# Explainable AI with Probabilistic Graphical Models

## A Study with Application and Evaluation on a Medical Domain

Master's Thesis submitted to the
Faculty of Informatics of the *Università della Svizzera Italiana*
in partial fulfillment of the requirements for the degree of
Master of Science in Artificial Intelligence

presented by
Thomas Francesco Tiotto

under the supervision of
Dr. Alessandro Facchini
co-supervised by
Dr. Alessandro Antonucci

September 2019

I certify that except where due acknowledgement has been given, the work presented in this thesis is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; and the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program.

Thomas Francesco Tiotto
Lugano, 6 September 2019

*To everyone who made me who I am.*
*To everyone who made me who I am not.*

Considerate la vostra semenza:
fatti non foste a viver come bruti,
ma per seguir virtute e canoscenza.

Dante Alighieri

# Abstract

Our societies are delegating an ever increasing number of decisions to artificially intelligent systems and the need to understand the rationale for these is becoming progressively more apparent. Especially in mission-critical domains as is the medical one, the demand for users to understand the *why* of an automated decision is crucial.

One of the main gaps in the current literature is the scarcity of *explainability* methods validated by real humans, in concrete settings. This thesis aims to address this omission by focusing on assessing the explanatory powers of Bayesian networks. Such an evaluation takes place in the medical domain and is done in collaboration with expert clinicians, employees of an institutional medical partner. To this end, a proof of concept Bayesian network-based system is developed, applied to a real medical data set and evaluated in its clinical relevance and explanatory powers by the expert users. The former is tested by executing a series of clinical questions on the system and comparing the outcome with that expected by the experts while the latter is evaluated by a think-aloud study and by a questionnaire.

The developed tool has proved its clinical relevance and ability to meaningfully interact with expert medical users. It is thus a step in the direction of validating the supposed explanatory powers of Bayesian networks, even if not all the characteristics that were expected to be important in making these models more explainable than other machine learning techniques, have been confirmed as such.

# Acknowledgements

First and foremost, my gratitude goes to my supervisors Dr. Alessandro Facchini and Dr. Alessandro Antonucci for being patient enough to guide me through all this process and for instilling me with the passion needed to push on through while having fun doing it. There is no way I could have made it without them and I thoroughly enjoyed all the time passed working together.

Likewise, Dr. Vittoria Martin from Istituto Cantonale di Patologia was invaluable in providing her honest support and often went beyond what could be expected. The same must be said of Dr. Ginevra Licandro, whose insights were essential in helping me to bridge the gap between the domains of computer science and medicine.

I naturally also want to thank my parents - Mary and Renato - and my girlfriend - Erin - for all the support and understanding they have overwhelmed me with, especially in this last period of my studies. Their fortitude and unwavering presence has helped in ferrying me across all rough waters.

# Contents

# Figures

# Tables

# Chapter 1

# Introduction

## 1.1 Context

While *artificial intelligence* (AI) - as a field - has existed for nearly seventy years [Moor, 2006], the concept of artificial intelligence dates back at least to ancient Greece. In early times, artificial intelligence, embodied in the dream of mechanical men, was part of the domain of myth; in the twentieth century, of that of science. During this last decade, artificial intelligence is no longer part of a single, specific domain, as it has materialised from Man's imagination, broken out of laboratories and has been given license to act in the world at large.

Artificial intelligence and specifically *machine learning*, the branch of AI that specialises in creating computer systems able to learn their programming from real-world examples instead of having to be explicitly coded by a human, have in the last two decades seen extraordinary success[1]. No sector of our economy has been left untouched by the recent and rapid rise of machine learning, which has been enabled by the rediscovery of neural networks, the availability of big data and increased access to parallel computing power[2]. Fields as diverse and critical as are government, healthcare, finance and bioinformatics have been revolutionised and the opportunity has been opened for new sectors - such as self-driving vehicles - to appear[3]. The increasing reliance of our civilisation on ever more complex machine learning-driven algorithms can only make us more worried about the ethical problems posed by such a series of circumstances. Our society has only very recently been confronted with the dilemma of assigning responsibility when a driverless car causes the death of a person[4]. This moral issue is only the tip of the iceberg, even when focusing exclusively on the automotive industry. As an increasing number of decisions are made in an automated way, with many of them significantly impacting both individuals and society at large, it is natural to stop and wonder which characteristics are desirable in the systems tasked with giving us these verdicts.

*Explainable AI* (xAI) is the sub-field of artificial intelligence that ideally should be found at the intersection between computer science, social sciences and philosophy and that should aim to define the desiderata of artificially intelligent systems and to develop methods to achieve

[1] See Shalev-Shwartz and Ben-David [2014] for an introduction to machine learning and Figure 1.1 to get an overview of the field of AI.

[2] A compelling talk, explaining the recent rise of ML, by AI pioneer Geoff Hinton can be found at `https://www.youtube.com/watch?time_continue=1&v=izrG86jycck`.

[3] For an overview of the impacts of AI on current society and labour market see Schwab [2017].

[4] `https://www.nytimes.com/interactive/2018/03/20/us/self-driving-uber-pedestrian-killed.html`

