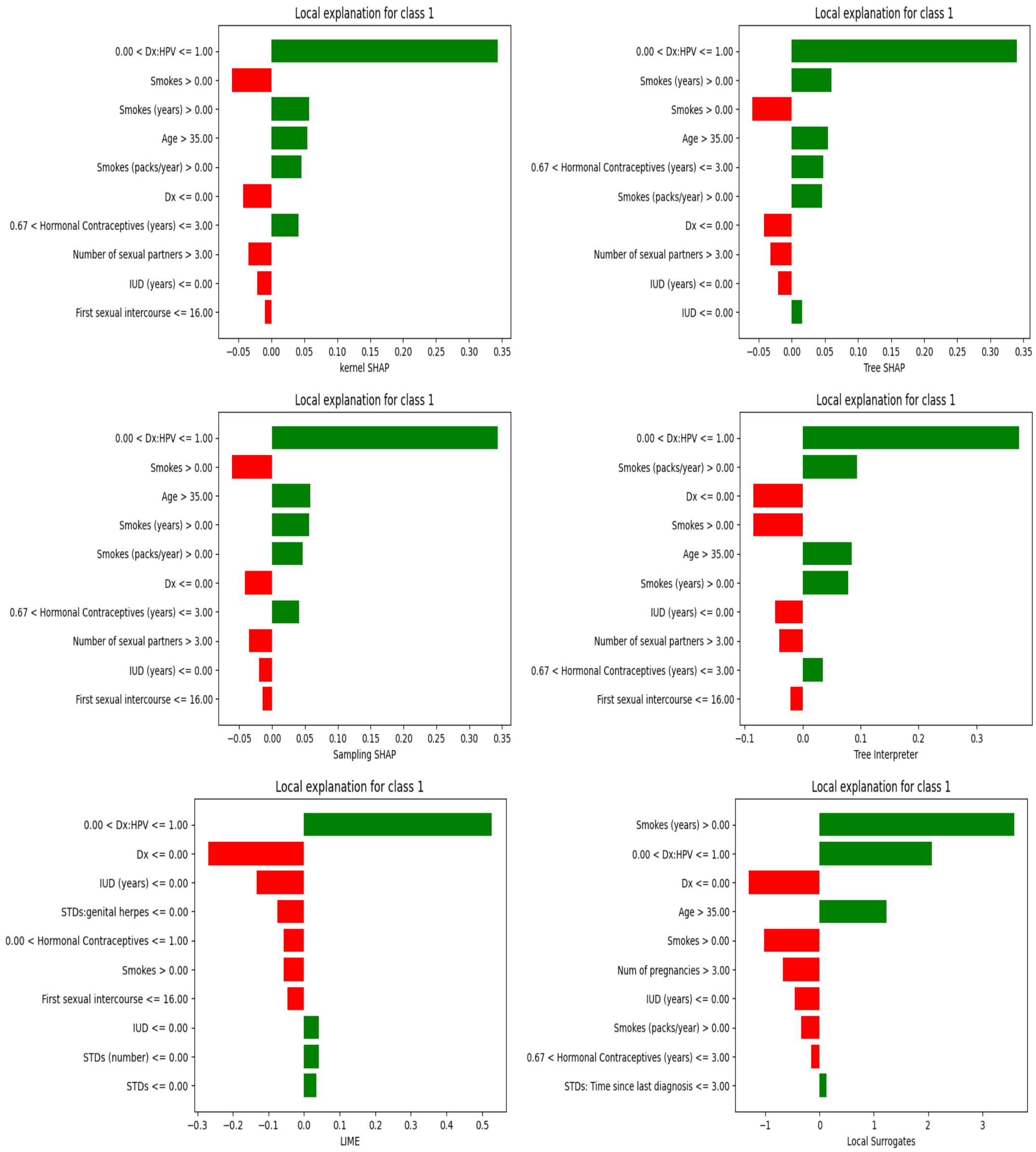

Figure 28: Feature importance attributions for Patient with id=222.

