Towards Reliable Post Hoc Explanations for Machine Learning on Tabular Data and their Applications

Célia Ayad

INSTITUT POLYTECHNIQUE DE PARIS

ÉCOLE POLYTECHNIQUE

IP PARIS

NNT : 2024IPPAX082

Thèse de doctorat

Towards Reliable Post Hoc Explanations for Machine Learning on Tabular Data

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École polytechnique

École doctorale n°626 : l'Ecole Doctorale de l'Institut Polytechnique de Paris (ED IP Paris)
Spécialité de doctorat: Mathématiques et Informatique

Thèse présentée et soutenue à **Palaiseau, le 25/09/2024**, par

Célia Wafa AYAD

Composition du Jury :

Albert Bifet Professor, Télécom Paris, IP Paris (LTCI)	Président
Alicia Troncoso Full professor, Universidad Pablo de Olavide (DSBD)	Rapporteur
Nistor Grozavu Full professor, CY Cergy Paris Université (ETIS)	Rapporteur
Jesse Read Full professor, École polytechnique, IP Paris (LIX)	Directeur de thèse
Benjamin Bosch Manager, Société Générale (MRM)	Invité

Abstract

As machine learning continues to demonstrate robust predictive capabilities, it has emerged as a very valuable tool in several scientific and industrial domains. Over recent years, machine learning models have proven to be indispensable across fields such as healthcare, finance, autonomous systems, and even climate modeling. These applications rely heavily on machine learning models to make accurate predictions, classify data, and optimize systems. However, as machine learning models evolve to achieve higher accuracy and better performance, they also become increasingly complex, requiring more parameters and more intricate architectures. In fact, some of the most accurate models, such as deep neural networks, can have millions of parameters and hidden layers that are difficult to interpret. This complexity can make the decision-making processes of machine learning models opaque, or what is commonly referred to as a "black box." In scenarios where these models are deployed to support decision-making in high-stakes areas like medical diagnosis or financial risk assessment, this lack of interpretability raises concerns. Understanding the inner workings of machine learning models has therefore become crucial. To establish trust in the predictions generated by these models, it is essential to provide insights into why a particular prediction is made. Trust and interpretability go hand in hand, especially in domains where decisions impact human lives or carry significant financial consequences. In response to these concerns, researchers in the field of explainable AI (XAI) have developed various explanation methods aimed at making machine learning models more transparent and interpretable. These explanation methods attempt to break down the complex processes within models and present them in ways that are comprehensible to users, including non-technical stakeholders. Explanation techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) have gained popularity because they provide post-hoc explanations that help clarify why a model arrived at a certain prediction. Despite these advances, explanation methods often fall short in accurately and consistently explaining model predictions in a manner that is intuitive to domain experts. This gap in effectiveness and usability makes it challenging for experts in fields like medicine, finance, and law to fully leverage these explanations for practical decision-making. It is crucial, therefore, to identify

the limitations and shortcomings of current ML explanations and work toward enhancing their reliability, interpretability, and ease of use. Moreover, given that many machine learning tasks are becoming increasingly data-intensive and the demand for machine learning integration is rising across industries, there is a growing need for explanation methods that not only provide transparency but do so in a way that is computationally efficient and cost-effective. Addressing these challenges will ensure that machine learning explanations can be trusted and widely adopted in real-world applications where decisions must be both accurate and justifiable. In this dissertation, we address these critical issues through two main research thrusts: First, we propose a comprehensive methodology for evaluating various explainability methods in the context of specific data properties, such as noise levels, feature correlations, and class imbalance. Our approach highlights how certain data characteristics can influence the effectiveness of different explainability techniques, offering practitioners and researchers a set of guidelines to help them choose the most suitable explainability method based on the specific characteristics of their datasets. By conducting extensive experiments across a range of datasets, we reveal where existing methods excel and where they fall short. In particular, we focus on use cases in healthcare, providing clinicians with personalized explanations for cervical cancer risk factors. These explanations are designed to align with the clinicians' desired properties, including ease of understanding, consistency across cases, and the stability of the explanations when input data changes slightly. This personalized approach ensures that domain experts can confidently use machine learning outputs to support their decision-making processes. Second, we introduce Shapley Chains, a novel explanation technique designed to address the lack of interpretability in multi-output predictions involving interdependent labels. In situations where labels depend on each other, such as sequential decisions in medical diagnoses or financial risk assessments, existing methods struggle to explain how features contribute to these chained predictions. Shapley Chains offer a new way to capture and explain the indirect contributions of features to subsequent labels in a prediction chain. For example, in healthcare, a feature such as patient age might not directly influence the final diagnosis, but it may have an indirect impact by affecting intermediate outcomes along the diagnostic chain. Shapley Chains allow users to trace these contributions throughout the sequence of predictions. Additionally,

we propose an enhancement to Shapley Chains called Bayes LIME Chains, which improves the robustness and reliability of the explanations by incorporating Bayesian inference techniques. This combination ensures that explanations remain consistent and reliable even in the presence of noisy or uncertain data.

Résumé

L'apprentissage automatique continue de démontrer de solides capacités prédictives et s'est révélé être un outil très précieux dans plusieurs domaines scientifiques et industriels. Ces dernières années, les modèles d'apprentissage automatique se sont révélés indispensables dans des domaines tels que la santé, la finance, les systèmes autonomes et même la modélisation climatique. Ces applications s'appuient largement sur les modèles d'apprentissage automatique pour faire des prédictions précises, classer les données et optimiser les systèmes. Cependant, à mesure que les modèles d'apprentissage automatique évoluent pour atteindre une plus grande précision et de meilleures performances, ils deviennent également de plus en plus complexes, nécessitant davantage de paramètres et des architectures plus complexes. En fait, certains des modèles les plus précis, tels que les réseaux neuronaux profonds, peuvent avoir des millions de paramètres et des couches cachées difficiles à interpréter. Cette complexité peut rendre les processus de prise de décision des modèles d'apprentissage automatique opaques, ou ce que l'on appelle communément une « boîte noire ». Dans les scénarios où ces modèles sont déployés pour soutenir la prise de décision dans des domaines à enjeux élevés comme le diagnostic médical ou l'évaluation des risques financiers, ce manque d'interprétabilité suscite des inquiétudes. Comprendre le fonctionnement interne des modèles d'apprentissage automatique est donc devenu crucial. Pour établir la confiance dans les prédictions générées par ces modèles, il est essentiel de fournir des informations sur les raisons pour lesquelles une prédiction particulière est faite. La confiance et l'interprétabilité vont de pair, en particulier dans les domaines où les décisions ont un impact sur la vie humaine ou ont des conséquences financières importantes. En réponse à ces préoccupations, les chercheurs dans le domaine de l'IA explicable (XAI) ont développé diverses méthodes d'explication visant à rendre les modèles d'apprentissage automatique plus transparents et interprétables. Ces méthodes d'explication tentent de décomposer les processus complexes au sein des modèles et de les présenter de manière compréhensible pour les utilisateurs, y compris les parties prenantes non techniques. Les techniques d'explication telles que SHAP (SHapley Additive exPlanations) et LIME (Local Interpretable Model-agnostic Explanations) ont gagné en popularité car elles fournissent des explications post-hoc qui aident à clari-

fier pourquoi un modèle est arrivé à une certaine prédiction. Malgré ces avancées, les méthodes d'explication ne parviennent souvent pas à expliquer avec précision et cohérence les prédictions du modèle d'une manière intuitive pour les experts du domaine. Ce manque d'efficacité et de facilité d'utilisation fait qu'il est difficile pour les experts de domaines comme la médecine, la finance et le droit d'exploiter pleinement ces explications pour la prise de décision pratique. Il est donc crucial d'identifier les limites et les lacunes des explications actuelles du ML et de travailler à améliorer leur fiabilité, leur interprétabilité et leur facilité d'utilisation. De plus, étant donné que de nombreuses tâches d'apprentissage automatique nécessitent de plus en plus de données et que la demande d'intégration de l'apprentissage automatique augmente dans tous les secteurs, il existe un besoin croissant de méthodes d'explication qui non seulement offrent une transparence, mais le font d'une manière efficace et rentable sur le plan informatique. Relever ces défis garantira que les explications de l'apprentissage automatique peuvent être fiables et largement adoptées dans des applications du monde réel où les décisions doivent être à la fois précises et justifiables. Dans cette thèse, nous abordons ces questions critiques à travers deux axes de recherche principaux : tout d'abord, nous proposons une méthodologie complète pour évaluer diverses méthodes d'explicabilité dans le contexte de propriétés de données spécifiques, telles que les niveaux de bruit, les corrélations de caractéristiques et le déséquilibre des classes. Notre approche met en évidence comment certaines caractéristiques des données peuvent influencer l'efficacité de différentes techniques d'explicabilité, offrant aux praticiens et aux chercheurs un ensemble de lignes directrices pour les aider à choisir la méthode d'explicabilité la plus appropriée en fonction des caractéristiques spécifiques de leurs ensembles de données. En menant des expériences approfondies sur une gamme d'ensembles de données, nous révélons où les méthodes existantes excellent et où elles échouent. En particulier, nous nous concentrons sur les cas d'utilisation dans le domaine de la santé, en fournissant aux cliniciens des explications personnalisées sur les facteurs de risque du cancer du col de l'utérus. Ces explications sont conçues pour s'aligner sur les propriétés souhaitées par les cliniciens, notamment la facilité de compréhension, la cohérence entre les cas et la stabilité des explications lorsque les données d'entrée changent légèrement. Cette approche personnalisée garantit que les experts du domaine peu-

vent utiliser en toute confiance les résultats de l'apprentissage automatique pour soutenir leurs processus de prise de décision. Deuxièmement, nous présentons les chaînes de Shapley, une nouvelle technique d'explication conçue pour remédier au manque d'interprétabilité dans les prédictions à sorties multiples impliquant des étiquettes interdépendantes. Dans les situations où les étiquettes dépendent les unes des autres, comme les décisions séquentielles dans les diagnostics médicaux ou les évaluations des risques financiers, les méthodes existantes ont du mal à expliquer comment les caractéristiques contribuent à ces prédictions enchaînées. Les chaînes de Shapley offrent une nouvelle façon de capturer et d'expliquer les contributions indirectes des caractéristiques aux étiquettes ultérieures dans une chaîne de prédiction. Par exemple, dans le domaine de la santé, une caractéristique telle que l'âge du patient peut ne pas influencer directement le diagnostic final, mais elle peut avoir un impact indirect en affectant les résultats intermédiaires tout au long de la chaîne de diagnostic. Les chaînes de Shapley permettent aux utilisateurs de retracer ces contributions tout au long de la séquence de prédictions. De plus, nous proposons une amélioration des chaînes Shapley appelées chaînes Bayes LIME, qui améliorent la robustesse et la fiabilité des explications en incorporant des techniques d'inférence bayésienne. Cette combinaison garantit que les explications restent cohérentes et fiables même en présence de données bruyantes ou incertaines.

Contents

List of Figures

List of Tables

List of Abbreviations

The following list describes several symbols, abbreviations, and acronyms that will be later used within the body of this thesis.

AI	Artificial Intelligence
BMA	Bayesian Models Averaging
BR	Binary Relevance
CC	Classifier chains
DiCE	Diverse Counterfactual Explanations
DL	Deep Learning
DNN	Deep Neural Networks
DT	Decision Trees
KNN	k-Nearest Neighbors
LIME	Local Interpretable Model-agnostic Explanations
LR	Logistic Regression
ML	Machine Learning
MLP	Multilayer Perceptron
RF	Random Forest
SHAP	Shapley Additive exPlanations
SVM	Support Vector Machine
TI	Tree Interpreter
XAI	Explainable AI

Chapter 1

Introduction

1.1 Context and Motivation

Artificial intelligence has become ubiquitous in our daily lives. From personalized recommendations on streaming platforms to autonomous vehicles on the road, AI systems are increasingly integrated into the very fabric of our existence.

Definition 1.1.1. "Artificial Intelligence refers to the simulation of human intelligence in machines that are programmed to think and mimic human actions." [70]

As AI technology advances, it becomes crucial to understand and govern every decision made by the machine learning (ML) systems. The urgency to achieve this understanding stems from the recognition that the capabilities of AI are evolving at an exponential rate, and we may soon reach a point where the transition into a more autonomous AI era becomes a reality.

Definition 1.1.2. "Machine Learning is the field of study that gives computers the ability to learn without explicitly being programmed." [82]

One of the significant challenges in this context is the "black box" nature of many ML models. Deep learning (DL) models for example, while highly effective, can be difficult to interpret. They make predictions and decisions based on complex interactions and the logic behind their conclusions may not be easily apparent. This lack of transparency raises concerns about the

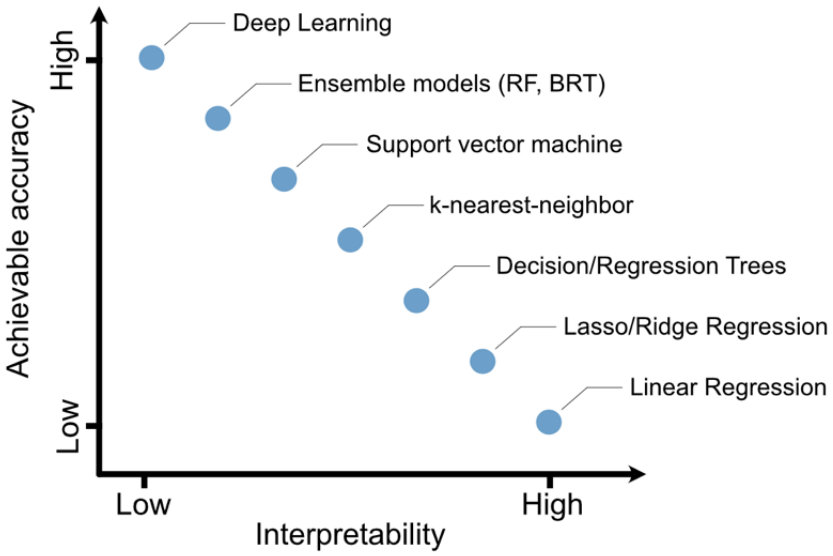


Figure 1.1: Trade-off between achievable accuracy and interpretability of ML and DL algorithms. DL can achieve the highest accuracy but shows the lowest interpretability [73].

risks associated with the decision-making, especially when these decisions impact critical domains like finance and healthcare.

Definition 1.1.3. "white box model refers to all recognized interpretable machine learning models, e.g. the models that are understandable for humans. The white box models are, for example: decision trees, linear models, rule based models, etc." [32]

Definition 1.1.4. "black box model refers to a machine-learning obscure model, whose internals are either unknown to the observer or they are known but uninterpretable by humans." [32]

Therefore, companies like Société Générale[1] are participating in the development [2] of transparent AI systems for several reasons, ranging from regulatory compliance to risk management, customer trust, ethical considerations and competitive advantage. Regulatory compliance requires banks to ensure transparency and explainability of their AI-enabled operations, reducing the risk of legal repercussions. Improving transparency also fa-

[1] https://www.societegenerale.com/en/responsability/ethics-and-governance
[2] https://github.com/MAIF/shapash

cilitates risk management by providing visibility into decision-making processes, enabling proactive identification and mitigation of potential biases or errors. Additionally, transparent AI systems build customer trust by providing clear explanations of how decisions are made, fostering strong relationships, and increasing customer engagement. From an ethical perspective, transparency ensures that AI-based decisions are consistent with societal values and ethical principles. Finally, transparent AI systems provide competitive advantage, drive innovation, attract customers and position banks as industry leaders.

To bridge the gap between the inherent complexity of machine learning models and human understanding, various approaches known as explainability and interpretability have been proposed [77, 56, 8]. These approaches encompass a diverse range of techniques and methods aimed at unveiling the black box nature of machine learning models, making their decisions comprehensible and transparent. These efforts are fundamental, as they enable the generation of interpretable explanations that shed light on the underlying logic of these models predictions.

Definition 1.1.5. "Interpretability is the ability to present the model in terms understandable by humans." [22]

Definition 1.1.6. "Explainability is an attribute of a machine-learning model that enables humans to understand the model's rationale for its outcome." [68]

Although a consensus on a singular definition of explainability remains elusive, and the distinction between explainability and interpretability is subject to ongoing debate [47, 6], the main objective shared by researchers in this field is to unlock the inner workings of black box models by addressing two fundamental questions: why and how, in order to provide a comprehensive understanding of the decisions made by machine learning models. Researchers are driven by the imperative need to generate explanations that can be conveyed in various forms, such as the importance [56] or effect [30] of each feature in the decision-making process or articulating decision rules that empower users to comprehend the model's reasoning.

Definition 1.1.7. "An explanation is additional meta information, generated by an external algorithm or by the machine learning model itself, to

describe the feature importance or relevance of an input instance towards a particular output classification." [21]

Various feature importance methods such as SHAP [56] and LIME [77] have been proposed to explain black box models including deep neural networks and ensemble trees such as Random Forest (RF) and XGBoost. These methods despite their differences in computing the feature importance distribution, have been widely used in real world applications such as in finance and healthcare. However, the difference in their attribution mechanisms [47] and the lack of ground truth feature importance to which the generated feature importance can be compared to and be validated, made these methods subject to criticism about their usefulness [81]. The feature independence and interactions hypothesis is one example of the limitations of these methods that is hard to satisfy in real-wold scenarios, specifically in a multi-output setting. These feature relationships can be represented with directed acyclic graphs such as classifier chains.

Definition 1.1.8. "Classifier Chains (CC) is an ensemble learning technique used in multi-label classification tasks, where multiple binary classifiers are trained in a chain-like fashion. Each classifier in the chain is responsible for predicting the presence or absence of a single label, and the predictions of earlier classifiers in the chain are used as input features for subsequent classifiers. Classifier Chains are effective for capturing label dependencies and have been shown to improve classification performance in scenarios where labels are correlated." [75]

1.2 Challenges of the Explainability Methods

Explainability methods in machine learning play a pivotal role in bridging the gap between the power and opacity of complex models, providing insights into their decision-making processes. However, these methods face several challenges as they struggle with inherent limitations and complexities. Some of these challenges include method diversity, feature interactions, absence of Ground Truth for Validation, scalability, regulatory compliance, robustness and security.

Method diversity. The field of explainability is diverse, with a multitude of techniques and approaches, each with its strengths and limitations. Disagreements among these methods on the ideal explanation or feature importance distribution can be confusing and challenge the standardization of best practices. Practitioners must carefully select the most appropriate method for their specific application, considering trade-offs between accuracy and comprehensibility.

Human-understandable and context and domain-specific explanations. Four, even when explanations are generated, ensuring that they are human-understandable and context and domain-specific remains a challenge. Balancing the depth of information and simplicity for users who may not have technical expertise is a fine line that must be walked carefully. For example, when interpreting medical diagnoses made by a machine learning model, providing explanations that are clinically meaningful and align with medical knowledge requires understanding not only the input features and model predictions but also the broader context of the patient's condition and relevant medical literature.

Feature Interactions. Many explainability methods make the simplifying assumption of feature independence, which often doesn't hold in real-world datasets. In reality, features can be correlated and exhibit complex interactions that are difficult to capture. Methods that fail to account for these interactions may provide incomplete or inaccurate explanations. Consider a machine learning model deployed by a bank to assess the creditworthiness of loan applicants. The model takes various features into account, such as income, credit history, debt-to-income ratio, and employment status, to make predictions about whether an applicant is likely to default on a loan.

In real-world scenarios, these features are often correlated and exhibit complex interactions. For example, applicants with higher incomes may also have better credit histories and lower debt-to-income ratios. Similarly, applicants who are employed full-time may be more likely to have stable incomes and lower default risks.

However, many explainability methods may make the simplifying assumption of feature independence, treating each feature as if it operates in

isolation from the others. This assumption may lead to incomplete or inaccurate explanations of the model's predictions. For instance, an explainability method that attributes a loan rejection to a low credit score without considering the applicant's high income and stable employment status may provide an incomplete explanation. In reality, the applicant's credit score may have been adversely affected by other factors, such as a recent job loss or unexpected medical expenses, which are not captured by the explanation.

No ground truth for validation. A fundamental challenge is the absence of a definitive ground truth for feature importance. In most cases, the true importance of features in a prediction is unknown, making it challenging to validate the accuracy of the explanations generated. This lack of validation can lead to debates over the efficacy and reliability of explainability methods.

Despite these challenges, the pursuit of more transparent and interpretable machine learning models is crucial for building trust, ensuring accountability, and advancing responsible AI applications in various fields. Overcoming these challenges will require continued research and collaboration to develop more robust and reliable explainability methods that can empower stakeholders and end-users to make informed and ethical decisions in the era of AI.

In order to contribute to this goal, we dedicate this thesis to explore the last three challenges faced by explainability methods and their implications for the field, by addressing two key research questions described in the next section (Section 1.3).

1.3 Research Questions

The focus of this thesis is on contributing to the ongoing efforts in the explainability field by addressing four challenges, including method diversity, the generation of human-understandable and context and domain-specific explanations, inclusion of feature/label interactions in the explanation design and evaluation metrics for XAI methods. Specifically in Sections 3.1, 3.2, 4.1 and 4.2, we sought to answer the following research questions:

The first question focuses understanding explanations diversity with respect to the data properties such as noise levels, feature correlations and

class imbalance.

> *RQ1: Can we discover a relationship between the local explanations computed in existing explainability methods and the data properties like noise level, feature correlations and class imbalance, draw conclusions about how each explainability method handles each of these data properties and make recommendation on when the user should trust or not these explanations based on the properties of their data?*

Several local explanation methods exist, each offering unique insights into the behavior of machine learning models on a per-instance basis. However, these methods often produce different explanations for the same prediction, leading to what is known as the disagreement problem.

While various research efforts have examined this issue, there remains a lack of assessment of these explanations based on underlying data properties. Yet, these properties, such as data distribution, feature correlations, and sample size, can significantly influence explanation generation. Investigating the relationship between explanations and data properties is crucial as it can provide valuable insights into the reliability, consistency, and robustness of local explanations. Understanding how different data characteristics impact explanation methods can lead to the development of more accurate and trustworthy interpretability techniques.

Additionally, we tackle the problem of human understandable and domain specific explanations by leveraging personalized, consistent and simple explanations of the risk factors of the cervical cancer and help clinicians to understand each patient specific risk factors. Cervical cancer stands as one of the most devastating and fatal cancers for women worldwide.

Despite numerous predictive models developed to identify women at risk, a comprehensive understanding of the underlying risk factors remains elusive. To address this gap, we use explainability methods as invaluable tools to assist clinicians in comprehending model predictions for cervical cancer on an individual patient level. By utilizing these methods, clinicians can gain insight into the factors contributing to a patient's risk, thereby enabling more informed decision-making regarding screening, prevention, and treatment strategies.

Understanding the individual risk factors associated with cervical cancer

is beneficial for both women and clinicians alike. For women, it empowers them with personalized information about their risk profile, allowing for proactive measures to mitigate risk and potentially prevent the development of cervical cancer. For clinicians, it facilitates more targeted and tailored approaches to patient care, leading to improved patient outcomes and overall healthcare delivery in the fight against cervical cancer.

The second question concerns the explanation methodology for multi-output tasks where the labels are interdependent.

> *RQ2: Can we design a post hoc and model agnostic local explainability method that can take into consideration label interactions when computing feature importance in a way to make the feature importance more complete compared to when the labels are independent and illustrate how these new feature importance can help understand the chaining of the labels ?*

Recent studies [2, 33, 94, 55] have shown that considering label interdependencies when predicting multi-output tasks yields superior performance compared to methods that neglect these dependencies. Despite the availability of numerous explainability methods capable of elucidating single or multi independent output predictions, many fail to account for label dependencies. Therefore, there is a pressing need for explainable artificial intelligence methods that incorporate label dependencies into explanation generation. These explanations, which we can term as indirect feature importance on label prediction, have the potential to enhance our understanding of model predictions like classifier chain predictions.

Classifier chains [75] are a technique used in multi-label classification where binary classifiers are trained sequentially, with each classifier considering the predictions of the previous ones as additional features. By computing indirect feature importance in the chaining of labels, these methods can provide insights into the factors influencing the prediction of each label in the chain, considering the dependencies between them, thus which chain order is best. This approach is beneficial as it enables a more comprehensive understanding of model behavior and facilitates informed decision-making in various applications, including multi-label classification tasks. Additionally, it empowers users to identify critical features that contribute to the prediction of specific labels within a chain, thereby improving multi-output

model interpretability and trustworthiness.

Furthermore, we suggest a way to make the explanations provided by Shapley Chains more reliable. We achieve this by measuring the uncertainty in how the explanations are generated, using local surrogate models. Shapley Chains uses local surrogate models such as Kernel SHAP and LIME to explain multi-output predictions by evaluating the importance of direct and indirect features on labels, and presents a robust tool for explaining complex models. Despite LIME's popularity, its instability often gives rise to inconsistent explanations.

Measuring the uncertainty around local explanations is crucial to building confidence in methods such as LIME and Kernel SHAP. Although Bayes LIME and Bayes SHAP attempt to address uncertainty, they lack a reliable assessment of uncertainty intervals and, if used for multiple-output explanations, they both ignore indirect label features . We propose Bayes LIME Chains to address the challenges mentioned above. The Bayes LIME make it possible to calculate the indirect importance of features as well as the measurement of credible intervals around these explanations. Unlike Bayes LIME and Bayes SHAP, we evaluate the comparison with the importance of ground truth features on synthetic datasets with multiple outputs.

1.4 Thesis Outline

The challenging properties of explanability methods as presented in Section 1.2, pose significant barriers on the usefulness of these methods and therefore the trust that the users may have in the field of explainable AI in general.

While several studies focus on explaining Deep Learning architectures in various tasks and domains of applications, there is yet few studies on when and for what contexts to use the explainbility methods designed to explain the predictions made by ensembles of trees such as Random forest and XGboost. At the same time, there is room for improvements in existing approaches, in terms of feature importance attribution, as well as objective metrics to validate the explanations.

Therefore, our study consists in providing understanding of the existing explainable methods with regards to data properties such as feature correlations, presence of irrelevant variables, noise and class imbalance on various synthetic and real world datasets, in assessing local explanations on the cer-

vical cancer prediction where human understanding is paramount, as well as designing and developing a new method that takes label interactions in the explanation design in order to ensure a complete representation of the feature/output interactions by computing the direct and indirect feature importance.

These indirect feature importance can be important in many cases, for example, in healthcare diagnosis using multi-label classification, direct feature importance can identify clinical biomarkers or symptoms directly associated with each diagnosed condition. Indirect feature importance can reveal how the presence of certain conditions or comorbidities influences the prediction of other related conditions, therefore indirect effects of some clinical biomarkers or symptoms on the related conditions, providing insights into disease interactions and patient health profiles.

We next present the organization of the presentation of the studies conducted in terms of this thesis, as well as the theoretical background information provided for understanding the topics discussed in each chapter.

Introduction and Background. In Chapter 1, we presented some of the most dominant challenges on explainability of machine learning models for real-world data. In this chapter, we also provide the thesis organization and an outline of the main chapters and topics that are discussed. In Chapter 2, we provide some key definitions and notations for the explainability field and an overview of existing methods. Those are important for understanding the background of the existing methods in this field and contain information about modules and properties to which we will refer in the main chapters that follow.

Chapter 3 presented in two section address the research question *RQ1*.

Feature Importance Depends on Properties of the Data: Towards Choosing the Correct Explanations for Your Data and Decision Tree based Model. In order to ensure the reliability of the explanations of machine learning models, it is crucial to establish their advantages and limits and in which case one should use each of these methods, especially with regards to the data properties. However, the current understanding of when and how each method of explanation can be used is insufficient and for which data properties each of these methods can be used. To fill this

gap, in Section 3.1 we perform an empirical evaluation by synthesizing multiple datasets with different properties, e.g., feature correlation and noise. Our main objective is to assess make recommendations of when each of the feature importance estimates provided by local explanation methods should be used, and when users should be careful and why.

Local Explainability Methods for Cervical Cancer Risk Assessment. 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. In Section 3.2, we use local explainability methods to assess then recommend which method a clinician should choose to explain and understand the cervical cancer risk factors for each patient based on their specific profiles and a set of fixed desired properties such as compactness and stability.

Lastly, Chapter 4 presented in two sections address the research question *RQ2*.

Shapley Chains: Local Explanations for Multi-output Decisions. in Section 4.1, we present Shapley Chains, which is a post-hoc model agnostic local explainability method designed to explain a multi-output classifier outputs using the Shapley value to compute feature importance. Shapley Chains attributes feature importance to all features that directly or indirectly contribute to the prediction of a given output, by tracking all the related outputs in the given chain order. Compared to existing methods such as Shapley flow that is restricted to causal graphs, we show a complete distribution of feature importance scores in multi-output synthetic and real-world datasets. Our method is model agnostic, meaning that it can be applied on any type of graphical model that represent complex feature and label interactions such as classifier chains.

Bayes LIME Chains. Shapley Chains incorporate label interdependence into the explanation design process to ensure that explanations reflect the interdependence of multiple outputs. This process has improved the explanation attribution in the context of multi-label and unveiled how the features contribute to the outputs directly and through subsequent related outputs

in different chain orders of the classifier chain. Therefore, in Section 4.2, we extend our method Shapley Chains to include a measurement of the uncertainty of the direct and indirect feature importance generated by Shapley Chains. Measuring uncertainty in the local explanations is a way to increase robustness and reliability of the generated explanations in general and in the direct and indirect feature importance computed by Shapley Chains in particular.

Conclusion. Chapter 5 summarizes our contributions, addresses the limitations of each proposed approach, and outlines perspectives for future work. In this concluding chapter, we reflect on the key findings and novel insights uncovered throughout our research journey and highlight the significance of our contributions to the field. Furthermore, we critically evaluate the limitations inherent in each proposed approach, acknowledging the challenges and constraints encountered during the research process.

1.5 Publications

The research conducted in this thesis has been submitted to and published in scientific journals within the machine learning field.

1. Ayad, Celia Wafa, Thomas Bonnier, Benjamin Bosch, and Jesse Read. "Shapley chains: Extending Shapley values to classifier chains." In International Conference on Discovery Science, pp. 541-555. Cham: Springer Nature Switzerland, 2022.

2. Ayad, Celia Wafa, Thomas Bonnier, Benjamin Bosch, Jesse Read, and Sonali Parbhoo. "Which Explanation Makes Sense? A Critical Evaluation of Local Explanations for Assessing Cervical Cancer Risk Factors." The Machine Learning for Healthcare Conference, 2023.

3. Ayad, Celia Wafa, Thomas Bonnier, Benjamin Bosch, and Jesse Read. "Feature Importance Depends on Properties of the Data: Towards Choosing the Correct Explanations for Your Data and Decision Trees based Models.", submitted to the Machine Learning Journal.

4. Ayad, Celia Wafa, Thomas Bonnier, Benjamin Bosch, and Jesse Read. "Bayes LIME Chains: Expanding the Scope of Shapley Chains with Bayesian Inference", to be submitted.

Chapter 2

Notation and Background

This section is dedicated to the notation that we use in the subsequent chapters and background information vital to understanding the intricacies of the research presented in this thesis.

2.1 Notations

This set of notations will be consistently used throughout the thesis to represent the key entities and mathematical operations in the context of machine learning.

- X is an input set of d-dimensional feature vectors;
- $\mathbf{x} \in X$ is an instance, described by a feature vector $\mathbf{x} = [x^1, \ldots, x^d]$
- $\mathbf{x}'$ is a perturbed input instance;
- Y is an output set of m-dimensional target vectors;
- D is the set of features;
- S is a random subset of D;
- Each instance $\mathbf{x} \in X$ is associated with an output vector $\mathbf{y} = [y^1, \ldots, y^m]$, $\mathbf{y} \in Y$;
- x_i denotes the i'th input instance $\mathbf{x}$;
- x^j denotes the j'th feature of input instance $\mathbf{x}$;

- x_i^j denotes the j'th feature of the i'th input instance $\mathbf{x}$;
- $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ is a dataset of n samples;
- $f : X \to Y$ is a predictive model that learns a single output in Y;
- $\mathcal{L}$ is a loss function;
- $g_\theta : X \to Y$ is an interpretable surrogate single-output model with parameters θ, such as a logistic regression or a decision tree;
- $h : X \to Y$ is a multi-output set of classifiers, $h = [h_1, h_2, \ldots, h_m]$;
- $\hat{\mathbf{y}} = h(\mathbf{x})$ is a prediction of multi-output model h for instance $\mathbf{x}$, $\hat{\mathbf{y}} = [\hat{y}^1, ..., \hat{y}^m]$;
- $P(\mathbf{y}|\mathbf{x}; \theta)$ is a conditional probability of the output $\mathbf{y}$ given the instance $\mathbf{x}$;
- $\mathbf{y}'$ is a perturbed output instance;
- $\mathcal{N}(\mu, \sigma^2)$ is a Gaussian distribution with mean μ and variance σ^2;
- $P(Y; \theta) = \prod_{k=1}^m P(y^k)$: Marginal output independencies;
- $P(Y|X; \theta) = \prod_{k=1}^m P(y^k|X, y^1, ..., y^{k-1})$: Conditional output interdependencies;
- ϕ_{x^j} denotes the feature contribution (also referred to as feature importance) of the feature x^j;
- $\phi_{x^j}(y^k)$ denotes the feature contribution (also referred to as feature importance) of the feature x^j to the prediction of the output y^k;
- $\phi^\star_{x^j}$ is the true value (ground truth) feature contribution of x^j to the prediction of the single output $\hat{\mathbf{y}}$;

2.2 Background

The background section of this thesis provides a foundational understanding of essential concepts and methodologies. We begin by examining decision tree-based models, highlighting their significance in classification and regression tasks. We then focus on multi-output learning, exploring how models

predict multiple variables simultaneously. Next, we investigate explainability methods, which explain black box models by providing interpretable insights. Finally, we discuss the evaluation of explanations, focusing on metrics and methodologies used to assess their quality and trustworthiness. This overview sets the stage for our contributions in subsequent chapters.

2.2.1 Decision Tree Methods

Tree-based models stand out as some of the most prevalent machine learning techniques that can be used both for regression and classification tasks. A decision tree [11] is a hierarchical graph structure and consists of nodes connected by directed edges. Every node may have outgoing edges connecting it to its children, known as the leaves. The top node where the first split takes place is called the root.

The decision tree is defined by a set of rules represented by the internal nodes of the tree. Each internal node tests a specific feature against a split criteria, directing the flow to the left or right child node based on the outcome of the test. Leaf nodes provide the final output, either a class label in classification or a continuous value in regression. Formally, the decision tree output is given by: The construction of the decision tree involves recursively partitioning the input space into regions, each associated with a unique set of rules. The decision tree is trained to optimize the purity of these regions in classification or the reduction of variance in regression, resulting in a predictive model that is interpretable and easy to understand. Fig 2.1 [1] illustrates a classification decision tree in which the task is to predict whether a day is suitable for playing outside based on three characteristics, namely humidity level, weather, and whether it is windy or not.

Decision trees offer an intuitive and interpretable framework for understanding the decision-making process, making them accessible even to individuals with limited expertise in the field. However, while these models provide transparency in their rule-based structure, they often fall short in accuracy, particularly when confronted with complex datasets or tasks with intricate feature-target relationships. As a solution, random forest, a popular ensemble method, has emerged as a preferred alternative to decision trees due to its superior performance in handling complexity.

[1] `https://www.baeldung.com/cs/decision-trees-vs-random-forests`

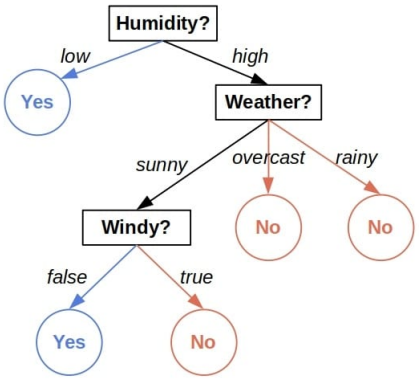


Figure 2.1: An example of a binary classification decision tree.

A random forest [10] is an ensemble learning method that constructs multiple decision trees. Each tree is built on a random subset of features, sampled with bootstrap sampling. The final prediction is a major vote for all tree predictions in the classification scenario or their mean average in the regression case. The combination of these randomization decision trees helps to decorrelate the individual trees, making the random forest less prone to overfitting and improving its overall predictive performance. While building multiple trees instead of a single one may seem more computationally expensive, taking a random subsample of features per tree alleviates this drawback. Nonetheless, random forest introduces its own layer of complexity, necessitating interpretation and understanding to effectively harness its full potential. This highlights the ongoing challenge in machine learning to strike a balance between model transparency and predictive power, ultimately emphasizing the importance of selecting the most suitable approach for the given task and dataset.

Figure 2.2 [2] illustrates the above example with three decision trees built on two features each. Here the overall prediction should be "yes" as the majority of the individual trees predict that the considered day is suitable to play outside.

2.2.2 Multi-Output Learning

In the field of machine learning, there exists a diverse range of models designed to tackle various prediction tasks. One fundamental distinction lies

[2] `https://www.baeldung.com/cs/decision-trees-vs-random-forests`

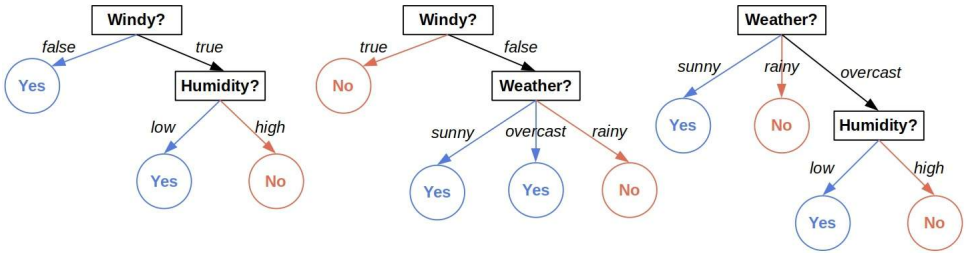


Figure 2.2: An example of a a random forest where each decision tree is built on 2 features.

in the number of outputs these models are capable of predicting. Some models are designed to predict only a single output variable, encapsulating tasks such as classification and regression. On the other hand, there are models explicitly engineered to handle scenarios where multiple output variables are involved. These multi-output models [100] are adept at capturing complex relationships and dependencies between input features and multiple target variables simultaneously, enabling tasks such as multi-label classification and regression.

One common example of multi-output classification is image tagging, where the task involves assigning multiple labels to an image. For instance, consider a scenario where an image classification model is trained to identify objects and activities within an image. Instead of predicting a single label for the entire image (e.g., "cat" or "dog"), the model may be required to predict multiple labels simultaneously (e.g., "cat," "outdoor," "playing"). In this case, the model is performing multi-output classification by predicting a binary or multi-class label for each distinct attribute or concept present in the image. This approach enables the model to capture the complex and diverse nature of visual content, providing more detailed and informative annotations for downstream tasks such as content-based image retrieval, image understanding, and automated tagging systems.

A multi-output binary classifier denoted as h learns a vector of base classifiers:

$$h(\mathbf{x}) = [h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_m(\mathbf{x})]$$

for a given instance $\mathbf{x}$ and returns a binary vector of predicted values:

$$\hat{\mathbf{y}} = [\hat{y}^1, \hat{y}^2, \ldots, \hat{y}^m]$$

The main challenge in multi-output learning lies in managing the inherent complexity of the output space and the interdependencies among the outputs. While each target can be addressed as an independent single-output problem often referred to as marginal learning (and as *binary relevance* [29, 94] in the case of classification), without explicitly considering the dependencies or relationships with other tasks, designing classifiers that learn jointly multiple outputs by incorporating these output dependencies makes it possible to better represent the relationships in the data (between outputs, therefore between features and outputs). In the first case, the m models are trained separately. This approach allows to different algorithms or hyperparameters to be used for each model, based on specific characteristics of each task. However, completely ignoring interdependencies between the targets can lead to suboptimal performance or prediction of impossible combinations [2]. While in the joint learning, the goal is to maximise model accuracy by considering all tasks or variables together, capturing the dependencies and interactions between them and maximizing the joint probability:

$$P(\mathbf{y}|\mathbf{x}) = \prod_{k=1}^{m} P(y^k|\mathbf{x}, y^1, ..., y^{k-1}) \tag{2.1}$$

Classifier chains. A classifier chain is one multi-output classification method that incorporates the chaining of the outputs in the learning of multiple classifiers (one classifier for each output, also referred to as *base classifier*). The choice of the base model depends on the characteristics of the problem and the desired performance.

The initial idea of the chaining approach, for classification [75], was to arrange per-target models in a chain, such that the previous labels are used to train each next model in the training phase and the output prediction of one model becomes an additional feature for the subsequent models in the prediction phase. Classifier Chains have proved to have high predictive performance and are widely known as one of the state-of-the-art techniques for multi-label modeling [75].