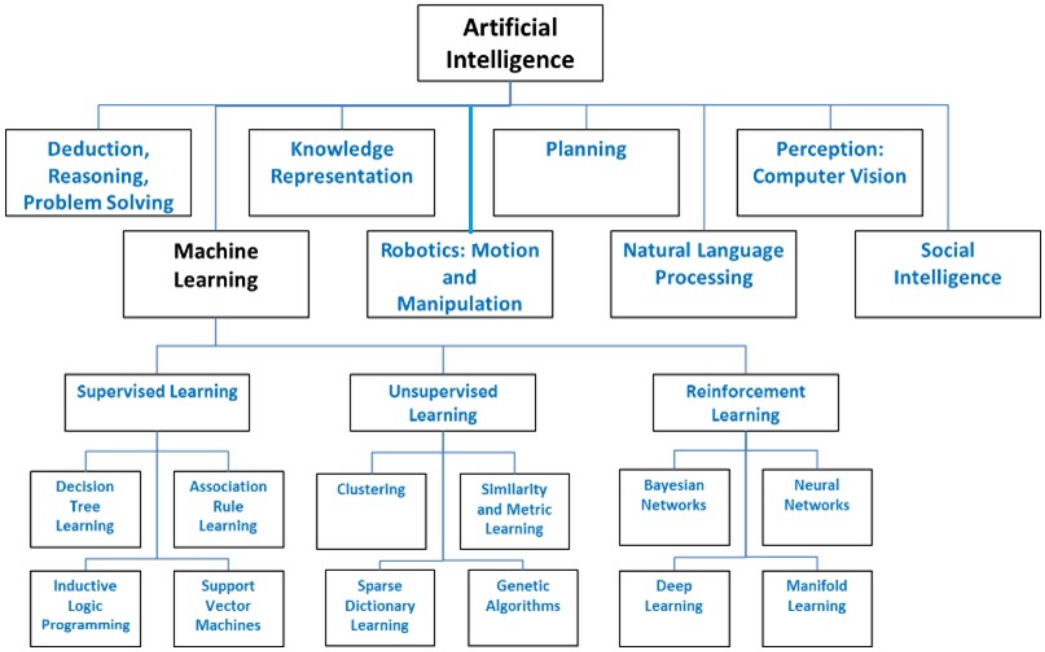



Figure 1.1. Dendrogram showing an overview of the field of AI with the position of ML emphasised [Sebastian Rudolph].

these. For example, how should a self-driving car behave when confronted with a real-world predicament analogous to the classic trolley problem - a situation where each course of action is liable to cause harm? On what basis should a person be denied a mortgage, access to university or a job interview? How can we be sure that the bias in the system is reduced to an acceptable level? How do we even define if the system is behaving morally? Would it currently be feasible for a person to appeal an automated decision they feel they had been harmed by?[5] The main strategy developed in order to achieve these goals is for the systems in question to be somehow made *explainable*. The problems in this process start at the very first step, given that within the xAI community, as noted by Doran et al. [2018], there is currently no unanimously agreed upon definition of *explainability* and consequently of the best way to achieve it in real systems. The review carried out in Section 2.2 highlights one of the fundamental problems of the field of xAI: that all researchers within the community claim their methods to be "explainable", but very few justify this with reasons grounded in the real world (as best summarised in the popular paper [Lipton, 2016]).

To solve this conundrum, various authors have tried to define taxonomies which classify systems based on their characteristics and how these relate to their perceived explainability. One of the most compelling of these is that proposed by Doran et al. [2018] and shown in Figure 1.2. The highest level in this classification is occupied by *explainable systems* i.e., those that emit explicit, human-understandable reasonings. This category appears somewhat nebulous in light of the previous paragraph: how should an explainable system be recognised if there is no good definition of explanation? A less elaborate, but still conceptually sound taxonomy

[5] For an overview of AI ethics see Bostrom and Yudkowsky [2011].

is the one proposed by Mittelstadt et al. [2019] who propose to classify systems based on the source, or *locus*, of their explainability. In this classification, models may be *ante-hoc* or *post-hoc* explainable; the former identifies systems whose explainability stems from some internal, inherent characteristic while the latter those whose source of explainability is to be found in some external behaviour, for example in the emission of extra symbols along the output. The two taxonomies are somewhat overlapping - even though the latter by Mittelstadt et al. [2019] does not consider *opaque* systems; the two classifications could be put into relationship by identifying *interpretable* with *ante-hoc* and *comprehensible* together with *explainable* with *post-hoc*. A third orthogonal classification forwarded by Doshi-Velez and Kim [2017] addresses how to evaluate the quality of an explanation. The first two taxonomies will be presented in more detail in Section 2.2 while the last in Section 2.4.

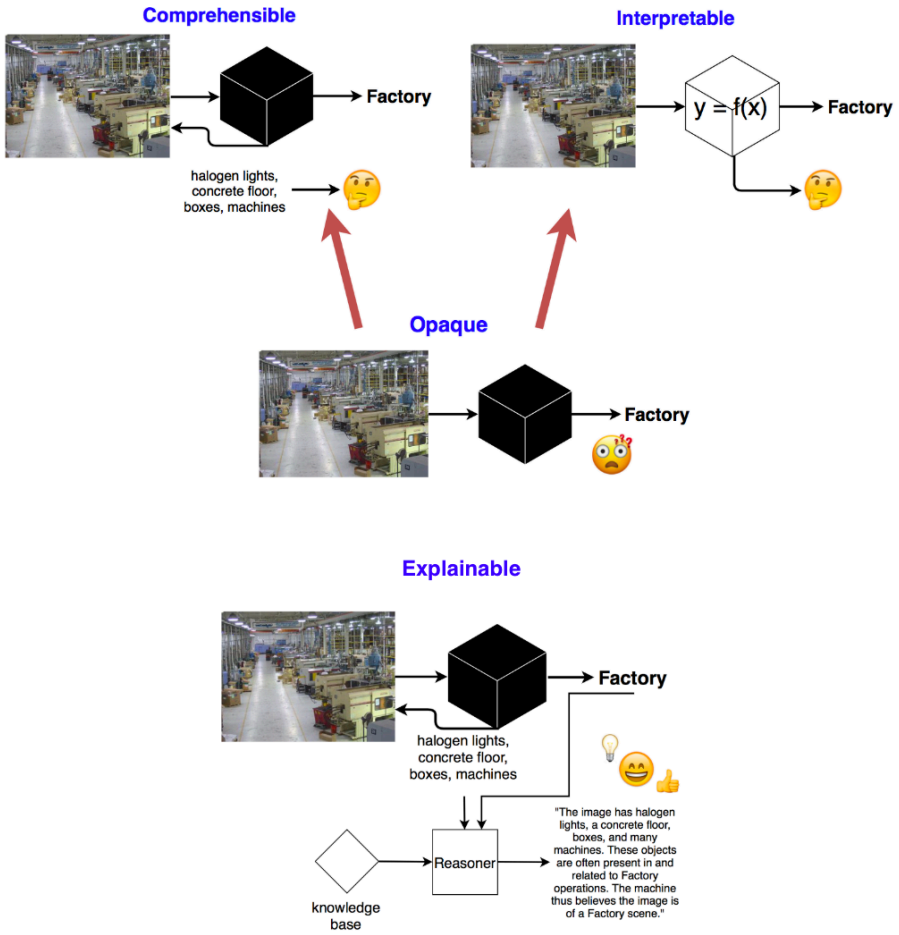



Figure 1.2. The relationship between opaque, interpretable, comprehensible and interpretable systems [adapted from [Doran et al., 2018]].