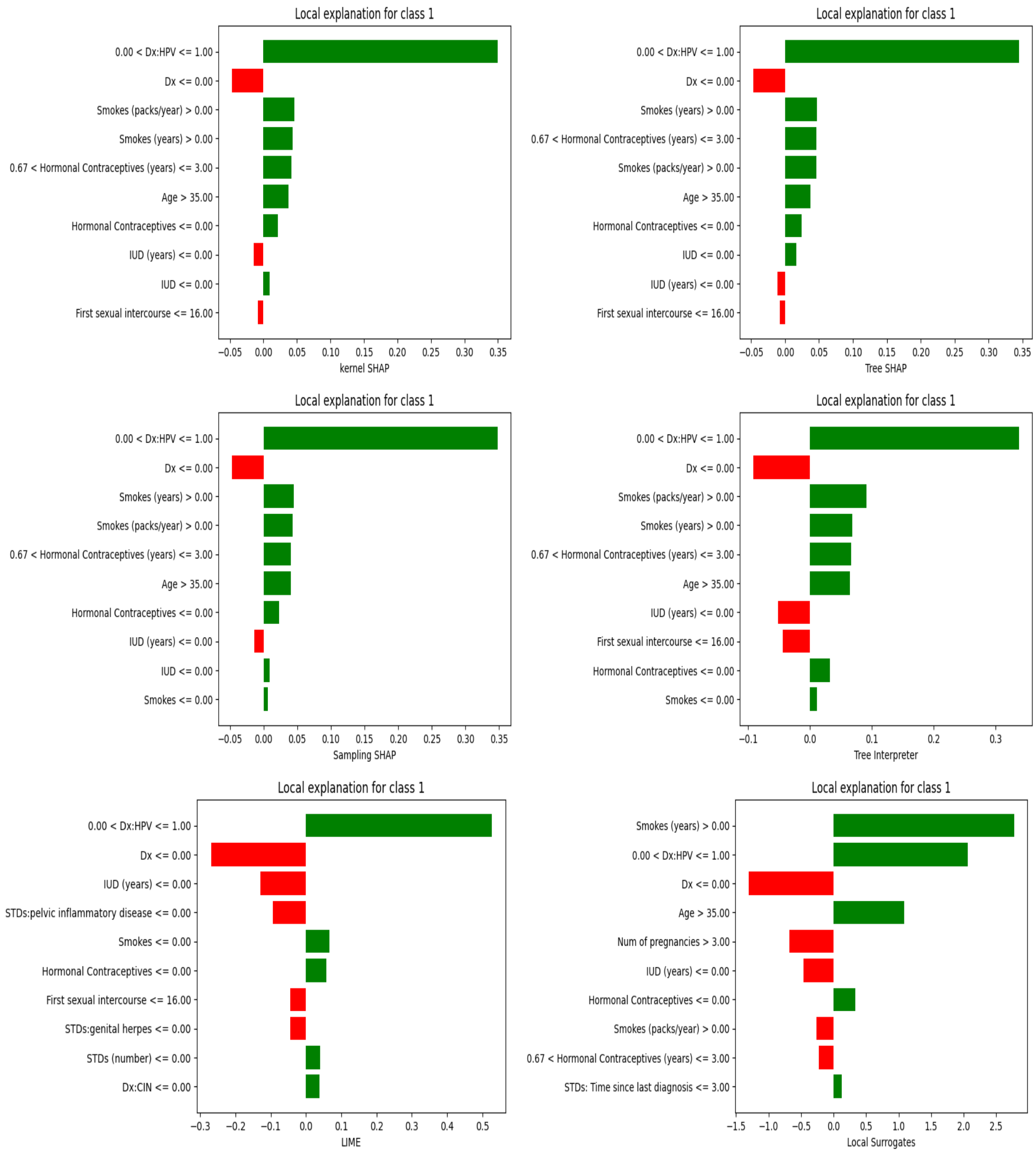

Figure 29: Feature importance attributions for Patient with id=141.