As opposed to independent modeling such as the binary relevance in

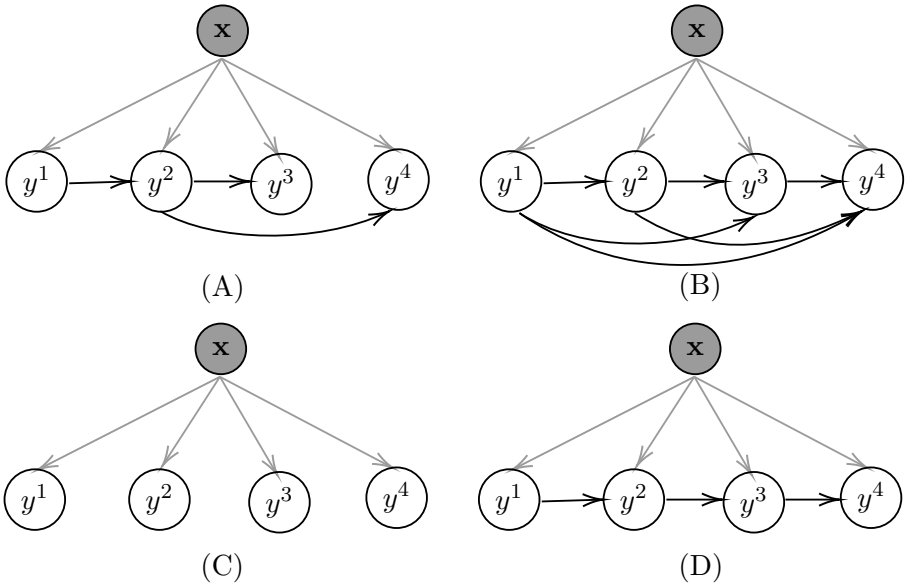


Figure 2.3: Different chain structures for a problem with $m = 4$ outputs.

classification, the chaining approach allows the model to capture the dependencies and interactions between the target variables. The chaining method is exactly an expression of Eq. 2.1, if expressed according to the chain rule of probability (i.e., Fig. 2.3 (B) as a probabilistic graphical model representation). That is one reason why conditional dependencies are interesting in this context. However, a classifier chain is not faithful to a 'proper' inference procedure, and rather takes a greedy approach to inference, plugging in predictions as observations; and proceeds much as a forward pass across a neural network. This creates some ambiguity between how much effect is gained from probabilistic dependence (as a probabilistic graphical model would) and feature effect (as one encounters via the latent layers of deep learning). Although discussion has been ongoing e.g., [76, 75], there is not yet a consistent understanding in practice of what role a prediction plays as a feature to another label.

The order of the chain has an impact on the model's ability to learn interdependencies between the targets and thus predictive performance. Different approaches have been suggested to optimize chain order including using correlation to build the best structure [60].

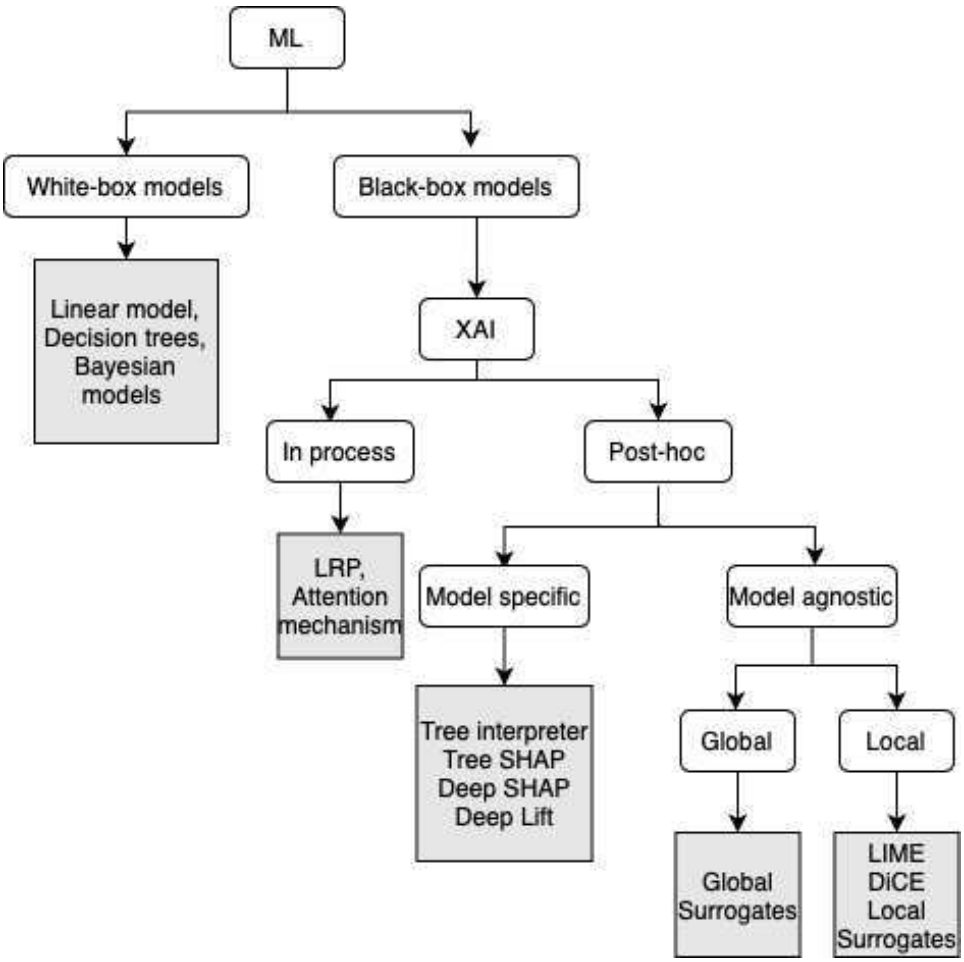


Figure 2.4: The taxonomy of XAI methods with examples.

2.2.3 Explainability Methods

In this section, we provide a comprehensive literature review on Explainable Artificial Intelligence. We begin by presenting a taxonomy of XAI approaches, categorizing these methods based on their underlying principles, applications, and interpretative paradigms. Subsequently, we focus on an in-depth analysis of the most prominent and state-of-the-art XAI approaches, highlighting both their strengths and limitations. Figure 2.4 illustrates the taxonomy of XAI methods.

Taxonomy of Explainability Methods

Various terms are employed to illustrate the significance or impact of input features on model predictions. These terms include feature importance, contribution, participation, and feature effect, each offering distinct insights into the relationship that the features may bring to the model decision making. To illustrate the significance of each of these terms, let's consider the example of a credit risk assessment model.

Feature importance quantifies the relative influence of individual features on the model's predictions. In this case, feature importance analysis might reveal that the borrower's credit score is the most important factor affecting the model's predictions of loan approval or rejection. This indicates that the credit score has the highest impact among all features considered in the model. By calculating the feature importance, one can conclude the feature ranking and then select the most important features, which can also be used as a dimensionality reduction tool.

On the other hand, feature contribution zooms in on the specific effect of a single feature on a particular prediction. For instance, suppose the model predicts loan approval for a specific applicant. The contribution of the borrower's credit score would indicate the extent to which it affects the likelihood of loan approval for that particular applicant. A higher credit score would likely contribute positively to the likelihood of loan approval, while a lower credit score might have a negative contribution.

Feature Participation assesses the involvement of each feature in the decision-making process of the model. In this case, the participation of the borrower's credit score would indicate its active role in the model's decision to approve or reject a loan application. A high participation value for the credit score suggests that it plays a significant role in determining the final decision.

Finally, feature effect can either show the single or the collective impact of multiple features on model predictions. Considering the borrower's credit score alongside other relevant features such as income, employment history, and debt-to-income ratio, the feature effect would illustrate how these factors collectively influence the model's predictions of loan approval or rejection. The feature effect provides insights into how changes in multiple features simultaneously impact the overall outcome of the model's predictions.

White box vs black box models. White box models as defined in 1.1.3, often referred to as interpretable models, are characterized by their transparency and ease of interpretation. These models include linear regression, decision trees, and logistic regression, among others. Their appeal lies in their simplicity, as they are built on straightforward, human-understandable principles. Some of their advantages include: transparency such that each feature's contribution to the model's output can be easily understood, interpretability which makes it easier to derive actionable insights and identify factors influencing predictions, trust that enables users to validate and verify the model's decisions, also crucial in domains like healthcare and finance, and finally their alignment with regulatory requirements, as they can provide clear explanations for their decisions.

On the other hand, black box models (Definition 1.1.4), such as deep neural networks, ensemble methods, and support vector machines, are known for their complexity and opacity. Their decisions are derived from intricate mathematical computations that are challenging for humans to intuitively grasp. These complex models are widely used for their high predictive accuracy, often outperforming white box models on complex tasks, and their ability to capture intricate relationships in data that might be beyond the capacity of simpler models, making them well-suited for image recognition, natural language processing, and other complex domains. These models can also generalize from data effectively, adapting to various patterns and making them versatile in a wide range of applications.

Post hoc vs in-process methods. While simple models such as decision trees offer some degree of interpretability, several efforts have proposed techniques that interpret the local and global decisions made by the black box models such as deep neural networks. These techniques can either include the interpretation process within the model learning also known as in-process methods, or explain its decision in a post hoc manner, meaning that these are applied after a machine learning model has been trained. These post hoc methods are often model-agnostic, meaning that they can be applied to a wide range of machine learning models without requiring changes to the model architecture and allow for analyzing existing models without the need to modify the training process, making them suitable for legacy systems. The generated post hoc explanations are typically tailored

for end-users, ensuring that the explanations are human-understandable and relevant to the specific context. Post hoc techniques include feature importance scores, perturbation-based methods, and surrogate models, based on the specific use case.

In-process methods, on the other hand, incorporate explainability considerations directly into the machine learning model's development and training process. The model is designed to be interpretable from the outset and embed transparency within its architecture, resulting in a model that is inherently easier to understand and explain. By avoiding the need for additional post hoc steps, in-process methods can be computationally more efficient and the explanations provided by such methods are consistent with the model's design and decision-making process, reducing the risk of explanation-model inconsistencies.

Model agnostic vs model specific methods. Model-agnostic methods are designed to provide explanations for a wide range of machine learning models, regardless of their architecture, complexity, or learning algorithm. These methods are often applied in a post hoc manner and can be used with any machine learning model, from decision trees to deep neural networks, without requiring modifications to the model itself. They are valuable for analyzing and explaining existing models, making them suitable for auditing or improving the transparency of legacy systems. Model-agnostic explanations are often designed with the end-users in mind, emphasizing human-understandable insights that enhance user trust and confidence.

On the other hand, Model-specific methods, as the name suggests, are tailored to a particular machine learning model or family of models. These methods are often developed in conjunction with the model's architecture or learning process and can build interpretability directly into the model's architecture, resulting in inherently interpretable models from the outset. By eliminating the need for additional post hoc steps, model-specific methods can be computationally efficient and more closely aligned with the model's design. Same as for in-process explainability methods, the explanations provided by model-specific methods are also consistent with the model's decision-making process, reducing the risk of inconsistencies between the model and its explanations.

Local vs global methods. Local methods focus on examining specific, localized aspects of data or model behavior for a specific instance. They are designed to provide insights into a single and specific data point, making them useful for fine-grained analysis, offering a detailed, close-up view of data patterns and model behavior in a specific context. They are valuable for assessing the sensitivity of a model's predictions to small changes in input data and provide insights into the model's stability and reliability. These methods are adept at detecting anomalies or outliers in a dataset, aiding in quality control and identifying data irregularities. The local feature importance analysis focuses on understanding how individual features impact model predictions in specific instance.

In contrast, global methods take a broader perspective, aiming to provide insights into the overall behavior and performance of data or models across the entire dataset. These methods analyze the data or model as a whole, offer a holistic view of data patterns and model performance, enabling a comprehensive assessment, are effective in recognizing recurring patterns and trends within data and are suitable for assessing a model's overall predictive accuracy, generalization, and robustness. The global feature importance analysis identifies trends and patterns in how features impact model predictions across the entire dataset.

Post Hoc Explainability

Features, or variables, encapsulate crucial information within datasets, influencing the performance and interpretability of models. The emergence of feature importance generation based methods has offered a profound avenue for comprehending the significance and impact of these features on model outputs. In this section, we present some of most popular methods for post hoc local and global explanations that can either be model agnostic or model specific.

Shapley Additive exPlanations (SHAP). SHAP introduced in [56], serve as a powerful tool for explaining the output of a machine learning model by attributing a value to each feature based on its contribution to individual predictions. These values are computed by considering all possible feature combinations, employing the Shapley value concept [79]. The general formula for computing the Shapley value of feature x^j can be expressed as

follows:

$$\phi_{x^j} = \sum_{S \subseteq D \setminus \{j\}} \frac{|S|!\ (|D| - |S| - 1)!}{|D|!} \left(f^*_{\mathbf{x} \in \mathbb{R}^{S \cup \{j\}}}(\mathbf{x}) - f^*_{\mathbf{x} \in \mathbb{R}^{S}}(\mathbf{x}) \right) \tag{2.2}$$

Here, $f^*_{\mathbf{x} \in \mathbb{R}^{S \cup \{j\}}}$ and $f^*_{\mathbf{x} \in \mathbb{R}^{S}}$ respectively denote predictions of the black box model for subsets with and without the feature x^j included.

The Shapley value use four axioms to serve as foundational principles guiding the fair attribution of feature importance. These axioms, rooted in cooperative game theory, ensure that attribution methods exhibit desirable properties such as consistency and fairness.

The Shapley value, a concept rooted in cooperative game theory, is governed by four axioms that define fair distributions of benefits within coalitions. Efficiency ensures that the total benefits produced by the grand coalition are equally distributed among its members. Symmetry mandates that players who contribute equally to all coalition subsets receive an equal share of benefits. Linearity allows for the additive combination of values from different coalition games when constructing a new cooperative game. Finally, the Dummy player axiom dictates that non-contributing players receive no benefits. While this axiom may raise ethical concerns in certain fields, such as economics, it poses no such issues in machine learning contexts where it is used to measure variable contributions to model predictions. These axioms collectively ensure that the Shapley value provides a unique and fair division of benefits within coalitions.

In order to understand the predictions of black box models, SHAP offers a wide range of explainers including Kernel SHAP, Sampling SHAP, Tree SHAP, and Deep SHAP, tailored to different model architectures. Among the multiple proposed explainers, Tree SHAP is a post hoc model specific method that is designed to explain the predictions of complex tree based model. It traverses the decision tree from root to leaf, computing feature contributions at each node based on the Shapley value concept, reflecting the difference between the model's output for the current instance and the output if that node were the root. These contributions are propagated back along the traversal path, considering feature interactions and splitting criteria. Aggregating contributions across all paths yields final feature importance values, offering insights into the impact of each feature on the

model's prediction. For ensemble models, Tree SHAP averages contributions across all trees to provide comprehensive explanations, facilitating a deeper understanding of model behavior and feature influence.

Deep SHAP, another technique of SHAP proposed to explain the intricate relationships between features and predictions in deep neural networks (DNNs). It navigates through the neural network layers, capturing feature attributions at each step by employing a combination of backpropagation and sampling techniques. It computes the contribution of each feature to the model's output by analyzing the changes in predictions when individual features are included or excluded. Through this process, Deep SHAP disentangles the complex interactions within DNNs, providing interpretable explanations for model predictions.

Furthermore, Kernel SHAP on the other hand, is a post hoc model agnostic explanation method allowing to compute local explanations for any complex model. It learns loally a surrogate model that mimics the black box model behavior on a locally generated new dataset around a given instance, that is also the istance we want explanations for. The weights of the surrogate linear model are then considered as the feature importance for that instance.

Lastly, Sampling SHAP is very similar to Kernel SHAP designed for high-dimensional datasets, offering a scalable solution for computing Shapley values in complex models. By leveraging sampling techniques, Sampling SHAP addresses the computational challenges associated with large feature spaces, ensuring efficient and practical computation of feature attributions. Like its counterpart, Sampling SHAP provides interpretable insights into the importance of features in machine learning models. With its ability to handle high-dimensional datasets, Sampling SHAP serves as a valuable tool for explaining the behavior of intricate models across diverse domains.

Local Interpretable Model-agnostic Explanations (LIME). LIME [77] is another local surrogate based methodology designed to provide interpretable approximations of the decision boundaries of complex black box models in the local vicinity (neighborhood) of a specific instance as shown in Figure 2.5. Let f denote the black box model of interest, and g represent the local interpretable model created by LIME. For a particular instance $\mathbf{x}$ in the input space, LIME seeks to approximate the behavior of f through a

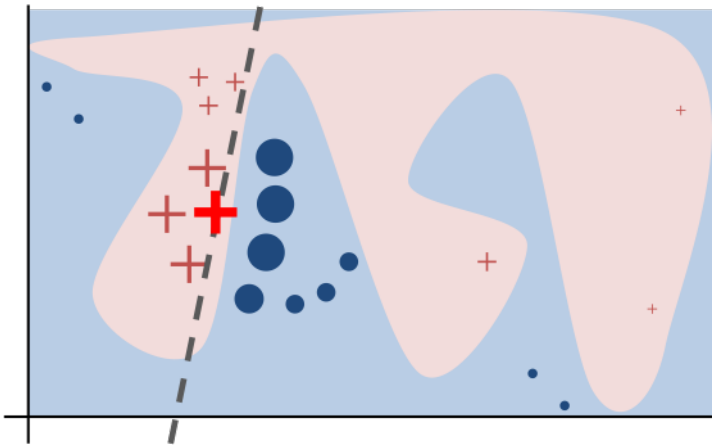


Figure 2.5: To explain a given instance, LIME trains a local linear model on a locally sampled data. The weights of the linear model are then considered as the explanations for the given instance [77].

simplified, interpretable model g that is locally faithful.

Formally, the local interpretable model g is obtained by solving the following optimization problem:

$$g(\mathbf{x}) = \arg\min_{k \in \mathcal{K}} \mathcal{L}(f(\mathbf{x}, k(x')) + \Omega(k) \tag{2.3}$$

where $\mathcal{K}$ is the set of possible interpretable models, $\mathcal{L}$ is a loss function measuring the dissimilarity between the predictions of the black box model f and the surrogate model $k \in \mathcal{K}$, x' is a set of perturbed instances generated around $\mathbf{x}$ to train the local model, and $\Omega(k)$ is a regularization term penalizing model complexity.

The primary objective of LIME is to generate a locally faithful and interpretable model that approximates the decision boundary of the black box model within a small neighborhood of the given instance while keeping the surrogate model as simple as possible. This facilitates a better understanding of the black box model's decision-making process in specific regions of the input space.

Global and Local Surrogates. Mimic explainer in *Interpret community*[3] is based on the idea of training global surrogate models to mimic blackbox models. A global surrogate model is an intrinsically interpretable model

[3] `https://interpret.ml`

that is trained to approximate the predictions of any black box model as accurately as possible, using one of the following interpretable models as your surrogate model: Light GBM (LGBM Explainable Model), Linear Regression (Linear Explainable Model), Stochastic Gradient Descent explainable model (SGD Explainable Model), and Decision Tree (Decision Tree Explainable Model).

Local surrogates [63] on the other hand, are interpretable models designed to approximate the behavior of a complex black box model within a local region of the input space. Let f represent the black box model of interest, and g denote the local surrogate model. For a specific instance x in the input space, the local surrogate model g is trained to approximate the response of the black box model f in the vicinity of $\mathbf{x}$.

Formally, the local surrogate model g is defined as:

$$g(\mathbf{x}) = \arg\min_{k \in \mathcal{K}} \mathcal{L}(f(\mathbf{x}), k(\mathbf{x'}))$$

Where $\mathcal{K}$ is the space of interpretable models, $\mathcal{L}$ is a loss function measuring the distance between the predictions of the black box model f and the surrogate models $k \in \mathcal{K}$, and $\mathbf{x'}$ denotes the set of sampled instances in the local neighborhood of $\mathbf{x}$ used to train the surrogate model.

Kernel SHAP, LIME, and local surrogates [4], share several fundamental similarities in their approaches to model interpretation. Firstly, all three methods aim to provide local explanations for individual predictions, enabling users to understand the model's decision-making process on a per-instance basis. Secondly, they adopt a model-agnostic perspective, allowing them to be applied to a wide range of machine learning models without relying on specific model structures. This flexibility makes them particularly useful in scenarios where the underlying model's complexity varies or is not fully understood. Thirdly, they employ local approximation techniques to explain model predictions, whether through kernel-based approximation (as in Kernel SHAP), generating interpretable surrogate models (as in LIME), or constructing local linear models (as in local surrogates). Despite their differences in implementation details, these methods share the common goal of enhancing model transparency and interpretability, enabling users to gain insights into model behavior at the individual prediction level.

[4] *interpret community package* `https://interpret.ml`

Counterfactual Explanations. While local explanations help understand why a model makes a particular decision, they do not explicitly reveal what needs to change to get a different outcome for a prediction. As a result, there are a growing number of methods that explain the decisions of these models to affected individuals and provide means for recourse [96].

DiCE [67] is a model-agnostic method for generating diverse and interpretable counterfactual explanations for individual predictions. It finds instances similar to original instance $\mathbf{x}$ (Figure 2.6 [5]), but with different predicted outcomes. Optimization requires minimizing a distance metric between the counterfactuals and $\mathbf{x}$, subject to constraints that ensure dissimilarity among generated counterfactuals. Counterfactuals are generated by perturbing the features of $\mathbf{x}$ while staying within the feasible range of feature values. The optimization problem can be formulated as:

$$\min_{x_i' \in \mathcal{X}} \delta(x_i', x_i) \in C(x_i') = y_i',\ x_i' \neq x_i$$

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

For example, recourse offers a person denied a loan by a credit risk model a reason for why the model made the prediction and what can be done to change the decision. Beyond providing guidance to stakeholders in model decisions, algorithmic recourse is also used to detect discrimination in machine learning models [34, 43, 84].

Tree Interpreter. Tree Interpreter [52] is a model-specific method for interpreting predictions of tree-based models, such as random forests and XGBoost. 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. As introduced in [71], 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

[5] `https://interpret.ml`

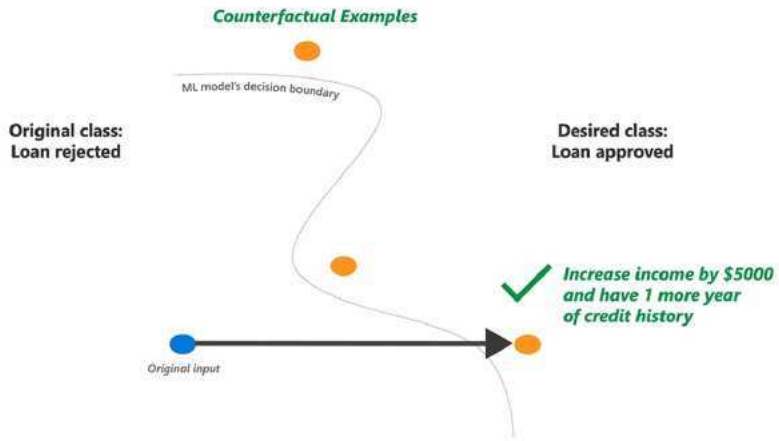


Figure 2.6: An example of counterfactual generation with DiCE for the loan approval for a given client. In order to approve the client's rejected loan, DiCE recommends increasing his income and waiting for one additional year of credit history.

node. The prediction can be decomposed into the sum of the feature contributions and the "bias" β (i.e. the mean of training set), and can be written down as:

$$f(\mathbf{x}) = \beta + \sum_{j=1}^{D} \phi_{x^j}$$

Unlike in the linear regression where the feature coefficients are fixed with a single constant for every feature that determines the contribution, the contribution of each feature ϕ_{x^j} in this tree prediction decomposition is not a single predetermined value, but depends on the rest of the feature vector which determines the decision path that traverses the tree and thus the contributions that are passed along the way.

2.2.4 Explanation Evaluation

In this section, we address the evaluation of explanations in the absence of ground truth, employing a diverse range of metrics to assess their properties. These metrics [16] reflects desired aspects such as local stability, faithfulness, and consistency.

The Local Stability Metric. evaluates how consistent explanations are for instances in the same neighborhood in a dataset. Essentially, if two instances have similar features and produce similar predictions, they should also receive similar explanations [72]. However, instances that are on the edge of a decision boundary may have different explanations, even if their features are alike. This is because many explanation methods base their explanations on the model's predictions. To assess stability, we employ the local Lipschitz metric [3], which measures the stability of explanations across instances in the same neighborhood. This helps us understand how robust the explanations are when the input data changes slightly.

$$\hat{L}(x_i) = \underset{x_j \in B_\epsilon(x_i)}{\operatorname{argmax}} \frac{\|\phi(x_i) - \phi(x_j)\|_2}{\|x_i - x_j\|_2} \tag{2.4}$$

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. [6] measures how simple an explanation is by looking at how many features are needed to explain a certain part of the model's prediction. We decide on a percentage of the prediction we want to explain, then see how many features it takes to explain that percentage. This helps us understand how straightforward the explanation is in terms of the number of features it uses to explain a specific part of the model's output.

RemOve And Retrain (ROAR). [39] 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. By systematically testing the impact of feature removal on model performance, it offers valuable information about the model's sensitivity to different features and its overall interpretability.

The Faithfulness Metric [47] is a crucial aspect of explainability evaluation, aiming to assess how faithfully an explanation method represents the true relationship between features and model predictions. It measures the

[6] `https://github.com/MAIF/shapash`

extent to which the explanations provided accurately reflect the underlying behavior of the model. One common approach to evaluating faithfulness involves comparing the explanations generated by a method against the ground truth feature importances, if available [1]. However, in many practical scenarios, accessing such ground truth information may be challenging or impossible. As an alternative, faithfulness can be evaluated based on the consistency and agreement of feature importance rankings and signs across different explanation methods. This approach focuses on determining whether the explanations generated by various methods align in terms of which features are considered important and how they contribute to predictions. Other measures of faithfulness from [1] can be viewed as variants of these.

The Consistency Metric. measures how much different explanation methods agree on the importance of features for the same data points. It calculates the distance between pairs of explanations using the l_2 distance [89]. When two methods provide similar explanations for the same data, it boosts the user's confidence in the model's predictions. This alignment in explanations makes it easier for users to trust the model's decisions.

Feature and Rank Agreements. [47] While feature agreement computes the fraction of common features between the sets of top-k features of two explanations, the rank agreement computes the fraction of features that are not only common between the sets of top-k features of two explanations, but also have the same position in the respective rank orders. Rank agreement is a stricter metric than feature agreement since it also considers the ordering of the top-k features.

Chapter 3

A Critical Evaluation of Local Explainability Methods

This chapter presents two important contributions to understanding and using local explainability methods in machine learning. First, we discover a relationship between the local explanations computed in existing explainability methods and the data properties like noise level, feature correlations and class imbalance. This helps us see how well they work in different situations. Second, we leverage personalized, consistent and simple explanations of the risk factors of the cervical cancer and help clinicians to understand each patient specific risk factors. By doing this, we can see how well the explanations help us understand the factors contributing to cervical cancer risk in individual cases. These contributions help us learn more about how local explainability methods work in different scenarios, both general and specific.

3.1 Feature Importance and Data Properties

In order to ensure the reliability of the explanations of machine learning models, it is crucial to establish their advantages and limits and in which case each of these methods. Several existing studies have addressed the challenges of post-hoc analysis in machine learning models, particularly when feature dependence is present. These studies recognize that real-world datasets often exhibit complex relationships among features, such as correlations or interactions, which can impact the interpretability and reliability of model

explanations. For instance, [89] discussed the implications of feature dependence on the consistency of explanation methods, highlighting the need for robust techniques that account for such dependencies. Additionally, [4] proposed methods for analyzing the stability of explanations in the presence of feature correlations, emphasizing the importance of understanding how explanations vary with changes in the input data. However, the current understanding of when and how each method of explanation can be used is insufficient. To fill this gap, we perform a comprehensive empirical evaluation by synthesizing multiple datasets with the desired properties. Our main objective is to assess the faithfulness, local stability and consistency of feature importance estimates provided by local explanation methods, which are used to explain predictions made by decision tree-based models. Analyzing the results obtained from synthetic datasets as well as publicly available binary classification datasets, we observe different magnitude and sign of the feature importance estimates generated by these methods. Moreover, we find that these estimates are sensitive to specific properties present in the data. Although some model hyper-parameters do not significantly influence feature importance assignment, it is important to recognize that each method of explanation has limitations in specific contexts. Our assessment highlights these limitations and provides valuable insight into the suitability and reliability of different explanatory methods in various scenarios.

3.1.1 Introduction

Decision tree based models such as random forest [10] are widely used machine learning algorithms in data science. Although deep learning has been increasingly popular, especially in domains such as computer vision and natural language processing, random forest, for example continues to be a competitive option on many kinds of tabular data in a diverse number of domains, including biology [49] and medicine [78], where interpretation is paramount. Small decision trees operating on understandable feature spaces are naturally interpretable, and although this interpretability is diluted across a large forest, it can be recovered in terms of feature importance, which is a major tool that can be used in practical applications for data understanding, model improvement, or model explainability. However, practitioners may lose trust in the importance scores provided for random forest [90], or simply be unable to use them to answer their research questions

from the feature importance result due to a number of reasons [93, 64], for example: (1) a relative lack of training examples leads to instability where the importance scores change due to only minor changes or additions to the dataset or hyper-parameters. (2) even with a large training set, multiple (possibly equivalent) feature scores can be presented. (3) the feature importance scoring mechanism is thrown off by particular properties of the data distribution such as noise, imbalance and feature type (in particular, the importance of continuous features is often over-estimated). (4) results where feature importance is assigned to spurious or even random features. Practitioners are thus often right to be reluctant to draw conclusions from or place trust in off-the-shelf feature-importance scorers, and we aim to remedy this to some extent with a benchmarking study.

To remedy this, researchers proposed explainability methods such as LIME [77] and SHAP [56] to explain black box models by attributing feature importance estimates as explanations of the model's predictions. While prior research [47, 7, 12, 9, 69] has already taken the first steps towards analyzing the disagreement of explanation methods for models such as deep neural networks, analyzing the behavior of the wide range of existing explanation methods for random forest or in general ensemble trees still insufficiently explored, with regard to particular data properties and model parameters [25].

Compared to other work which is either model agnostic focused or deep neural networks specific, we study the explainability methods suited to explain decision tree based models. Some of these methods are specific to tree ensembles and the rest are general model-agnostic (which, thus, can also be applied to random forest). We do so with extensive experiments on synthetic alongside real-world datasets, and certain manipulations thereof, which we carry out to isolate and identify aspects which lead to particular results insofar as feature importance. This provides a more thorough understanding, which we use to highlight some limitations of existing methods, and formulate a number of recommendations for practitioners. The contribution of this work is twofold:

- Conducting a thorough evaluation of various explainability methods in the context of specific data properties, such as noise levels, feature correlations, and class imbalance, elucidating their strengths and limitations.

- Offering valuable guidance for practitioners and researchers on selecting the most suitable explainability method based on the characteristics of their dataset.

3.1.2 Explainability Benchmarking Frameworks

The landscape of explainable artificial intelligence has witnessed a surge in research efforts aimed at understanding and evaluating the diverse methodologies employed for interpreting complex machine learning models. Several survey and benchmarking papers, including XAI-survey [9] and BenchXAI [54], have played a crucial role in shedding light on the disagreement problem within existing explainability methods [47, 69, 12, 35, 95]. Notably, these contributions have been important to the understanding of the challenges and nuances associated within the field of machine learning explainability.

While the majority of existing benchmarks have primarily focused on explaining neural networks for text and image data with feature importance generation methods such as [40, 7, 9, 103, 106, 36], the research community has introduced several frameworks to facilitate the transparent evaluation of explainability methods. Examples include OpenXAI [1], Captum [46], Quantus [37], and many others such as [31, 50]. In addition, [95] introduced a quantitative framework with specific metrics for assessing the performance of post-hoc interpretability methods, particularly in the context of time-series classification. This research provides a targeted approach to evaluating the temporal aspects of interpretability. These frameworks aim to provide a structured approach to assess the effectiveness and reliability of various explainability techniques.

The evaluation of post-hoc interpretability methods for ensemble trees predictions with respect to different data properties is crucial for understanding the robustness and reliability of these methods across various real-world scenarios. Despite the growing interest in interpretability, there remains a gap in understanding how these methods perform under diverse data conditions. This gap is significant because real-world datasets often exhibit varying properties such as noise levels, feature correlations, and class imbalances, which can influence the effectiveness of interpretability techniques. For instance, consider the scenario where a bank utilizes an ensemble tree model to assess credit risk. In such cases, the interpretability of the model's predictions is crucial for regulatory compliance and risk management. How-

ever, banking datasets often exhibit complex characteristics, such as high dimensionality, imbalanced classes, and correlations between financial variables. These data properties can significantly impact the performance of interpretability methods, potentially leading to misinterpretations or unreliable insights. This work addresses this gap by investigating how existing interpretability methods designed for ensemble trees predictions behave under different data conditions.

3.1.3 Synthetic Data Generation

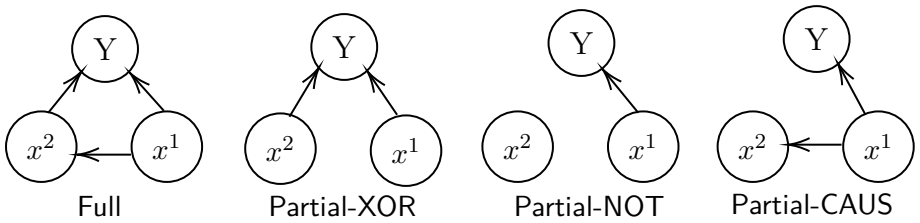


Figure 3.1: Bayesian networks that we use as a schema to generate synthetic data, illustrating one full and three partial factorizations of $P(X, Y)$.

Considering only two features, we could represent the concept as a Bayesian network (Fig. 3.1 illustrates).

$$P(X, Y) = P(x^1, x^2, Y) = P(Y|x^1, x^2)P(x^2|x^1)P(x^1)$$

i.e., a Full factorization of the joint, and thus we could consider the following properties (as nodes and edges):

1. $P(x^1)$: specifying the type of the feature x^1;
2. $P(x^2|x^1)$: the amount of conditional dependence of x^2 on x^1; and
3. $P(Y|x^1, x^2)$: the amount and type of correlation between features and target, revealing the special case of $P(Y|x^1, x^2) = P(Y)$ when there is no correlation.

Partial-XOR and Partial-NOT exhibit feature independence (features are independent from each other – when the target is observed), and both features are required to make a perfect prediction for Partial-XOR, and only x^1 is required to perfectly predict Partial-NOT; in this case deterministic.

In real-world settings, Full represents the case where one feature is related to another feature and both participate to make the prediction of the output, for example in ADULT INCOME dataset (that we later include in our experiments) the feature OCCUPATION is correlated to AGE and both predict the output INCOME. Partial-CAUS on the other hand, may represent a causal relationship between one feature and the outcome through a chain of causality; or an indirect correlation of one feature on the output through another feature (or even multiple features), forming a chain of correlations.

Full and Partial-CAUS are specific cases of respectively Partial-XOR and Partial-NOT when x^1 is dependent on x^2, thus for the rest of this section, we denote XOR to refer to both Full and Partial-XOR, and use NOT to refer to Partial-CAUS and Partial-NOT. The data can be generated as:

$$P(X) \sim \mathcal{N}(\mu, \Sigma) \tag{3.1}$$

Where $\mathcal{N}$ is a bi-variate normal distribution and Σ is the covariance matrix, and the amount of correlation between the two random variables x^1 and x^2 is denoted as ρ. In order to introduce noise, we invert ϵ percentage of predictions $\hat{y}$ and we keep the rest unchanged.

Ground truth feature importance. We use $\phi^*_X(f^*)$ to denote the ground truth feature importance that are given by the true model f^* to which we compare $\phi_X(f)$, the feature importance estimates that is generated by each of the local explainability methods to explain the predictions of the learned model f. Intuitively, the true model f^* can be illustrated with a d-depth decision tree. With $d = 2$ for XOR dataset variants (the first split on x^1 and the second on x^2) and $d = 1$ for NOT dataset variants (only one split on x^1).

When $\epsilon = 0$, the ground truth feature importance $\phi^\star_X$ for all variants of XOR datasets are fixed as $\phi^\star_{x^1} = \phi^\star_{x^2} = .5$, because both x^1 and x^2 are *necessary* to make the prediction of XOR. The amount of the correlation ρ between x^1 and x^2 doesn't affect the importance as both are *necessary* to make the prediction of XOR. Meanwhile, only x^1 is *necessary* to make the prediction of NOT, thus $\phi^\star_{x^1} = 1$ and $\phi^\star_{x^2} = \rho$, because when x^2 is correlated to x^1, x^2 have an indirect influence estimated by ρ to predict NOT.

On the other hand, when $\epsilon \neq 0$, $\phi^\star_{x^1} = \phi^\star_{x^2} = .5 * \epsilon$ for XOR dataset

variants, and $\phi^{\star}_{x^1} = 1 * \epsilon$ and $\phi^{\star}_{x^2} = \rho * \epsilon$ for NOT dataset variants.

For the XOR function, both $x^1$and x^2 are equally important in determining the output, and their importance scores should ideally converge to 0.5 when considering a large amount of data points. Consider a decision tree model that aims to predict the XOR function using features x^1 and x^2. For simplicity, let's assume that the decision tree splits on both x^1 and x^2 at each level. The decision tree's predictions can be expressed as:

$$\hat{y} = f(x^1, x^2)$$

Now, let's define the feature importance scores (ϕ_{x^1} and ϕ_{x^2}) using the Gini impurity criterion, a common metric for decision trees:

$$\phi_{x^1} = \sum_{\text{nodes splitting on } x^1} \text{Gini decrease at the node}$$

$$\phi_{x^2} = \sum_{\text{nodes splitting on } x^2} \text{Gini decrease at the node}$$

In a large dataset, the decision tree will be able to accurately capture the XOR relationship, and both x^1 and x^2 should contribute equally to the impurity decrease, leading to similar importance scores. For a balanced decision tree, these Gini decreases would be distributed among the splits involving x^1 and x^2. In the limit of a large dataset, we would expect:

$$\lim_{\text{large dataset}} \phi_{x^1} = \lim_{\text{large dataset}} \phi_{x^2} = 0.5$$

This indicates that, as the dataset size increases, the decision tree's feature importance for predicting the XOR function would converge to 0.5 for both x^1 and x^2, reflecting their equal importance in determining the output.

3.1.4 Empirical Setup

To carry out our experiments, we demonstrate our findings on four real-world datasets: Heart Diagnosis, Cervical Cancer, Adult Income and German Credit Risk. These datasets include properties such as feature interactions (dependence or independence), noise, random irrelevant variables and class imbalance.

We generate 24 synthetic datasets (Figures 3.2 and 3.3) expressing different combinations of these properties by varying several parameters such as the correlation amount of the normal distribution from which the data points are drawn and the probability of each class, thus, the amount of generated noise and class imbalance. Each dataset is divided to 80% for training and 20% for testing. We report the results of the feature importance estimates on the test set. We compute feature importance estimates on the true model f^* and learned model f, so that we can compare the generated feature importance estimates to their ground truth values. Finally, we analyze the advantages and limitations of each explainability method on the synthetic datasets and we run larger experiments on the above real-world datasets from the UCI repository [23].

Datasets, Models and Metrics

Datasets. Figures 3.2 and 3.3 show the generated datasets by varying the parameters as in Eq 3.1.3: $\mu \in \{[0,1],[1,0]\}$, $\epsilon \in \{0,.25,.5\}$, and $\Sigma \in \{[[1,0],[0,1]],[[1,.1],[.1,1]],[[1,.9],[.9,1]],[[1,1],[1,1]]\}$.

In addition, Table 3.1 summarizes the properties of the four real-world datasets that we use to demonstrate our findings.

Dataset	#instances	#features	% discrete	% continuous	imbalance
Heart Diagnosis	303	13	43	57	yes
Cervical Cancer	858	35	62	38	yes
Adult Income	32561	11	65	35	no
German Credit Risk	1 000	23	70	30	yes

Table 3.1: Summary of the real-world datasets we include in our experiments.

Models. For the synthetic datasets, we compute the feature importance scores of the learned model f on datasets with 1 000 instances. The learned model f can be either a decision tree or a random forest. On the other hand, we use the random forest model with parameters learned using grid search and evaluated with 10-fold cross-validation for each of the real-world datasets. Table 3.2 summarizes the performances and the feature importance of the decision tree and the random forest models for the generated datasets.