The purpose of this brief introduction to explainable artificial intelligence, is to help in understanding that many of the problems that the field aims to tackle are hard *per-se* and may not have a unique optimal solution. This is because such issues are not only engineering problems, but exist at the intersection between man and machine and as such cannot be solved using only engineering methods[6]. This lack of awareness for the centrality of the human element of an explanation is one of the main pitfalls of the xAI community, since most of its researchers come from the *hard-science domains* of computer science, mathematics and statistics. There is little hope of defining the desiderata for intelligent systems without the guidance that can only come from philosophy, because of its millennia-long tradition in dealing with ethical issues. There is also no way to satisfactorily move towards and evaluate these desiderata without resorting to the well-established literature and methods of the social sciences[7], as the human element is inherent in any explanation. An attempt to define how a computer should relate itself to its user could be assimilated with the notion of "reinventing the wheel", as the field of human computer interaction has many such techniques already. Although an endeavour still pursued by many xAI researchers, it should be clear that when the human - and particularly the ethical - domain are part of the equation, it is impossible *by definition* to find an optimal and unique solution[8].

Effectively, the biggest gap that can be identified is a dearth of explainability methods that have been validated not only by domain experts, but even by real humans. Many researchers seem content with claiming *formal explainability* and neglect the human element of the explanation. An explanation, by its vert nature, involves an *explainer*, the machine, and an *explainee*, us humans. Up till now, it seems that xAI is content with only explaining the machine but, in doing so, overlooks the fact that it may be offering the users no explanation at all[9].

## 1.2 Problem and Significance

The biggest problem of AI is no longer its perceived utility, as this has mostly been solved by its recent successes, but its capacity to *elicit the trust of users*. The creators of an automated system should be able to make it be trusted in a manner proportional to the criticality of its application. [Gilpin et al., 2018], [Abdul et al., 2018] The potential lack of trust felt by users stems from the difficulty in *verifying the system's outputs*; if no rationale can be inferred for why a given ML model made a certain decision, there is also no way to understand if these outputs conform to our moral norms[10]. As discussed in Section 2.3, no explicit justification for the link between the explainability and the *trustability* of a model has been found in the reviewed literature; nonetheless, it seems quite natural to infer that a person would not trust decisions made on an unknown rationale. Unfortunately, the *explainability* and performance of machine learning models are usually inversely proportional, as is shown in Figure 1.3.

There are many examples of modern methods - such as boosted trees, random forests, bagged trees, kernelized-SVMs - that show the tendency outlined in Figure 1.3, but it is best exemplified by *deep neural networks* (DNNs). Deep neural networks are machine learning models constructed by stacking many layers of artificial neurons; these systems currently offer state of

[6] This is what is meant by Doshi-Velez and Kim [2017] when they talk about "incompleteness in the problem formalization".

[7] For a good example of how this may work, see Stumpf et al. [2009].

[8] See the concept of "wicked problem" in [Rittel, Horst, 1973].

[9] See Mittelstadt et al. [2019] for a critique of the field of xAI, based on this lack of awareness.

[10] See Gilpin et al. [2018] and Abdul et al. [2018] for a discussion on trust in AI.

the art performance on a variety of tasks but are among the least interpretable systems due to the fact that they represent information in an *implicit* (non-symbolic) and *distributed* manner. Some older models, like decision trees or rule-based methods, are inherently more interpretable due to their simplicity and the fact that they can explicitly demonstrate their reasoning steps, but are less accurate and flexible than more modern techniques. [Biran and McKeown, 2017] (as exemplified in Figure 1.3)

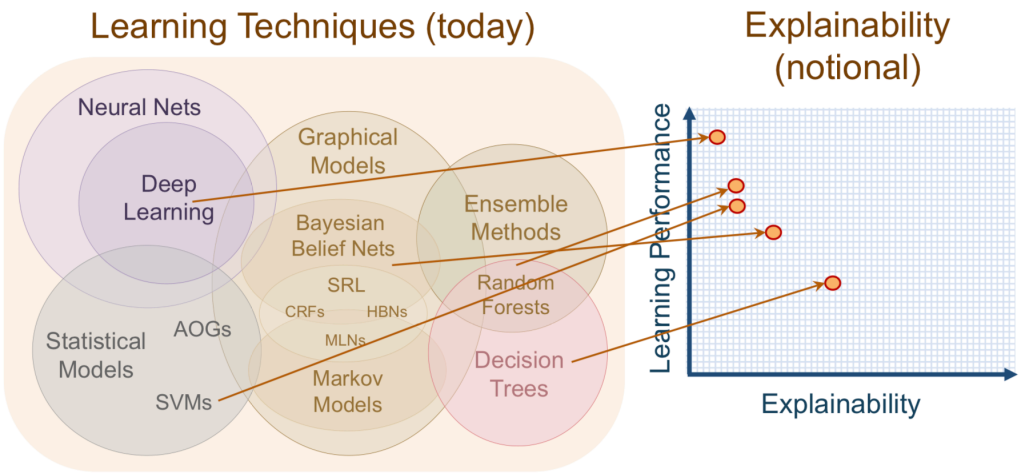



Figure 1.3. Mapping showing the trade-off between performance and interpretability of contemporary and older machine learning models [Gunning, 2017].

The runaway success obtained by modern machine learning in a variety of domains, on a spectrum that goes from oceanography to social work, has created the desire to also apply these methods to mission-critical and traditionally more entrenched fields. A perfect example of an area exhibiting both these characteristics is that of medicine[11]. The first successful artificially intelligent systems date back to the 1970s and 1980s, these were based on *symbolic methods* integrated with *knowledge-bases*. These systems were, by design, capable of providing an explanation for their reasoning and were thus accepted by the medical community in an implementation known as *expert systems*, which aimed to aid in the diagnosis of disease[12]. The insufficient ability of modern AI methods in being able to provide a justification for their reasoning has stunted their acceptance in the field of medicine, regardless of their superior performance and accuracy.

One modern ML model that has seen a modicum of success in the medical domain is that of *Bayesian networks* (BNs) (defined in much greater detail in Section 3.5), a graphical and computationally efficient way of representing dependencies between variables of interest[13]. The graphical component is given by the fact that each variable is represented by a node of a directed acyclic graph (DAG) (Definition 3.19), with the edges connecting them modelling their dependencies. The efficiency stems from the fact that the graph structure imposes a factorisation of the joint probability space and thus allows each variable's values be calculated using

[11] For an overview of deep neural networks applied to medicine see Gitter and Greene.
[12] For an overview of expert systems see Liao [2005].
[13] For an overview of BNs in the medical domain see Lucas [2001].

only those of its parents. Bayesian networks may be uniquely suited to providing effective explanations by virtue of their inherent characteristics, as is discussed in detail in Section 2.5 below.