E.6. Patient with Age=74

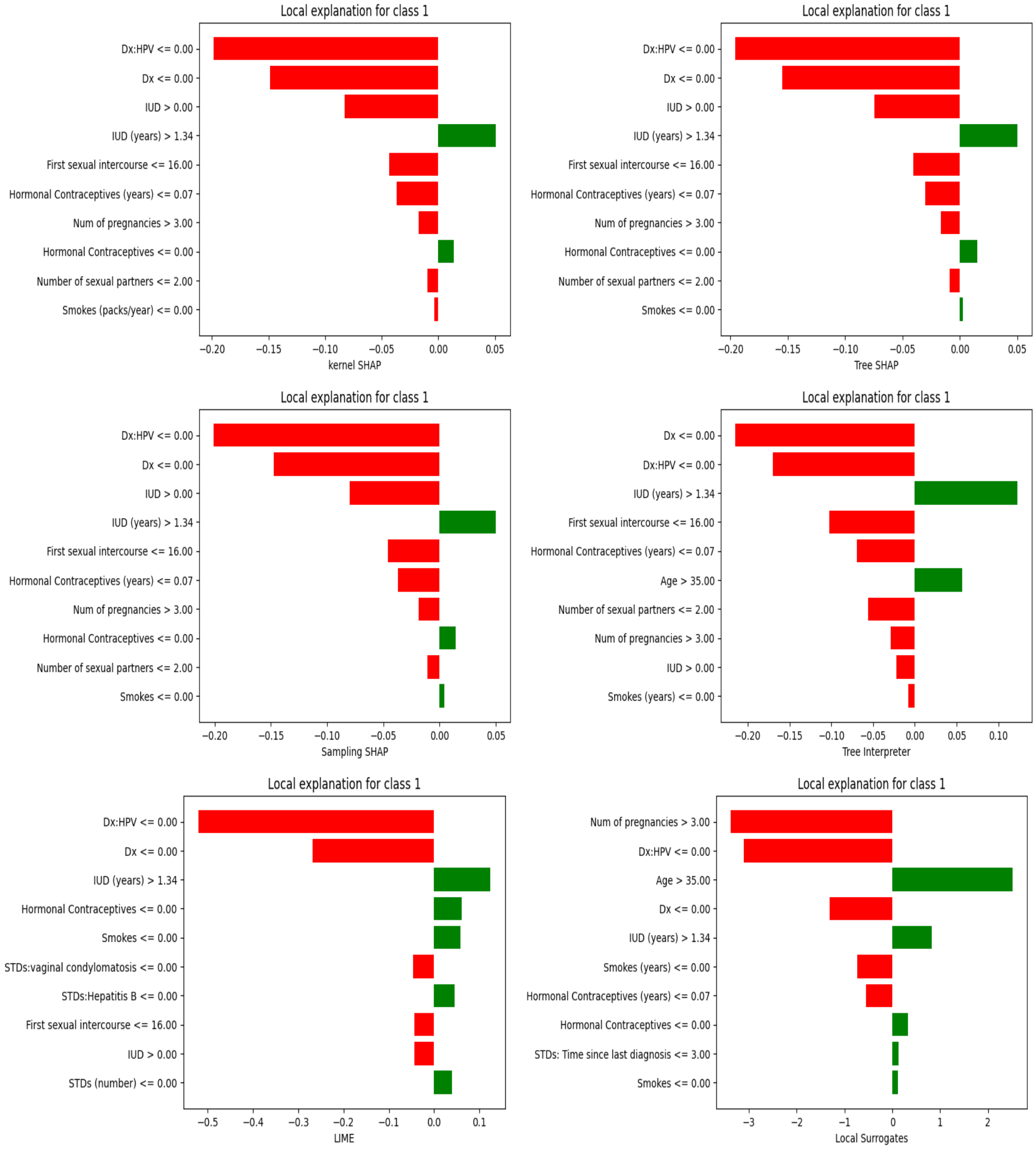

Figure 30: Feature importance attributions for Patient with id=57.