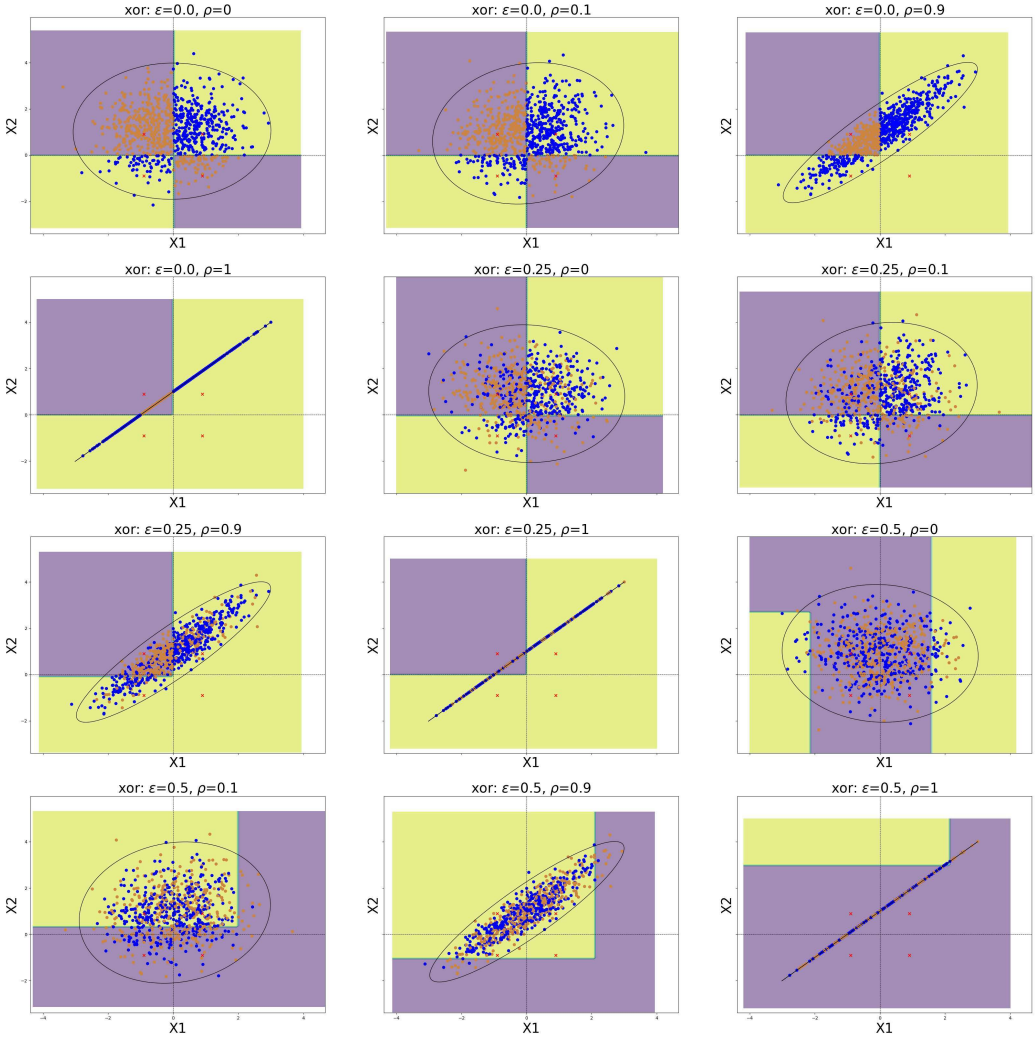


Figure 3.2: Synthetic data XOR with decision boundaries. $X \sim \mathcal{N}(\mu, \Sigma)$. Each dataset expresses a different combination of properties.

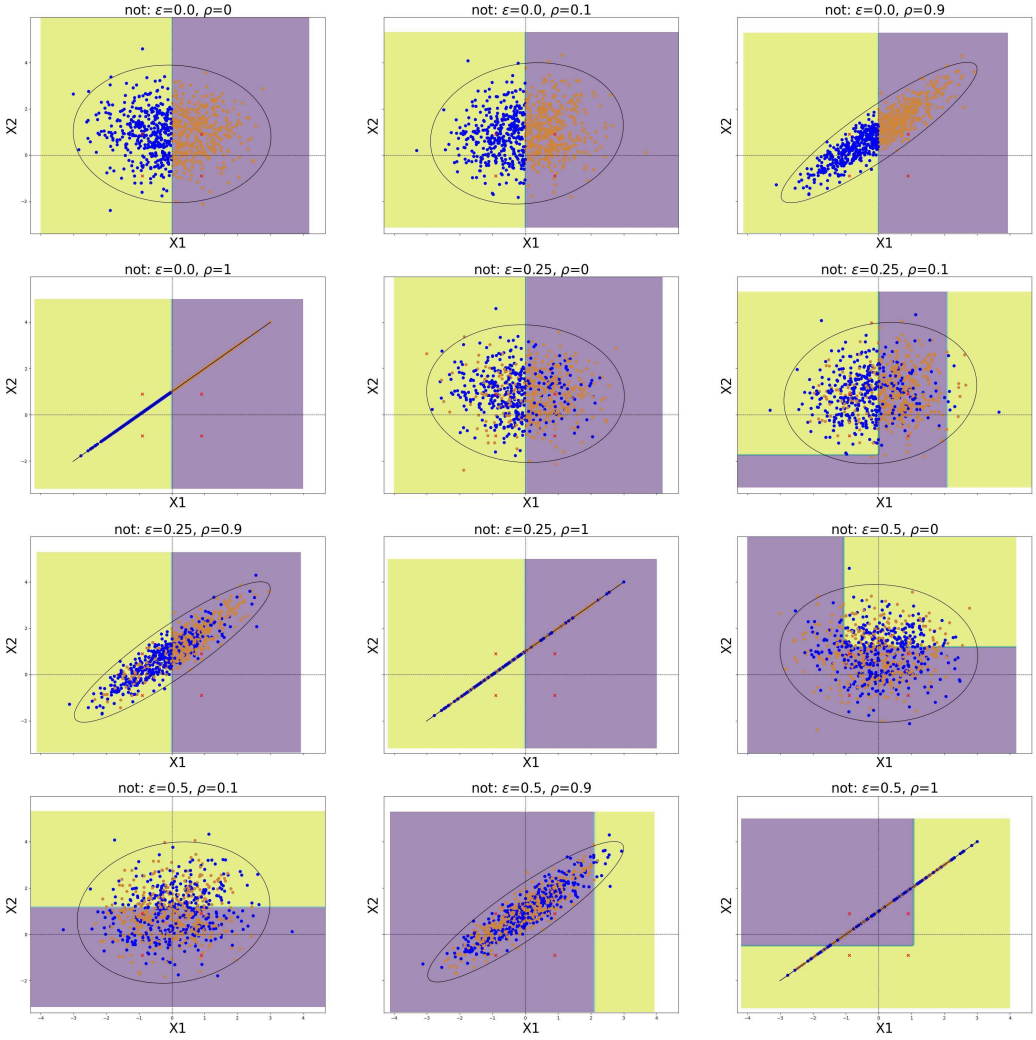


Figure 3.3: Synthetic data NOT with decision boundaries. $X \sim \mathcal{N}(\mu, \Sigma)$. Each dataset expresses a different combination of properties.

Decision function	ϵ	ρ	DT Accuracy	DT feature importance	RF Accuracy	RF feature importance
XOR	0.00	0.00	1.00	ϕ_{x^1}=0.44, ϕ_{x^2}=0.56	0.85	ϕ_{x^1}=0.69, ϕ_{x^2}=0.31
XOR	0.00	0.10	1.00	ϕ_{x^1}=0.41, ϕ_{x^2}=0.59	0.90	ϕ_{x^1}=0.63, ϕ_{x^2}=0.37
XOR	0.00	0.90	1.00	ϕ_{x^1}=0.51, ϕ_{x^2}=0.49	1.00	ϕ_{x^1}=0.55, ϕ_{x^2}=0.45
XOR	0.00	1.00	1.00	ϕ_{x^1}=0.5, ϕ_{x^2}=0.5	1.00	ϕ_{x^1}=0.53, ϕ_{x^2}=0.47
XOR	0.25	0.00	0.70	ϕ_{x^1}=0.47, ϕ_{x^2}=0.53	0.61	ϕ_{x^1}=0.62, ϕ_{x^2}=0.38
XOR	0.25	0.10	0.72	ϕ_{x^1}=0.4, ϕ_{x^2}=0.6	0.62	ϕ_{x^1}=0.64, ϕ_{x^2}=0.36
XOR	0.25	0.90	0.72	ϕ_{x^1}=0.51, ϕ_{x^2}=0.49	0.72	ϕ_{x^1}=0.56, ϕ_{x^2}=0.44
XOR	0.25	1.00	0.71	ϕ_{x^1}=0.51, ϕ_{x^2}=0.49	0.71	ϕ_{x^1}=0.53, ϕ_{x^2}=0.47
XOR	0.50	0.00	0.52	ϕ_{x^1}=0.83, ϕ_{x^2}=0.17	0.50	ϕ_{x^1}=0.57, ϕ_{x^2}=0.43
XOR	0.50	0.10	0.53	ϕ_{x^1}=0.69, ϕ_{x^2}=0.31	0.56	ϕ_{x^1}=0.57, ϕ_{x^2}=0.43
XOR	0.50	0.90	0.49	ϕ_{x^1}=0.64, ϕ_{x^2}=0.36	0.46	ϕ_{x^1}=0.53, ϕ_{x^2}=0.47
XOR	0.50	1.00	0.54	ϕ_{x^1}=0.55, ϕ_{x^2}=0.45	0.47	ϕ_{x^1}=0.53, ϕ_{x^2}=0.47
NOT	0.00	0.00	1.00	ϕ_{x^1}=1.0, ϕ_{x^2}=0.0	1.00	ϕ_{x^1}=0.84, ϕ_{x^2}=0.16
NOT	0.00	0.10	1.00	ϕ_{x^1}=1.0, ϕ_{x^2}=0.0	1.00	ϕ_{x^1}=0.83, ϕ_{x^2}=0.17
NOT	0.00	0.90	1.00	ϕ_{x^1}=1.0, ϕ_{x^2}=0.0	1.00	ϕ_{x^1}=0.61, ϕ_{x^2}=0.39
NOT	0.00	1.00	1.00	ϕ_{x^1}=1.0, ϕ_{x^2}=0.0	1.00	ϕ_{x^1}=0.49, ϕ_{x^2}=0.51
NOT	0.25	0.00	0.72	ϕ_{x^1}=1.0, ϕ_{x^2}=0.0	0.72	ϕ_{x^1}=0.77, ϕ_{x^2}=0.23
NOT	0.25	0.10	0.69	ϕ_{x^1}=0.99, ϕ_{x^2}=0.01	0.72	ϕ_{x^1}=0.8, ϕ_{x^2}=0.2
NOT	0.25	0.90	0.71	ϕ_{x^1}=1.0, ϕ_{x^2}=0.0	0.71	ϕ_{x^1}=0.63, ϕ_{x^2}=0.37
NOT	0.25	1.00	0.72	ϕ_{x^1}=0.97, ϕ_{x^2}=0.03	0.72	ϕ_{x^1}=0.49, ϕ_{x^2}=0.51
NOT	0.50	0.00	0.50	ϕ_{x^1}=0.39, ϕ_{x^2}=0.61	0.48	ϕ_{x^1}=0.55, ϕ_{x^2}=0.45
NOT	0.50	0.10	0.58	ϕ_{x^1}=0.0, ϕ_{x^2}=1.0	0.57	ϕ_{x^1}=0.5, ϕ_{x^2}=0.5
NOT	0.50	0.90	0.50	ϕ_{x^1}=0.69, ϕ_{x^2}=0.31	0.46	ϕ_{x^1}=0.53, ϕ_{x^2}=0.47
NOT	0.50	1.00	0.48	ϕ_{x^1}=0.49, ϕ_{x^2}=0.51	0.50	ϕ_{x^1}=0.51, ϕ_{x^2}=0.49

Table 3.2: Parameterization and performances of the decision tree (DT) and the random forest (RF) for the 24 generated datasets with 1.000 instances. Maximum depth of both DT and RF is set to 2.

Metrics. To evaluate the quality of the feature importance estimates attributed by the methods in Section 2.2.3, we compare the feature importance estimates to the ground truth feature importance in Section 3.1.3 of the synthetic datasets because the ground truth feature importance estimates in the real-world datasets are hard to obtain. We also evaluate the stability, compactness, consistency, feature and rank agreements for the synthetic and real-world datasets.

3.1.5 Experiments

Synthetic Datasets

Figures 3.4 and 3.5 show the normalized feature importance estimates attributed by the selected explainability methods. After the normalization of the absolute importance of x^1 and x^2, their contributions sum to one. We

perform the normalization to faithfully compare the feature attributions to their ground truth values.

Explainability methods based on learning surrogate models overestimate the importance to irrelevant variables, Tree interpreter is sensitive to noise and SHAP explainers always favor one feature over the other. Overall, all explainers except local surrogates overestimate the importance of x^1 over x^2 across the XOR datasets. Also, none of these methods perfectly matches ground truth feature importance on average across all datasets. Moreover, LSurro and LIME feature importance attributions are the least affected by noise and feature correlation. Indeed, LSurro and and LIME attribute comparable importance to x^1 and x^2 for XOR and NOT dataset variants, and both overestimate the importance of unimportant features (such as x^2 in case of NOT). Notably, TI is the most affected by noise, that is confirmed in its decomposition of the the feature and noise contributions to the prediction. Additionally, feature correlations increase the importance and instability of x^2 importance in XOR datasets attributed by SHAP explainers, and noise lowers the importance of x^1 and x^2 for all the explainers. Finally, SHAP explainers and TI have the highest variance of feature importance estimates in the NOT datasets.

SHAP explainers yield very comparable explanations. Figure 3.6 shows the faithfulness of the explanations to the ground truth measured by mean consistency and mean feature agreements across the XOR and NOT generated datasets.

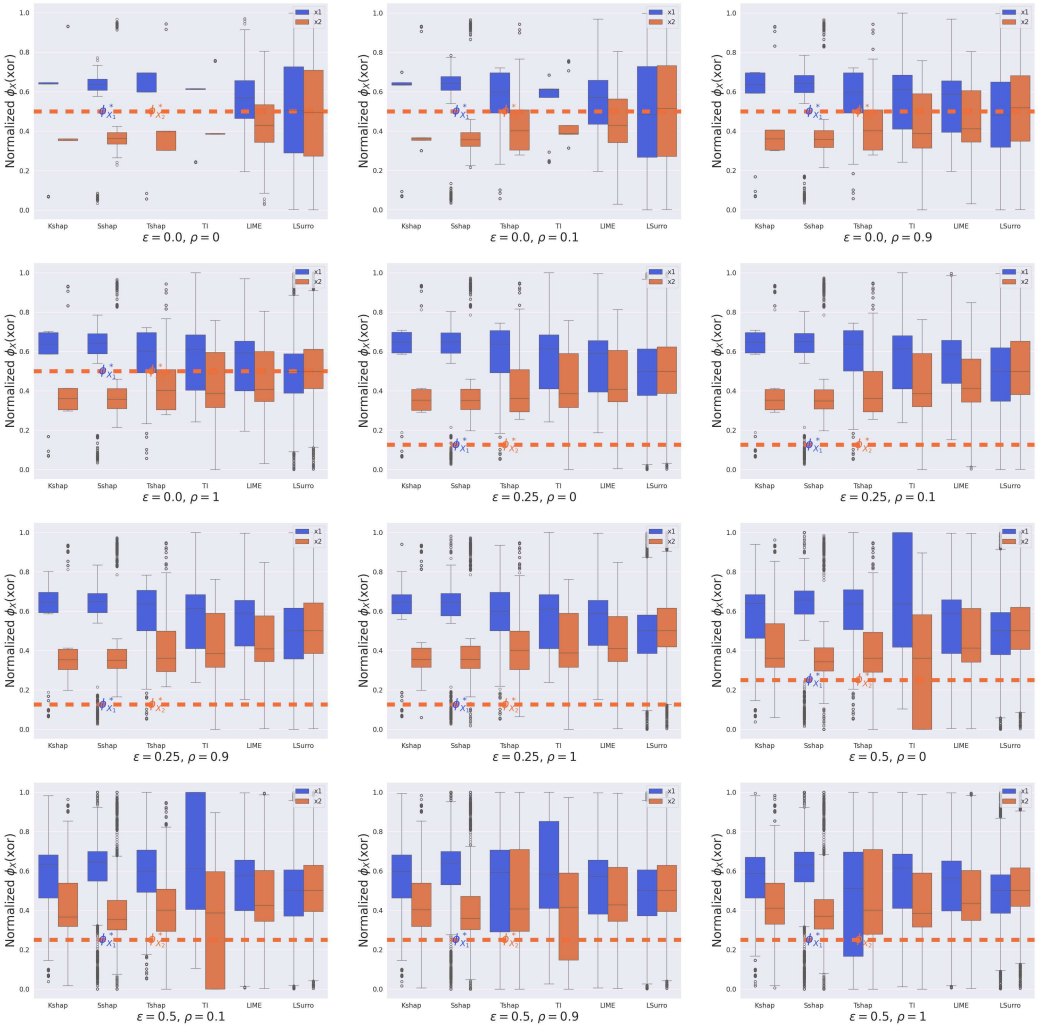


Figure 3.4: Normalized feature importance estimates of the XOR datasets. These feature importance estimates are obtained for the decision trees trained on datasets with 1 000 instances.

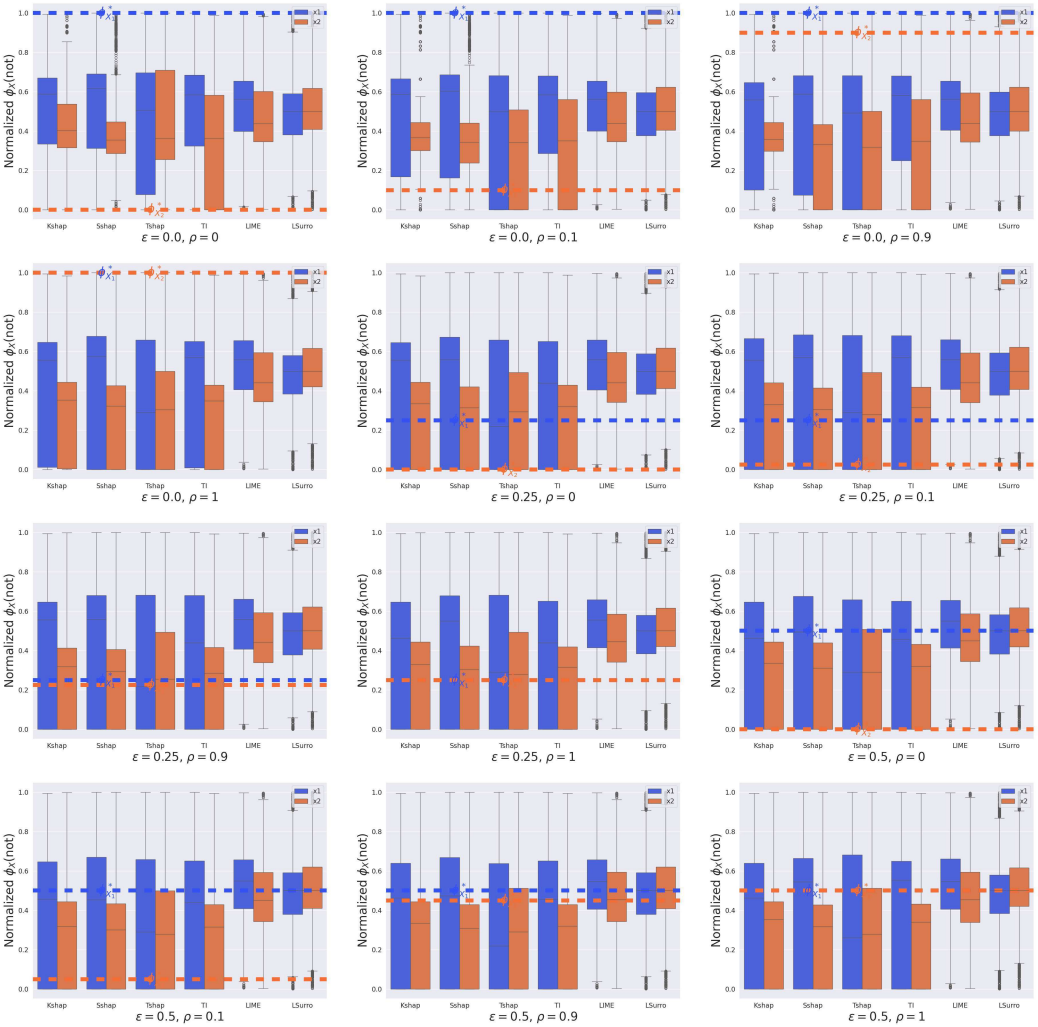


Figure 3.5: Normalized feature importance estimates of the NOT datasets. These feature importance estimates are obtained for the decision trees trained on datasets with 1 000 instances.

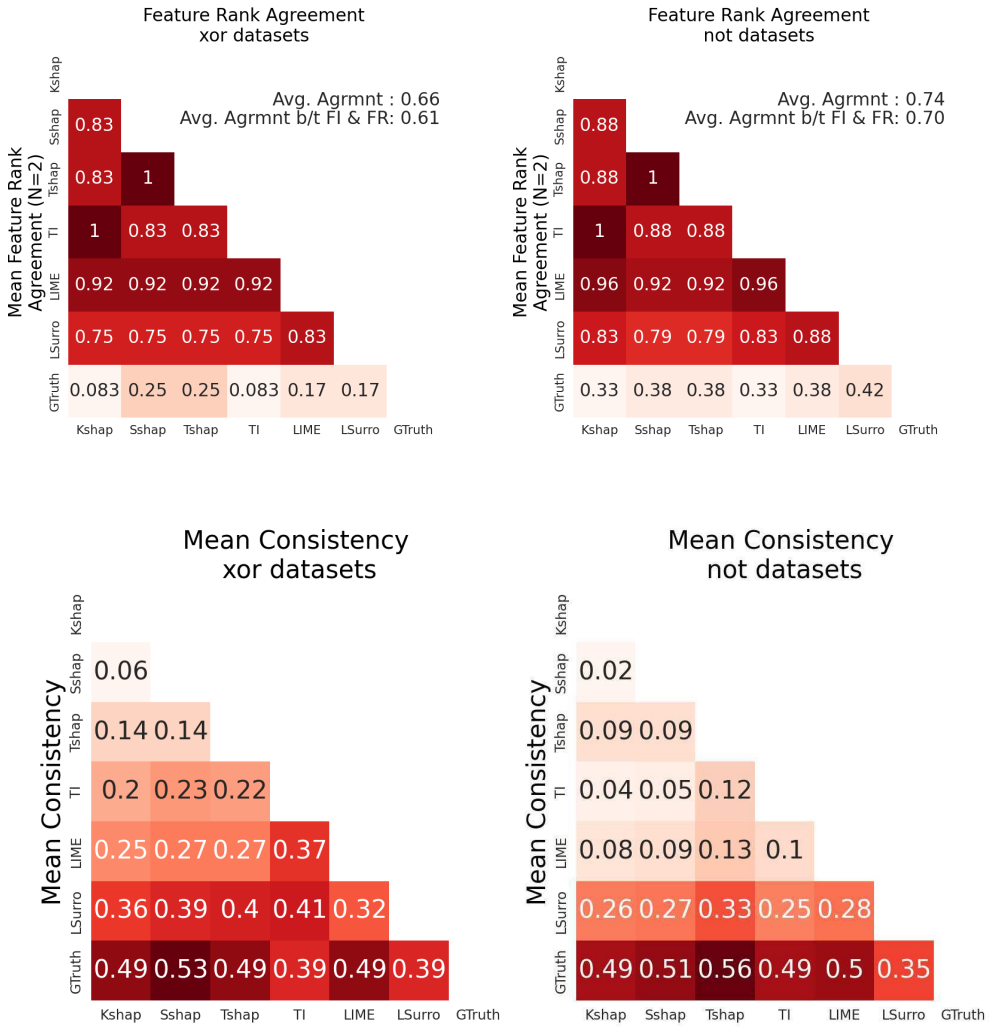


Figure 3.6: Mean consistency, mean feature agreements for XOR and NOT datasets. Consistency is expressed in l_2 distance (the lower the better). Feature agreement measures the fraction of common features between the sets of top-k features of the two rankings (the higher the better).

SHAP explainers yield consistent explanations due to the same feature importance attribution mechanism they all employ. However their explanations are the most inconsistent with respect to the ground truth values. Furthermore and for both XOR and NOT datasets, on average the fraction of common feature importance between TI and KShap and between SShap

and TShap match perfectly.

function	Methods	# Features for 90% approximation	Distance with 1 feature(%)	Mean consistency	Mean Stability (10)
XOR	Kshap	1.00	0.10	0.21	0.21
XOR	Sshap	1.00	0.11	0.23	0.24
XOR	Tshap	1.00	0.10	0.24	0.25
XOR	TI	1.00	0.17	0.26	0.24
XOR	LIME	1.00	0.16	0.28	0.33
XOR	LSurro	1.33	0.43	0.32	0.15
NOT	Kshap	1.00	0.03	0.14	0.16
NOT	Sshap	1.00	0.03	0.15	0.20
NOT	Tshap	1.00	0.03	0.19	0.19
NOT	TI	1.00	0.02	0.15	0.19
NOT	LIME	1.00	0.04	0.17	0.27
NOT	LSurro	1.17	0.30	0.25	0.08

Table 3.3: Compactness (represented in distance reached with fewer features and nmber of features needed to achieve 90% of the model performance) and stability of the explanations for the XOR datasets. Tshap and TI are the most stable explainers for XOR and LSurro uses only one feature to make nearly half of the prediction of NOT.

LSurro is the most locally stable, overestimates unimportant features and achieves better model accuracy with less features. Table 3.3 shows mean consistency, mean stability and compactness across the XOR and NOT datasets. For XOR and NOT datasets respectively, Kshap is the most consistent to the rest of the explanatory methods on average. Additionally, on average, LSurro generates the most locally stable explanations in XOR and NOT datasets, achieves higher model estimation and often consider both features as important fr both datasets.

3.1.6 Real-World Data

For the real-world datasets the ground truth feature importance is unavailable, we perform evaluation of the different metrics in Section 2.2.4.

Local surrogates achieves 100% of model accuracy with 5 features on Adult Income dataset. Figure 3.7 shows feature agreements for Adult Income Income dataset. Kshap and Sshap have exactly the same top-10 feature attributions and ranking. TI and Tshap share the same set of top-10 features. The rest of the explainers share 90% of the top 10 most important features. LIME share the lowest of top-10 important features with

TI, Kshap and Sshap on the ranking of the top 10 features. Additionally, Table 3.4 illustrates the compactness, mean consistency and stability of the different methods on the ADULT INCOME Income dataset. LSurro explains 100% of model output with only 5 features. Kshap, Tshap and TI are the most stable for this dataset and LIME have the highest mean consistency across all the datasets.

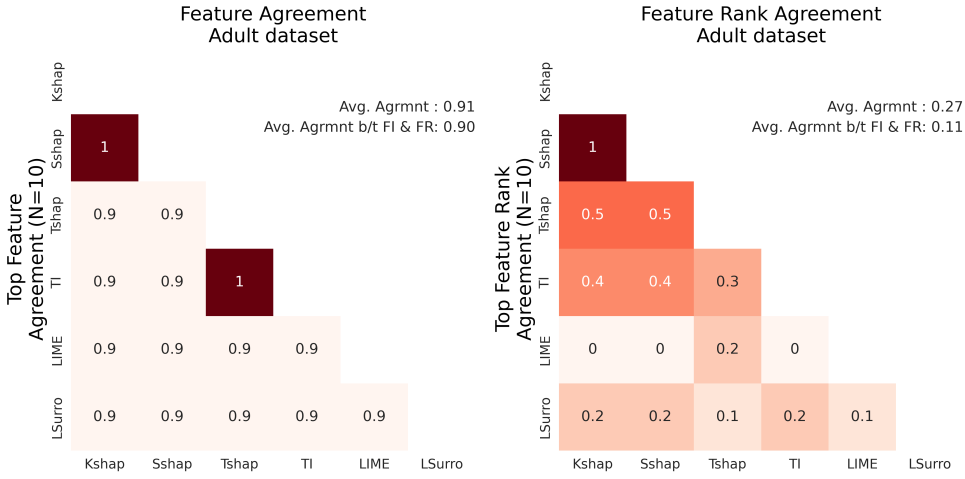


Figure 3.7: Feature and rank agreements for ADULT INCOME Income dataset.

Methods	# Features for 90% Accuracy	Accuracy with 5 feature(%)	Mean consistency	Mean Stability
Kshap	**1**	09	0.43	**0.00**
Tshap	**1**	09	0.43	**0.00**
Sshap	**1**	08	0.43	0.01
LIME	**1**	07	**0.15**	0.07
TI	**1**	10	0.5	**0.00**
LSurro	3	**100**	0.7	0.01

Table 3.4: Compactness, mean consistency and stability for ADULT INCOME dataset.

SHAP explainers generate the most consistent explanations for German Credit Risk dataset. Figure 3.8 and Table 3.5 show the computed metrics on GERMAN CREDIT RISK dataset. Overall, all the explainers achieve 90% of model accuracy with only 1 feature and SHAP explainers are the most consistent among the methods. LIME is the most stable among the explainers. Moreover, SHAP explainers have same top-10 most important

features as they all use the same mechanism of computation of the Shapley values the feature importance estimates, LIME have only 6 top features in common with SHAP explainers, and Kshap and Tshap share 7 features with the same rankings.

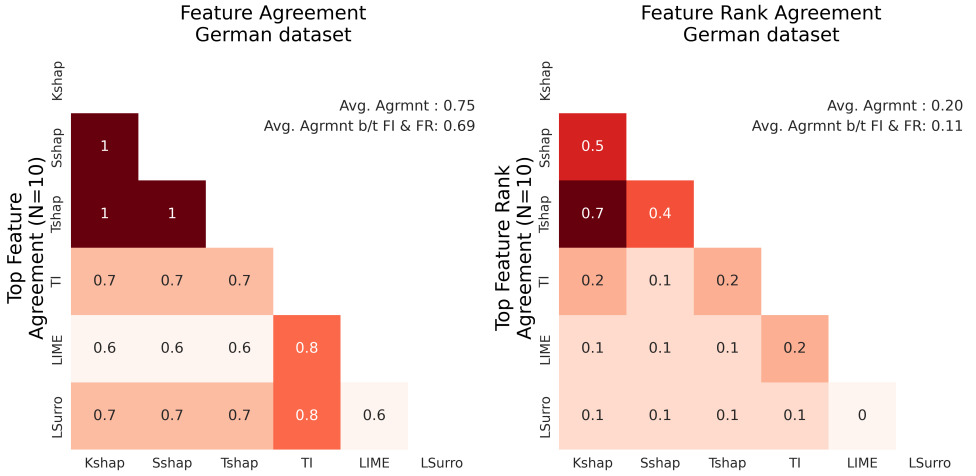


Figure 3.8: Feature and rank agreements for German Credit Risk dataset.

Methods	# Features for 90% Accuracy	Accuracy with 5 feature(%)	Mean consistency	Mean Stability
Kshap	**1**	12	**0.45**	0.03
Tshap	**1**	13	**0.45**	0.03
Sshap	**1**	12	**0.45**	0.03
LIME	**1**	41	1.18	**0.00**
TI	**1**	13	0.54	0.02
LSurro	**1**	**57**	0.74	0.03

Table 3.5: Compactness, mean consistency and stability for German Credit Risk dataset.

SHAP explainers and TI share the top-10 most important feature for Heart Diagnosis dataset. In Figure 3.9 and Table 3.6, SHAP explainers and TI share the top-10 most important feature, contrary to LIME which doesn't share same ranking of the top features with other explainers. Kshap and TI have exactly the same top-10 features and in the same rankings. On the other hand, LSurro estimates 100% of model accuracy with only 5 features and SHAP explainers and TI generate the most stable and consistent explanations.

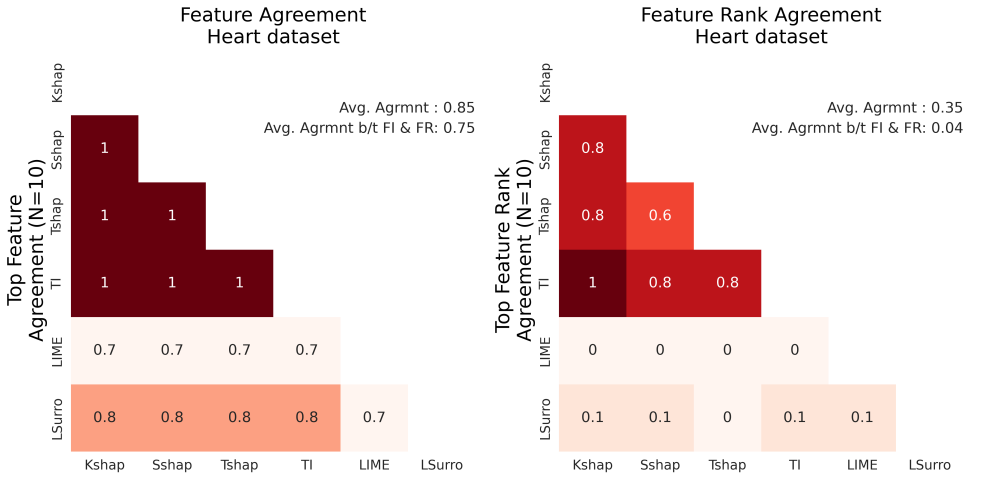


Figure 3.9: Feature and rank agreements for HEART DIAGNOSIS dataset.

Methods	# Features for 90% Accuracy	Accuracy with 5 feature(%)	Mean consistency	Mean Stability
Kshap	**1**	18	**0.4**	**0.00**
Tshap	**1**	21	**0.4**	**0.00**
Sshap	**1**	18	**0.4**	**0.00**
LIME	**1**	31	1.2	0.03
TI	**1**	25	0.45	**0.00**
LSurro	6	**100**	0.63	0.03

Table 3.6: Compactness, mean consistency and stability for HEART DIAGNOSIS dataset.

Local surrogates are the most consistent and explains 100% of model output with 5 features for Cervical Cancer dataset. Figure 3.10 and Table 3.7 illustrate above metrics on the CERVICAL CANCER dataset. Tshap and Kshap share the top 10 features. LIME and Sshap have the lowest top features in common. LIME and LSurro have no comparable rankings of the the features with TI, although they both learn a surrogate linear model in the neighborhood of each instance but use different mechanisms of the generation of the local neighborhood, which can explain the disagreement in their features importance estimates. Additionally, LSurro is the most consistent and explains 100% of model output with 5 features, and SHAP explainers have the highest mean stability.

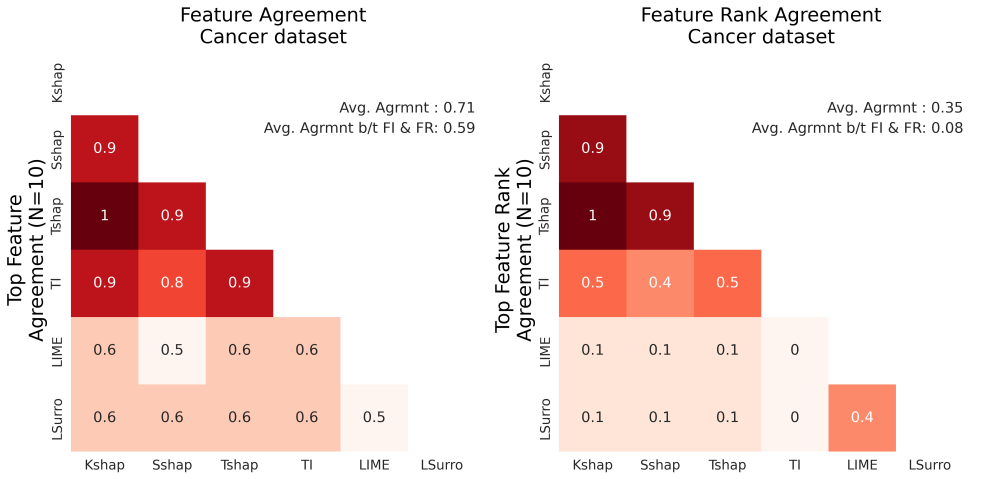


Figure 3.10: Feature and rank agreements for CERVICAL CANCER dataset.

Methods	# Features for 90% Accuracy	Accuracy with 5 feature(%)	Mean consistency	Mean Stability
Kshap	**1**	12	1.44	**0.34**
Sshap	**1**	12	0.87	**0.34**
Tshap	**1**	12	0.36	**0.34**
LIME	**1**	45	0.66	0.46
TI	**1**	19	0.59	0.42
LSurro	3	**100**	**0.22**	0.54

Table 3.7: 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. LSurro have the highest mean stability, while SHAP variants have the highest mean consistency.

3.1.7 Conclusion

The experiments on the synthetic and real-world datasets showed that feature importance attribution can be affected by multiple factors such as data properties, the black box model and the assumptions on which the explainable method is built to attribute feature contributions. We restricted our study to the first factor with a focus on some data properties, decision tree based models, on tabular data and for a binary classification task. For the matter of simplicity and ease of understanding of the model's behavior, we restricted our generation model to two variables in order to easily track the

feature interactions, which can be a limit in real-world scenarios because of the need to to handle high-dimensional data in many situations.

- For datasets with irrelevant variables, avoid using LSurro and LIME because both overestimate the importance of unimportant features.
- We recommend to avoid using TI for highly noisy datasets because TI is the most unstable compared to other explainers in such datasets. This can be justified by the decomposition of of the feature importance used by TI, which allocates importance to the noise.
- Kernel, Sampling and Tree SHAP explainers give very similar explanations, thus we recommend using Sshap or Tshap for faster computations and adaptability for decision trees.

Perspectives. Future work should further focus on each single explainability method separately to be able to explore in depth the effect of its assumptions and its inner workings for a single model parameter and on one specific data property on the feature importance attribution. Also, it is of interest to assess these feature importance estimates on other data parameters such as the number of instances in the test set and the number of features. Our work can be applied on other tasks such as regression and multi-class classification, to image and text data.

3.2 Application on Cervical Cancer Risk Assessment

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 section, 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.

3.2.1 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 [92]. 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 [44]. 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, [74, 48] compare the performances of random forests, deep learning and Naive Bayes for predicting cervical cancer risk, while [59] 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 [14, 51, 20, 19, 15] 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, [62] 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 [47, 7, 12, 9, 69].

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.

3.2.2 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 [14, 51, 20, 19, 15] 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 [102] which are applied on tabular and image datasets to predict cervical cancer risk. Similarly, [62] 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 black box machine learning models 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*. [77] 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, [86] and [45] output a simple program or an influence function, respectively. Other approaches have used input gradients to characterize local logic [57, 83]. However, such local explanations often do not match with human notions of contexts [61]: 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. [1, 17, 107] have identified and described important properties and desiderata for explanations. Among these, [107] provide an overview of various metrics for evaluating explanation types. [53] conduct a user survey to build a taxonomony of desired properties of explanations; [98] 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* [58]. Others such as [17] 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 [17] in the context of cervical cancer risk assessment. Unlike [1], 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. [99]) which we want to avoid. Instead, we compute the ROAR metric to measure the impact of removing top features on the model performance.

3.2.3 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 3.11. 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 preferable to others for cervical cancer risk assessment.

Problem Setup

We propose a multistage analysis pipeline. First, we test the performances of supervised learning models $f : \mathbb{R}^{N \times D} \rightarrow \{0, 1\}$ for predicting cervical cancer risk. From these models, we find the model f^* that best predicts a patient's risk of cervical cancer based on their data. Assume $\mathcal{L}$ denotes the loss function used to train f, $\mathbf{x}$ represent an instance of interest and ψ

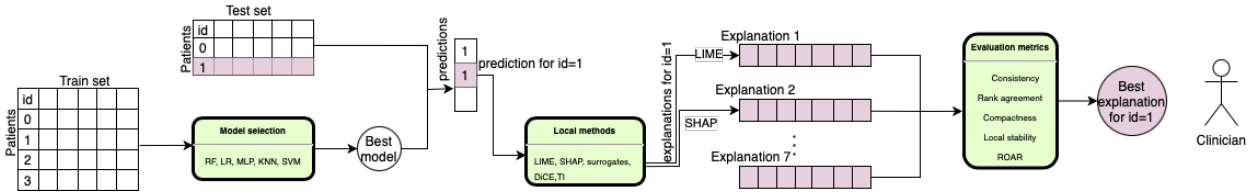


Figure 3.11: Ilustration of the two stages pipeline. (1) choosing the best model and (2) selecting the best explanation for individual patients.

represent a regularization term added to $\mathcal{L}$ to prevent overfitting.

$$\mathcal{L}(y, f(\mathbf{x})) := -\frac{1}{N}\sum_{i=1}^{N} y_i \log f(x_i) + (1 - y_i)\log(1 - f(x_i)) + \psi(x_i), \quad (3.2)$$

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:

$$f^*(x_i) = \min_f \mathcal{L}(y_i, f(x_i)). \quad (3.3)$$

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 section are detailed in the background section. We focus on these methods as they are most widely used across several healthcare applications.