In a high-stakes domain such as the medical one, it would be unthinkable for a doctor to trust the predictions of an AI system *a priori*. Any decision with profound moral implications - such as prescribing or interrupting the treatment of a patient - would first have to be validated by a human; who would need to understand the rationale behind the machine's output to be able to do so. The feasibility of carrying out this validation is dependent on the degree of interpretability of the model that made the decision and, unfortunately, the lack of explainability methods tested in the real world are one of the main gaps identified in the field xAI. BNs are no exception because, as noted by Timmer et al. [2015], their underlying formalism makes them appear akin to "black-boxes" to domain experts, who are usually not well-versed in statistical reasoning. Among other professionals, doctors certainly cannot be expected to double as machine learning experts. Therefore, the onus of developing comprehensible models falls squarely on the researchers in the xAI community. Through a *process of real-world validation and testing*, such researchers need to strive to develop methods that are not only provably correct but, just as importantly, confirmed in their capacity of relating efficiently to their users.

Explainability is necessary for a ML system's outputs to be verified and this, in turn, is a prerequisite for them to be applied in mission-critical domains. Furthermore, explainability is also essential to be able to extract knowledge from data. The amount of information that a machine can process is many orders of magnitude greater than that inspectable by any human; this may allow for a computer to spot new patterns in the data, which would otherwise remain undetectable when observing only a limited amount of samples. The ability to understand the mapping from the model's inputs to its outputs can be seen as *understanding the system's "reasoning" process* and could thus lead to new insights or to the confirmation of existing theories. This is because understanding the process the model uses to give a certain output can make the machine's *deep/vertical* analytical power available to our more *general/horizontal* capacities. This is also noted *en passant* by Doshi-Velez and Kim [2017] when they state that "humans' goal is to gain knowledge. We do not have a complete way of stating what knowledge is; *thus the best we can do is ask for explanations we can convert into knowledge*".

Within this context, the main focus of this thesis is to address one of the most severe gaps in the current xAI literature: the lack of validation of machine learning systems with real expert users in concrete situations. The work will focus in particular on the medical area and will attempt to assess the effectiveness of a Bayesian network-based system by means of an evaluation carried out in collaboration with expert clinicians in a real work setting, over a period of time. The main hypothesis under investigation is if Bayesian networks may be inherently suited to being made explainable, compared to other ML models. Another objective is also to lay out the methodological groundwork for future research aimed at addressing the lack of *human-centred evaluations*, that was found missing in the xAI literature.

## 1.3 Response

In order to contextualise the current work, Chapter 2 will investigate the notion of *explainability*, its importance and how to evaluate it, as defined in the current xAI literature. In particular, the explainability of Bayesian networks will be reviewed in detail as this model will be the basis for the methods carried out in this thesis.

To address the gaps identified in the previous section and in the following chapter, the work conducted in this thesis will concentrate on explainability in the medical domain and will present both a practical and a theoretical part. The methods will include the implementation of a Bayesian network-based system, inspired by the work by Butz et al. [2018] (see Section 2.6), and its subsequent evaluation. This recent xAI paper never provided any results for the compelling methods it presented, thus a proof of concept system implementing them, together with novel extensions, will be developed. The system will be a means of exploring the efficacy of the explanatory modes for BNs - as identified by Lacave and Díez [2002] - mainly the *graphical*, *linguistic* and, in particular, the *dialogical*. Their taxonomy for BN explanations, together with the psychological characteristics of an explanation [Miller, 2018], will be the framework against which the methods developed in Chapters 4 and evaluated in Chapter 5 will be measured. The explainability of the work will be validated by real medical experts, in a concrete setting, over a period of time; this will also be a means to provide a validation for the methods of the initial paper [Butz et al., 2018] and, more in general, of Bayesian networks as a whole. The system will be a proxy to explore the explainability of BNs that should, in theory, be well adapted to giving highly effective explanations (as discussed in Section 2.5). The hope is also to set a methodological precedent for other *application-grounded evaluations* (see [Doshi-Velez and Kim, 2017] and Section 2.4), with the aim of helping to reduce the gap in the xAI literature of the absence of actual human evaluation of explainability.

The methodology that will be used to evaluate these hypotheses will start with the study of a specific, recent xAI paper by Butz et al. [2018], whose method will become the basis for the work carried out. Then a proof of concept system will be developed, coded in Python and that will implement the learning and updating of a Bayesian network (see Subsections 3.5.2, 3.5.3 and 3.5.4) together with standard methods (see Subsection 4.3.2) and novel algorithms (see Subsection 4.4.1). The system's main aim will be to *support medical decision making* by establishing a dialogue with the domain expert user and to evaluate the usefulness, in terms of explainability, of various other interaction modes such as conditional probability and most probable explanation queries (see Subsection 3.5.4); various algorithms and interaction paradigms will be developed to do so, as detailed in Section 4.4. Finally, this system will be evaluated by expert pathologists at *Istituto Cantonale di Patologia* of Locarno (Switzerland), the partner institution who will collaborate on the testing and evaluation of the software tool that will be developed. The data set that will be used to learn the Bayesian network, composed by the clinical profiles of over 3000 breast cancer patients in the Swiss canton of Ticino, will also be contributed by the institute. The evaluation will be done through the use of interviews, questionnaires and observation of the experts at work, with the objective of *clinically validating* the system and *assessing its capacity to relate to the medical professionals* in a meaningful manner.

The objective of such a methodology will be to both validate the claims made by Butz et al. [2018] and, more generally, to investigate the assertions made in the literature regarding the explainability of Bayesian networks (see Section 2.5); at the same time, the aim is to also identify which characteristics of these models may be the most important in enabling their *comprehensibility*.

# Chapter 2

# Literature Review

## 2.1 Introduction

The aim of this chapter is to carry out a review of recent, relevant literature with the objective of clarifying the main concepts relating to the field of explainable AI, to gain a picture of how researchers have approached these and, most importantly, to identify current trends and gaps.