Appendix F. Explanation Difference between Smoking and Non Smoking Patients

F.1. Patients with Smokes=1

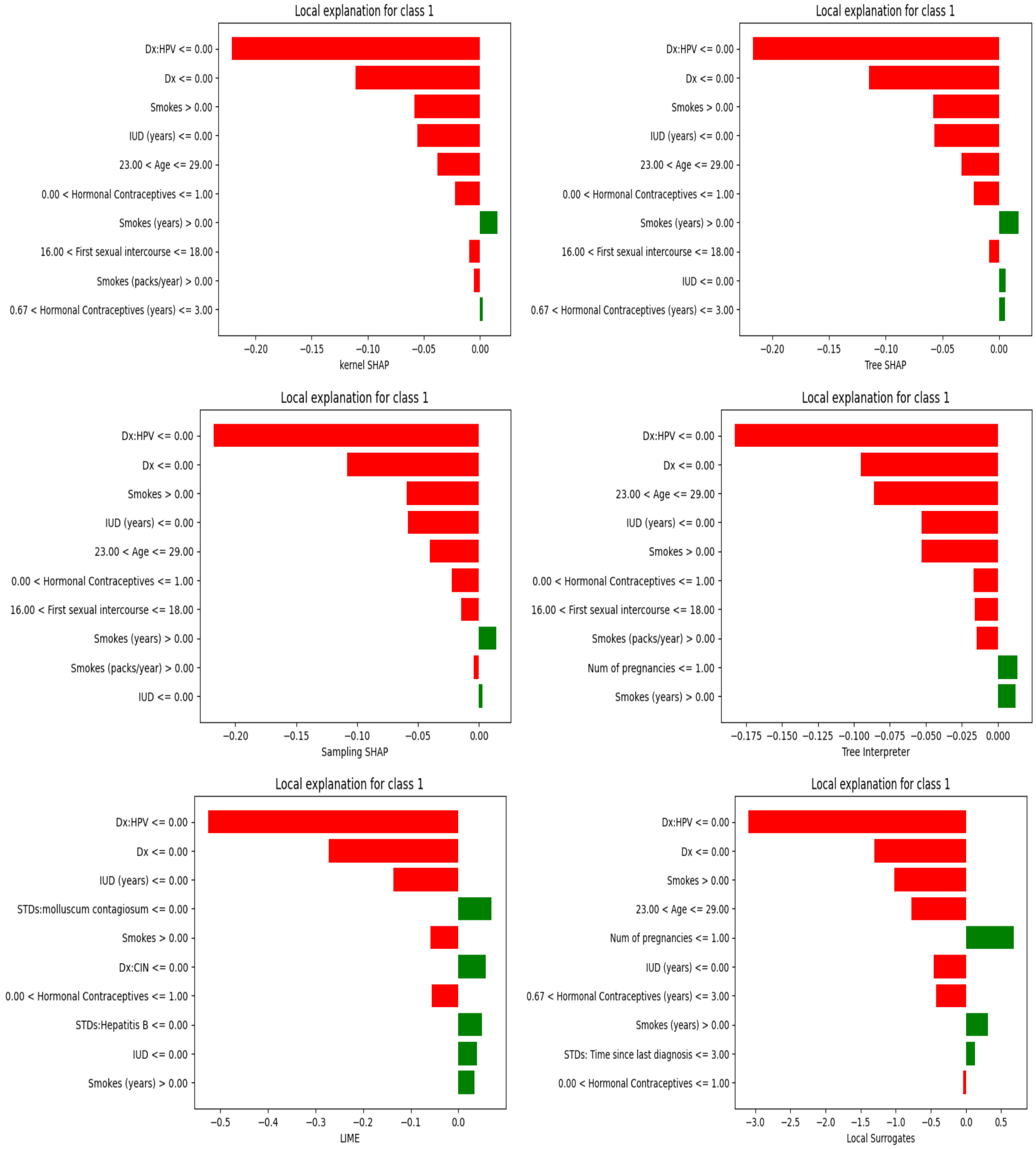

Figure 31: Feature importance attributions for Patient with id=30.