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 metrics in Background section, based on those in [17]. 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 (for instance, [42, 5]).

A number of works deem stability, consistency, compactness and faithfulness as important facets of interpretability for healthcare domains overall. These include [5] and [42]. 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.

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].

3.2.4 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 from the UCI repository [24] to predict whether a female has high or low risk of getting diagnosed with cervical cancer. This data contains 858 female patients

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

Algorithm 1 A Local Feature Contribution Assessment Framework

Data: $f \in F, g \in G, m \in M, w \in W$

▷ F:models, G: explainable models, M: explainability metrics and W: weights of the explanations

Result: θ^*

$\theta \leftarrow \emptyset$

$\mathcal{D} \leftarrow ADASYN(\mathcal{D})$

$X_{train}, y_{train}, X_{test}, y_{test} \leftarrow split(\mathcal{D})$

while f *in* F **do**

  $f(\mathbf{x}) = \text{argmin}\, \mathcal{L}(\mathbf{y}, f(\mathbf{x}))$ ▷ Learn the best model.

  remove f from F

end

$f^* \leftarrow \min_f \mathcal{L}(\mathbf{y}, f(\mathbf{x}))$

while $g \in G$ **do**

  append featureImportance(g, X_{test}) to θ ▷ featureImportance computes feature importance of the test set using the explainable model g.

end

while $m \in M$ **do**

  $scores \leftarrow evaluate(\theta, m)$ ▷ Evaluate each explainable model with each metric.

end

$\theta^* \leftarrow \max(\sum_{i=1}^{M} scores_i * w_i)$ ▷ w_i are defined by the clinician for the desired explanation properties.

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

characterized by 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 3.8.

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 B.1 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 3.2. 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

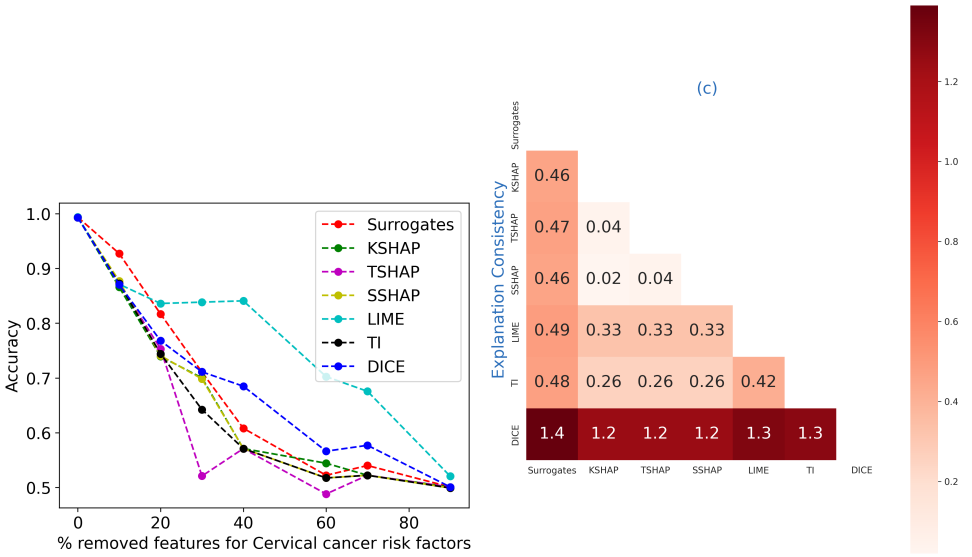
SHAP (KSHAP) and Sampling SHAP (SSHAP), Tree Interpreter, DiCE and Local Surrogates. These provide local explanations for individual instances, highlighting the contribution of each feature towards the model's prediction. We apply these methods to the test set.

3.2.5 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 3.12. 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 withSHAPpredominantly 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. UnlikeSHAPmethods, 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 3.9 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.

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 B.1 in



(a) ROAR performances

(b) Consistency of the XAI methods

Figure 3.12: (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.

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 3.9: 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, whileSHAPvariants have the highest mean consistency.

Appendix B. The results for Patient 0 and 1 are shown in Figures 3.13 and B.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 B.1 in Appendix B, who is diagnosed as not having cancer. These results are shown in Figure B.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,SHAPexplainers 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 3.15. 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.

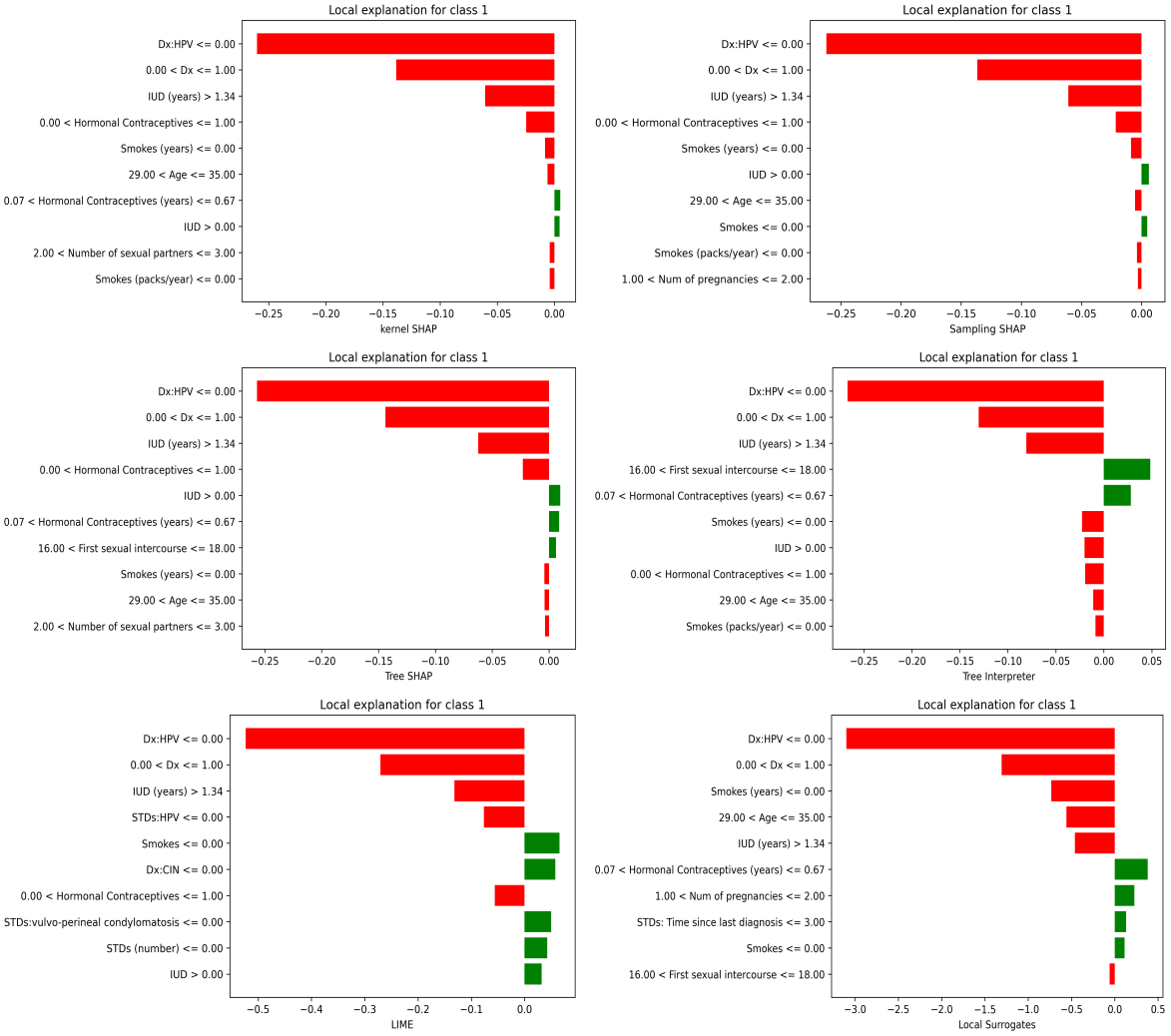


Figure 3.13: Patient 0 diagnosed as not having cancer (Dx:Cancer=0).

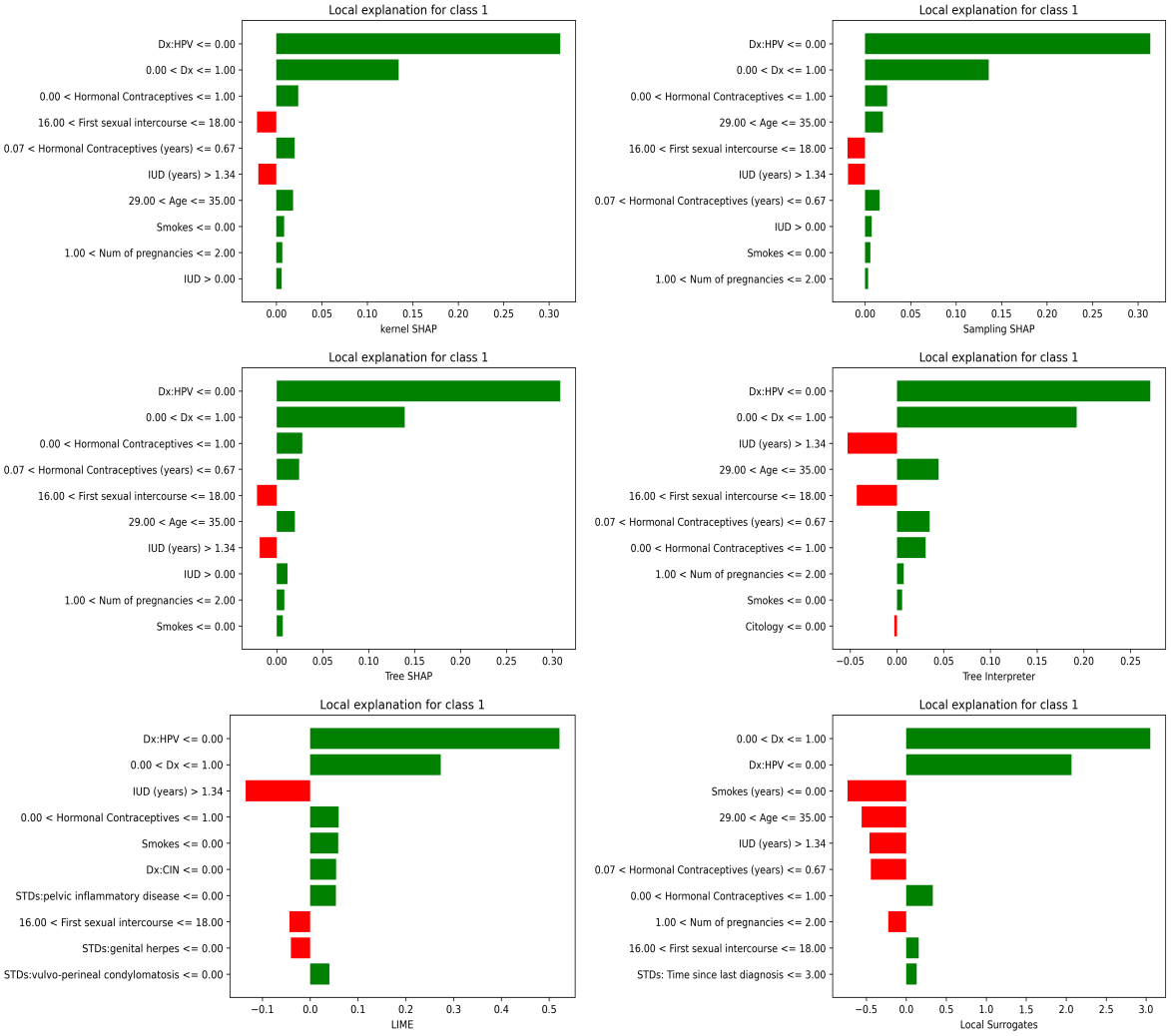
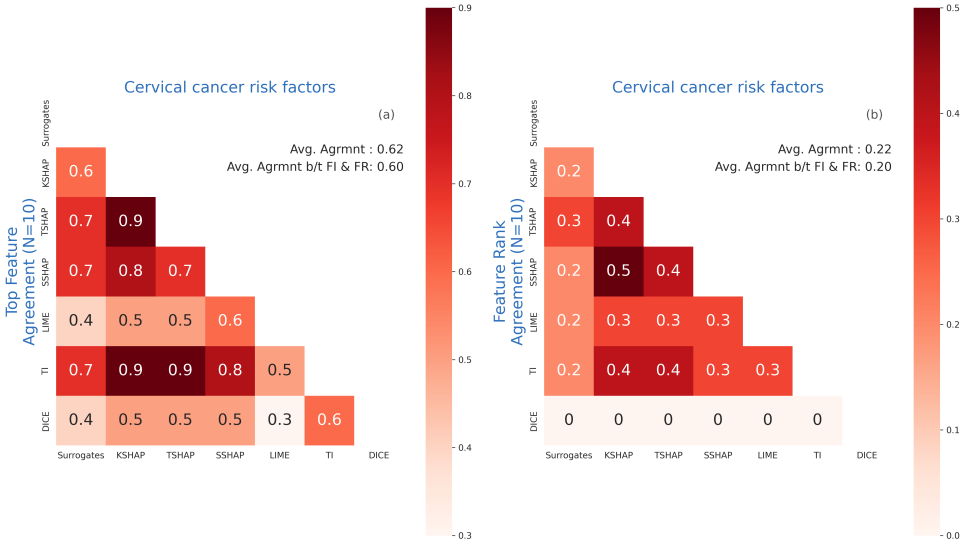
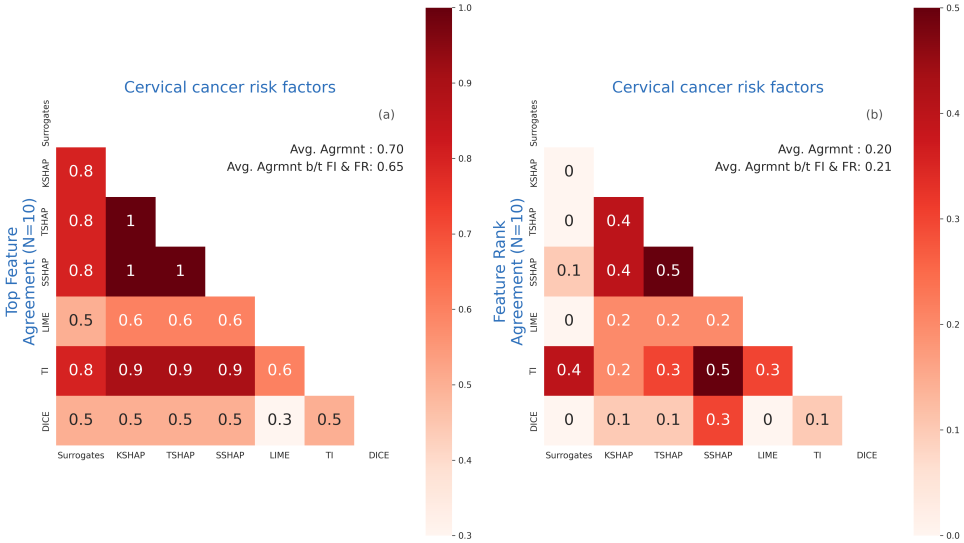


Figure 3.14: Patient 1 diagnosed with cancer (Dx:Cancer=1).



(a)



(b)

Figure 3.15: Feature and rank agreements for the Patients 0 and 1. For each patient in sub-figures, (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.

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. These results are shown in Figure B.1 in Appendix B and Figure 3.16 respectively. Figure B.1 in Appendix B. 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 surrogates 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 3.16 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 [18, 13]. 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 [13] . Yet for Patient 1, we see contraception may play a role, which is not always the case for other patients in the cohort (Appendix B for details)

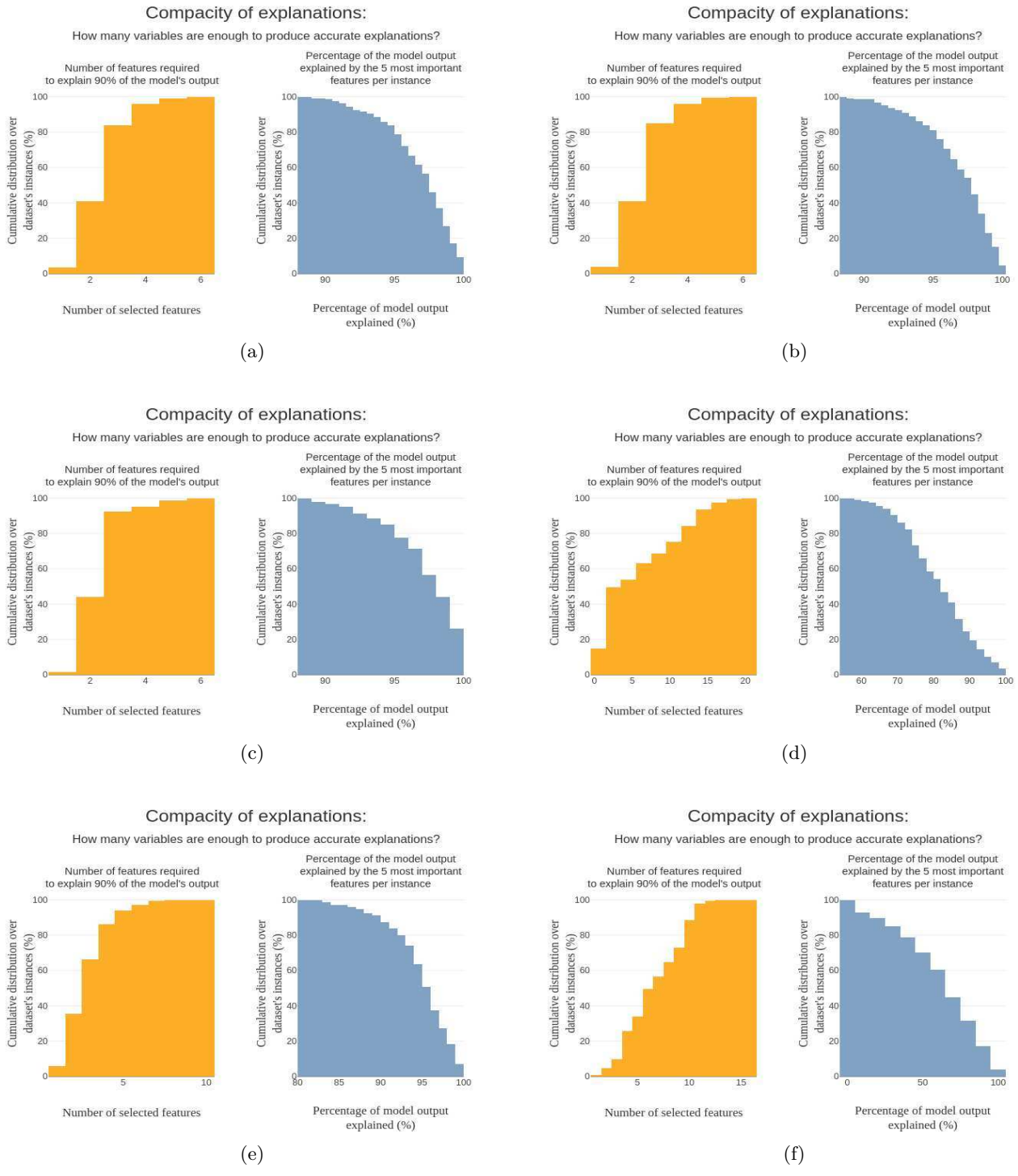


Figure 3.16: Compactness of the explanations generated by (a) Kernel SHAP, (b) Sampling SHAP, (c) Tree SHAP, (d) LIME,(e) Tree Interpreter, (f) Local surrogates.

3.2.6 Discussion

In this section, 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.

3.2.7 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 work, 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.

Chapter 4

Feature Importance Chains

4.1 Computing Indirect Feature Importance with Shapley Chains

In spite of increased attention on explainable machine learning models, explaining multi-output predictions has not yet been extensively addressed. Methods that use Shapley values to attribute feature contributions to the decision making are one of the most popular approaches to explain local individual and global predictions. By considering each output separately in multi-output tasks, these methods fail to provide complete feature explanations. We propose Shapley Chains to overcome this issue by including label interdependencies in the explanation design process. Shapley Chains assign Shapley values as feature importance scores in multi-output classification using classifier chains, by separating the direct and indirect influence of these feature scores. Compared to existing methods, this approach allows to attribute a more complete feature contribution to the predictions of multi-output classification tasks. We provide a mechanism to distribute the hidden contributions of the outputs with respect to a given chaining order of these outputs. Moreover, we show how our approach can reveal indirect feature contributions missed by existing approaches. Shapley Chains help to emphasize the real learning factors in multi-output applications and allows a better understanding of the flow of information through output interdependencies in synthetic and real-world datasets.

4.1.1 Introduction

A multi-output model predicts several outputs from one input. This is an important learning problem for decision-making involving multiple factors and complex criteria in the real-world scenarios, such as in healthcare, the prediction of multiple diseases for individual patients. Classifier chains [76] is one such approach for multi-output classification, taking output dependencies into account by connecting individual base classifiers, one for each output. The order of output nodes and the choice of the base classifiers are two parameters yielding different predictions thus different explanations for the given classifier chain.

To address the lack of transparency in existing machine learning models, solutions such as SHAP [56], LIME [77], DEEPLIFT [85] and Integrated Gradients [91] have been proposed. Using Shapley values [80] is one approach to attribute feature importance in machine learning. The framework SHAP [56] provides Shapely values used to explain model predictions, by computing feature marginal contributions to all subsets of features. This theoretically well founded approach provides instance-level explanations and a global interpretation of model predictions by combining these local (instance-level) explanations.

However, these methods are not suitable for multi-output configurations, especially when these outputs are interdependent. In addition, the SHAP framework provides separate feature importance scores only for independent multi-output classifiers. By assuming the independence of outputs, one ignores the indirect connections between features and outputs, which leads to assigning incomplete feature contributions, thus an inaccurate explanation of the predictions.

Fig. 4.1 is a graphical representation of a classifier chain: patients with two conditions, obesity (y^{OB}) and psoriasis (y^{PSO}), given four features: genetic components (X^{GC}), environmental factors (X^{EF}), physical activity (X^{PA}) and eating habits (X^{EH}). From a clinical point of view, all factors X are associated with both conditions Y, obesity and psoriasis. However, since obesity is a strong feature for predicting psoriasis [41] (indeed, a motivating factor for using such a model is that predictive accuracy can be improved by incorporating outputs as features), it may mask the effects of other features. Namely, X_{PA} and X_{EH} will be found by methods as SHAP applied to each

output separately to have zero contribution towards predicting y^{PSO}, and one might interpret that psoriasis is mainly affected by factors which cannot be modified by the patient (environment and genetics). The *indirect* effects (physical activity and eating habits) will not be detected or explained.

We propose Shapley Chains to address this limitation of incomplete attribution of feature importance in multi-output classification tasks by taking into account the relationships between outputs and distributing their importance among the features with respect to a given order of these outputs. Calculating the Shapley values of outputs helps to better understand the importance of the chaining that connects these outputs and to visualize this relationship impact on the prediction of subsequent outputs in the chain. For these subsequent outputs, the computation of the Shapley values of the associated outputs shows the indirect influence of some features through the chain, which is generally not intuitive and missed by existing work. Our method will successfully explain these *indirect* effects. By attributing importance to the features X^{PA} and X^{EH}, Shapley Chains will help doctors to emphasize the importance of eating healthy and practicing physical activities in order to prevent and better cure psoriasis instead of blaming only genetics and exterior environmental factors.

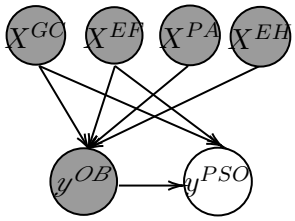


Figure 4.1: An example of a multi-output task: predicting Y-outputs from X-features. A classifier chain uses the first output y^{OB} as an additional feature to predict the second output y^{PSO}.

This work addresses the problem of attributing feature contributions in multi-output classification tasks with classifier chains when outputs are interdependent. Our contribution in this work is resumed to :

- We propose Shapley Chains, a novel post-hoc model agnostic explainability method designed for multi-output classification task using classifier chains.

- Shapley Chains attribute feature importance to all features that directly or indirectly contribute to the prediction of a given output, by tracking all the related outputs in the given chain order.
- Compared to existing methods, we show a more complete distribution of feature importance scores in multi-output synthetic and real-world datasets.

We devote Section 4.1.2 related work. In Section 4.1.3, we detail our proposed method Shapley Chains. Finally in Section 4.1.4, we run experiments on a synthetic and real-world datasets. The results of our method compared to SHAP values applied to independent classifiers are then discussed.

4.1.2 Related Work

The explainability of machine learning is an active research topic in the recent years. Several contributions have been made to explain single-output models and predictions. Inspecting feature importance scores of existing models is an intuitive approach that has served for many studies. These feature importance scores are either derived directly from feature weights in a linear regression for instance, or learned from feature permutations based on the decrease in model performance. Other more complex methods like LIME [77] learn a surrogate model locally (around a given instance) in order to explain the predictions of the initial model with simple and interpretable models like decision trees. On the other hand, DeepLift [85], Integrated gradient [91] and LRP [65] are some neural network specific methods proposed to explain deep neural networks. The SHAP framework is one popular method attributing Shapley values as feature contributions. It provides a wide range of model-specific and model-agnostic explainers. Researchers have also proposed other Shapley value inspired methods incorporating feature interactions in the explanation process. For example, asymmetric Shapley values [27] incorporates causal knowledge into model explanations. This method attributes importance scores to features that do not directly participate in the prediction process (confounders), but fails to capture all direct feature contribution. On the other hand, On-manifold Shapley values [26] focus on better representing the out of coalition feature values but provides misleading interpretation of feature contributions. Wang et al. [101] have

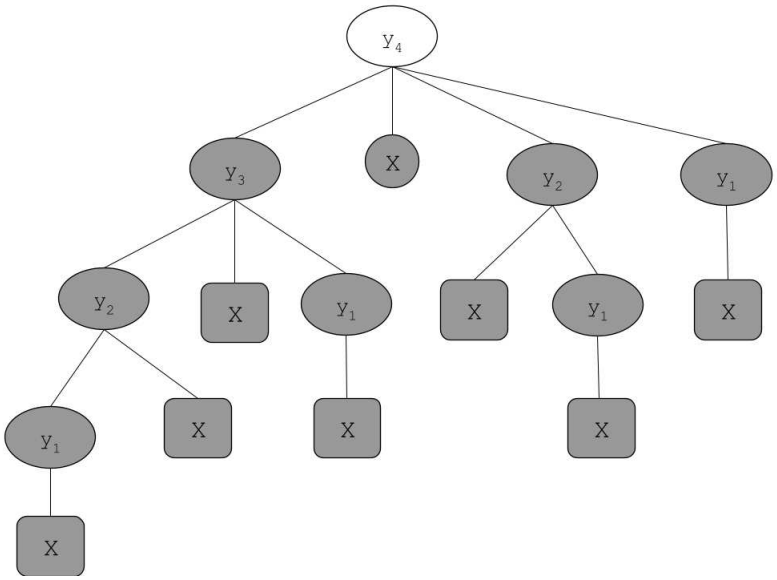


Figure 4.2: Representation of direct and indirect contributions for a dataset with 4 outputs (y^1, y^2, y^3 and y^4). For example: the 4th output y^4 has 7 indirect Shapley values (7 paths ending with square leaves) and one direct Shapley value (one path ending with a circle leaf).

proposed Shapley Flow, providing both direct and indirect feature contributions when a causal graph is provided. Assuming the causal graph is available and accurate for real-world data sets limits the applicability of this method. These methods significantly contributed to advancing the explainability of machine learning models but none of them have tackled multi-output problems, more specifically when outputs are interdependent. Shapley Chains address this limitation.

4.1.3 Proposed Method: Shapley Chains

In this section, we introduce our approach to compute direct and indirect feature Shapley values for a classifier chain model. Note that our proposed method is model-agnostic, meaning that our computations do not depend directly on the chosen base learner used by the classifier chain.

We want to compute feature contributions to the prediction of each output $y^j \in Y$ for each instance $\mathbf{x}$. For example, Fig. 4.2 shows the direct and indirect contributions of x^i to predict output y^4 given in Fig. 2.3 (B). In the

next two sections, we detail the computations of the Shapley value of each feature to predict each output. We refer to these Shapley values as direct and indirect feature contributions.

Direct contributions. The direct contributions are computed for features and outputs as in Eq. 2.2. Consider again the example of patients with the two conditions: psoriasis and obesity. For both y^{OB} and y^{PSO}, we use the framework SHAP in order to compute the Shapley value of each feature : X^{GC}, X^{EF}, X^{PA} and X^{EH}. This will attribute non zero Shapley values to X^{GC} and X^{EF} to predict y^{OB} and y^{PSO} separately. On the other hand, X^{EF} and X^{PA} will have non-zero Shapley values to predict y^{OB} and zero values for the prediction of y^{PSO}. The classifier chain method will add y^{OB} to the feature set to predict y^{PSO}. By running the SHAP framework on this new set, y^{OB} will have a non zero Shapley value because it is dependent to y^{PSO}. This Shapley value will be attributed to the features that are correlated to y^{OB}. The attribution mechanism of direct feature (and output) contributions can be generalized to the classifier H with m base classifiers as shown in Algorithm 2.

Algorithm 2 Computing direct feature contributions

procedure DICONTRIBUTION(X, Y, H) ▷ features, outputs, classifier chain model
  $j \leftarrow 1$, $\Phi \leftarrow \emptyset$
while $j < m$ **do**
  $i \leftarrow 1$
  while $i < d$ **do**
    $\phi_{x^i}(y^j) \leftarrow \mathsf{SHAP}(X, y^j, H)$ ▷ Shapley values of inputs w.r.t. y^j
    append $\phi_{x^i}(y^j)$ to Φ
    $i \leftarrow i + 1$
  end
  append y^j to X
  $j \leftarrow j + 1$
  end
  return Φ ▷ Φ is the set of the direct feature importance of features and the labels that were included as features.
end procedure

For the first output y^1, we calculate the Shapley value of each feature according to Eq. 2.2, as done in the SHAP framework. This marginal value of all possible subsets to which the feature can be associated to is the feature's

contribution to predict the first output y^1. For the second output y^2, we append the predictions y^1 made by the first classifier h_1 to the features set, and we train a second classifier h_2 to learn the second output y^2. We again use the SHAP framework to assign Shapley values to features and the first output y^1. Here, the feature set includes the first prediction. We perform the same steps for each remaining output. At each step, we calculate the Shapley values for features and previous predicted outputs that are linked via the chaining to the current output. At the final step, the feature set will contain the n features and $m-1$ outputs: $X = \{x^1, x^2, ..., x^n, y^1, y^2, ..., y^{m-1}\}$.

Indirect contributions. The indirect contribution $\Phi_{x^i{}_{indirect}}(y^j)$ of x^i to predict y^j is the weighted sum of the direct contributions of all $y^k \in Y$ that are chained to y^j. $\Phi_{x^i{}_{indirect}}(y^j)$ is computed according to the Eq. 4.1.

$$\Phi_{x^i{}_{indirect}}(y^j) = \sum_{k=1}^{j-1} \Phi_{y^k}(y^j) \cdot Z_k(x^i) \tag{4.1}$$

where $j > 1$ and the function $Z_k(x^i)$ computes the weight vector for all paths from output y^k down to x^i. For $k > 1$ and $Z_1(x^i) = W(y^1, x^i)$, $Z_k(x^i)$ is recursively computed as follows:

$$Z_k(x^i) = \sum_{l=1}^{k-1} W(y^k, y^{k-l}) \cdot Z_{k-l}(x^i) + W(y^k, x^i) \tag{4.2}$$

where $W(y^k, y^{k-l})$ is the corresponding weight of y^{k-l} to predict the next output y^k (the direct contribution of y^{k-l} to predict y^k). And, $W(y^k, x^i)$ is the weight of x^i to predict y^k (the direct contribution of x^i to predict y^k). The weights $W(y^k, y^{k-l})$ and $W(y^k, x^i)$ are calculated according to:

$$W(y^k, .) = \frac{|\Phi_.(y^k)|}{\left(\sum_{q=1}^{n} |\Phi_{x^q}(y^k)| + \sum_{p<k} |\Phi_{y^p}(y^k)|\right)} \tag{4.3}$$

where $\Phi_{x^q}(y^k)$ is the direct contribution, as in Eq. 2.2; of each feature x^q to predict y^k. $p < k$ means the output p is chained to the output j forming a directed acyclic graph illustrated in Fig. 2.3.

The overall feature Shapley values for a classifier chain are obtained by marginalizing over all possible output chain structures. Specifically, the contribution of feature x^i to the prediction of output y^j within a given

chaining order c is computed as:

$$\phi_{x^i}(y^j) = \frac{1}{|\mathsf{C}|} \sum_{c \subseteq \mathsf{C}} \phi_{x^i}(y^j)^c$$

In this equation, $\phi_{x^i}(y^j)^c$ represents the contribution of feature x^i to the prediction of y^j with respect to the chaining order c and C is the possible chain orders. We want to consider the average feature importance if the chain order is not fixed. By reporting feature contributions for each chain structure independently and demonstrating the impact of different chaining orders through marginalization, we aim to elucidate the significance of considering interdependencies among outputs in feature attribution, as discussed in Section 4.1.4.

For instance, in order to have a complete fair distribution of feature importance for the prediction of y^{PSO}, we compute the indirect Shapley values of the features X^{PA} and X^{EH}. We do so by distributing the direct Shapley value of y^{OB} computed previously to the four features. By the distribution operation, we mean the multiplication of the direct Shapley value of each feature by the direct Shapley value of y^{OB}, divided by the sum of the Shapley values of all features for to predict the same output (y^{OB}).

We generalize this mechanism in Algorithm 3 of calculating indirect Shapley values to the chain structure in Fig. 2.3 (B). The first output y^1 has always zero indirect Shapley values because there is no output that precedes it in the chaining. Thus, for the rest of this section, we compute feature indirect contributions for $y^j \in \{y^2, y^3, ..., y^m\}$. For each output y^j, there exists one direct path to the features thus one direct feature contributions and $2^j - 1$ indirect paths for each feature.

One should notice that for the matter of the simplicity of understanding, we take the absolute value in Eq. 4.3. Thus, all the contributions will be positive. These absolute values can be replaced by the raw Shapley values in order to keep the positive or negative sign of feature contributions. Keeping the sign helps to understand if the feature penalizes or is in favor of the prediction.

In classifier chains, the sequence in which labels are considered plays a crucial role in optimizing model performance. Traditionally, labels are classified based on their meaning or their association with other labels. Several strategies exist for determining this ordering, including randomization,

Algorithm 3 Computing indirect feature contributions to predict a given y^j with $j > 1$

procedure INCONTRIBUTION(X, Y, Φ) ▷ inputs, outputs, Shapley values of features and outputs
  $k \leftarrow 1$
while $k < j$ **do**
  $l \leftarrow 1$
  while $l < k$ **do**
    compute $W(y^k, y^{k-l})$
    $i \leftarrow 1$
    while $i < d$ **do**
      compute $W(y^k, x^i)$ in Eq. 4.3
      compute $Z_k(x^i)$ in Eq. 4.2
      $i \leftarrow i + 1$
    end
    $l \leftarrow l + 1$
  end
  $k \leftarrow k + 1$
  end
  return $\Phi_{indirect}$ in Eq. 4.1 ▷ returning indirect feature contributions.
end procedure

frequency-based ordering, or the use of correlation matrices between features and labels and between labels. Direct feature importance reflects correlations between features and labels, while indirect feature importance reflects correlations between the labels. Therefore, leveraging the concept of indirect feature importance offers a promising approach to help choose the sequence (order) in which these labels are learned.

4.1.4 Experiments

In order to assess the importance of the features that is attributed by our proposed framework[1] to explain their contributions to predict multiple outputs with a classifier chain, we run experiments on both synthetic and real-world datasets: a XOR data that we describe next, and the ADULT INCOME dataset from the UCI repository [23]. Here, we rely on human explanation to validate our results.

[1] https://github.com/cwayad/Shapleychains

Synthetic Datasets

To demonstrate our work, we first run experiments on a multi-output synthetic dataset containing two features (x^1 and x^2) and three outputs (AND, OR and XOR) corresponding to the logical operations of the same names performed on x^1 and x^2. We split this dataset to 80% for the training and 20% for the test of our classifier.

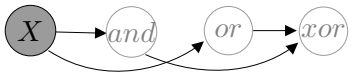
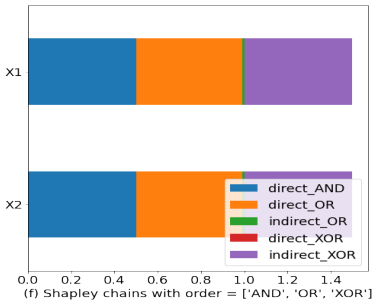


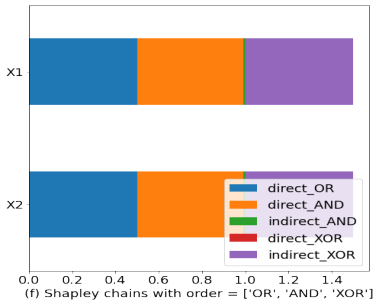
Figure 4.3: The classifier chain structure for XOR data. X is the set of features x^1 and x^2. AND, OR and XOR are the outputs for which we want to compute direct and indirect feature importance and uncertainty intervals.

Next, we construct a classifier chain with the chaining order illustrated in Fig. 4.3. We use a logistic regression as the base learner. Our method is model agnostic meaning that it can be applied to a classifier chain with any other base learners. The use of the logistic regression as the base learner to predict XOR is justified by the accuracy that this model achieves compared to other classifiers like decision trees. The classifier chain is trained on the train set using x^1 and x^2 to predict AND and OR separately. Then, we append these two predicted outputs to the features set in order to predict XOR. Here, the order in which we predict AND and OR does not change our method's behavior.

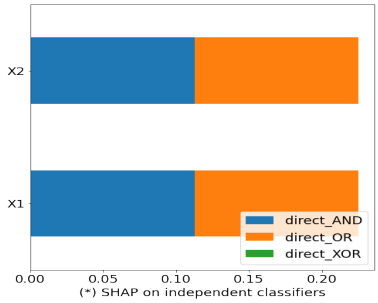
To explain the influence of x^1 and x^2 on the prediction of XOR, we compared the application of the framework SHAP on each classifier independently and Shapley Chains on the trained classifier chain. We report our analysis on the test data. The results of the comparison shown in Fig. 4.4 indicate that the output chaining propagates the contributions of x^1 and x^2 to predict XOR via AND and OR. Specifically, Fig. 4.4(a) and Fig. 4.4(b) illustrate that our method detects the indirect contributions of x^1 and x^2 (indirect_xor) to predict XOR thanks to the chaining of AND and OR to XOR implemented with the classifier chain model, which tracks down all feature contributions through the chaining of outputs. Furthermore, Fig. 4.4(a) and Fig. 4.4(b) confirm that predicting OR before AND or vice versa does not affect the feature contributions attribution, which confirms the chain structure for this data. On the other hand, these contributions of x^1 and x^2



(a) Chain order= [AND, OR, XOR]



(b) Chain order= [OR, AND, XOR]



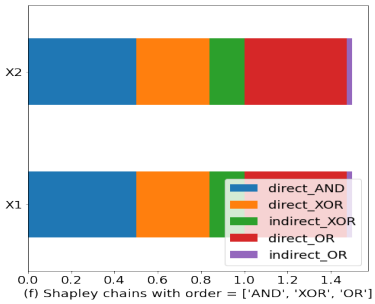
(*) Independent SHAP

Figure 4.4: A comparison of SHAP applied on independent classifiers and Shapley Chains. From the left to the right. (a) and (b) Normalized direct and indirect feature contributions made by Shapley Chains to predict AND, OR and XOR for chain orders [AND, OR, XOR] and [OR, AND, XOR]. ($*$) Independent SHAP assigns contributions to x^1 and x^2 only to predict AND and OR outputs and completely misses their contributions to predict XOR. Absent colors refer to null Shapley values.