The chapter is organised as follows:

- Section 2.2 aims to clarify the concept of *explainability*, which is central to the field of explainable AI and to every further discussion.
- Section 2.3 investigates which may be the reasons for the importance assigned to explainability by our contemporary societies.
- Section 2.4 is an overview of the main methods that have been proposed to measure explainability.
- Section 2.5 will focus on an assessment of the concepts discussed in the previous sections, as applied specifically to Bayesian networks.
- Section 2.6 offers a review of a recent paper by Butz et al. [2018] and connects it to the aforementioned notions. The reason for this analysis, is because the paper constitutes an important reference for the work carried out in this thesis since it will be the starting point for the developed methodology.

Throughout the chapter, various gaps that are present in the literature will be identified and assessed; these will be summarised coherently in Section 2.7.

## 2.2 Explainability

Doran et al. [2018], carried out a frequency analysis of explanation terms within documents from relevant research communities. They highlight how different circles have different approaches to the concept of explainability and how, even within the same group, terms are used interchangeably. In particular, they note the overloading of the notion of *explanation* with that of *interpretability*; a concept that is often defined within the xAI community as necessary for,

but distinct from, explainability. The use of *interpretable*, as signifying the property belonging to a system whose inner workings are accessible, can be found, for example, in the recent paper [Gilpin et al., 2018]. In other recent works the two terms are conflated, for example in [Mittelstadt et al., 2019], [Guidotti et al., 2018] and in the influential work by Doshi-Velez and Kim [2017]. This seems to prove the point made by Lipton [2016] in the widely-cited paper "The Mythos of Model Interpretability", that "the task of interpretation appears underspecified. Papers provide diverse and sometimes non-overlapping motivations for interpretability and offer myriad notions of what attributes render models interpretable".

Most works, even those that blend the notions of interpretability and explainability, seem to agree on the end-goal that implementing such a concept should have; that is, to "summarize the reasons for neural network behavior, gain the trust of users, or produce insights about the causes of their decisions" [Gilpin et al., 2018] by being able to "explain or to present in understandable terms to a human" [Doshi-Velez and Kim, 2017]. Where the consensus diverges, is in defining what constitutes an explanation and in fixing the desiderata that it may have. Mittelstadt et al. [2019] identify two broad classes of interpretability/explainability within the literature: *transparency* (also called *ante-hoc* explainability) and *post-hoc* interpretability. The former type deals with the internal workings of a system while the latter applies to its external behaviour. Lipton [2016] identifies three explanations that can make a model transparent: a mechanistic understanding of the workings of the system in its entirety, of the individual components or of the algorithm. A system may be made post-hoc interpretable by way of natural language explanations, visualisations or interactive interfaces, among others. These methods often do not precisely clarify the exact working of a model, but "they may nonetheless confer useful information for practitioners and end-users of machine learning". Biran and McKeown [2017] note how the transparent or *white-box* paradigm was sufficient for classic rule-based models but - with the advent of contemporary machine learning models - is no longer useful. They argue that it is nowadays unreasonable to expect that any domain expert be able to understand a prediction if they are not also a machine learning specialist. To address this issue, they propose a *natural language generation system*; that is, a *post-hoc* explanation in the categorisation by Mittelstadt et al. [2019].

A widely-recognised feeling, closely connected with the already identified lack of shared working definitions, seems to be that researchers of explainable AI are ignoring the enormous corpus of existing work in the fields of philosophy, psychology, cognitive and social sciences and human-computer interaction. This feeling of disconnect is echoed by Gilpin et al. [2018] who point out how philosophical texts have long debated what constitutes an explanation and by Mittelstadt et al. [2019] who explicitly say how "many different people, be they lawyers, regulators, machine learning specialists, philosophers, or futurologists, are all prepared to agree on the importance of explainable AI [...] very few stop to check what they are agreeing to, and to find out what explainable AI means to other people involved in the discussion". The fact that explainable AI researchers seem to be intent on "reinventing the wheel" is stated most strongly by Miller [2018], whose paper "Explanation in Artificial Intelligence: Insights from the Social Sciences" is based on the premiss that "most of the research and practice in this area seems to use the researchers' intuitions of what constitutes a 'good' explanation" and argues for the adoption of the existing research in the social sciences. The author's views are well summarised by the position xAI is set to occupy in Figure 2.1. The feeling is that explainable AI researchers are calling their methods an *explanation* based on purely personal views and are thus building explanations that only work for themselves; in other words, "the inmates are running the asylum" [Miller et al., 2017]. The following quote from Guidotti et al. [2018] seem

to perfectly sum up the state of the research in the field:

> It is evident that the research activity in this field completely ignored the importance of studying a general and standard formalism for defining an explanation, identifying which are the properties that an explanation should guarantee, e.g., soundness, completeness, compactness and comprehensibility. Concerning this last property, there is no work that seriously addresses the problem of quantifying the grade of comprehensibility of an explanation for humans, although it is of fundamental importance.
>
> [Guidotti et al., 2018, pag. 37]

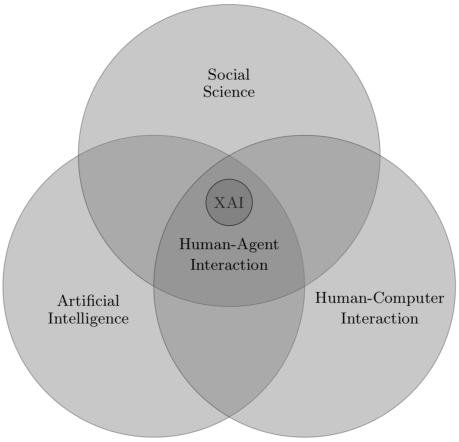



Figure 2.1. Venn diagram showing where the field of xAI should ideally be positioned [Miller, 2018].

To remedy this state of affairs, there have been a number of works, such as [Doshi-Velez and Kim, 2017], which have attempted to define what an explanation means and to reach a consensus on it. The most compelling attempt is in a paper by Doran et al. [2018] (already mentioned in Section 1.1) where the authors try to synthesise the current state of affairs into a taxonomy of models:

- *Opaque systems*: these are systems that offer no insight into how their internal workings transform the inputs symbols, usually real-world data, into some output, for example labels that classify the data or predictions of unseen cases. All closed-source algorithms fall under this definition.
- *Interpretable systems*: this is the vastest category, as the characteristic of these systems is *transparency* i.e., their inner workings are accessible but the onus of comprehensibility falls completely onto the user. The classical example is that of neural networks, where the mapping from inputs to outputs is inspectable by the user who can, theoretically and depending on her skill, interpret them. In the case of neural networks this may be a daunting task due to their distributed and non-symbolic nature; in other classes of machine learning models, for example Bayesian networks, the task may be easier as discussed in Section 2.5.