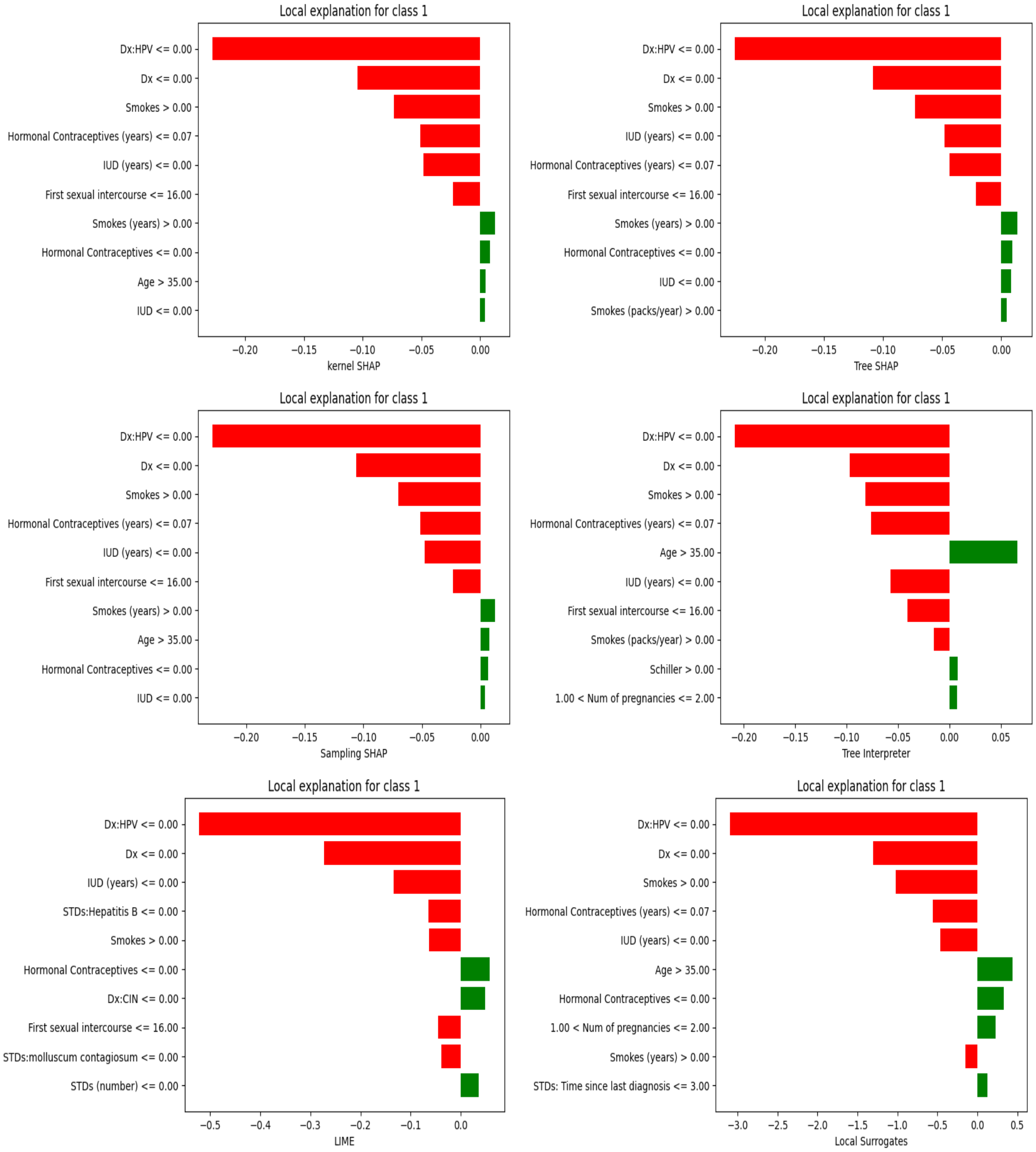

Figure 32: Feature importance attributions for Patient with id=4.

F.2. Patients with Smokes=0

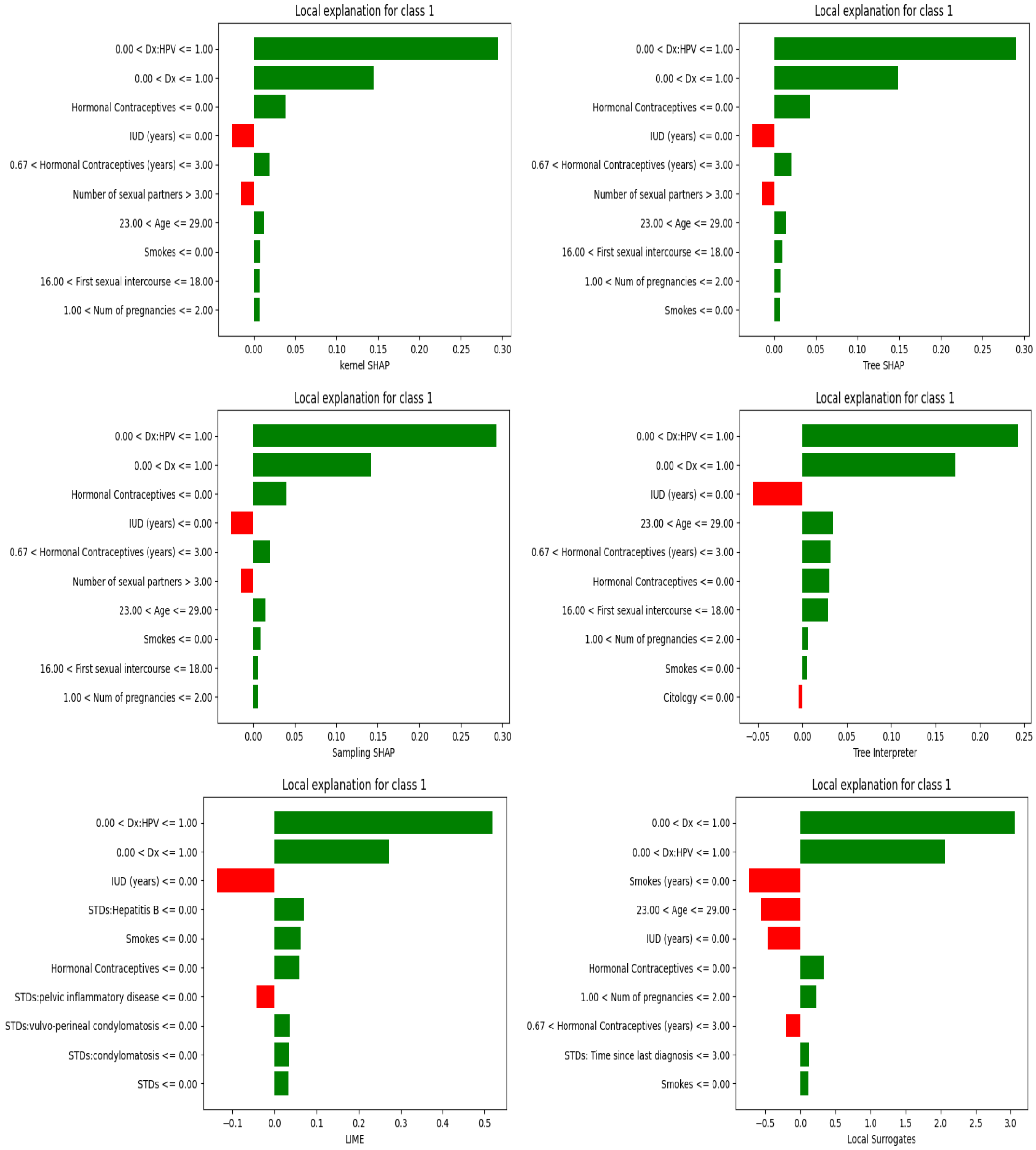

Figure 33: Feature importance attributions for Patient with id=0.