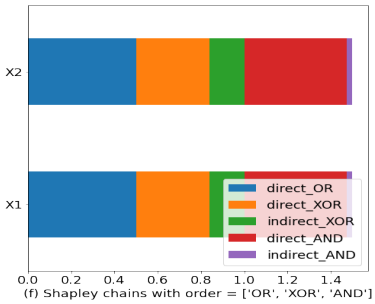
are completely neglected by the SHAP framework on independent classifiers (Fig. 4.4($*$)).

Impact of the chaining order on the classifier chain explainability. In order to measure the impact of the chaining order on the explainability of our classifier chain model with Shapley Chains, we performed analysis on the $3! = 6$ possible output chaining orders in the synthetic dataset (scenarios (a) and (b) in Fig. 4.4 and scenarios (c), (d), (e) and (f) in Fig. 4.5).

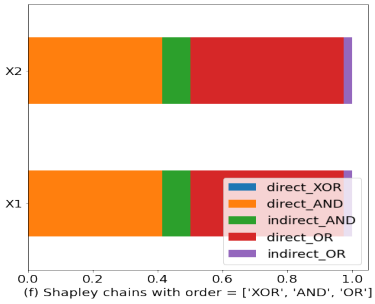
The information known to the classifier chain when training each output changes depending on the order of these outputs. For instance, in scenarios a and b (Fig. 4.4), we first learn the two outputs AND and OR using x^1 and x^2 features. XOR is then predicted using AND and OR. Here, in both scenarios, both features x^1 and x^2 contribute indirectly (through AND and



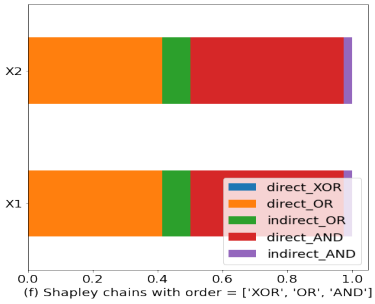
(c) Chain order= [AND, XOR, OR]



(d) Chain order= [OR, XOR, AND]



(e) Chain order= [XOR, AND, OR]



(f) Chain order= [XOR, OR, AND]

Figure 4.5: Possible output chaining orders for XOR data. Normalized total feature contributions (direct and indirect Shapley values) for c, d, e and f.

OR) to predict XOR. Meanwhile in the scenario c (or d), the model relies on AND (or OR), x^1 and x^2 to predict XOR. We observe that x^1 and x^2 have direct and indirect contributions, meaning that the classifier chain relies partially on these two features to predict XOR (direct contributions of x^1 and x^2), and on AND (indirect contributions of x^1 and x^2 via AND). The last two scenarios e and f show no contribution of x^1 and x^2 to predict XOR, which is explained by the fact that using only these two features, the model can not predict XOR without having the information about the dependencies of XOR to AND and OR.

These results show that the chain order of AND, OR and XOR outputs has an important role in the explainability of the classifier chain, because feeding different inputs to the classifier chain yields different predictions, thus different Shapley values are attributed to the features. x^1 and x^2 importance scores can either be derived from a direct inference of XOR output only if there is additional information on output dependencies (for example AND is linked to XOR) or by extracting it from the chain that links

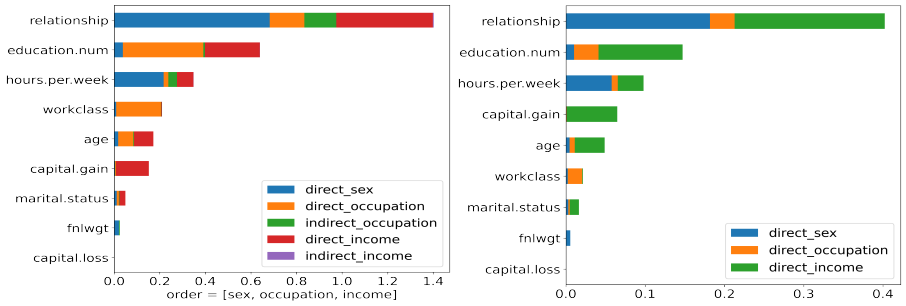
AND and OR to XOR. In the absence of all output dependencies of AND or OR to XOR, the model completely ignores the importance of features x^1 and x^2 in the prediction of XOR.

Explaining Adult Income with Shapley Chains

We run Shapley Chains on the UCI Adult Income dataset. This dataset contains over 32500 instances with 15 features. We first perform a nominal encoding on *workclass*, *marital status* and *relationship* features. We remove *race*, *education* and *native country* and normalize the dataset with the min/max normalizer. Next, we split it into two subsets, using 80% for the training and the remaining 20% for testing. We evaluated the hamming loss of a classifier chain with different base learners and we kept the best base classifier, the logistic regression in this case.

In order to explain feature contributions to the predictions of the three outputs *sex*, *occupation* and *income*, we compared the results of Shapley Chains against classic Shapley values applied on separate logistic regression classifiers for different chain orders. Fig. 4.6 shows graphical representation of normalized and stacked feature contributions when applying Shapley Chains on our data set (Fig. 4.6.(a)), and stacked feature contributions from independent logistic regression classifiers (Fig. 4.6.(b)). In both cases, the magnitude of the feature contributions is greater in Shapley Chains compared to independent Shapley values, which confirms our initial hypothesis of some contributions are missed by SHAP framework, and these contributions can be detected when we take into account output dependencies. For example, the number of hours worked in a week (*hours.per.week*) has a more important indirect contribution to predict individual's *occupation* than a direct contribution. This is explained by the fact that *sex* is related to *occupation*, and this relationship is propagated to the features by Shapley Chains. *relationship* is another example of Shapley Chains detecting indirect feature contributions to predict *occupation*. Furthermore, feature rankings are different in Shapley Chains. For example, the ranking of *capital.gain* comes in the fourth position (before *workclass*) using SHAP applied to independent classifiers. In our method, this feature's ranking is always less important (according to different chaining orders) than *workclass* to predict *sex*, *occupation* and *income* which makes more sens to us.

We also tested the impact of different chain orders of these three outputs

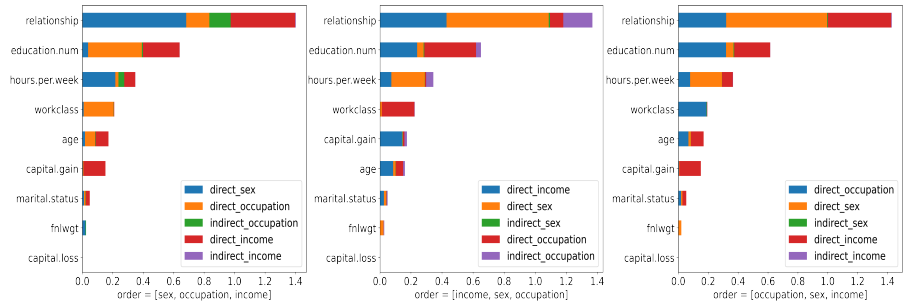


(a) Shapley Chains (b) SHAP on independent classifiers

Figure 4.6: (a) Direct and indirect Shapley values on Adult Income data: we normalize and stack each feature's direct and indirect contributions to each output. *sex* has only direct contributions because it is the first output we predict in this chain order. (b) Stacked Shapley values of independent classifiers on Adult Income data.

on the feature importance attribution. Fig. 4.7 illustrates three different chaining orders. Each different order allows each classifier to use different prior knowledge to learn these outputs. For example in Fig. 4.7(b), we first predict *income* and *sex* and we use this information to predict *occupation*. Intuitively, *occupation* is correlated to individual's *sex* and *income*. The classifier chain uses this information provided to the third classifier to predict *occupation*. Here, Shapley Chains attribute more importance to the factors that predict both *income* and *sex*, when predicting *occupation*. Shapley Chains preserve the order of feature importance scores across all the chaining orders in general, but the magnitude of each feature's importance differs from one chain to another. This is due to the prior knowledge that is fed into the classifier when learning each output. In addition, these feature importance scores are always more important in Shapley Chains compared to Shapley values of independent classifiers for all chain orders.

Shapley chains attributes Shapley Flow on the direct and indirect feature importance across the synthetic and real world datasets. Table 4.1 shows the mean distance to ground truth of the feature importance attributed by Shapley Chains, Shapley Flow, on manifold SHAP and independent SHAP. Only Shapley Flow and Shapley Chains compute the direct and indirect feature importance. Shapley Chains outperforms Shapley Flow across the synthetic and real-world datasets.



(a) (b) (c)

Figure 4.7: Stacked direct and indirect feature importance for 3 different chain orders over Adult Income data.

Method	ADULT(D)	ADULT(I)	XOR(D)	XOR(I)
Shapley Chains	3.42	0.22	2.46	3.50
Shapley Flow	3.66	2.31	10.00	3.45
Independent	1.90	NA	10.05	NA
On-manifold	4.69	NA	5.0	NA

Table 4.1: Mean distance to ground truth (lower distances represent more similar explanations to the ground truth). Similar to [101], we compare the distance to ground truth. The ground truth for the real world datasets is the weights of the local linear model trained on 10k instances. Shapley chains attribution is the most similar to ground truth for the direct and indirect feature importance across the datasets.

4.1.5 Conclusion

In this work, we presented Shapley Chains, a novel method for calculating feature importance scores based on Shapley values for multi-output classification with a classifier chain. We defined direct and indirect contribution and demonstrated on synthetic and real-world data how the attribution of indirect feature contribution to the prediction is more complete with Shapley Chains. Our method helps practitioners to better understand hidden influence of the features on the outputs by detecting indirect feature contributions hidden in output dependencies. Although the rankings of feature importance are not always different from independent feature importance scores, the magnitude of these scores is always important in Shapley Chains, which is more important to look at in applications that are sensitive to the magnitude of these importance scores rather than their rankings. By

extending the Shapley value to feature importance attribution of classifier chains, we make use of output interdependencies that is implemented in classifier chains in order to represent the real learning factors of a multi-output classification task.

To extend this work, Shapley Chains could be evaluated on multi-output regression tasks. Exploring the relationship's type between the outputs, and studying weather Shapley Chains preserve all these relationships when attributing feature contributions is another open question of our work.

4.2 Quantifying Uncertainty with Bayes LIME Chains

As machine learning models become more prevalent in various fields, the need for interpretability and decision understanding of these models becomes more critical. Local explanations offer valuable insights into the decision-making process of machine learning models at the local level, i.e. for individual predictions. However, ensuring the reliability, robustness, and stability of these explanations remains a challenge. Although methods like Bayes LIME, Bayes SHAP, and Bay LIME have introduced Bayesian frameworks to enhance the reliability of these explanations, they are limited in their applicability to multi-output settings involving interdependent labels. Quantifying these uncertainties is crucial to enhance the reliability of explanations and making informed decisions based on them. Meanwhile, Shapley Chains captures the indirect effects of a feature on the prediction of interdependent labels, its lack a mechanism for quantifying the associated uncertainties. To address this gap, we propose Bayes LIME Chains, a novel method aimed at quantifying the uncertainty surrounding the indirect contribution of each feature to the prediction of interdependent labels. Our findings suggest that the uncertainty intervals computed using Bayes LIME Chains are reliable and provide a comprehensive assessment of the uncertainty associated with feature importance estimates. These results underscore the effectiveness of our approach in generating explanations with quantified uncertainties.

4.2.1 Introduction

One way to gain insights into a model's behavior is by examining the contribution of individual features to the predicted output. LIME [77] and Kernel SHAP [56] are two popular methods to quantify the impact of a feature on

the model's prediction for a given instance. These two methods learn an interpretable surrogate model in the neighborhood of the instance to explain. While such methods have many advantages, prior research showed that the generated explanations lack desired properties such as the local stability, compactness, inconsistency and reliability [47, 97, 81, 28, 38].

To address these challenges, researchers introduced Bayesian frameworks like Bayes SHAP, Bayes LIME [87], and Bay LIME [105]. These methods aim to provide local explanations while also quantifying the uncertainty associated with those explanations. They allow users to specify the desired level of uncertainty and optimize the number of perturbations to achieve accurate explanations. However, they are not well-suited for multi-output tasks whith interdependent labels.

The interpretation of feature contributions becomes more complex when dealing with interdependent multiple labels. In such cases, it becomes necessary to take into consideration hidden dependencies among features and outputs. To remedy this, researchers have proposed Shapley Chains[8] that computes direct and indirect feature impacts in a chain of interdependent labels. Although the authors demonstrate the effectiveness of their approach on synthetic and real-world datasets, showing that it can identify important features that are missed by explainability methods designed for single output tasks and provide valuable insights into the relationships between features and labels, this method fail to provide insights on the reliability of the feature importance attribution.

This information is valuable, especially in high-stakes applications such as medical diagnosis or financial risk assessment, where incorrect explanations can have serious consequences. Therefore, it's essential to understand both direct and indirect feature importances to gain insights into how each feature contributes to the predictions of each output label. For example, the administration of one drug in an intensive care unit setting can indirectly impact the administration of a second drug due to various factors, including drug interactions, organ function, allergies, adverse effects, and the patient's overall condition. In our example, it is crucial for healthcare professionals to carefully evaluate these factors by knowing reliable indirect feature importance and allow them to identify specific risk factors or interventions for each condition independently, in order to ensure patient safety and optimize treatment outcomes.

Therefore, in this section, we propose Bayes LIME Chains, a novel method to measure uncertainty around the local indirect feature contributions to predict interdependent labels as a measure to assess the reliability of the explanation to predict the multiple interdependent labels. We use synthetic datasets where ground truth is known in order to evaluate the feature importance and their credible intervals.

4.2.2 Related work

Several methods have been proposed to measure uncertainties around local explanations in machine learning models. Bayes LIME and Bayes SHAP, introduced by Slack et al. [88], leverage Bayesian frameworks to generate local explanations with associated uncertainties. These methods provide valuable insights into the reliability and robustness of local explanations, enhancing the interpretability of machine learning models. Additionally, Bay LIME, proposed by Zhao et al. [104], offers a Bayesian approach to LIME, enabling the quantification of uncertainties in local explanations. While these methods contribute to a deeper understanding of the uncertainties inherent in direct local explanations, there is currently no established mechanism for evaluating the quality or reliability of indirect feature importance estimates across diverse datasets in both single and multi-output settings.

4.2.3 Proposed Method: Bayes LIME Chains

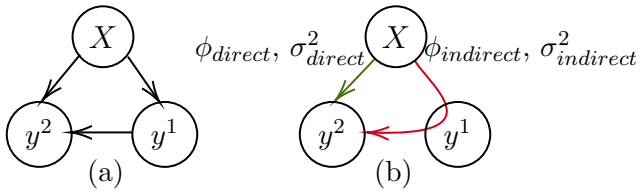


Figure 4.8: (a) An example of a 2-output task with interdependent labels. (b) Direct and indirect importance of X to predict the label y^2

.

Let's consider an instance $\mathbf{x}$ for which we aim to provide an explanation of its prediction made by the black box model f, along with an assessment of the uncertainty associated with this explanation. Local explainability models such as LIME and Kernel SHAP utilize surrogate linear models g to

emulate the behavior of the black box model f in the vicinity of the instance $\mathbf{x}$ for explanation purposes. The surrogate linear model g is trained on a sampled dataset $\mathbf{x}'$, with the width parameter $\pi_x(\mathbf{x}')$ being a user-specified value. In order to measure the uncertainty of the explanations, Bayes LIME employs a Bayesian linear model [66] trained on the sampled dataset to compute feature importance ϕ along with their uncertainties. The model prediction y for a given instance $\mathbf{x}'$ in the sampled dataset is expressed as:

$$y|\mathbf{x}', \phi, \epsilon \sim \phi^T \mathbf{x}' + \epsilon \tag{4.4}$$

The prior distribution of the feature importance vector ϕ can be written as:

$$\phi|\sigma^2 \sim \mathcal{N}(0, \sigma^2 I) \tag{4.5}$$

The error term ϵ and the width of the neighborhood around instance $\mathbf{x}$:

$$\epsilon \sim \mathcal{N}(0, \sigma^2/\pi_x(\mathbf{x}')); \pi_x(\mathbf{x}') = \exp(-dist(x, \mathbf{x}')^2/\sigma^2) \tag{4.6}$$

And the prior distribution of the variance:

$$\sigma^2 \sim \text{Inv} - \chi^2(n_0, \sigma_0^2) \tag{4.7}$$

Where n_0 and σ_0 are set to 10^{-6} to ensure an uninformative prior and $dist$ is distance metric which can be l_2 or cosine.

On the other hand, Shapley Chains enables the computation of both direct and indirect feature importance in multi-output tasks using techniques like Kernel SHAP or LIME. To leverage the benefits of both approaches (Bayes LIME and Shapley Chains), we propose a novel method called Bayes LIME Chains. This method aims to provide explanations and their uncertainties for both direct and indirect feature importance in multi-output tasks. Therefore, we measure the uncertainties of the indirect feature importance by distributing the uncertainties associated with direct label importance when used as a feature to predict subsequent labels in the chain among the set of features involved in predicting that label. This can be written as:

The direct contribution and uncertainty of the features on the each label independently are computed as in Bayes LIME (Equations. 4.4,4.5,4.6 and 4.7), and the indirect contribution of the features $\Phi_{indirect}(y^j)$ to predict

label y^j is computed as in Shapley Chains, i.e. by computing the direct label importance on subsequent labels and distributing them on the features. On the other hand, for each label y^j in the chain, the indirect feature uncertainty measured with $\sigma^2_{indirect}(y^j)$ is computed as follows using the same mechanism in Equations 4.1, 4.2 and 4.3:

$$\sigma^2_{indirect}(y^j) = \sum_{k=1}^{j-1} \sigma^2(y^k) \cdot U_k \tag{4.8}$$

$j > 1$ and U_k computes the weight vector for all paths from output y^k down to the features. For $k > 1$ and $U_1 = W(y^1)$, U_k is recursively computed as follows:

$$U_k = \sum_{l=1}^{k-1} W(y^k, y^{k-l}) \cdot U_{k-l} + W(y^k) \tag{4.9}$$

$W(y^k, y^{k-l})$ is the corresponding weight of y^{k-l} to predict the next output y^k (the direct contribution of y^{k-l} to predict y^k). And, $W(y^k)$ is the weight of the features to predict y^k (the direct contribution of the features to predict y^k). The weights $W(y^k, y^{k-l})$ and $W(y^k)$ are calculated according to:

$$W(y^k, .) = \frac{|\sigma^2(y^k)|}{\left(\sum_{q=1}^{n} |\sigma^2_{x^q}(y^k)| + \sum_{p=1}^{k} |\sigma^2_{y^p}(y^k)|\right)} \tag{4.10}$$

Where $\sigma^2_{x^q}(y^k)$ is the direct uncertainty interval (in Eq. 4.7) of each feature x^q to predict y^k. $\sigma^2_{y^p}(y^k)$ is the direct uncertainty interval of each label y^p that is now considered as a new feature to predict y^j.

4.2.4 Experiments

To demonstrate the efficacy of our work, we use the synthetic and the Adult Income datasets and set up the experiments such as in Section 4.1.4.

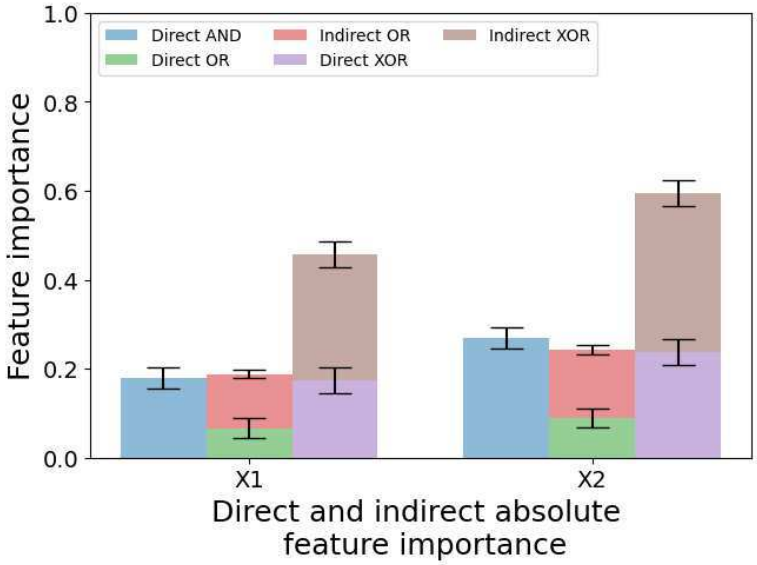


Figure 4.9: Direct and indirect feature importance of the features x^1 and x^2 to predict the labels AND, OR and XOR, with the uncertainty on each contribution. Unlike LIME, which only calculates direct feature contributions, Bayes LIME chains assign equal total contributions (combining direct and indirect effects) of features to predict each label, reflecting the ground truth.

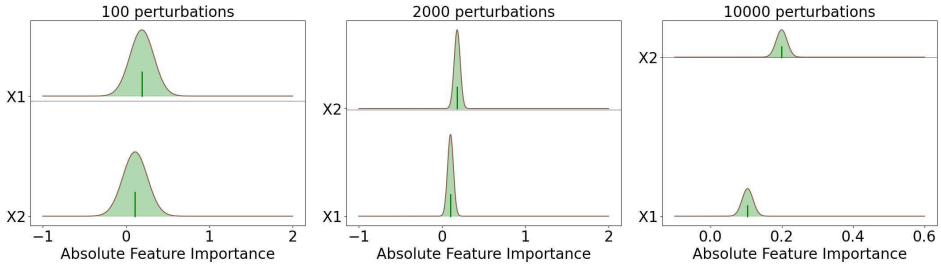


Figure 4.10: [87] An example of an explanation attributed by Bayes LIME Chains to a given instance in the XOR test set to explain the indirect impact of the feature to predict XOR label (corresponds to the brown illustration in Fig. 4.9). The vertical lines illustrate the indirect feature importance (red represents negative effect, green represents positive) and the shaded region visualizes the indirect uncertainty estimated by Bayes LIME Chains. The uncertainty intervals computed on different numbers of perturbations confirm that x^1 and x^2 are equivally important to predict the label XOR.

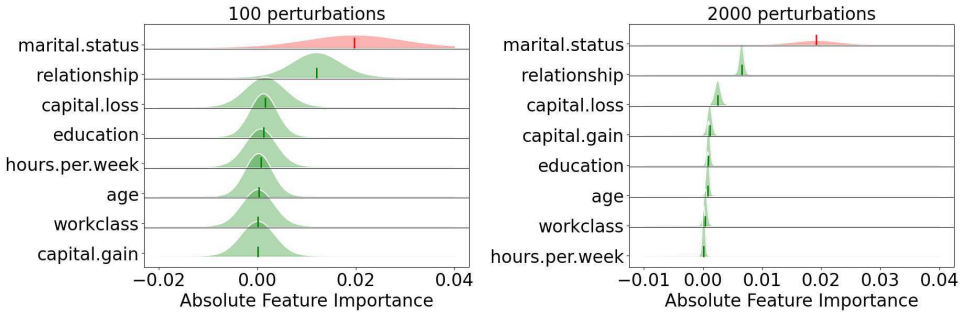


Figure 4.11: [87] An illustration of an explanation provided by Bayes LIME Chains for a specific instance within the Adult test set. The overlapping uncertainty intervals in the explanation generated with 100 perturbations indicate challenges in discerning the most influential feature. Conversely, narrower uncertainty intervals observed in the explanation generated with 2000 perturbations highlight marital status and relationship as the primary influential features.

Quality of the uncertainty intervals

Figure 4.12 illustrates the uncertainty intervals computed for the indirect feature importance of the XOR label using Bayes LIME Chains. Our analysis demonstrates that the uncertainty intervals are robust and encompass the ground truth feature importance values(as in Section 3.1.3) for most features. In contrast, for the real-world dataset, we conducted experiments with a larger neighborhood of instances to explain ($N = 10k$) and considered the feature importance obtained from Bayes LIME Chains as the ground truth. Figure 4.13 presents the uncertainty intervals computed for the indirect feature importance of the Occupation label. The results indicate that the uncertainty intervals remain robust and effectively capture the variability in the feature importance estimates.

4.2.5 Conclusion

In conclusion, we have introduced a novel method called Bayes LIME Chains for computing reliable direct and indirect feature importance in multi-label classification tasks with interdependent labels. This work is crucial for generating robust explanations that aid in understanding model predictions and building trust in the decision-making process. Through experiments

Method	Direct	Indirect
XOR(100)	98.0	91.0
XOR(2K)	95.0	94.0
Adult(100)	97.5	96.8
Adult(2K)	95.8	90.5

Table 4.2: Similar to [87], we assess the calibration of the credible intervals, by computing the percentage of instances where the 95% credible intervals of the direct and indirect feature importance to predict the XOR and INCOME, computed using 100 and 2K perturbations, encapsulate their true values (determined from 10k perturbations). Higher values indicate better calibration. Bayes LIME chains yield well-calibrated intervals despite the number of perturbations.

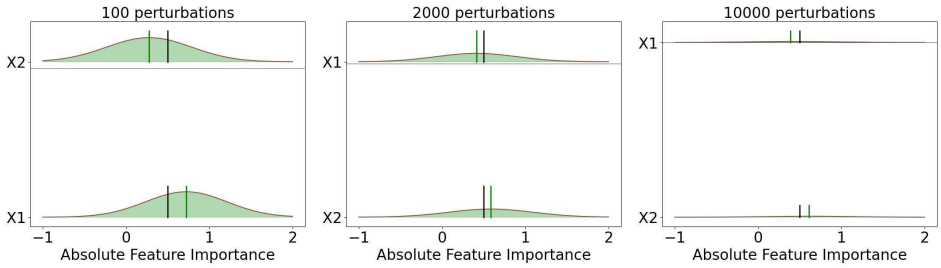


Figure 4.12: [87] Normalized indirect feature importance and uncertainty to predict XOR compared to the ground truth. The black vertical lines correspond to the ground truth for the XOR dataset. Since both features are important and necessary for the prediction of XOR label, they share equal importance (.5 for each). For different numbers of perturbations, both features are around .5 importance.

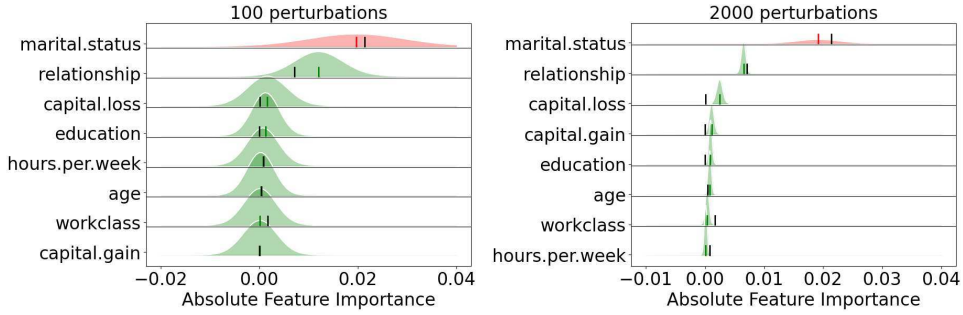


Figure 4.13: [87] Indirect feature importance and uncertainty compared to the ground truth. The black vertical lines correspond to the ground truth for the Adult dataset. The ground truth feature importance is obtained by running Bayes LIME Chains on 10k perturbations and is often included in the estimated intervals.

Method	Bayes LIME Chains	Shapley Flow	Independent	On-manifold
Adult(D)	97.5	40.6	13.0	47.3
Adult(I)	86.2	53.8	NA	NA
XOR(D)	98.0	83.3	87.5	77.9
XOR(I)	91.0	66.7	NA	NA

Table 4.3: Similar to [87], we assess the calibration of the credible intervals, by computing the percentage of instances where the 95% credible intervals of the direct and indirect feature importance to predict the XOR and Adult Income. Higher values indicate better calibration. Bayes LIME chains yield well-calibrated intervals across the datasets.

conducted on synthetic and real-world datasets, we have demonstrated the effectiveness of our approach in providing explanations with associated uncertainties. However, it is important to acknowledge the limitations of our work, particularly the absence of ground truth for real-world datasets against which the direct and indirect feature importance and credible intervals could be compared. Moving forward, future research could explore extending this methodology to other local explanation models such as Kernel SHAP and local surrogates. Additionally, incorporating Bayesian model averaging techniques for heterogeneous data could be an interesting avenue for further investigation.

Chapter 5

Conclusion

In conclusion, this thesis has made significant contributions to the field of reliable XAI. Our work has provided novel insights into advantages and limitations of existing methods, shedding light on previously unexplored areas such as finding where these methods work or fail with respect to data properties. Our work also provided new reliable explanation methodologies for multi-output tasks by including label interdependencies in the feature importance attribution.

Limitations However, it's important to acknowledge the limitations of each proposed approach. For instance, in the first chapter we evaluate the local explanation methods on synthetic datasets with two features. While the aim was to simplify the feature interactions and focus on specific cases for different parameters, this over simplification might be inadequate to represent the complex feature interactions that are present in real-world datasets. Moreover, we restricted our evaluation to decision tree based models such as random forest, although this can be applied to other decision tree based models such as XGboost and other complex models like deep neural networks. In addition, the Shapley chain method, while applicable to any type of label interdependencies and doesn't necessitate a prior causal knowledge on which label affects the others, it doesn't permit to clarify these type of label interdependencies. Also, the computation of the Shapley value for multi-output is very time and memory consuming. Moreover, Bayes LIME Chains is tested on synthetic datasets where the ground truth is available, it is hard to make same conclusions on real-world datasets where ground truth

is missing, unless we consider the parameters of the best fit local linear model, which is true to the model and not to the data itself. Despite these constraints, our research has laid the groundwork for future investigations in this domain.

Perspectives Moving forward, there are several promising avenues for future work. One such direction could involve the assessment of explanations should be extended to other types of data such as images and text. Additionally, extend it to other more complex synthetic datasets to include many features that can be more representative of real world scenarios, to other tasks such as multi-class classification and regression, to other decision trees based models such as XGBoost and other blackbox models such as deep neural networks. Another might focus on exploring the direct and indirect feature importance computed by Shapley Chains on specific label interdependencies such as causal relationships and correlations, and to explore the mays that the total feature importance can help understand the best order to lean each label in the chain in order to maximise the performance for the given task (for instance, accuracy or Hamming loss for the multi-output classification). Additionally, future work should focus on optimizing the time complexity of the compuatation of the Shapley Chains. On the other hand, Bayes LIME Chains can also be applied to other data modalities such as images and text and extended to other local surrogate XAI based methods such as Local surrogates. Finally, future work can also build more robust and reliable explanations by exploring other ways of measuring the uncertainty intervals such as using Bayesian models averaging (BMA) instead of single linear models.

By addressing these avenues, we aim to further advance the field and continue to contribute to the ongoing discourse on reliable local explanations for single and multi-output tasks on tabular data.

Bibliography

[1] Agarwal, C., Krishna, S., Saxena, E., Pawelczyk, M., Johnson, N., Puri, I., Zitnik, M., Lakkaraju, H.: Openxai: Towards a transparent evaluation of model explanations. Advances in Neural Information Processing Systems **35**, 15784–15799 (2022)

[2] Alvares-Cherman, E., Metz, J., Monard, M.C.: Incorporating label dependency into the binary relevance framework for multi-label classification. Expert Systems with Applications **39**(2), 1647–1655 (2012)

[3] Alvarez-Melis, D., Jaakkola, T.S.: On the Robustness of Interpretability Methods (Jun 2018)

[4] Alvarez-Melis, D., Jaakkola, T.S.: On the robustness of interpretability methods. arXiv preprint arXiv:1806.08049 (2018)

[5] Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V.I.: Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC medical informatics and decision making **20**(1), 1–9 (2020)

[6] Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information fusion **58**, 82–115 (2020)

[7] Attanasio, G., Pastor, E., Di Bonaventura, C., Nozza, D.: Ferret: A Framework for Benchmarking Explainers on Transformers (Aug 2022)

[8] Ayad, C.W., Bonnier, T., Bosch, B., Read, J.: Shapley chains: Extending shapley values to classifier chains. In: International Conference on Discovery Science. pp. 541–555. Springer (2022)

[9] Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D., Rinzivillo, S.: Benchmarking and survey of explanation methods for black box models **37**(5), 1719–1778

[10] Breiman, L.: Random Forests. Machine Learning **45**(1), 5–32 (Oct 2001)

[11] Breiman, L.: Classification and regression trees. Routledge (2017)

[12] Camburu, O.M., Giunchiglia, E., Foerster, J., Lukasiewicz, T., Blunsom, P.: Can I Trust the Explainer? Verifying Post-hoc Explanatory Methods (Dec 2019)

[13] 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)

[14] Chadaga, K., Prabhu, S., Sampathila, N., Chadaga, R., KS, S., Sengupta, S.: Predicting cervical cancer biopsy results using demographic and epidemiological parameters: a custom stacked ensemble machine learning approach. Cogent Engineering **9**(1), 2143040 (2022)

[15] Chekin, N., Ayatollahi, H., Karimi Zarchi, M.: A Clinical Decision Support System for Assessing the Risk of Cervical Cancer: Development and Evaluation Study. JMIR Medical Informatics **10**(6), e34753 (Jun 2022)

[16] Chen, Z., Subhash, V., Havasi, M., Pan, W., Doshi-Velez, F.: What Makes a Good Explanation?: A Harmonized View of Properties of Explanations

[17] Chen, Z., Subhash, V., Havasi, M., Pan, W., Doshi-Velez, F.: Does the explanation satisfy your needs?: A unified view of properties of explanations. arXiv preprint arXiv:2211.05667 (2022)

[18] Cohen, P.A., Jhingran, A., Oaknin, A., Denny, L.: Cervical cancer. The Lancet **393**(10167), 169–182 (2019)

[19] Conceição, T., Braga, C., Rosado, L., Vasconcelos, M.J.M.: A Review of Computational Methods for Cervical Cells Segmentation and Abnormality Classification. International Journal of Molecular Sciences **20**(20), 5114 (Oct 2019)

[20] Curia, F.: Cervical cancer risk prediction with robust ensemble and explainable black boxes method. Health and Technology **11**(4), 875–885 (Jul 2021)

[21] Das, A., Rad, P.: Opportunities and challenges in explainable artificial intelligence (xai): A survey. arXiv preprint arXiv:2006.11371 (2020)

[22] Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)

[23] Dua, D., Graff, C.: UCI machine learning repository (2017), `http://archive.ics.uci.edu/ml`

[24] Fernandes, K., Cardoso, J.S., Fernandes, J.: 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. pp. 243–250. Springer (2017)

[25] Flora, M., Potvin, C., McGovern, A., Handler, S.: Comparing explanation methods for traditional machine learning models part 1: An overview of current methods and quantifying their disagreement. arXiv preprint arXiv:2211.08943 (2022)

[26] Frye, C., de Mijolla, D., Begley, T., Cowton, L., Stanley, M., Feige, I.: Shapley explainability on the data manifold (Dec 2021)

[27] Frye, C., Rowat, C., Feige, I.: Asymmetric Shapley values: Incorporating causal knowledge into model-agnostic explainability (Dec 2021)

[28] Ghorbani, A., Abid, A., Zou, J.: Interpretation of neural networks is fragile. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 3681–3688 (2019)

[29] Godbole, S., Sarawagi, S.: Discriminative methods for multi-labeled classification. In: Pacific-Asia conference on knowledge discovery and data mining. pp. 22–30. Springer (2004)

[30] Greenwell, B.M., Boehmke, B.C., McCarthy, A.J.: A simple and effective model-based variable importance measure. arXiv preprint arXiv:1805.04755 (2018)

[31] Guidotti, R.: Evaluating local explanation methods on ground truth. Artificial Intelligence **291**, 103428 (2021)

[32] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) **51**(5), 1–42 (2018)

[33] Guo, Y., Gu, S.: Multi-label classification using conditional dependency networks. In: IJCAI Proceedings-international joint conference on artificial intelligence. vol. 22, p. 1300 (2011)

[34] Gupta, V., Nokhiz, P., Roy, C.D., Venkatasubramanian, S.: Equalizing recourse across groups. arXiv preprint arXiv:1909.03166 (2019)

[35] Han, T., Srinivas, S., Lakkaraju, H.: Which Explanation Should I Choose? A Function Approximation Perspective to Characterizing Post hoc Explanations (Jun 2022)

[36] Han, T., Srinivas, S., Lakkaraju, H.: Which explanation should i choose? a function approximation perspective to characterizing post hoc explanations. Advances in Neural Information Processing Systems **35**, 5256–5268 (2022)

[37] Hedström, A., Weber, L., Bareeva, D., Motzkus, F., Samek, W., Lapuschkin, S., Höhne, M.M.C.: Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations (Feb 2022)

[38] Hooker, G., Mentch, L.: Please stop permuting features: An explanation and alternatives. arXiv preprint arXiv:1905.03151 **2** (2019)

[39] Hooker, S., Erhan, D., Kindermans, P.J., Kim, B.: A Benchmark for Interpretability Methods in Deep Neural Networks (Jun 2018)

[40] Ismail, A.A., Gunady, M., Corrada Bravo, H., Feizi, S.: Benchmarking Deep Learning Interpretability in Time Series Predictions. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6441–6452. Curran Associates, Inc.