- *Comprehensible systems*: systems in this category emit additional symbols together with their outputs with the explicit intent of giving the user the means to interpret and understand the automated decisions. The additional symbols may be visualisations, natural language text or any other means of demystifying the output. These extra symbols would need to be graded based on the user's expertise, as comprehension is a property that involves both man and machine. An example of such a system would be an image classifier that, together with its output, highlighted the parts of the image it used to make its decision.

- *Explainable systems*: the highest level in the taxonomy includes those models that emit an explicit *explanation* i.e., a human-understandable line of reasoning.

It is recognised that comprehensibility depends not only on the system's characteristics, but also on the user's ability and knowledge, thus implicitly accepting the view that xAI should include elements from the social sciences. *Comprehensibility* and *interpretability* are considered separate concepts as comprehension requires transparency but interpretation does not, given that the user may reason only over the emitted extra symbols. This notion of comprehensibility is expanded into that of *real* explainability, based on a notion of "ability to formulate, for the user, a line of reasoning that explains the decision making process of a model using human-understandable features of the input data" [Doran et al., 2018].

## 2.3 Importance of Explainability

As noted by Edwards and Veale [2018], "businesses and governments are increasingly deploying machine learning (ML) systems to make and support decisions that have a crucial impact on everyday life" so, as Gilpin et al. [2018] say, "it becomes necessary for these mechanisms to explain themselves". This feeling of urgency and purpose is echoed throughout the reviewed literature; it seems that even if researchers and the field as a whole cannot agree on a definition of explainability (as discussed in Section 2.2), there is a keen awareness of the need for models to be explainable. This intense urge to define explainability and, at the same time, to try and create models exhibiting this property, may be counterproductive as the field risks fragmenting into a series of diverging strands, as noted by Abdul et al. [2018] and visualised in Figure 2.2.

There are a myriad of reasons brought forth as a justification for the development of explainable AI, and we will review these in the following paragraphs, but it seems timely to start with one in particular: the need introduced by the European Union's broad General Data Protection Regulation (GDPR). The GDPR was approved by the European Parliament in 2016 and came into effect in 2018. More than one author cites an urgency to conform to this regulation (for example Doshi-Velez and Kim [2017], Gilpin et al. [2018]), most likely referring to Article 22 of the regulation that, supposedly, mandates for a "right to an explanation" from algorithms. While algorithmic explainability is undoubtedly a commendable objective, it may be the case that this particular reason to strive for it be a false one. Edwards and Veale [2018] posit that Article 22 of the GDPR actually does not contain the publicised right to an explanation but is "merely a right to stop processing unless a human is introduced to review the decision on challenge" and as the authors point out, there are, nowadays at least, very few systems without a human in the loop. Secondly, there is no mandate for the "explanation" to be understandable by humans, so the result obtained might actually be no explanation at all. If the analysis of Edwards and Veale [2018] were correct, then the urgency advocated by many researchers on

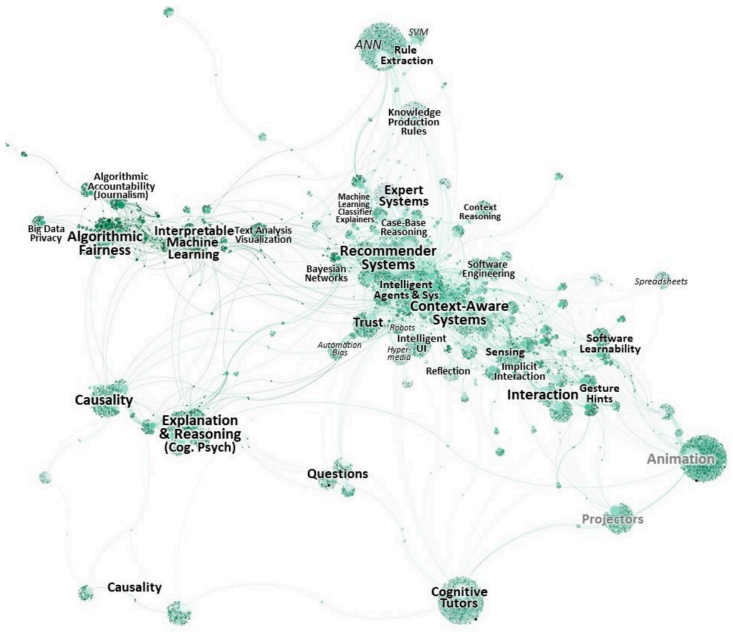



Figure 2.2. Citation network that is emblematic in showing the breadth of research strands in the field of xAI [Abdul et al., 2018].

the grounds of having to conform to the GDPR would, in reality, turn out to be based on no valid reason.

A second motive brought forward for the need for explainability, for example by Gilpin et al. [2018] and Abdul et al. [2018], is that comprehensible models are much more likely, or even necessary, to engender users' trust. While this may very probably be the case, no motives are given for why this should be the purpose of explainable AI and not just a desirable by-product of obtaining explainability. A reasoning for why trust may stem from explainability can be found in [Kyrimi and Marsh, 2016] where it is claimed that "the lack of trust may be due to the difficulty of understanding how a prediction is inferred from the given data. As Aristotle wrote 'we do not have knowledge of a thing until we have grasped its why, that is to say, its explanation'. Hence, explaining a model's reasoning - its inference - could increase trustworthiness." What could be assumed, is that *trust is a prerequisite for explainability* because if the human does not believe the outputs of the machine, there is little hope for an explanation to come into being.