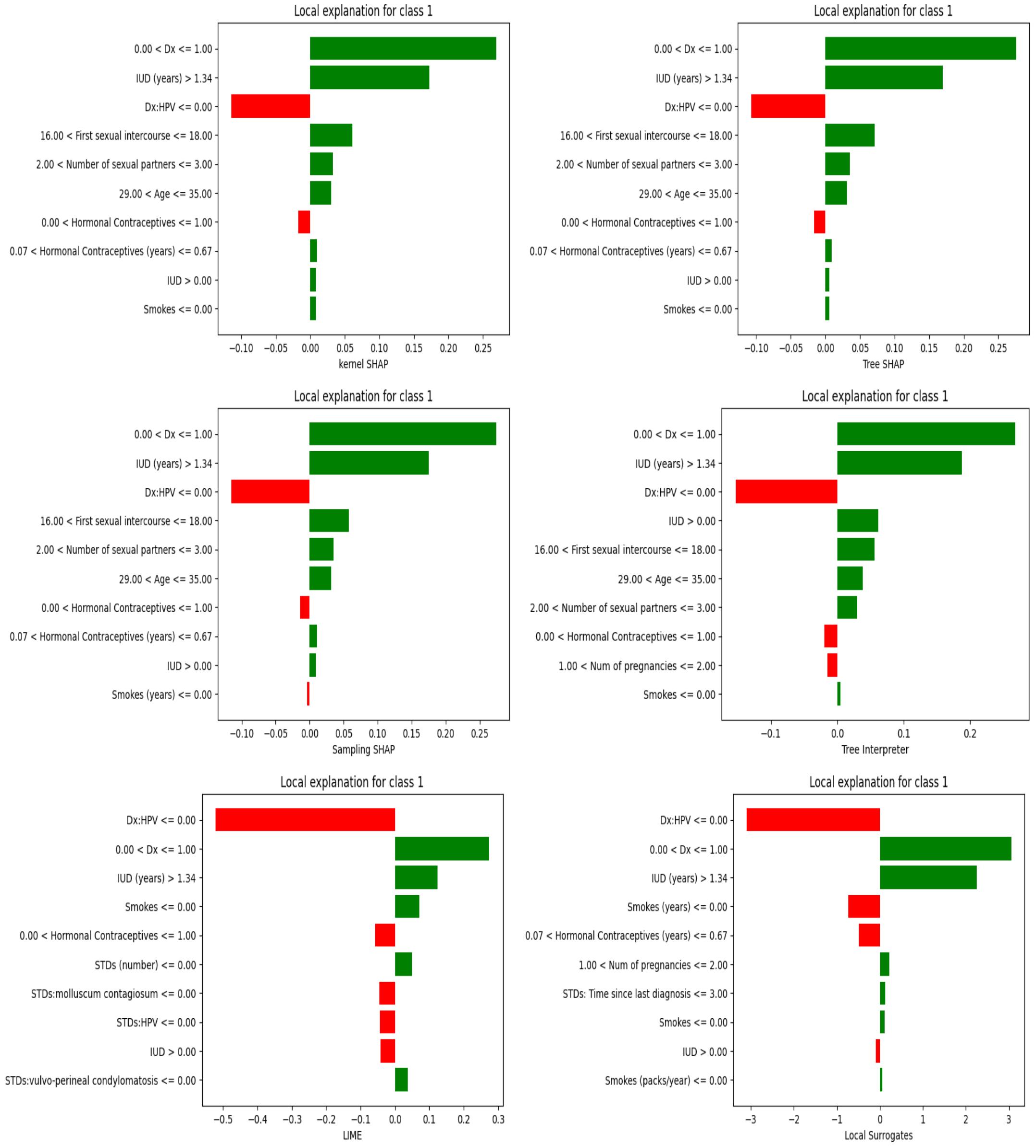

Figure 34: Feature importance attributions for Patient with id=1.

Appendix G. Second Best Model: MLP

We tested the second best model (MLP) in terms of AUROC and accuracy. Results of the feature importance show small changes in feature ranking and sign. For example, kernel SHAP for an MLP shows that for Patient 1, the age of first sexual activity is positively correlated with the prediction, while for a random forest this is negative. Sampling SHAP ranks age in the top 4 features for the random forest for Patient 1 while age is not in the top 10 features of the MLP. Similar insights can be found on other patients.