[41] Jensen, P., Skov, L.: Psoriasis and Obesity. Dermatology (Basel, Switzerland) **232**(6), 633–639 (2016)

[42] Jung, J., Lee, H., Jung, H., Kim, H.: Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. Heliyon (2023)

[43] Karimi, A.H., Barthe, G., Schölkopf, B., Valera, I.: A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. arXiv preprint arXiv:2010.04050 (2020)

[44] Kashyap, N., Krishnan, N., Kaur, S., Ghai, S.: Risk factors of cervical cancer: a case-control study. Asia-Pacific journal of oncology nursing **6**(3), 308–314 (2019)

[45] Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. arXiv preprint arXiv:1703.04730 (2017)

[46] Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., Reblitz-Richardson, O.: Captum: A unified and generic model interpretability library for PyTorch (Sep 2020)

[47] Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., Lakkaraju, H.: The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective (Feb 2022)

[48] Kruczkowski, M., Drabik-Kruczkowska, A., Marciniak, A., Tarczewska, M., Kosowska, M., Szczerska, M.: Predictions of cervical cancer identification by photonic method combined with machine learning. Scientific Reports **12**(1), 3762 (2022)

[49] Kulmala, L., Read, J., Nöjd, P., Rathgeber, C.B., Cuny, H.E., Hollmén, J., Mäkinen, H.: Identifying the main drivers for the production and maturation of scots pine tracheids along

a temperature gradient. Agricultural and Forest Meteorology **232**(January), 210–224 (2017), http://www.sciencedirect.com/science/article/pii/S0168192316303677

[50] Le, P.Q., Nauta, M., Van Bach Nguyen, S.P., Schlötterer, J., Seifert, C.: Benchmarking explainable ai-a survey on available toolkits and open challenges. In: International Joint Conference on Artificial Intelligence (2023)

[51] Lee, E., Jung, S.Y., Hwang, H.J., Jung, J.: Patient-Level Cancer Prediction Models From a Nationwide Patient Cohort: Model Development and Validation. JMIR Medical Informatics **9**(8), e29807 (Aug 2021)

[52] Li, X., Wang, Y., Basu, S., Kumbier, K., Yu, B.: A Debiased MDI Feature Importance Measure for Random Forests. arXiv:1906.10845 [cs, stat] (jun 2019), http://arxiv.org/abs/1906.10845, arXiv: 1906.10845

[53] Liao, Q.V., Zhang, Y., Luss, R., Doshi-Velez, F., Dhurandhar, A.: 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. vol. 10, pp. 147–159 (2022)

[54] Liu, Y., Khandagale, S., White, C., Neiswanger, W.: Synthetic Benchmarks for Scientific Research in Explainable Machine Learning (Nov 2021)

[55] Luaces, O., Díez, J., Barranquero, J., del Coz, J.J., Bahamonde, A.: Binary relevance efficacy for multilabel classification. Progress in Artificial Intelligence **1**, 303–313 (2012)

[56] Lundberg, S., Lee, S.I.: A Unified Approach to Interpreting Model Predictions. arXiv:1705.07874 [cs, stat] (Nov 2017)

[57] Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov), 2579–2605 (2008)

[58] Marcinkevičs, R., Vogt, J.E.: Interpretability and explainability: A machine learning zoo mini-tour. arXiv preprint arXiv:2012.01805 (2020)

[59] Mehmood, M., Rizwan, M., Gregus ml, M., Abbas, S.: Machine learning assisted cervical cancer detection. Frontiers in public health **9**, 788376 (2021)

[60] Melki, G., Cano, A., Kecman, V., Ventura, S.: Multi-target support vector regression via correlation regressor chains. Information Sciences **415**, 53–69 (2017)

[61] Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence (2018)

[62] Mohanty, A., Mishra, S.: A Comprehensive Study of Explainable Artificial Intelligence in Healthcare. In: Mishra, S., Tripathy, H.K., Mallick, P., Shaalan, K. (eds.) Augmented Intelligence in Healthcare: A Pragmatic and Integrated Analysis, pp. 475–502. Studies in Computational Intelligence, Springer Nature, Singapore (2022)

[63] Molnar, C.: Interpretable Machine Learning. 2 edn. (2022), `https://christophm.github.io/interpretable-ml-book`

[64] Molnar, C., Gruber, S., Kopper, P.: Limitations of Interpretable Machine Learning Methods (2020)

[65] Montavon, G., Binder, A., Lapuschkin, S., Samek, W., Müller, K.R.: Layer-Wise Relevance Propagation: An Overview. In: Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.R. (eds.) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, vol. 11700, pp. 193–209. Springer International Publishing, Cham (2019)

[66] Moore, A.: Locally weighted bayesian regression (1995)

[67] Mothilal, R.K., Sharma, A., Tan, C.: Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations (Dec 2019)

[68] National Institute of Standards and Technology (NIST): Explainable artificial intelligence (XAI) definitions. Special Publication 500-325, U.S. National Institute of Standards and Technology

[69] Neely, M., Schouten, S.F., Bleeker, M.J.R., Lucic, A.: Order in the Court: Explainable AI Methods Prone to Disagreement (Jul 2021)

[70] Oxford Languages: Artificial intelligence

[71] Palczewska, A., Palczewski, J., Robinson, R.M., Neagu, D.: Interpreting random forest models using a feature contribution method. 2013 IEEE 14th International Conference on Information Reuse & Integration (IRI) pp. 112–119 (2013), `https://api.semanticscholar.org/CorpusID:16311309`

[72] Petsiuk, V., Das, A., Saenko, K.: RISE: Randomized Input Sampling for Explanation of Black-box Models (Sep 2018)

[73] Pichler, M., Hartig, F.: Machine learning and deep learning – a review for ecologists (04 2022)

[74] Ratul, I.J., Al-Monsur, A., Tabassum, B., Ar-Rafi, A.M., Nishat, M.M., Faisal, F.: 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). pp. 1–4. IEEE (2022)

[75] Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. Machine Learning **85**(3), 333–359 (Dec 2011)

[76] Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier Chains: A Review and Perspectives. Journal of Artificial Intelligence Research **70**, 683–718 (Feb 2021)

[77] Ribeiro, M.T., Singh, S., Guestrin, C.: "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. pp. 1135–1144. KDD '16, Association for Computing Machinery, New York, NY, USA (Aug 2016)

[78] Ricordeau, J., Lacaille, J.: Application of Random Forests to Engine health Monitoring (Sep 2010)

[79] Roth, A.E.: The Shapley value: essays in honor of Lloyd S. Shapley. Cambridge University Press (1988)

[80] Rozemberczki, B., Sarkar, R.: The Shapley Value of Classifiers in Ensemble Games (Jun 2021)

[81] Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence **1**(5), 206–215 (2019)

[82] Samuel, A.L.: Some studies in machine learning using the game of checkers. IBM Journal of research and development **3**(3), 210–229 (1959)

[83] Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-cam: Why did you say that? arXiv preprint arXiv:1611.07450 (2016)

[84] Sharma, S., Henderson, J., Ghosh, J.: Certifai: A common framework to provide explanations and analyse the fairness and robustness of black-box models. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. pp. 166–172 (2020)

[85] Shrikumar, A., Greenside, P., Kundaje, A.: Learning Important Features Through Propagating Activation Differences (Oct 2019)

[86] Singh, S., Ribeiro, M.T., Guestrin, C.: Programs as black-box explanations. arXiv preprint arXiv:1611.07579 (2016)

[87] Slack, D., Hilgard, A., Singh, S., Lakkaraju, H.: Reliable post hoc explanations: Modeling uncertainty in explainability. Advances in neural information processing systems **34**, 9391–9404 (2021)

[88] Slack, D., Hilgard, S., Singh, S., Lakkaraju, H.: Reliable Post hoc Explanations: Modeling Uncertainty in Explainability

[89] Slack, D., Hilgard, S., Singh, S., Lakkaraju, H.: Reliable Post hoc Explanations: Modeling Uncertainty in Explainability (Nov 2021)

[90] Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T.: Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics **8**(1), 25 (Jan 2007)

[91] Sundararajan, M., Taly, A., Yan, Q.: Axiomatic Attribution for Deep Networks (Jun 2017)

[92] Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: 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)

[93] Terence, P., Prince, G.: Beware Default Random Forest Importances. http://explained.ai/decision-tree-viz/index.html

[94] Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal of Data Warehousing and Mining (IJDWM) **3**(3), 1–13 (2007)

[95] Turbé, H., Bjelogrlic, M., Lovis, C., Mengaldo, G.: Evaluation of post-hoc interpretability methods in time-series classification. Nature Machine Intelligence **5**(3), 250–260 (2023)

[96] Ustun, B., Spangher, A., Liu, Y.: Actionable recourse in linear classification. In: Proceedings of the conference on fairness, accountability, and transparency. pp. 10–19 (2019)

[97] Visani, G., Bagli, E., Chesani, F.: Optilime: Optimized lime explanations for diagnostic computer algorithms. arXiv preprint arXiv:2006.05714 (2020)

[98] Vitali, F.e.a.: 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. vol. 346, p. 235. IOS Press (2022)

[99] Wachter, S., Mittelstadt, B., Russell, C.: Bias Preservation in Machine Learning: The Legality of Fairness Metrics under EU Non-Discrimination Law **123**, 735 (2021)

[100] Waegeman, W., Dembczyński, K., Hüllermeier, E.: Multi-target prediction: a unifying view on problems and methods. Data Mining and Knowledge Discovery **33**, 293–324 (2019)

[101] Wang, J., Wiens, J., Lundberg, S.: Shapley Flow: A Graph-based Approach to Interpreting Model Predictions. In: Proceedings of The

24th International Conference on Artificial Intelligence and Statistics. pp. 721–729. PMLR (Mar 2021)

[102] Wang, R., Armin, M.A., Denman, S., Petersson, L., Ahmedt-Aristizabal, D.: Towards interpretable attention networks for cervical cancer analysis. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 3613–3616. IEEE (2021)

[103] Yang, M., Kim, B.: Benchmarking attribution methods with relative feature importance. arXiv preprint arXiv:1907.09701 (2019)

[104] Zhao, X., Huang, W., Huang, X., Robu, V., Flynn, D.: BayLIME: Bayesian Local Interpretable Model-Agnostic Explanations

[105] Zhao, X., Huang, X., Robu, V., Flynn, D.: Baylime: Bayesian local interpretable model-agnostic explanations. In: Conference on Uncertainty in Artificial Intelligence (2020), https://api.semanticscholar.org/CorpusID:227334656

[106] Zhong, Z., Liu, Z., Tegmark, M., Andreas, J.: The clock and the pizza: Two stories in mechanistic explanation of neural networks. arXiv preprint arXiv:2306.17844 (2023)

[107] Zhou, J., Gandomi, A.H., Chen, F., Holzinger, A.: Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics **10**(5), 593 (2021)

APPENDIX A. FEATURE IMPORTANCE DEPENDS ON DATA PROPERTIES

Appendix A

Feature Importance depends on Data Properties

A.1 DT and a RF for the 48 generated datasets with 50 000 instances

Decision function	ϵ	ρ	DT Accuracy	feature importance by DT	RF Accuracy	feature importance by RF
NOT	0.00	0.00	1.00	ϕ_{x^1}=1.0, ϕ_{x^2}=0.0	1.00	ϕ_{x^1}=0.68, ϕ_{x^2}=0.32
NOT	0.00	0.10	1.00	ϕ_{x^1}=1.0, ϕ_{x^2}=0.0	1.00	ϕ_{x^1}=0.7, ϕ_{x^2}=0.3
NOT	0.00	0.90	1.00	ϕ_{x^1}=1.0, ϕ_{x^2}=0.0	1.00	ϕ_{x^1}=0.51, ϕ_{x^2}=0.49
NOT	0.00	1.00	1.00	ϕ_{x^1}=0.0, ϕ_{x^2}=1.0	1.00	ϕ_{x^1}=0.4, ϕ_{x^2}=0.6
NOT	0.00	0.00	1.00	ϕ_{x^1}=1.0, ϕ_{x^2}=0.0	1.00	ϕ_{x^1}=0.65, ϕ_{x^2}=0.35
NOT	0.00	0.10	1.00	ϕ_{x^1}=1.0, ϕ_{x^2}=0.0	1.00	ϕ_{x^1}=0.7, ϕ_{x^2}=0.3
NOT	0.00	0.90	1.00	ϕ_{x^1}=1.0, ϕ_{x^2}=0.0	1.00	ϕ_{x^1}=0.52, ϕ_{x^2}=0.48
NOT	0.00	1.00	1.00	ϕ_{x^1}=0.0, ϕ_{x^2}=1.0	1.00	ϕ_{x^1}=0.4, ϕ_{x^2}=0.6
NOT	0.25	0.00	0.90	ϕ_{x^1}=1.0, ϕ_{x^2}=0.0	0.90	ϕ_{x^1}=0.67, ϕ_{x^2}=0.33
NOT	0.25	0.10	0.90	ϕ_{x^1}=1.0, ϕ_{x^2}=0.0	0.90	ϕ_{x^1}=0.7, ϕ_{x^2}=0.3
NOT	0.25	0.90	0.90	ϕ_{x^1}=1.0, ϕ_{x^2}=0.0	0.90	ϕ_{x^1}=0.51, ϕ_{x^2}=0.49
NOT	0.25	1.00	0.90	ϕ_{x^1}=0.0, ϕ_{x^2}=1.0	0.90	ϕ_{x^1}=0.4, ϕ_{x^2}=0.6
NOT	0.25	0.00	0.89	ϕ_{x^1}=1.0, ϕ_{x^2}=0.0	0.89	ϕ_{x^1}=0.6, ϕ_{x^2}=0.4
NOT	0.25	0.10	0.90	ϕ_{x^1}=1.0, ϕ_{x^2}=0.0	0.90	ϕ_{x^1}=0.7, ϕ_{x^2}=0.3
NOT	0.25	0.90	0.90	ϕ_{x^1}=1.0, ϕ_{x^2}=0.0	0.90	ϕ_{x^1}=0.52, ϕ_{x^2}=0.48
NOT	0.25	1.00	0.90	ϕ_{x^1}=0.0, ϕ_{x^2}=1.0	0.90	ϕ_{x^1}=0.4, ϕ_{x^2}=0.6
NOT	0.5	0.00	0.50	ϕ_{x^1}=0.0, ϕ_{x^2}=1.0	0.50	ϕ_{x^1}=0.37, ϕ_{x^2}=0.63
NOT	0.5	0.10	0.50	ϕ_{x^1}=0.66, ϕ_{x^2}=0.34	0.50	ϕ_{x^1}=0.43, ϕ_{x^2}=0.57
NOT	0.5	0.90	0.51	ϕ_{x^1}=1.0, ϕ_{x^2}=0.0	0.50	ϕ_{x^1}=0.34, ϕ_{x^2}=0.66
NOT	0.5	1.00	0.49	ϕ_{x^1}=0.74, ϕ_{x^2}=0.26	0.49	ϕ_{x^1}=0.39, ϕ_{x^2}=0.61
NOT	0.5	0.00	0.50	ϕ_{x^1}=0.66, ϕ_{x^2}=0.34	0.50	ϕ_{x^1}=0.37, ϕ_{x^2}=0.63
NOT	0.5	0.10	0.51	ϕ_{x^1}=0.0, ϕ_{x^2}=1.0	0.50	ϕ_{x^1}=0.39, ϕ_{x^2}=0.61
NOT	0.5	0.90	0.50	ϕ_{x^1}=1.0, ϕ_{x^2}=0.0	0.49	ϕ_{x^1}=0.4, ϕ_{x^2}=0.6
NOT	0.5	1.00	0.49	ϕ_{x^1}=0.6, ϕ_{x^2}=0.4	0.49	ϕ_{x^1}=0.32, ϕ_{x^2}=0.68

Table A.1: Parameterization and performances on the test set of a decision tree and a random forest for the 48 generated datasets with 50 000 instances. The feature importance estimates of the decision tree converge to the ground truth feature importance estimates when number of generated instances = 50 000. DT and RF learning are extremely affected by the noise.

Decision function	ϵ	ρ	DT Accuracy	feature importance by DT	RF Accuracy	feature importance by RF
XOR	0.00	0.00	1.00	ϕ_{x^1}=0.48, ϕ_{x^2}=0.52	0.97	ϕ_{x^1}=0.53, ϕ_{x^2}=0.47
XOR	0.00	0.10	1.00	ϕ_{x^1}=0.46, ϕ_{x^2}=0.54	0.94	ϕ_{x^1}=0.53, ϕ_{x^2}=0.47
XOR	0.00	0.90	1.00	ϕ_{x^1}=0.52, ϕ_{x^2}=0.48	1.00	ϕ_{x^1}=0.38, ϕ_{x^2}=0.62
XOR	0.00	1.00	1.00	ϕ_{x^1}=0.48, ϕ_{x^2}=0.52	1.00	ϕ_{x^1}=0.35, ϕ_{x^2}=0.65
XOR	0.00	0.00	1.00	ϕ_{x^1}=0.54, ϕ_{x^2}=0.46	0.99	ϕ_{x^1}=0.13, ϕ_{x^2}=0.87
XOR	0.00	0.10	1.00	ϕ_{x^1}=0.53, ϕ_{x^2}=0.47	0.84	ϕ_{x^1}=0.13, ϕ_{x^2}=0.87
XOR	0.00	0.90	1.00	ϕ_{x^1}=0.48, ϕ_{x^2}=0.52	0.96	ϕ_{x^1}=0.32, ϕ_{x^2}=0.68
XOR	0.00	1.00	1.00	ϕ_{x^1}=0.48, ϕ_{x^2}=0.52	1.00	ϕ_{x^1}=0.35, ϕ_{x^2}=0.65
XOR	0.25	0.00	0.90	ϕ_{x^1}=0.46, ϕ_{x^2}=0.54	0.90	ϕ_{x^1}=0.53, ϕ_{x^2}=0.47
XOR	0.25	0.10	0.90	ϕ_{x^1}=0.47, ϕ_{x^2}=0.53	0.88	ϕ_{x^1}=0.54, ϕ_{x^2}=0.46
XOR	0.25	0.90	0.90	ϕ_{x^1}=0.52, ϕ_{x^2}=0.48	0.90	ϕ_{x^1}=0.38, ϕ_{x^2}=0.62
XOR	0.25	1.00	0.90	ϕ_{x^1}=0.48, ϕ_{x^2}=0.52	0.90	ϕ_{x^1}=0.35, ϕ_{x^2}=0.65
XOR	0.25	0.00	0.90	ϕ_{x^1}=0.54, ϕ_{x^2}=0.46	0.78	ϕ_{x^1}=0.17, ϕ_{x^2}=0.83
XOR	0.25	0.10	0.90	ϕ_{x^1}=0.53, ϕ_{x^2}=0.47	0.76	ϕ_{x^1}=0.14, ϕ_{x^2}=0.86
XOR	0.25	0.90	0.90	ϕ_{x^1}=0.48, ϕ_{x^2}=0.52	0.87	ϕ_{x^1}=0.32, ϕ_{x^2}=0.68
XOR	0.25	1.00	0.90	ϕ_{x^1}=0.49, ϕ_{x^2}=0.51	0.90	ϕ_{x^1}=0.35, ϕ_{x^2}=0.65
XOR	0.5	0.00	0.50	ϕ_{x^1}=0.0, ϕ_{x^2}=1.0	0.50	ϕ_{x^1}=0.39, ϕ_{x^2}=0.61
XOR	0.5	0.10	0.50	ϕ_{x^1}=1.0, ϕ_{x^2}=0.0	0.49	ϕ_{x^1}=0.37, ϕ_{x^2}=0.63
XOR	0.5	0.90	0.50	ϕ_{x^1}=0.0, ϕ_{x^2}=1.0	0.50	ϕ_{x^1}=0.36, ϕ_{x^2}=0.64
XOR	0.5	1.00	0.50	ϕ_{x^1}=0.63, ϕ_{x^2}=0.37	0.50	ϕ_{x^1}=0.32, ϕ_{x^2}=0.68
XOR	0.5	0.00	0.50	ϕ_{x^1}=0.4, ϕ_{x^2}=0.6	0.49	ϕ_{x^1}=0.44, ϕ_{x^2}=0.56
XOR	0.5	0.10	0.50	ϕ_{x^1}=0.33, ϕ_{x^2}=0.67	0.50	ϕ_{x^1}=0.35, ϕ_{x^2}=0.65
XOR	0.5	0.90	0.50	ϕ_{x^1}=0.0, ϕ_{x^2}=1.0	0.50	ϕ_{x^1}=0.37, ϕ_{x^2}=0.63
XOR	0.5	1.00	0.51	ϕ_{x^1}=0.69, ϕ_{x^2}=0.31	0.51	ϕ_{x^1}=0.37, ϕ_{x^2}=0.63

Table A.2: Parameterization and performances on the test set of a decision tree and a random forest for the 48 generated datasets with 50 000 instances. The feature importance estimates of the decision tree converge to the ground truth feature importance estimates when number of generated instances = 50 000. DT and RF learning are extremely affected by the noise.

APPENDIX B. CERVICAL CANCER RISK FACTORS

Appendix B

Cervical Cancer Risk Factors

B.1 Contribution of the Data Points to the Prediction Making.

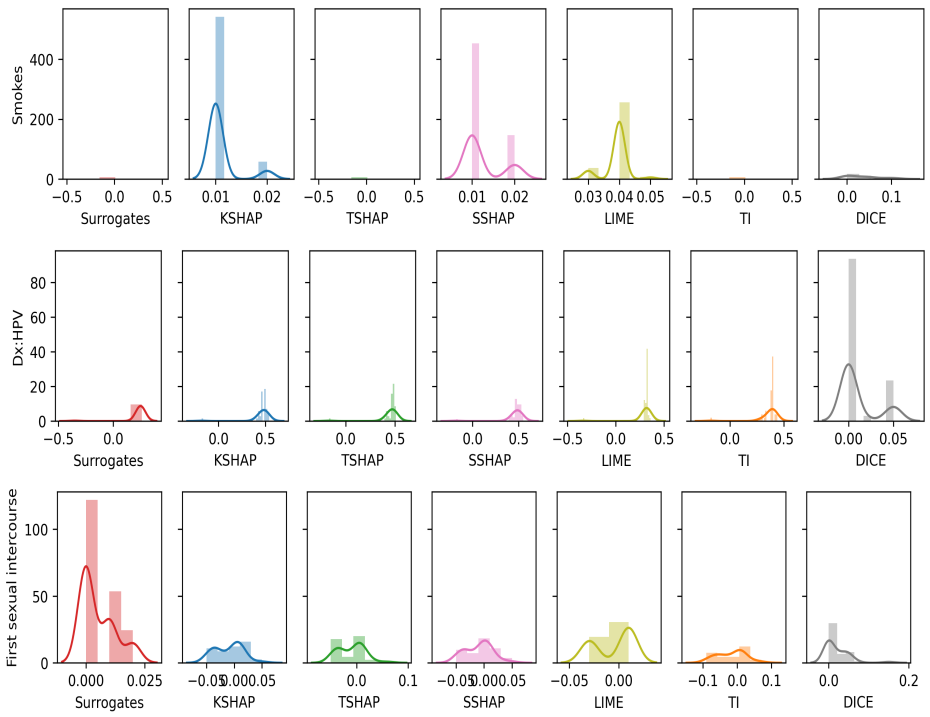


Figure B.1: Local variability of the feature importance for Patient 1. We compute the feature importance of 30 nearest neighbors to this patient, among these we keep the patients having the same explanations and predictions and compute the distance between each data point and Patient 1. Small distances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

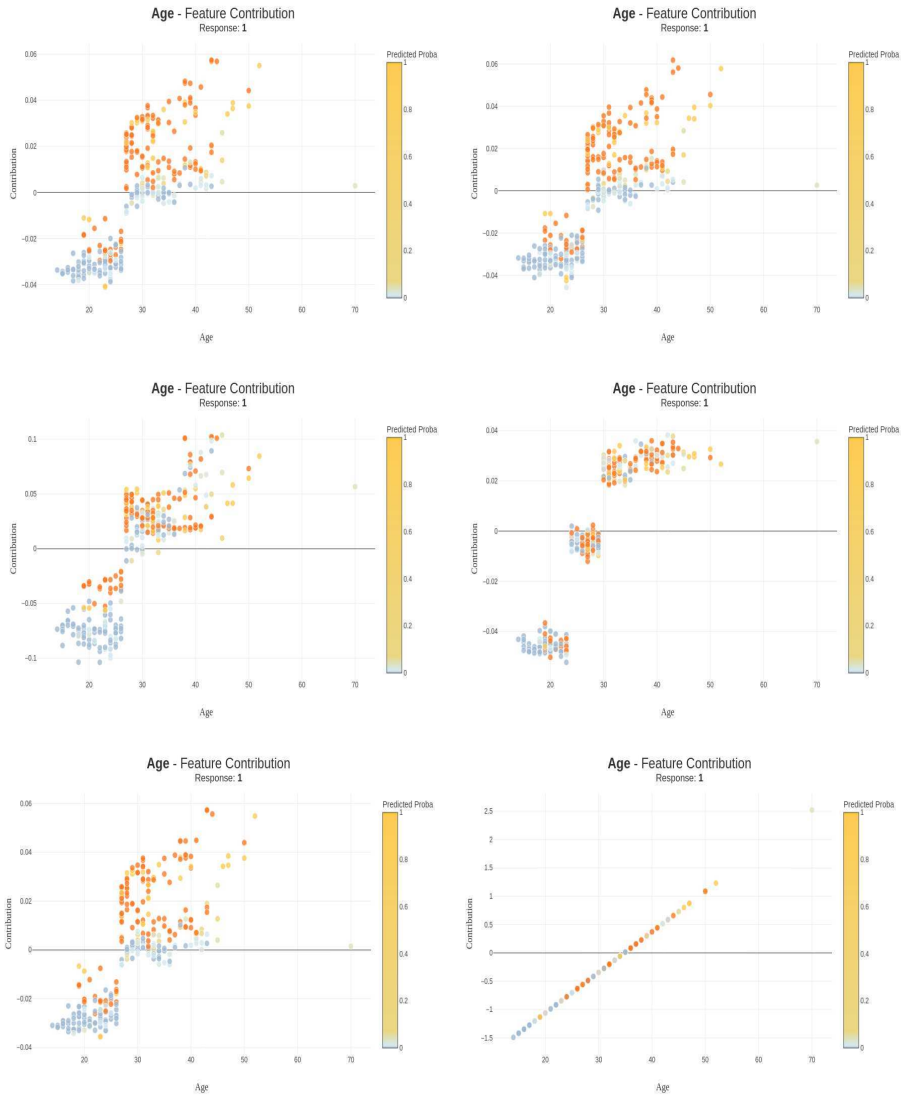


Figure B.2: Contribution of the Age of all patients to the class 1 (diagnosed with Cancer).

B.2 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 B.1: Summary statistics of four different patients diagnosed with cervical cancer relative to the mean of the population.

B.2.1 Patient 2 with Age =14 and Dx:Cancer= 0

Feature importance attributions

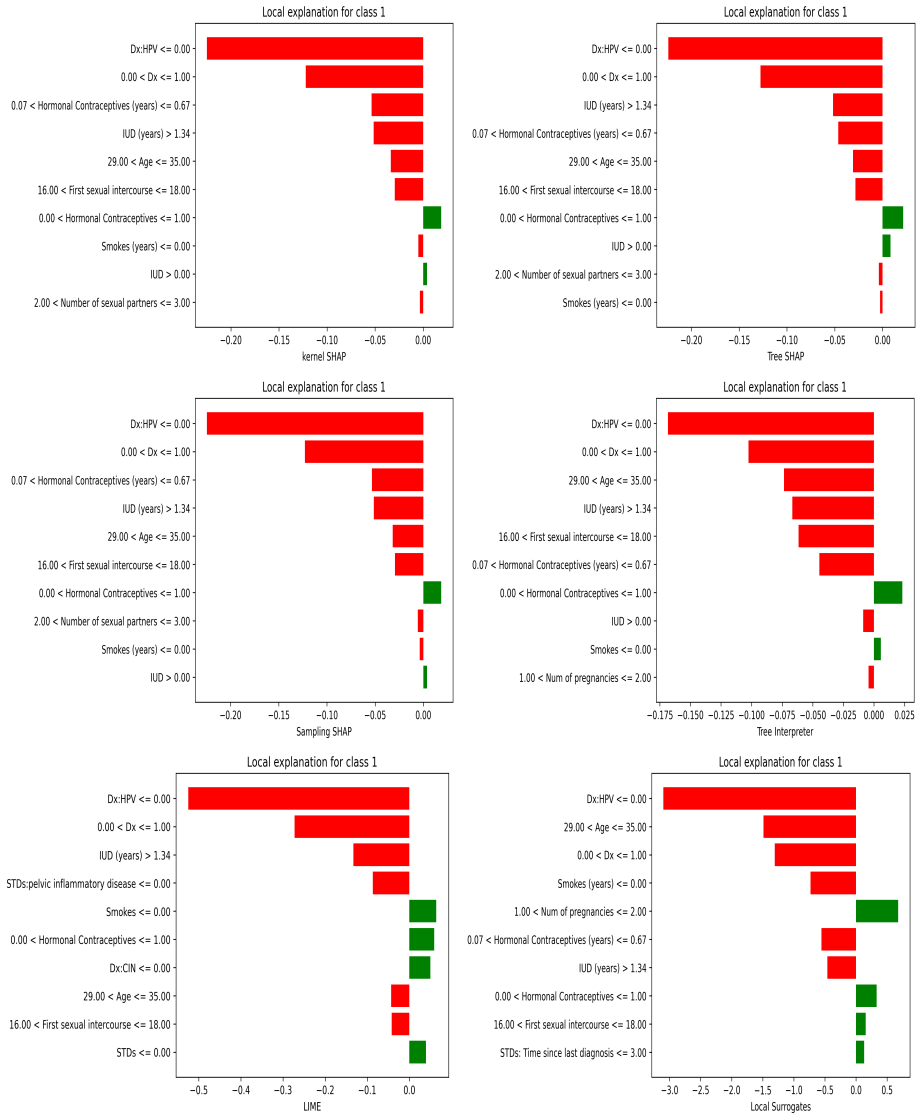


Figure B.3: Feature importance attribution for Patient 2, with Age =14 and Dx:Cancer= 0.

Stability of the features in the neighborhood

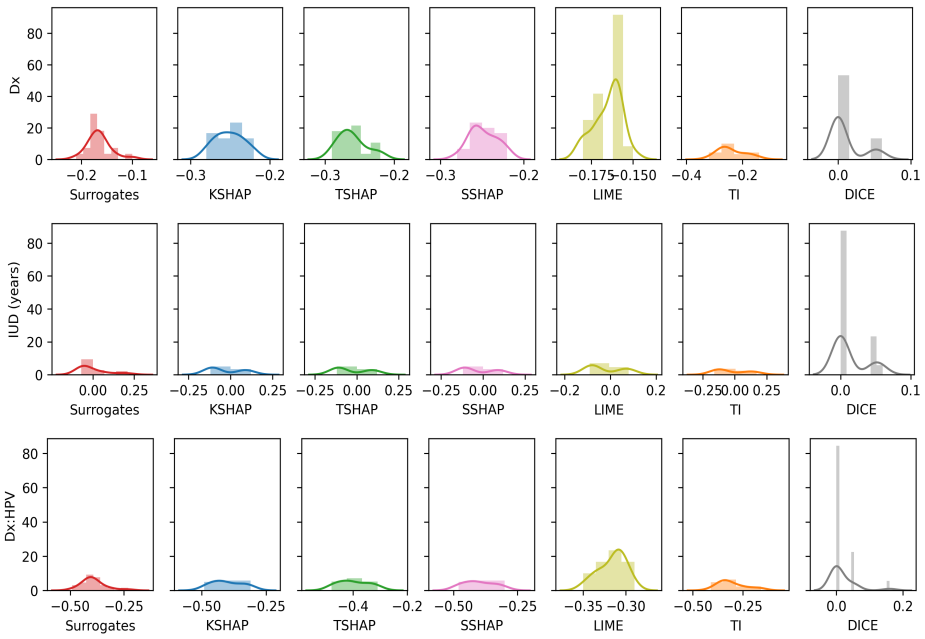


Figure B.4: Local variability of the feature importance for Patient with id=2. We compute the feature importance of 30 nearest neighbors to this patient, among these we keep the patients having the same explanations and predictions and compute the distance between each data point and Patient with id=2. Small distances mean more stable explanations. For example, LIME is the most stable method for feature Dx of the Patient with id=2.

Rank and feature agreements

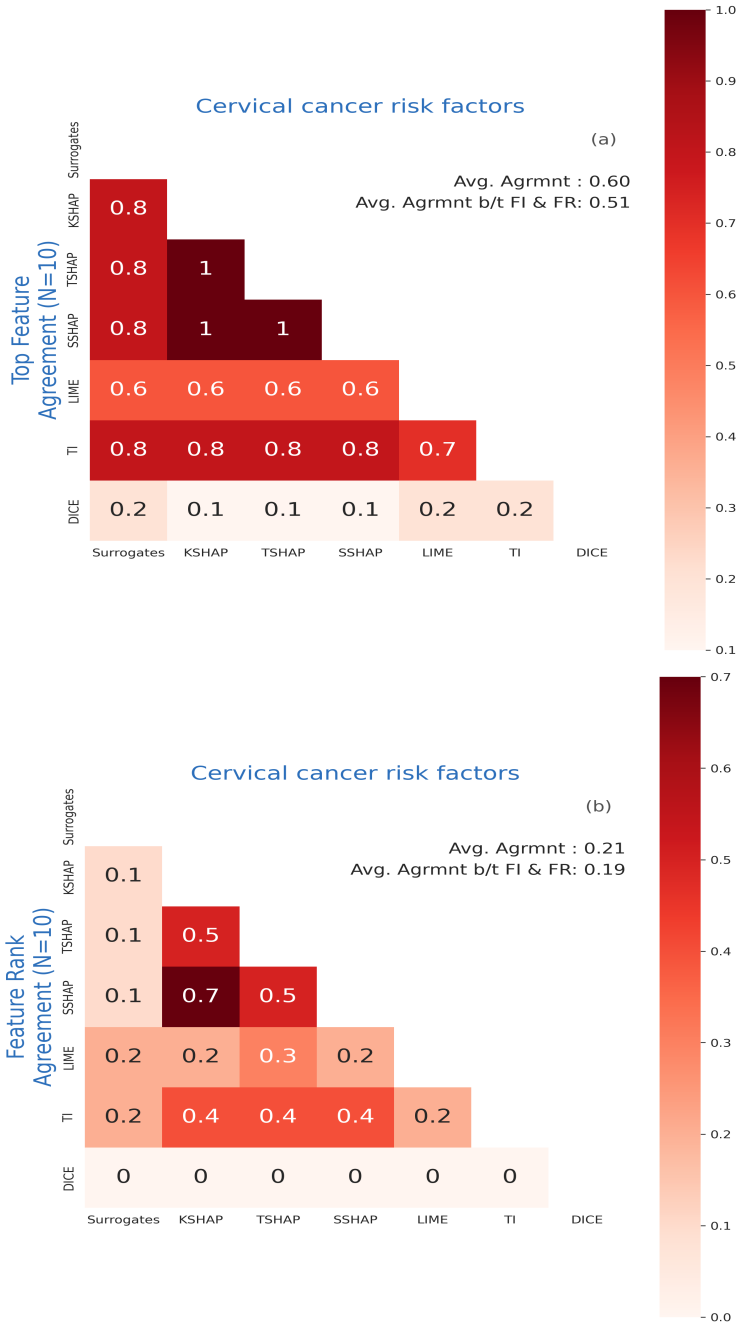


Figure B.5: Feature agreement for patient 2 (Age=14).

B.2.2 Patient 3 with Age =70 and Dx:Cancer= 0

Feature importance attributions

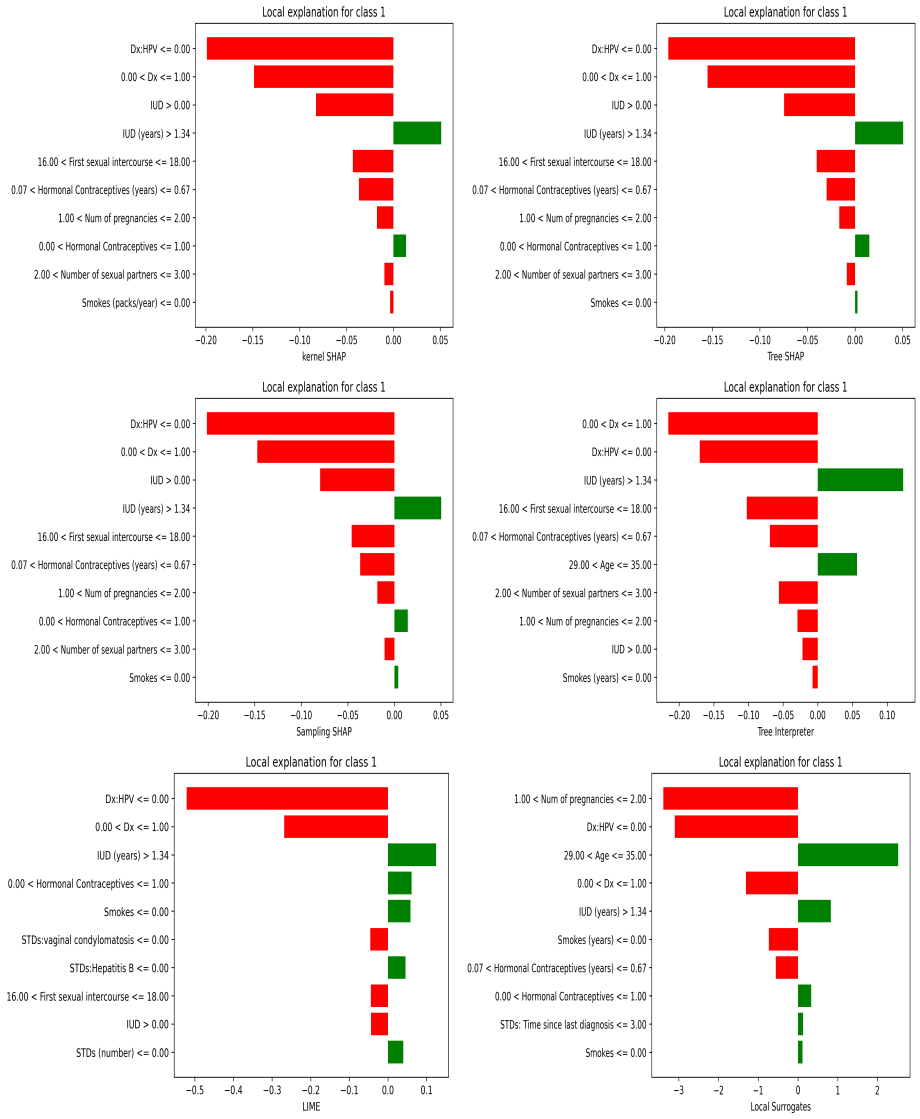


Figure B.6: Feature importance attributions for Patient 3, with Age =70 and Dx:Cancer= 0.

Stability of the features in the neighborhood

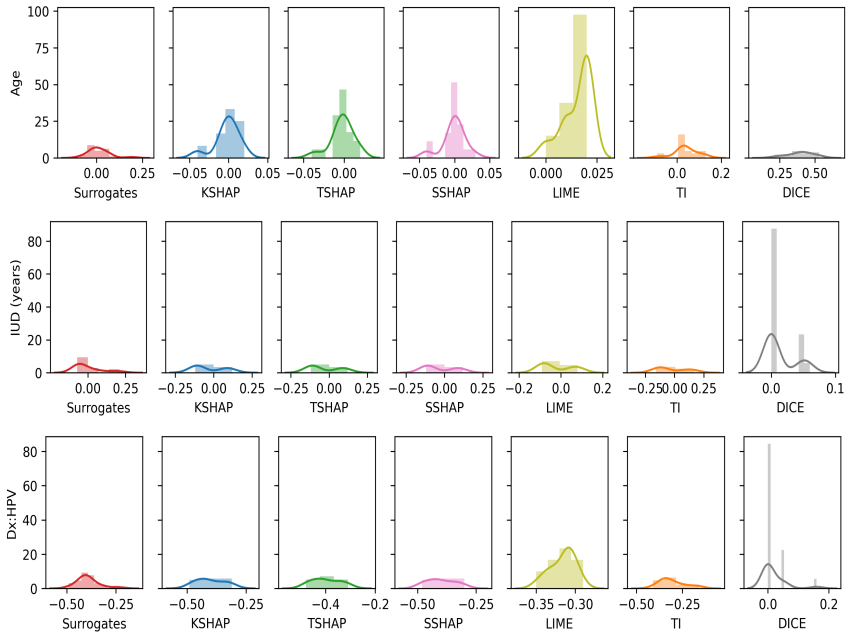


Figure B.7: Local variability of the feature importance for Patient with id=3. We compute the feature importance of 30 nearest neighbors to this patient, among these we keep the patients having the same explanations and predictions and compute the distance between each data point and Patient with id=3. Small distances mean more stable explanations. For example, LIME is the most stable method for feature Dx:HPV of the Patient with id=3.

Rank and feature agreements

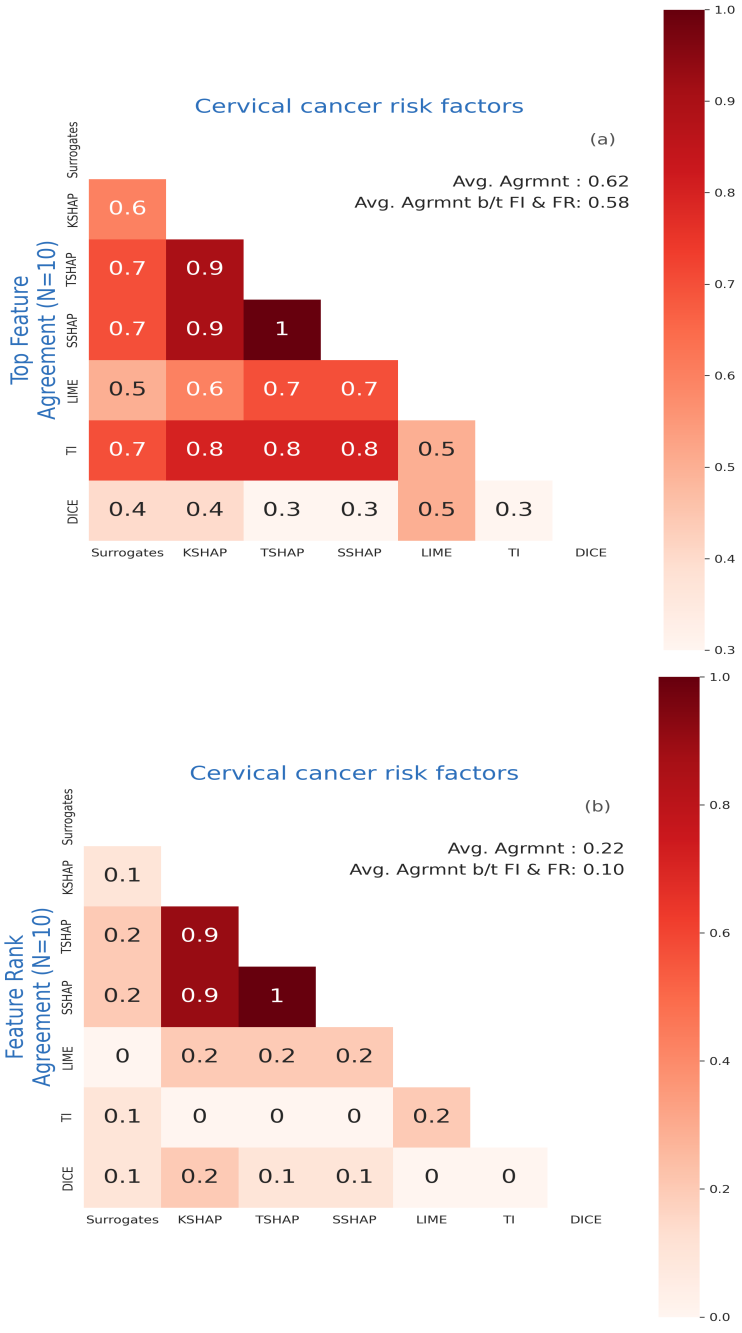


Figure B.8: Feature agreement for patient 3 (Age=70).