Other authors, for example, Doshi-Velez and Kim [2017] and Guidotti et al. [2018], frame the issue as one of moral necessity. One need not look far to find examples of ML models displaying *covert bias* or making decisions we would regard as unethical; a more in-depth investigation would reveal that this was the case even before the popularisation of "black box" models as are deep neural networks. Guidotti et al. [2018] give a reasonably comprehensive list of classic cases that show the risks inherent in not having comprehensible AI. The oldest of these dates back to the 1970s and 1980s and tells of a system used to screen job applicants that was still seen to discriminate against minorities, even though programmed to ignore people's ethnicity. In the same vein but much more recently, the American Military discovered that their computer vision system, developed to differentiate between enemy and friendly tanks automat-

ically, had poor accuracy because it had learned to use the background information of the test set photos instead of the pixels representing the actual vehicles. Both these cases exemplify how an algorithm may make "wrong" inferences based on spurious or latent information that was already in the data set but that no human could have imagined being relevant. Other failures epitomise how a model may learn our own social prejudices; for example, a recent Princeton study [Caliskan et al., 2017] proved how models trained on web text corpora showed marked bias (towards race, gender ...) that reflected the ones present in our own society. Based on their findings, the authors went so far as to suggest that transparency would not be enough to uproot bias, since the very *semantics* of language reflects prejudice latent in our culture.

The driving motivation and sense of urgency present in all of the reviewed literature is quite certainly tied to the renewed interest and applicability of AI. Preece [2018] claims that the interest for explainability is naturally linked to that in AI; if this were true, then it would confirm that a need for transparency is implicit in the field itself. This would validate the assertion made by Doshi-Velez and Kim [2017], that the need for an explanation stems from an *incompleteness in the formalisation*. What is meant by this is that optimising for certain objectives may introduce an *unquantifiable bias* into the system, which is very different from mere *uncertainty*. Mere uncertainty can be rigorously quantified, formalised and reasoned upon by using probability theory; unquantifiable bias is the result of an *incompleteness in formalisation* of the problem that the ML system is tasked with. For example, this is the case when a system is coded to pursue soft objectives such as *ethics* as such an end-goal may be too abstract and nebulous to be completely formalised. It might also be the case that a particular objective is far too complex for all its possible outcomes to be exhaustively enumerated. A good characterisation of objectives leading to incompleteness could be given by applying the concept of *wicked problem* introduced by Rittel, Horst [1973] when analysing the issues arising in social planning. The first defining characteristic of a wicked problem is that the mere definition of the problem in hand is the wicked problem itself, since even its description is dependent on one's idea for solving it; there is no definite locus that one can point to as the source of the problem. Other defining features are the absence of *stopping and objective evaluation criteria* and the fact that each solution to a wicked problem is essentially "one-shot" and unique. Even the set of possible solutions and causes are not predetermined and non-stationary. The authors noticed how these problems presented a series of defining characteristics, that they were able to generalise into the notion of *wicked problem*. A classic example are the issues that arise when trying to solve a social planning problem: various parties will have competing objectives and different ideas to obtain them, so no unique best solution is possible; problems are "at best, only re-solved" [Rittel, Horst, 1973]. These situations generally tend to arise when dealing with human values, as there is always a degree of ethical relativism that makes it difficult to develop and evaluate solutions. Lipton [2016] is also aware of this and states that "the demand for interpretability arises when there is a mismatch between the formal objectives of supervised learning and the real world costs in a deployment setting". In the presence of such unquantifiable uncertainty, unavoidable in many of the important applications of AI, requiring the resulting model to be open to inspection would make the "gaps in problem formalization visible to us" and thus enable us to apply our best human judgement to evaluate them and their consequences [Doshi-Velez and Kim, 2017]. No wicked problem has a solution that is either true or false, but only good or bad; necessarily, the only judge for this can be a human.

## 2.4 Evaluation of Explainability

As concluded in Section 2.3, the fact that ML models operate on incomplete assumptions makes it a necessity to have some form of evaluation of their performance. As Lipton [2016] states, "it turns out that many situations arise when our real world objectives are difficult to encode as simple real-valued functions" and this could lead to evident difficulties in optimisation with respect to soft concepts such as ethics and legality which are, however, of paramount importance. Being able to evaluate an automated explanation lets us "serve those objectives that we deem necessary but struggle to model formally".

Unfortunately, the finding outlined in Section 2.2 of there being no consensus on the definition of explainability also necessarily entails that there is no agreed-upon methodology to evaluate such a property. Doshi-Velez and Kim [2017] note as much when they comment "unfortunately, there is little consensus on what interpretability in machine learning is and how to evaluate it for benchmarking". This makes perfect sense because trying to evaluate something without first having defined it, is worse than trying to hit a moving target. Once again, the feeling shared among many authors is that "inmates are running the asylum".

Some authors have tried to put some order in the barrage of proposed methods with Doshi-Velez and Kim [2017] providing one of the most compelling attempts. These authors set out to outline a taxonomy, having noted a "lack of rigour" and how current interpretability approaches usually fall into two categories: *interpretability in the context of an application and interpretability via a quantifiable proxy*. The former approach assumes that "if the system is useful in either a practical application or a simplified version of it, then it must be somehow interpretable"; the latter sees researchers claim that a model class is interpretable and then present algorithms to optimise within that class. In their words, both classes rely on a notion of "you'll know it when you see it". The taxonomy the authors propose is laid out in Figure 2.3 and borrows from methods already standard in human-computer interaction and visualisation; the guiding ideal is that "evaluation of applied work should demonstrate success in the application" and thus the best kind of evaluation is the one that involves humans the most:

- *Functionally-grounded Evaluations*: at the lowest level of their taxonomy are those methods requiring no human-in-the-loop and that evaluate the quality of an explanation given by a system by using some proxy measure; the advantage is the low cost, but the tradeoff is a lack of specificity. A proxy measure that has already been human-validated as regards its explainability, for example a decision tree, a set of rules or a linear model [Guidotti et al., 2018], may be substitute to estimate the explainability of more complex systems.
- *Human-grounded Evaluation*: the second level of evaluation involves humans, albeit not domain expert ones, carrying out simplified versions of the target application; this kind of setup enables the testing of more general notions of explainability.
- *Application-grounded evaluation*: this is considered the gold standard evaluation; the authors claim that there is no better way to evaluate explainability that having a domain expert test it in the context of a real task. In their words, "the best way to show that the model works is to evaluate it with respect to the task": "for example, a visualization for correcting segmentations from microscopy data would be evaluated via user studies on segmentation on the target image task; a homework-hint system is evaluated on whether the student achieves better post-test performance. Specifically, we evaluate the quality of an explanation in the context of its end-task, such as whether it results in better identification of errors, new facts, or less discrimination".