G.1. Feature importance for Patient 1

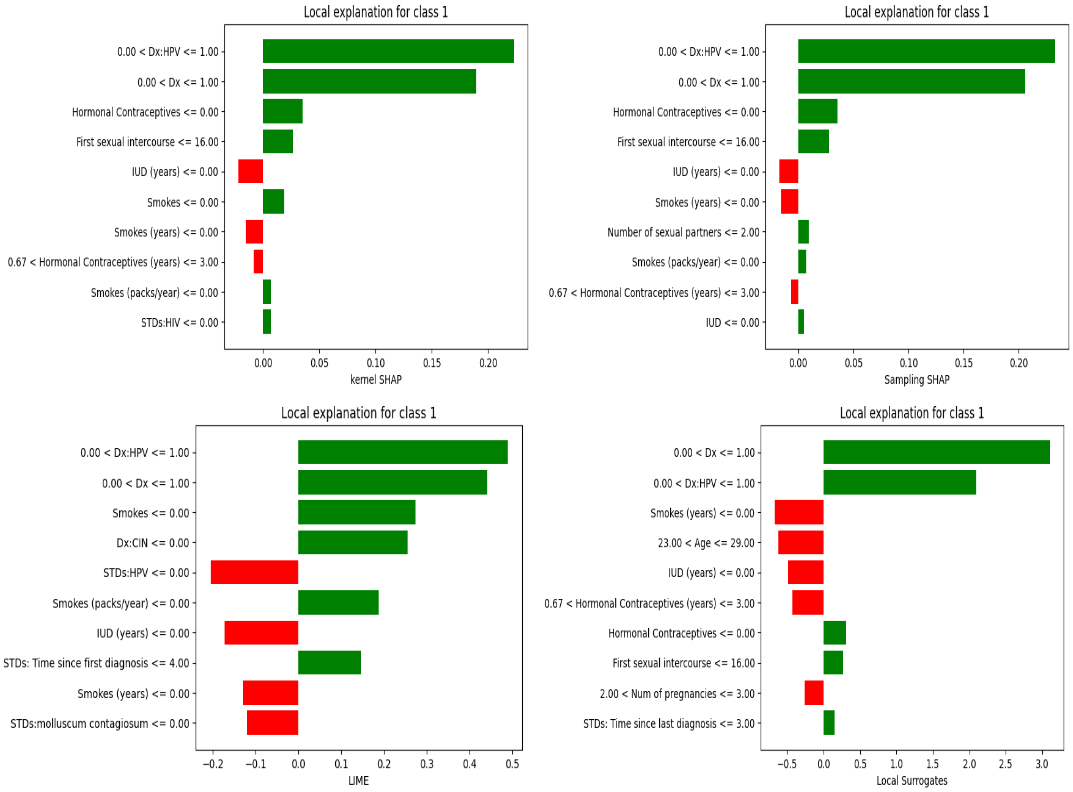

Figure 35: Feature importance attributions for Patient 1. The predictions are made by the MLP model.