B.3 Fooling Explanations with Random Variables

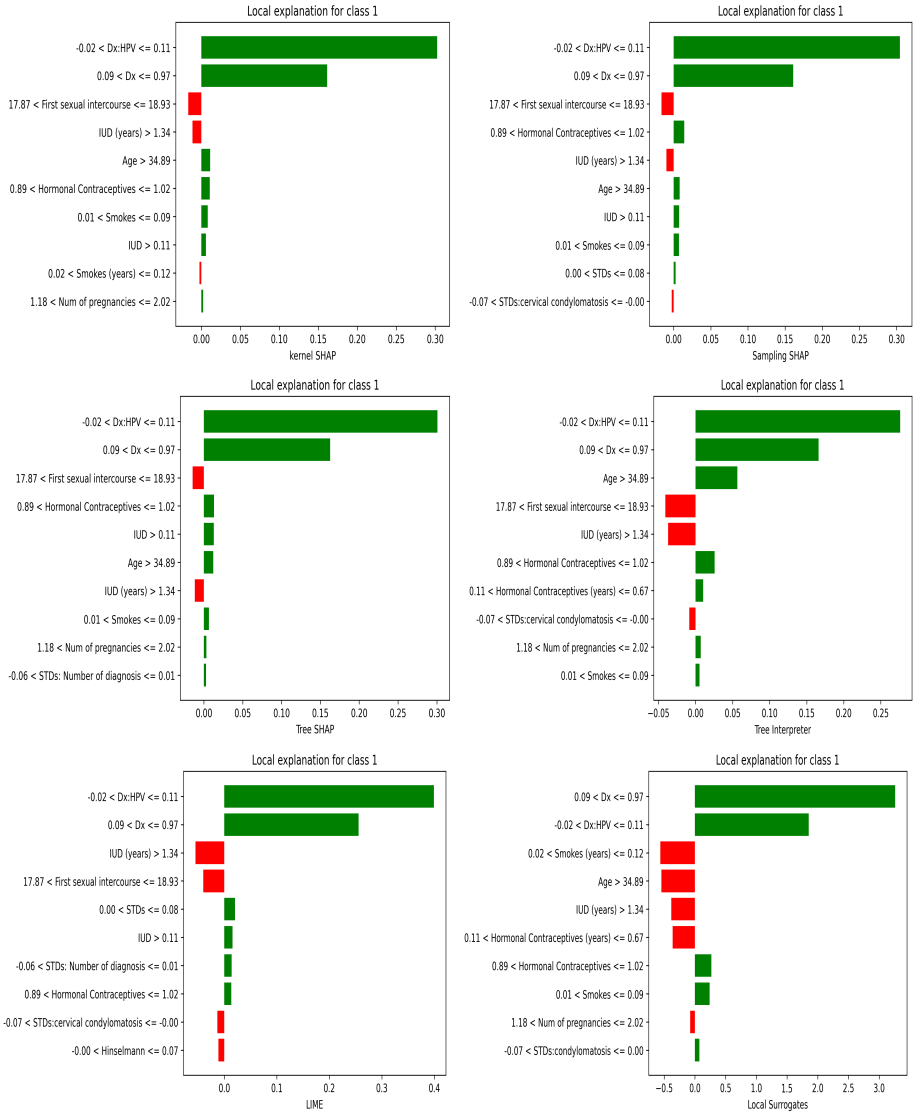


Figure B.9: Adding a binary, a continuous random variables and noise to the features ($\epsilon \in \mathcal{N}(0, .1)$).

B.4 Effect of Changing a Feature Value on the Explanations for Patient 1.

B.4.1 Changing Age from 27 to 80

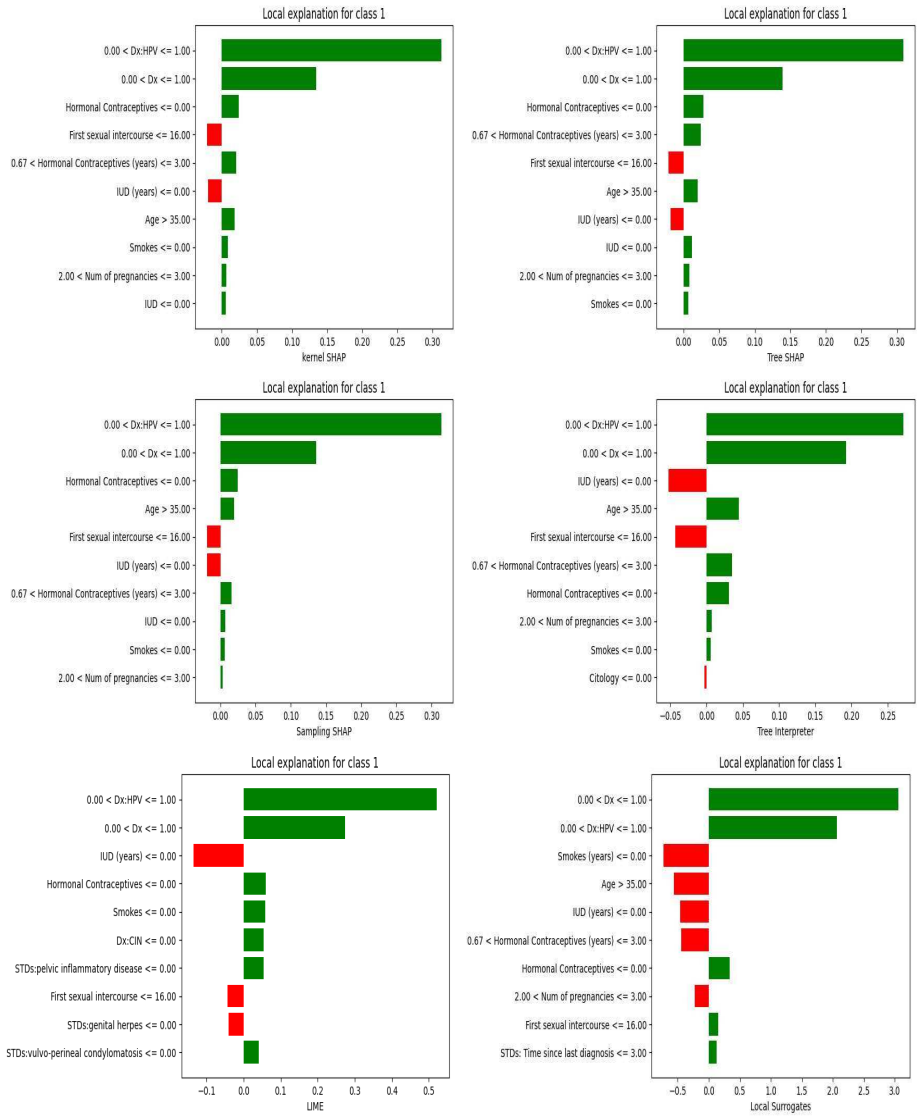


Figure B.10: Feature importance attributions for Patient with id=291 and Age = 80.

B.4.2 Changing Number of pregnancies from 3 to 0

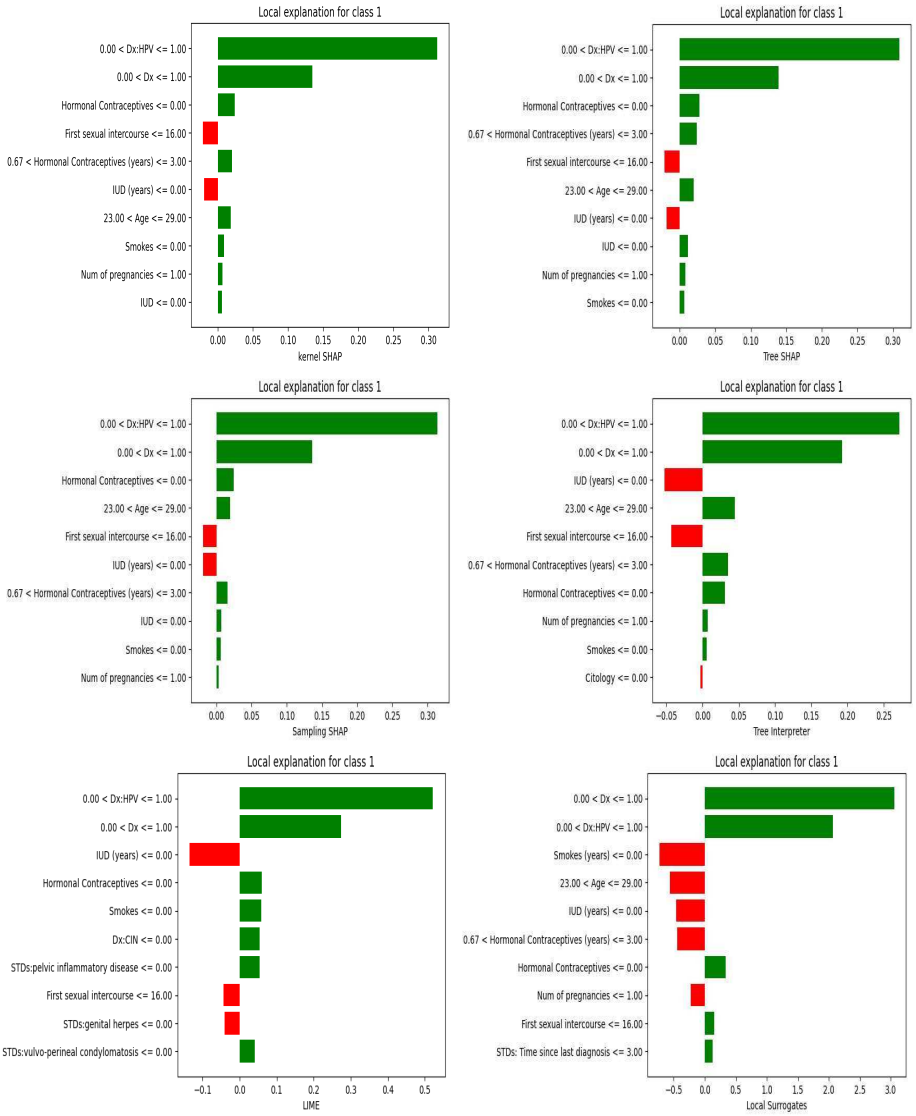


Figure B.11: Feature importance attributions for Patient with id=291 and Number of pregnancies = 0.

B.4.3 Changing Smokes from 0 to 1

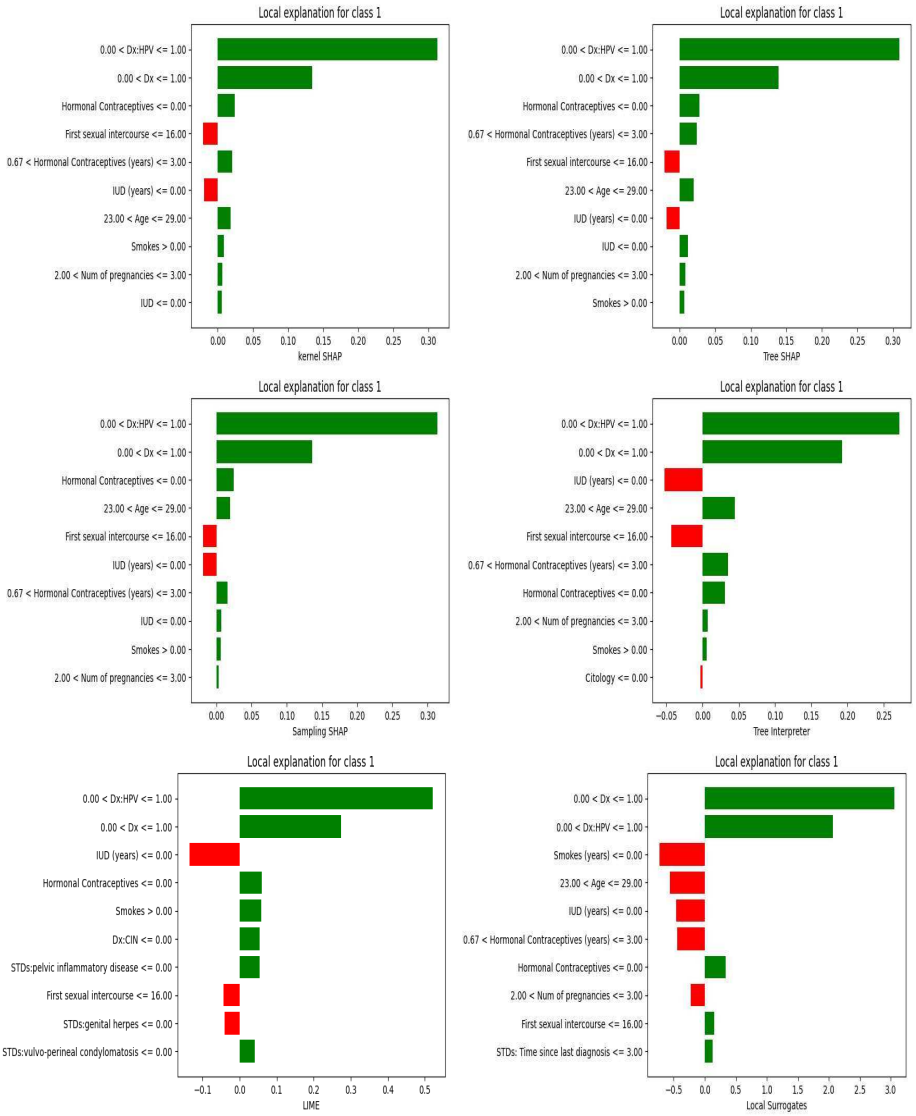


Figure B.12: Feature importance attributions for Patient with id=291 and Smokes = 1.

B.4.4 Changing Number of sexual partners from 2 to 60

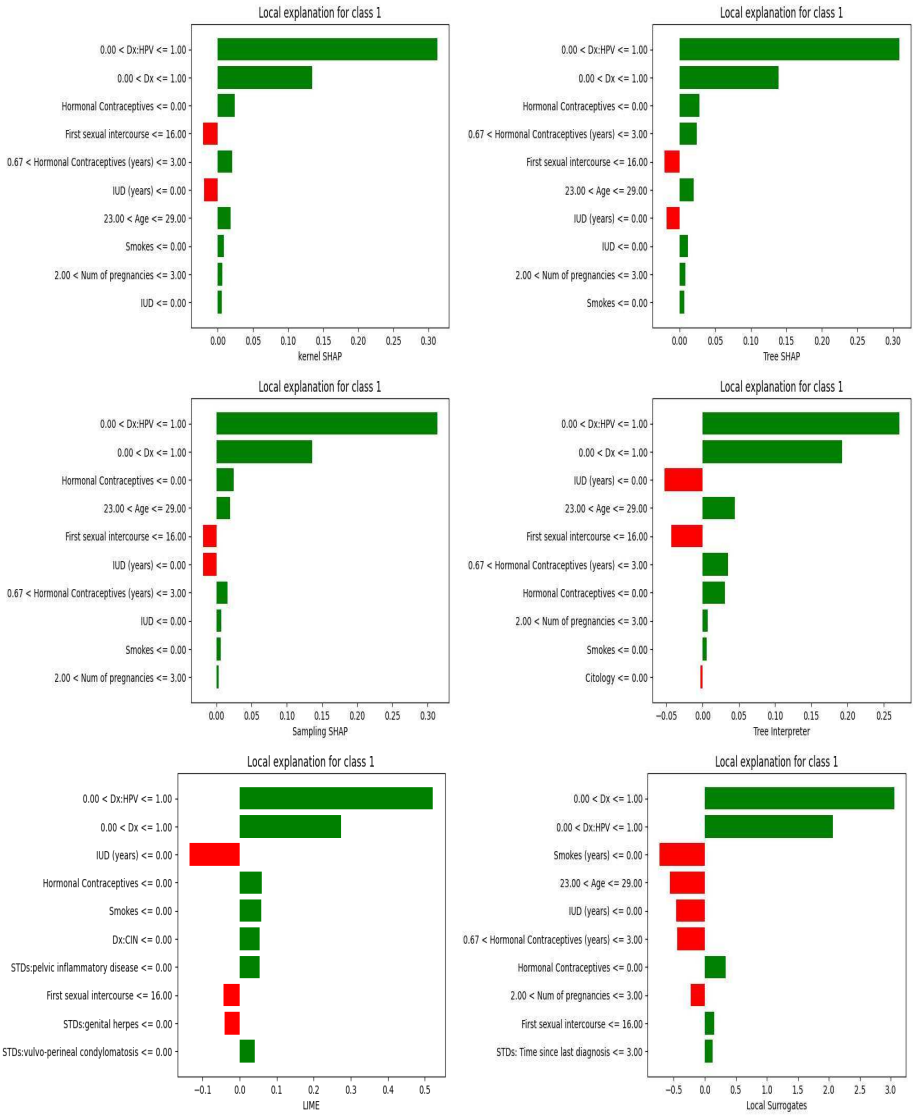


Figure B.13: Feature importance attributions for Patient with id=291 and Number of sexual partners = 60.

B.4.5 Changing First sexual intercourse from 14 to 40

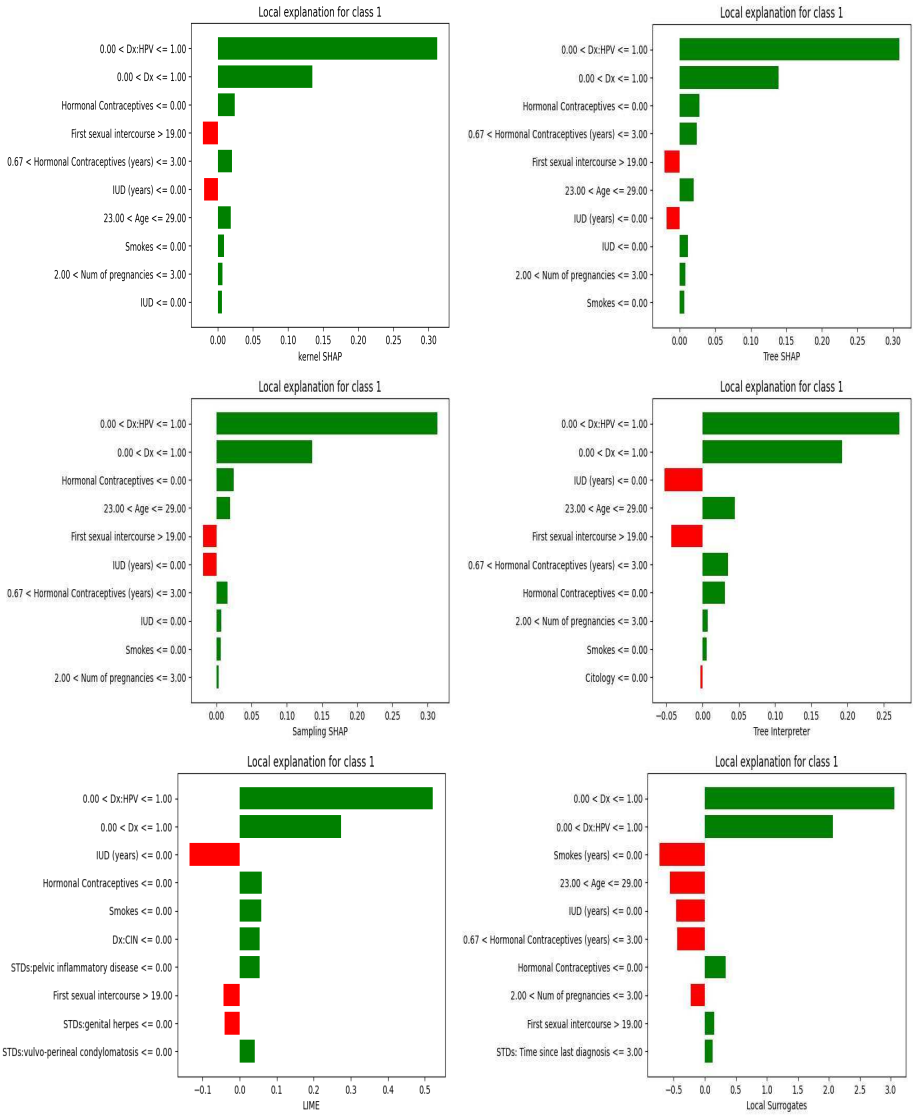


Figure B.14: Feature importance attributions for Patient with id=291 and First sexual intercourse = 40.

B.5 Difference in the Explanations for Different Age Categories

B.5.1 Patients with Age 14 and 19

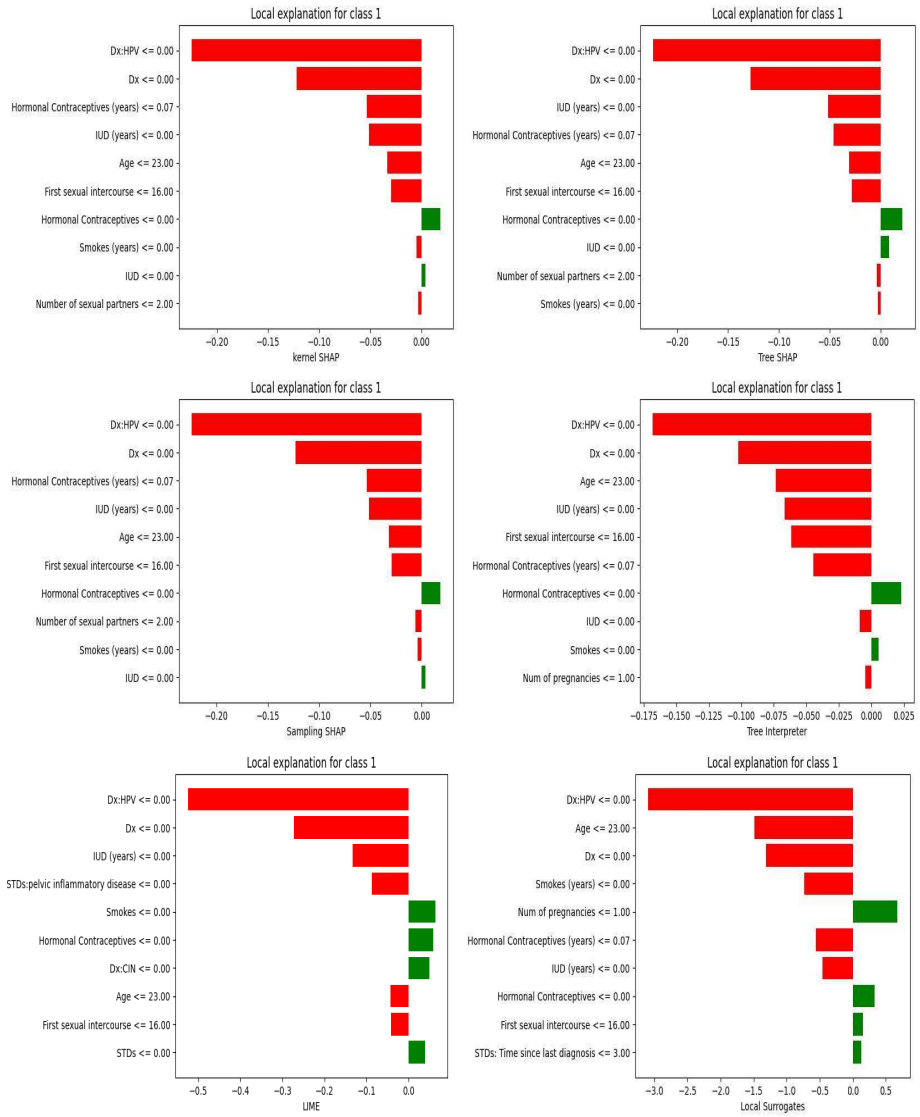


Figure B.15: Feature importance attributions for Patient with id=29.

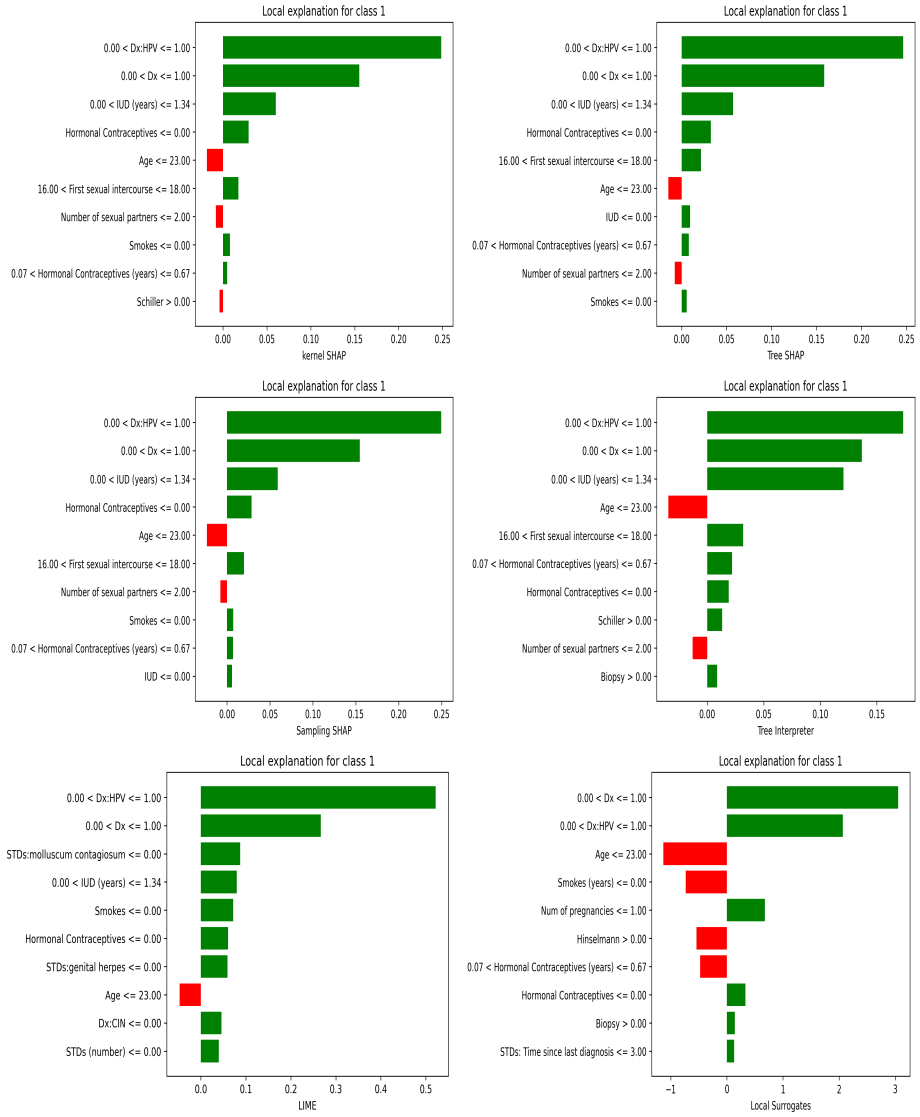


Figure B.16: Feature importance attributions for Patient with id=19.

B.5.2 Patients with Age=24

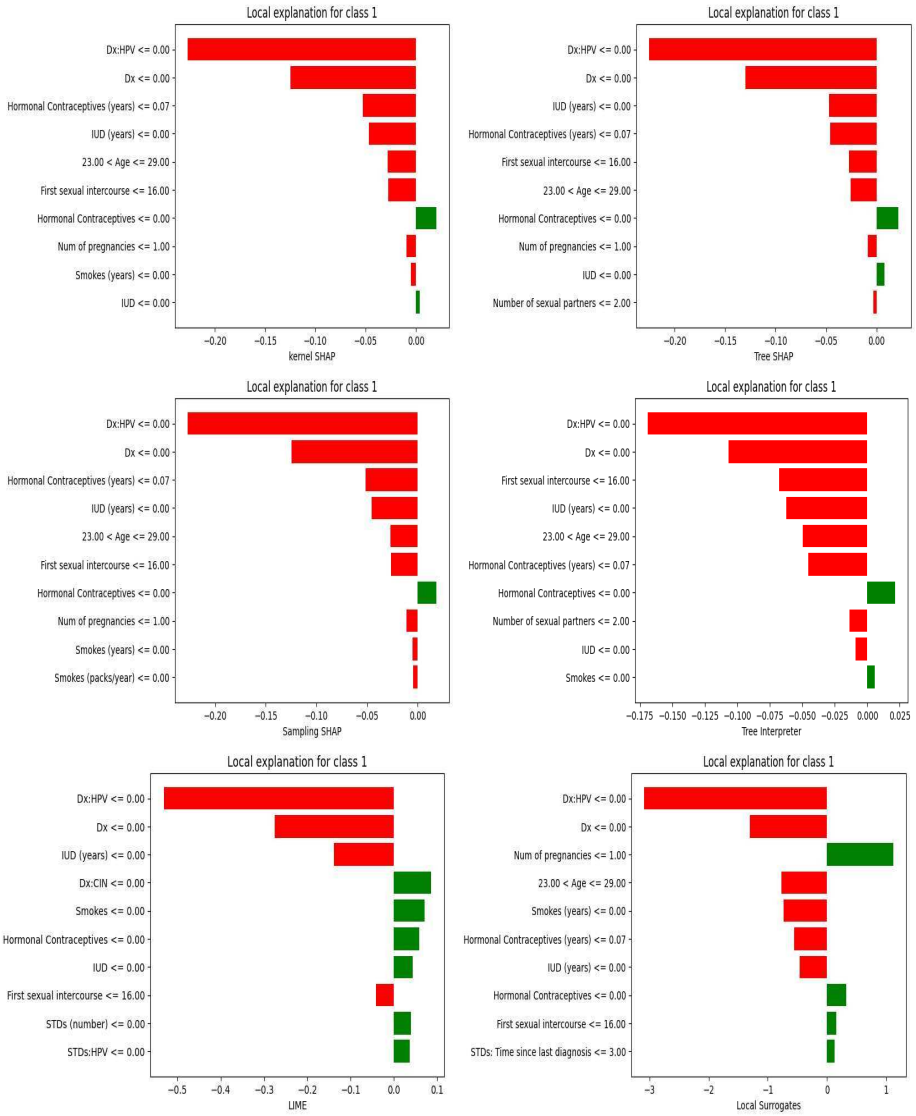


Figure B.17: Feature importance attributions for Patient with id=11.

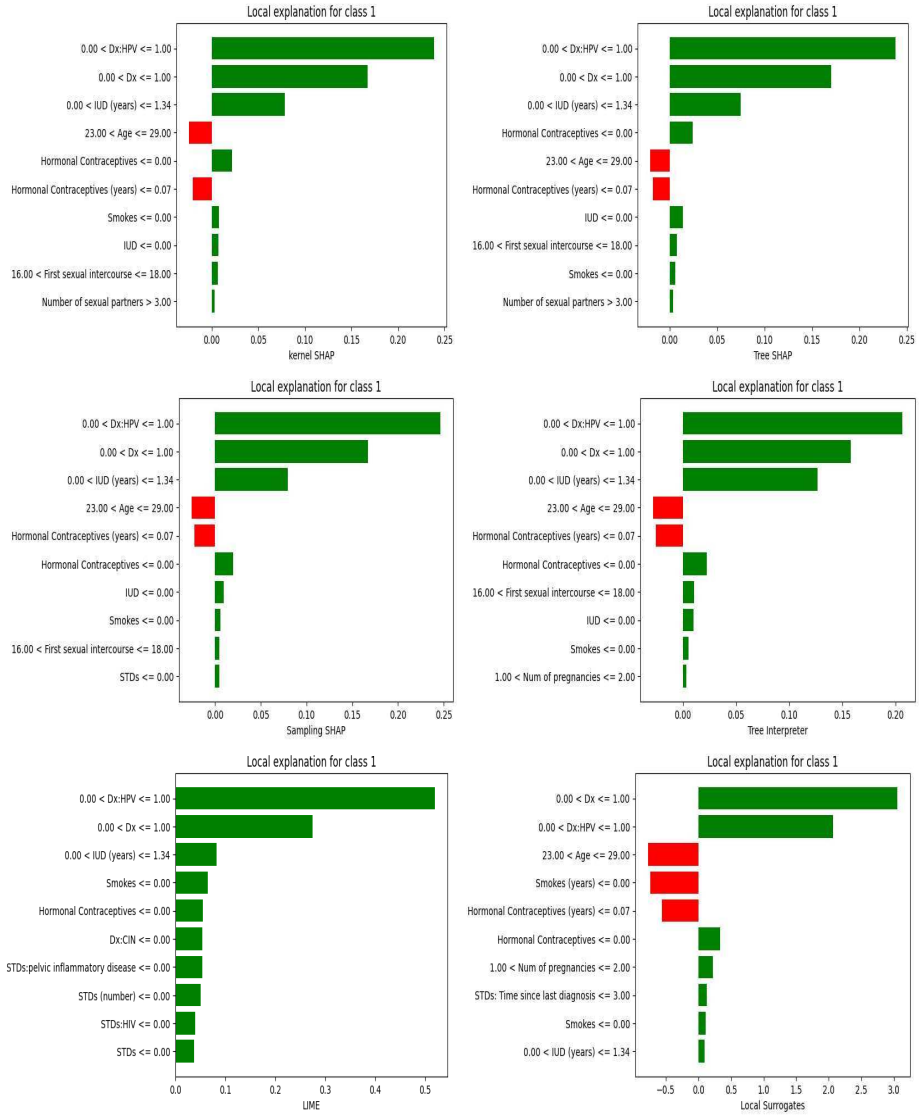


Figure B.18: Feature importance attributions for Patient with id=136.

B.5.3 Patients with Age=34

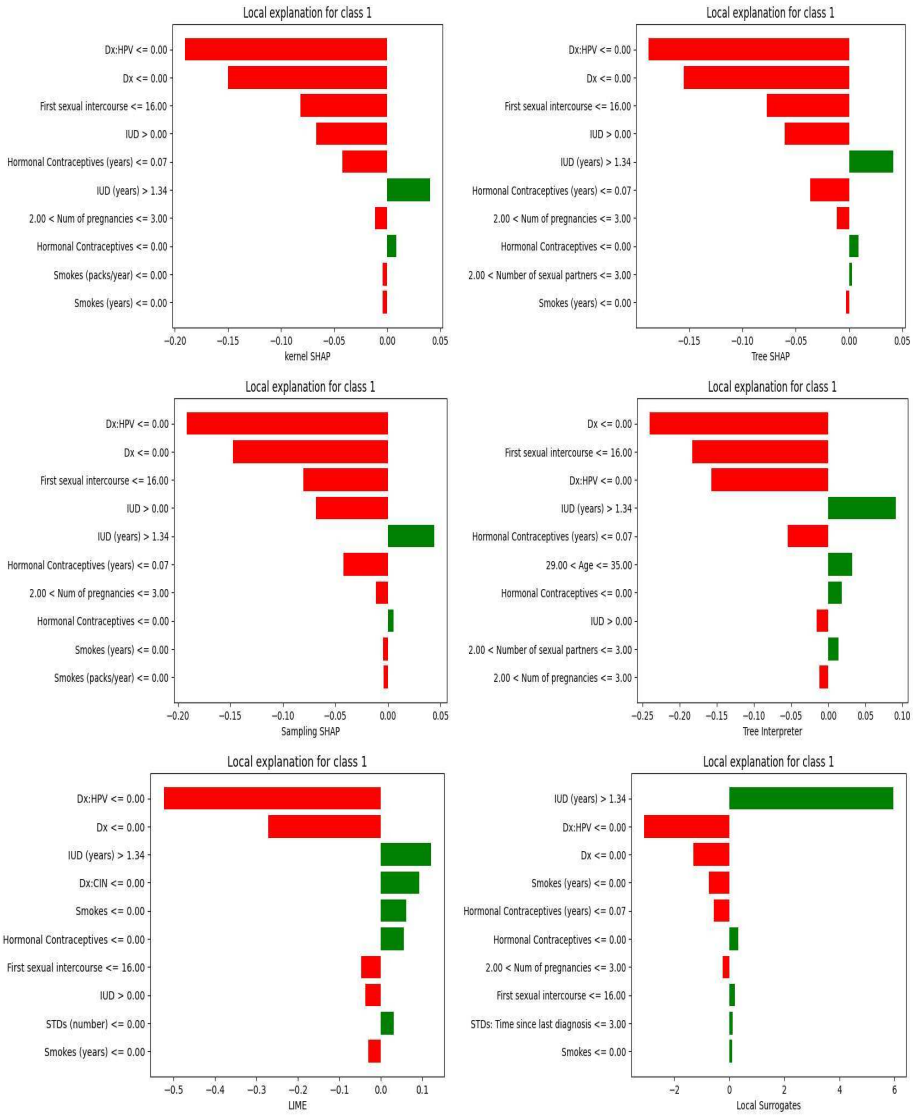


Figure B.19: Feature importance attributions for Patient with id=6.

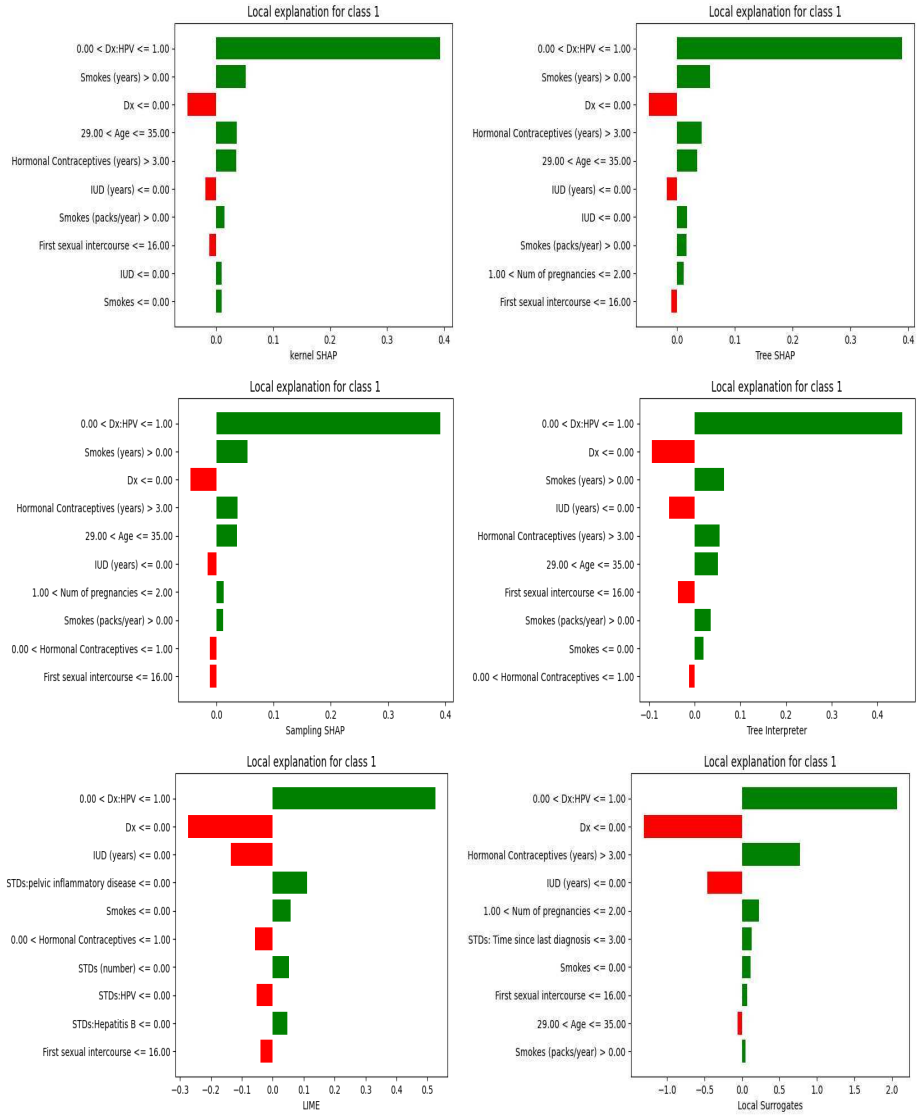


Figure B.20: Feature importance attributions for Patient with id=307.