G.2. ROAR and Consistency of the explanations

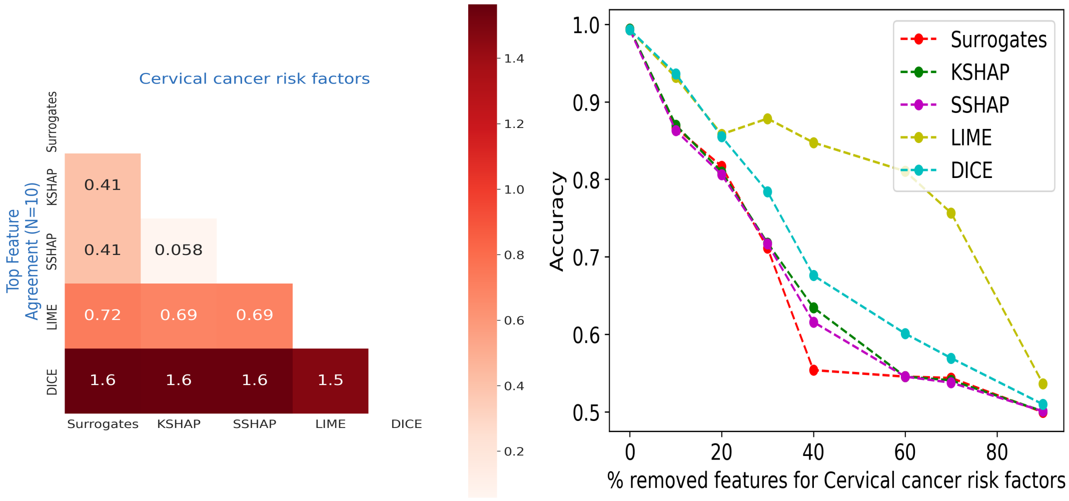

Figure 36: (a) RAOR. (b). Consistency

G.3. Faithfulness: Rank and feature agreements

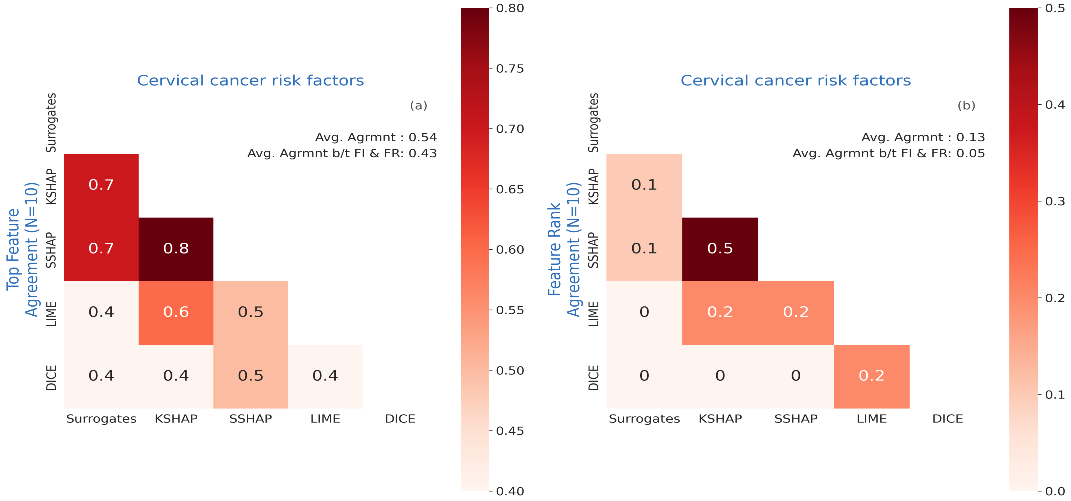

Figure 37: Feature agreement for Patient 1.

G.4. Compactness, global stability and consistency

Methods	# Features for 90% Accuracy	Accuracy with 5 features(%)	Mean Stability	Mean Consistency
Surrogates	2	**100**	**0.22**	0.62
KSHAP	**1**	18	1.05	**0.54**
SSHAP	**1**	16	1.66	**0.54**
LIME	**1**	**100**	0.52	0.72
DICE	8	**100**	2.04	1.23

Table 4: Compactness, stability and consistency of the explanation for Patient 1 for the prediction made by the MLP.

Appendix H. Comparing Different Patients with Different Risk Factors

Methods	# Features for 90% Accuracy	Accuracy with 5 features(%)	Stability	Mean Feature Agree	Mean Rank Agree
Surrogates	3	23	0.02	0.80	0.10
KSHAP	**1**	02	0.03	**1.00**	**1.00**
TSHAP	**1**	00	0.03	**1.00**	0.50
SSHAP	**1**	20	0.03	**1.00**	0.70
LIME	**1**	23	**0.01**	0.60	0.20
TI	**1**	03	0.02	0.80	0.40
DICE	2	**50**	0.03	0.10	0.00

Table 5: Patient with ID: 29, Age: 14, First Sexual Intercourse: 14, Number of Sexual Partners: 2, Number of pregnancies : 1, and Smokes: 1.

Methods	# Features for 90% Accuracy	Accuracy with 5 features(%)	Stability	Mean Feature Agree	Mean Rank Agree
Surrogates	3	64	0.02	0.80	0.40
KSHAP	**1**	06	0.03	**1.00**	**1.00**
TSHAP	**1**	04	0.03	**1.00**	0.80
SSHAP	**1**	05	0.03	**1.00**	**1.00**
LIME	**1**	23	**0.01**	0.60	0.20
TI	**1**	07	0.02	0.90	0.40
DICE	4	**70**	0.03	0.30	0.10

Table 6: Patient with ID: 285, Age: 26, First Sexual Intercourse: 16, Number of Sexual Partners: 10, Number of pregnancies : 1 and Smokes: 0.

Methods	# Features for 90% Accuracy	Accuracy with 5 features(%)	Stability	Mean Feature Agree	Mean Rank Agree
Surrogates	4	**100**	0.02	0.6	0.20
KSHAP	**1**	06	0.03	**1.00**	**1.00**
TSHAP	**1**	05	0.02	**1.00**	**1.00**
SSHAP	**1**	07	0.02	**1.00**	0.80
LIME	**1**	03	**0.01**	0.60	0.30
TI	**1**	10	0.02	0.90	0.40
DICE	4	60	0.03	0.40	0.00

Table 7: Patient with ID: 263, Age: 29, First Sexual Intercourse: 10, Number of Sexual Partners: 4, Number of pregnancies: 5 and Smokes: 0.