B.5.4 Patients with Age=44 and 45

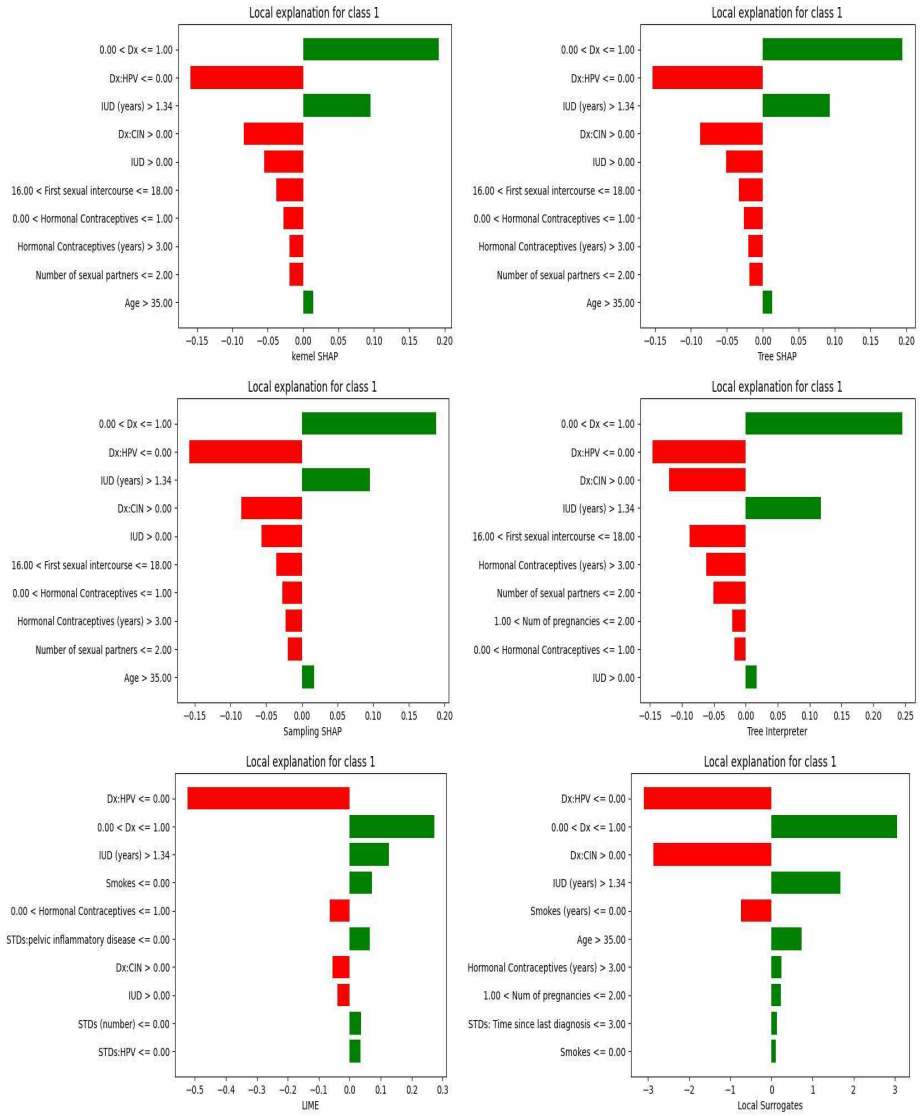


Figure B.21: Feature importance attributions for Patient with id=45.

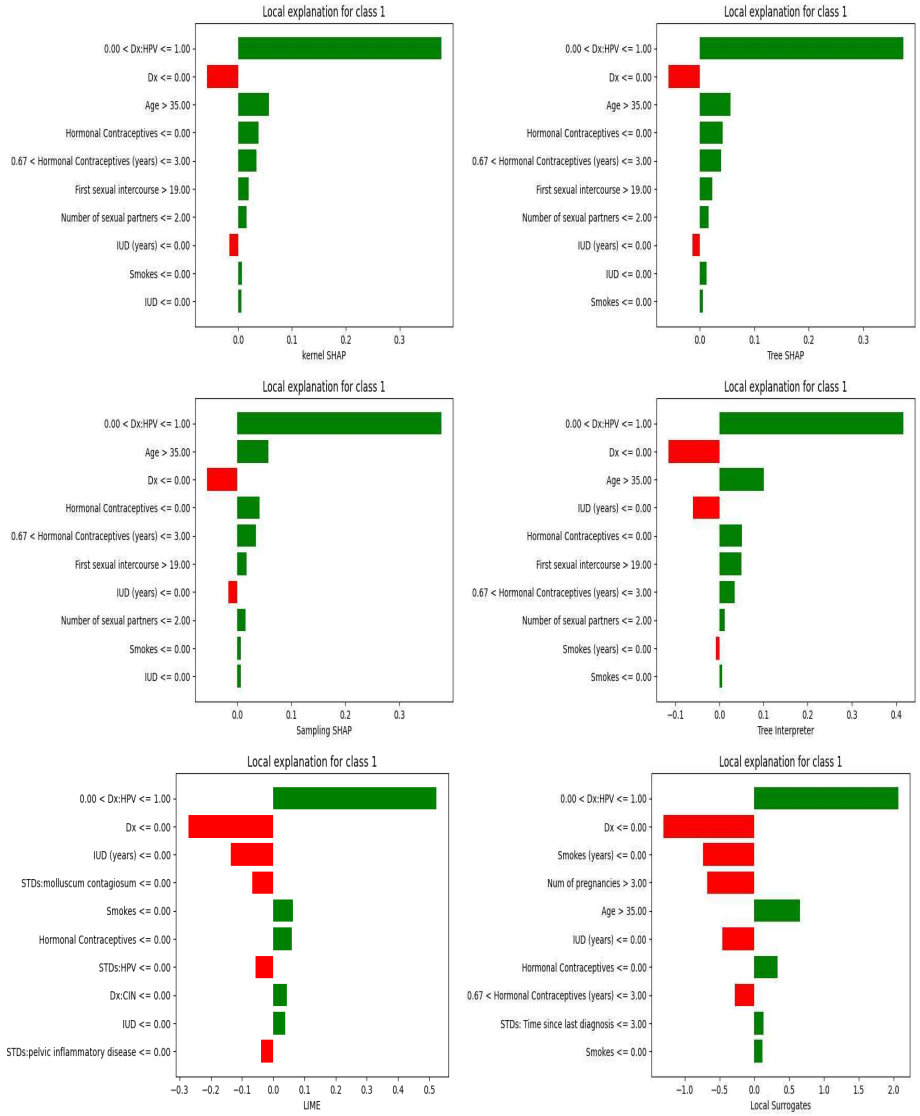


Figure B.22: Feature importance attributions for Patient with id=158.

B.5.5 Patients with Age=54

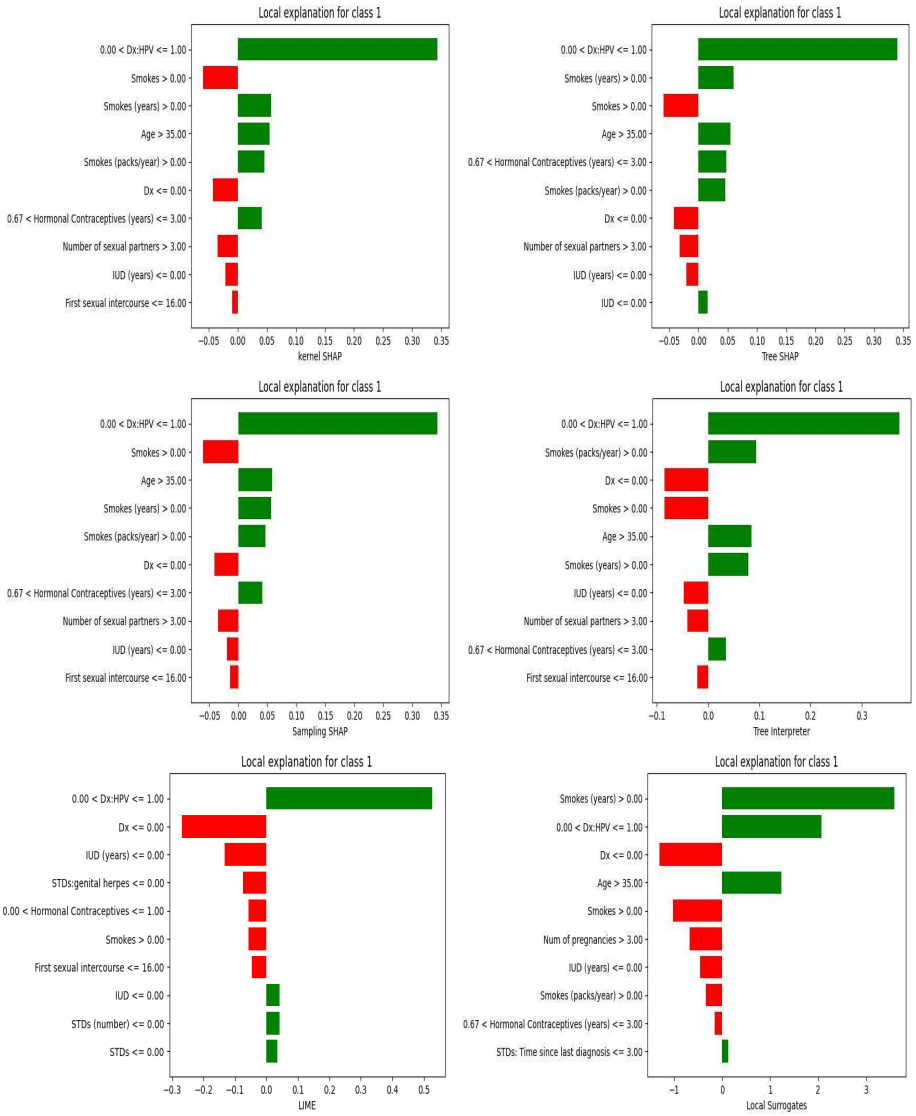


Figure B.23: Feature importance attributions for Patient with id=222.

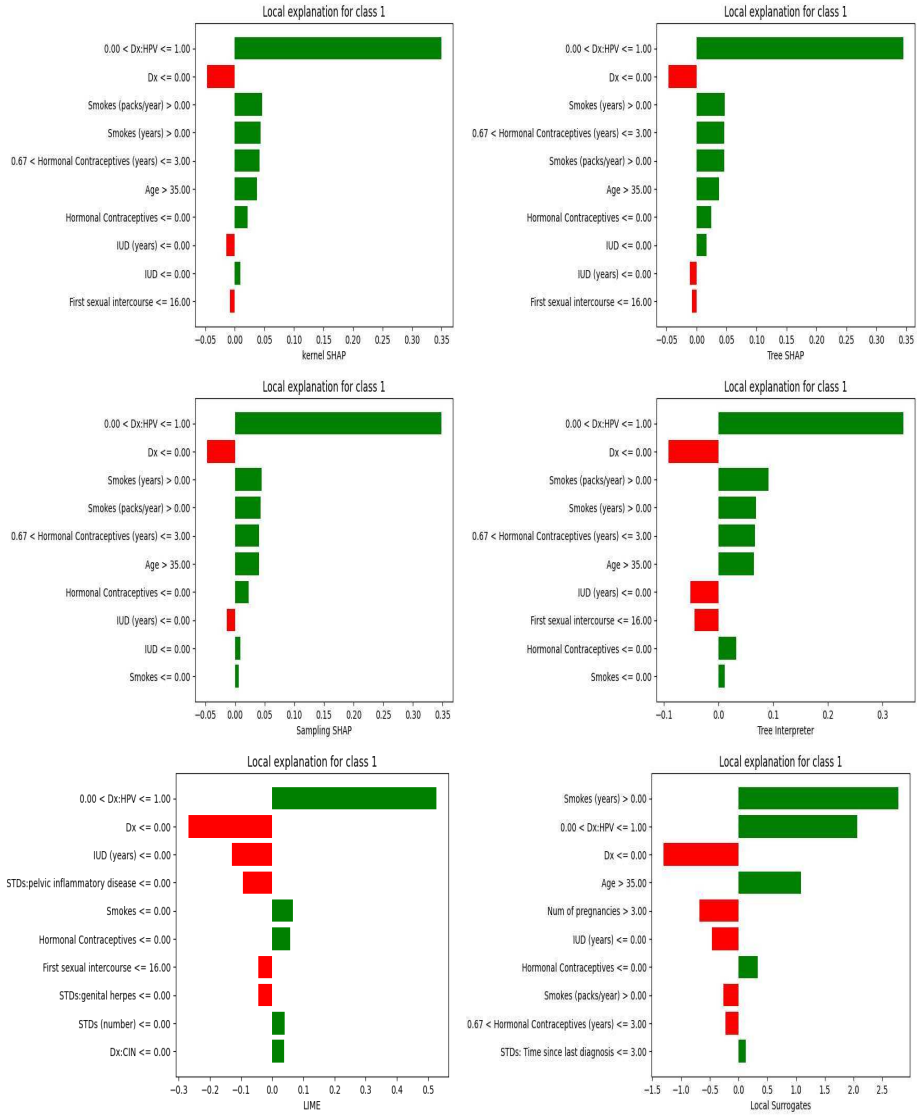


Figure B.24: Feature importance attributions for Patient with id=141.

B.5.6 Patient with Age=74

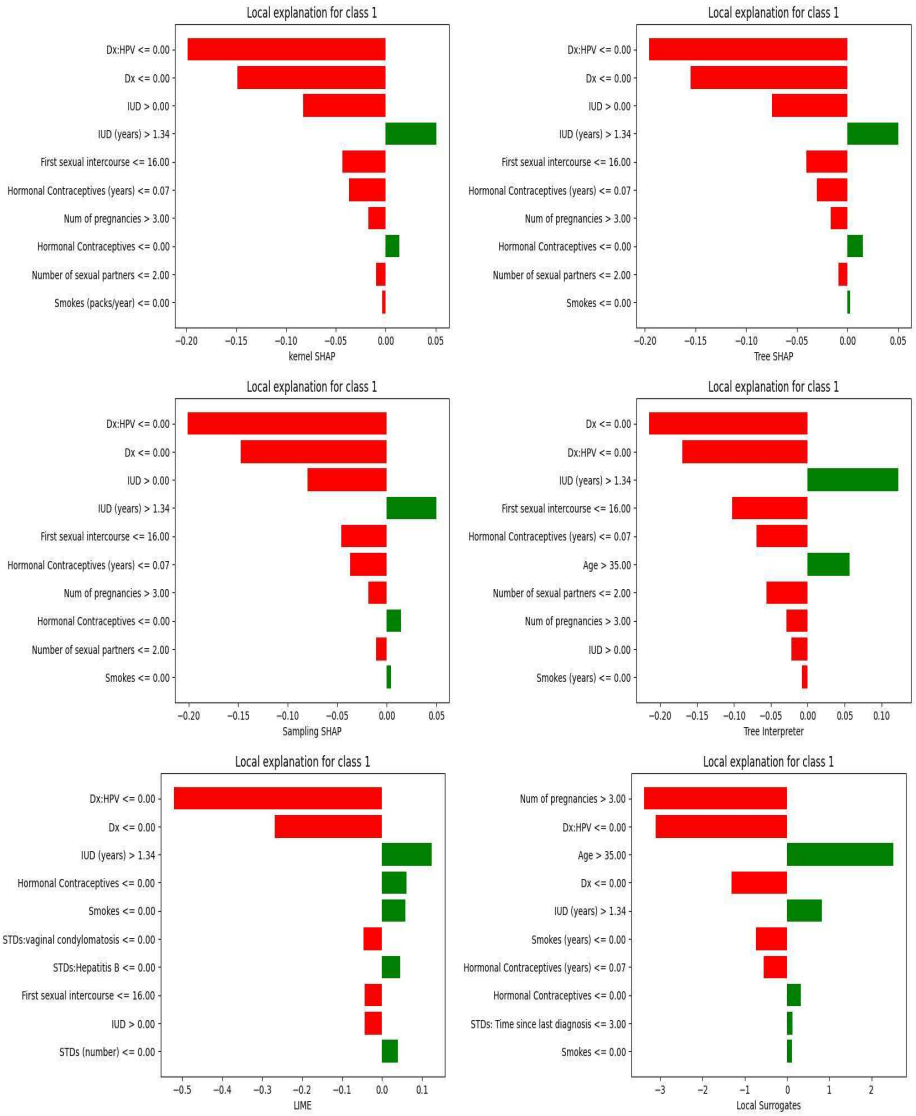


Figure B.25: Feature importance attributions for Patient with id=57.

B.6 Explanation Difference between Smoking and Non Smoking Patients

B.6.1 Patients with Smokes=1

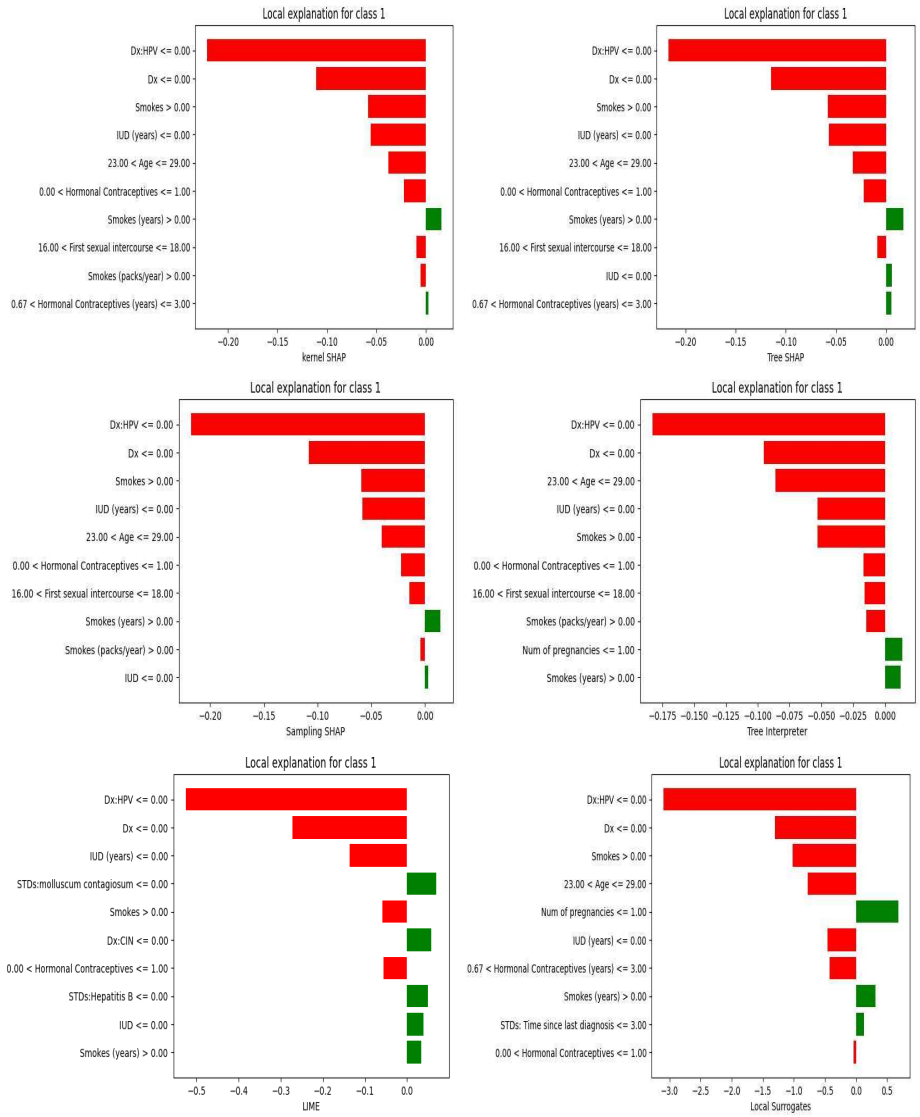


Figure B.26: Feature importance attributions for Patient with id=30.

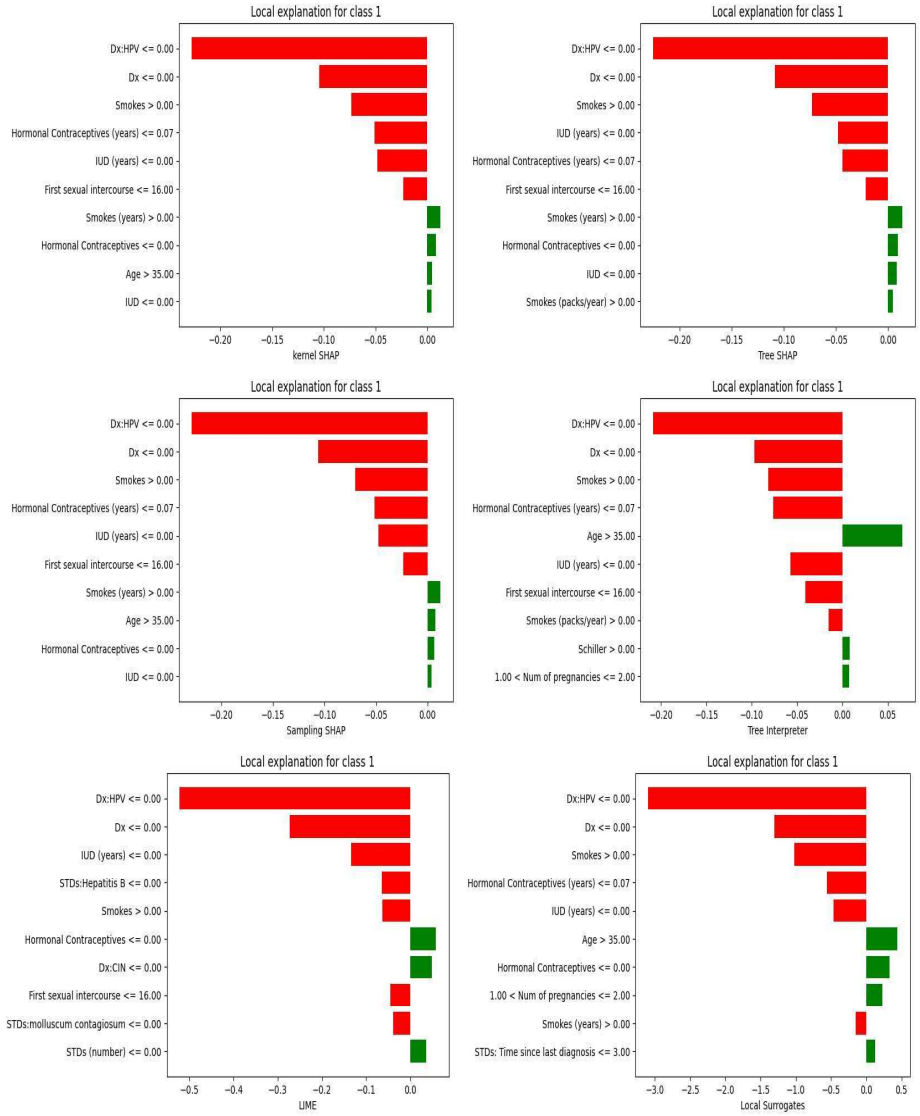


Figure B.27: Feature importance attributions for Patient with id=4.

B.6.2 Patients with Smokes=0

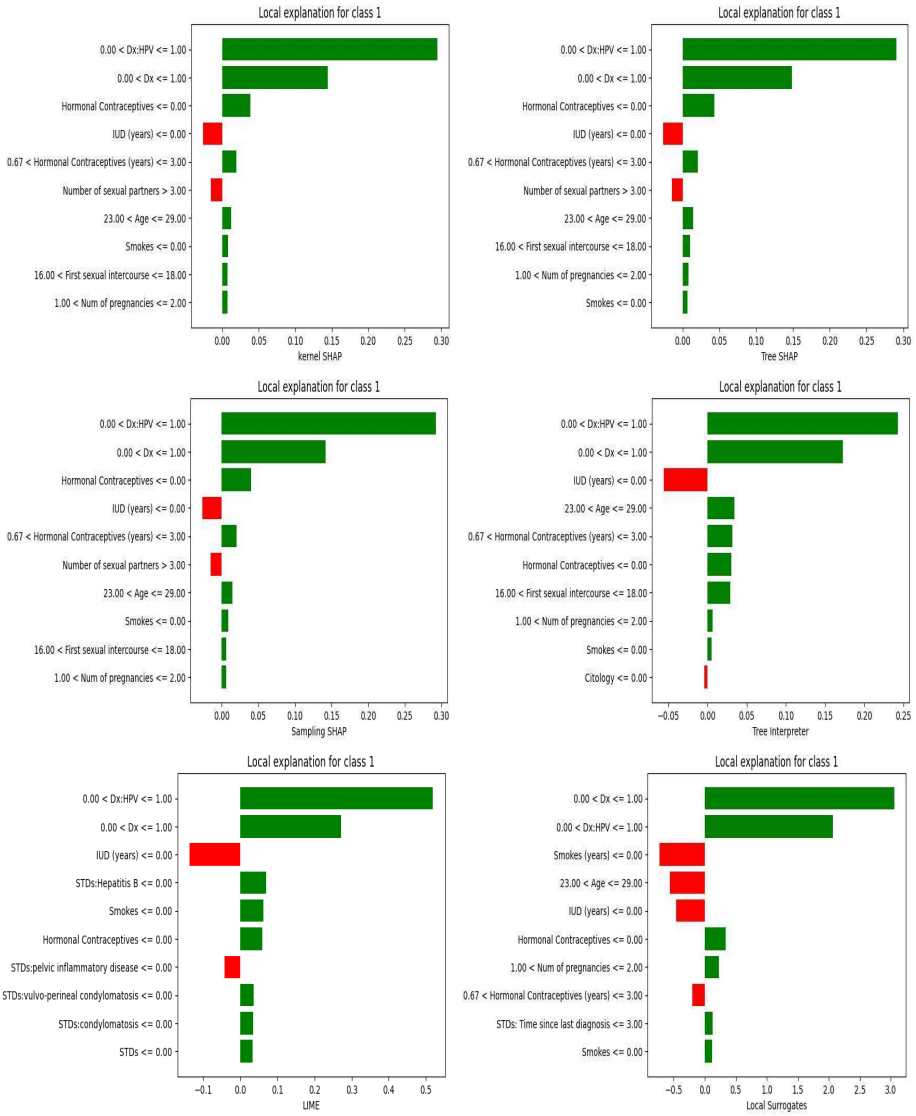


Figure B.28: Feature importance attributions for Patient with id=0.

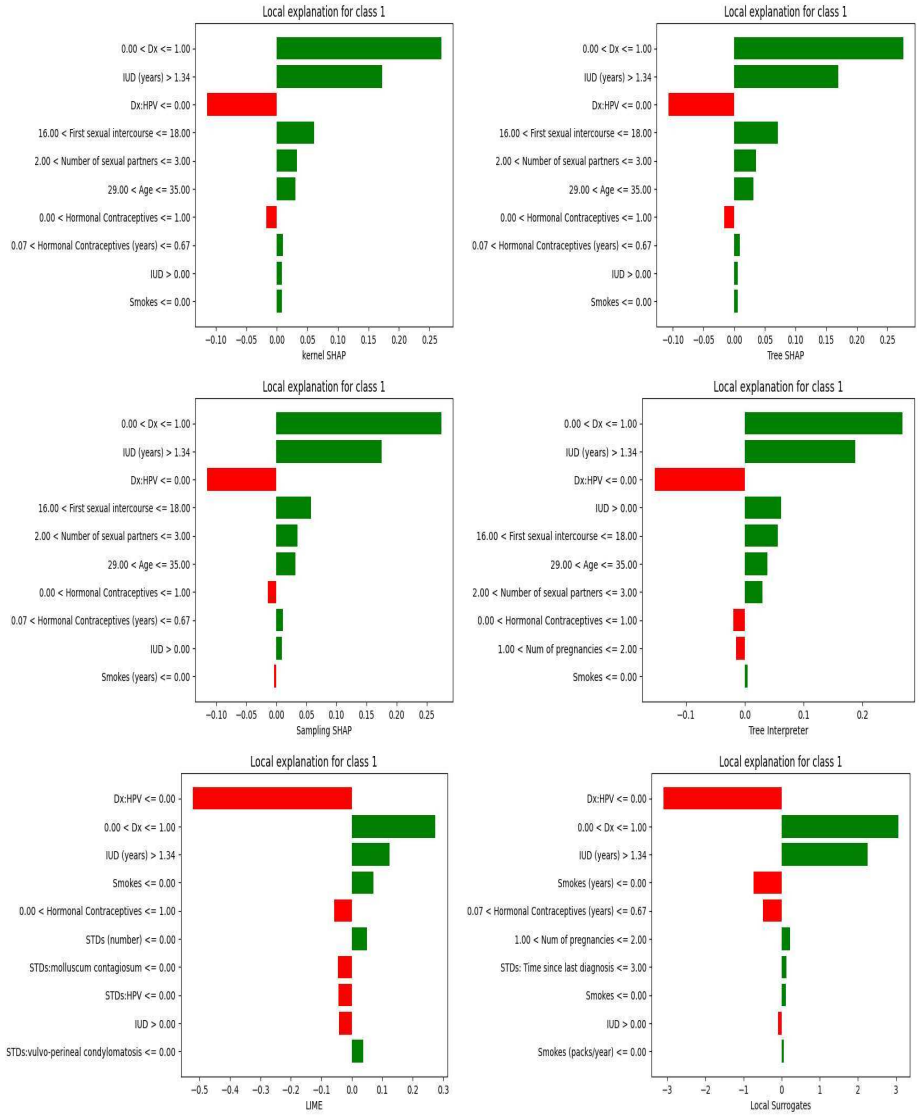


Figure B.29: Feature importance attributions for Patient with id=1.

B.7 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.

B.7.1 Feature importance for Patient 1

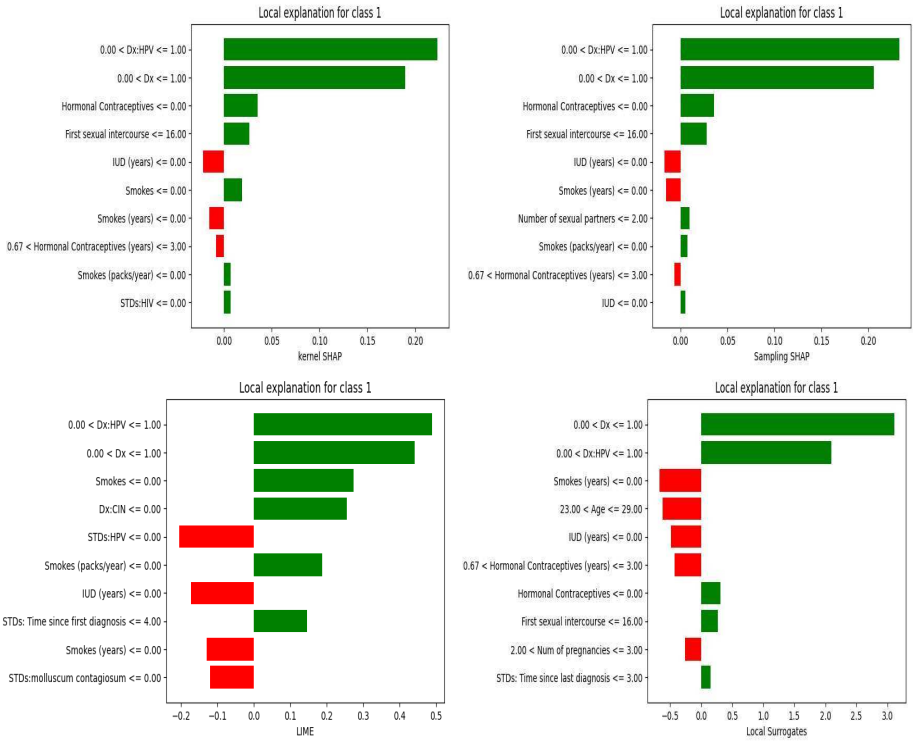


Figure B.30: Feature importance attributions for Patient 1. The predictions are made by the MLP model.

B.7.2 ROAR and Consistency of the explanations

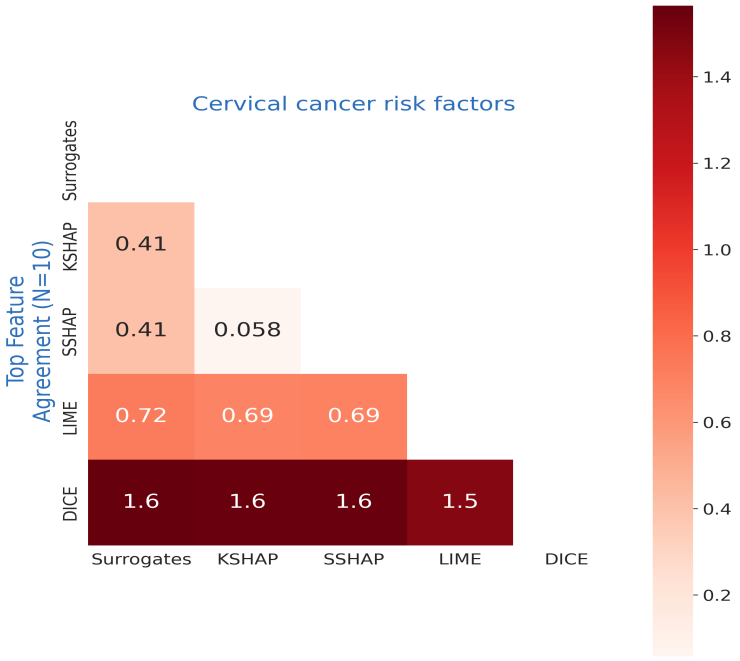


Figure B.31: Consistency

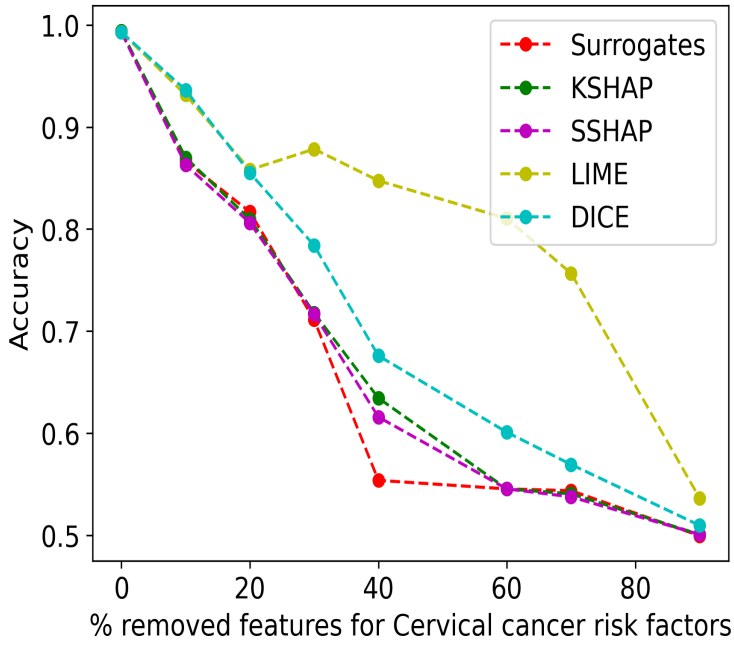


Figure B.32: ROAR

B.7.3 Faithfulness: Rank and feature agreements

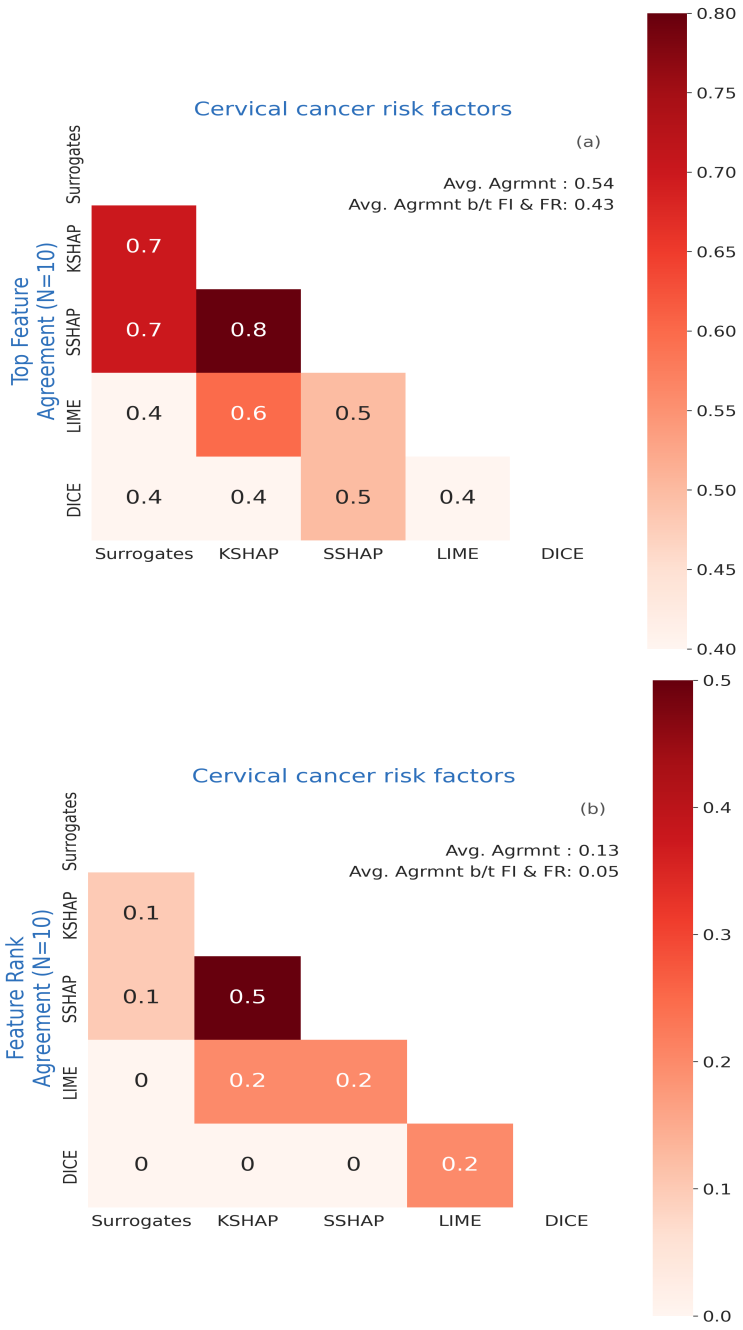


Figure B.33: Feature agreement for Patient 1.

B.7.4 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 B.2: Compactness, stability and consistency of the explanation for Patient 1 for the prediction made by the MLP.

B.8 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 B.3: 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 B.4: 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 B.5: Patient with ID: 263, Age: 29, First Sexual Intercourse: 10, Number of Sexual Partners: 4, Number of pregnancies: 5 and Smokes: 0.

ECOLE DOCTORALE

Titre : Vers des Explications Post Hoc Fiables pour l'Apprentissage Automatique sur les Données Tabulaires

Mots clés : Apprentissage Automatique, intelligence Artificielle, Apprentissage Statistique

Résumé :L'apprentissage automatique continue de démontrer de solides capacités prédictives, ce qui en fait un outil précieux dans les domaines scientifiques et industriels. Cependant, à mesure que les modèles deviennent de plus en plus complexes, il est essentiel de comprendre leur fonctionnement et de renforcer la confiance dans leurs prédictions, en particulier dans des domaines critiques comme la santé et la finance. Les chercheurs ont développé des méthodes d'explication pour rendre les modèles ML plus transparents, mais ceux-ci ne parviennent souvent pas à expliquer clairement les prédictions, ce qui limite leur utilisation pour les experts du domaine.
Dans cette thèse, nous nous concentrons sur deux axes de recherche :

Tout d'abord, nous proposons un cadre pour évaluer les méthodes d'explicabilité basées sur des caractéristiques de données spécifiques (par exemple, le bruit, les corrélations de caractéristiques, le déséquilibre des classes), offrant des conseils sur la sélection de la meilleure méthode pour différents ensembles de données. De plus, nous fournissons aux cliniciens des explications personnalisées des facteurs de risque de cancer du col de l'utérus, adaptées pour être compréhensibles et cohérentes.
Deuxièmement, nous introduisons les chaînes de Shapley, une nouvelle technique d'explication pour les prédictions à sorties multiples avec des étiquettes interdépendantes, et les chaînes LIME de Bayes pour améliorer sa robustesse.

Title : Towards Reliable Post Hoc Explanations for Machine Learning on Tabular Data

Keywords : Machine Learning, Artificial Intelligence, Statistical Learning

Abstract : As machine learning continues to demonstrate robust predictive capabilities, it has become a valuable tool across scientific and industrial fields. However, as models grow in complexity, understanding their workings and building trust in their predictions, particularly in critical areas like healthcare and finance, is essential. Researchers have developed explanation methods to make ML models more transparent, but these often fail to explain predictions clearly, limiting their usability for domain experts.
In this dissertation, we focus on two research directions:

First, we propose a framework to evaluate explainability methods based on specific data characteristics (e.g., noise, feature correlations, class imbalance), offering guidance on selecting the best method for different datasets. Additionally, we provide clinicians with personalized explanations for cervical cancer risk factors, tailored to be understandable and consistent.
Second, we introduce Shapley Chains, a novel explanation technique for multi-output predictions with interdependent labels, and Bayes LIME Chains to enhance its robustness.