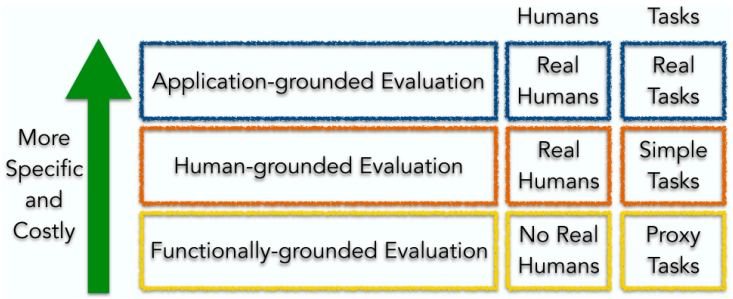



Figure 2.3. Taxonomy of methods for the evaluation of explanations [Doshi-Velez and Kim, 2017].

Guidotti et al. [2018] outline a taxonomy developed along a different axis; specifically they identify:

- *methods to explain black box models*;
- *methods to explain black box outcomes*;
- *methods to inspect black boxes and methods to design transparent boxes*.

An essential factor that the authors identify is that *evaluation is a graded notion*; as also noted by Gilpin et al. [2018]: different users, with different expertise and background, may rate the quality of the same explanations very differently. Another point they make - which is quite novel - is that *time should also be part of an evaluation*: depending on time available a user may prefer a more straightforward or more elaborate explanation. They end by noting that *very few studies take user expertise and the time taken to understand the proposed explanation into account*. The concept of *time to understand an explanation* will be part of the methods of this thesis.

The neglect of the human side of explanations is also lamented by Abdul et al. [2018] who see researchers focusing on creating *mathematically explainable models* at the expense of ones that are usable and practical in real-world situations. Again, the underlying issue seems to be that xAI researchers are either unaware or are ignoring the sizeable corpus of research available in cognitive psychology, human-computer interfaces and philosophy. The authors see the opportunity for HCI to bridge the gap between models and users by way of an interactive approach, as opposed to the mainstream static explanations being proposed in the literature. An interactive explanation may take the form of a dialogue or of various visualisation techniques; the defining characteristic of such an evaluation modality is that it lets the user freely explore the system's behaviour. Guidotti et al. [2018], while discussing the types of data used in ML models, lend credence to the adequacy of this output modality via the statement that "other forms of data which are very common in daily human life are images and texts. They are perhaps for the human brain even more easily understandable than tables".

Guidotti et al. [2018], though, take the view that evaluation of model *comprehensibility* should be equated to its *complexity*, which is not an opinion that often appears in the reviewed literature. Basically, the authors are advocating for the use of complexity - the number of identifiable elements in a particular class of model, for example the number of weights in a neural network or of rules in an expert system - as a proxy measure for explainability, if we are framing

the issue using the taxonomy proposed by Doshi-Velez and Kim [2017] (see Figure 2.3). This may very well be a valid approach, but there is no supporting evidence for it in the paper itself.

A useful reference for how to set up an experiment falling into the class of either *human-grounded or application-grounded evaluation*, can be found in Stumpf et al. [2009]. In this work the authors set up "three experiments to understand the potential for rich interactions between users and machine learning systems"; the first, and most relevant, was a *think-aloud study* that investigated "how machine learning systems should explain themselves to end users, and what kinds of improvement feedback end users might give to the machine learning systems". These studies are interesting as a blueprint for future human-centred evaluations, of the type whose absence is being lamented by many authors.

Mittelstadt et al. [2019] summarise the existing critiques to offer a clear and direct evaluation of the field of explainable AI as a whole, when they state that:

> no matter the approach taken in xAI, reflexivity *[taking account of itself or of the effect of the personality or presence of the researcher on what is being investigated, clarification not by authors]* is needed to ensure the community actually works towards its normative and practical goals to render models holistically transparent or provide high-quality post-hoc interpretations of model behaviour. Critical questions must be repeatedly asked and answered. For example, will the methods developed make machine learning models more interpretable? More trustworthy to users? More accountable? And to whom will explanations be accessible, comprehensible, and useful? Answering such questions requires considering the methods developed in xAI in the context of prior work in fields addressing such normative and social questions. Local and approximation models may in fact resemble existing, well-known approaches to explanations in the 'explanation sciences', which would provide insight.
>
> Mittelstadt et al. [2019, pag. 3]

They then conclude by stating that "xAI generally avoids the challenges of testing and validating approximation models, or fully characterising their domain". From the review of the current state-of-the-art carried out in this section and Sections 2.2 and 2.3, these could both be seen as entirely valid criticisms. It really seems that the field of xAI as a whole should try and reposition itself as suggested in Figure 2.1 and not attempt to build methods from first principles, many of which may be outside the domain of expertise of the researchers proposing them.

## 2.5 Explainability in Bayesian networks

Bayesian networks are a popular class of probabilistic models which has enjoyed widespread appeal as a machine learning method, especially in the field of medicine. The classic "Asia" toy example of a BN is shown in Figure 2.4; this simple BN is composed of eight nodes, arranged in a parent → child relationship; the visualisation makes clear how each one is associated with a *probability distribution*, whose values only depend on the parent nodes (Definition 3.4). The popularity of BNs in the area of medicine may be because the formalism (see Section 3.5) "offers a natural way to represent the uncertainties involved in medicine when dealing with diagnosis, treatment selection, planning, and prediction of prognosis. This is due to the fact that the influences and probabilistic interactions among variables can be described readily in a

BN" [Lucas, 2001]; that is, even if the BN model is *complete*, in the sense that every possible probabilistic statement can be computed in it, it is also easy to combine multiple variables of interest into composite statements. Thus, unlike other popular ML algorithms which may have higher learning performance, BNs enable *reasoning* on the model (for example by using the algorithms presented in Subsection 3.5.4). Another attractive feature of BNs is their relatedness to the class of *causal networks*, popularised by the groundbreaking work of Pearl and Dechter [1988]; for all intents and purposes, a causal network is simply a Bayesian network where all the relationships represent a causal effect. Nonetheless, *BNs are not considered as inherently interpretable by the literature*, and thus, a series of methods were developed to address this shortcoming. Timmer et al. [2015] note this in the introduction to their paper by stating that "for non-statistical experts, however, Bayesian networks may be hard to interpret. Especially since the inner workings of Bayesian networks are complicated they may appear as black box models", "the interpretation of BNs is a difficult task, especially for domain experts who are not trained in probabilistic reasoning".

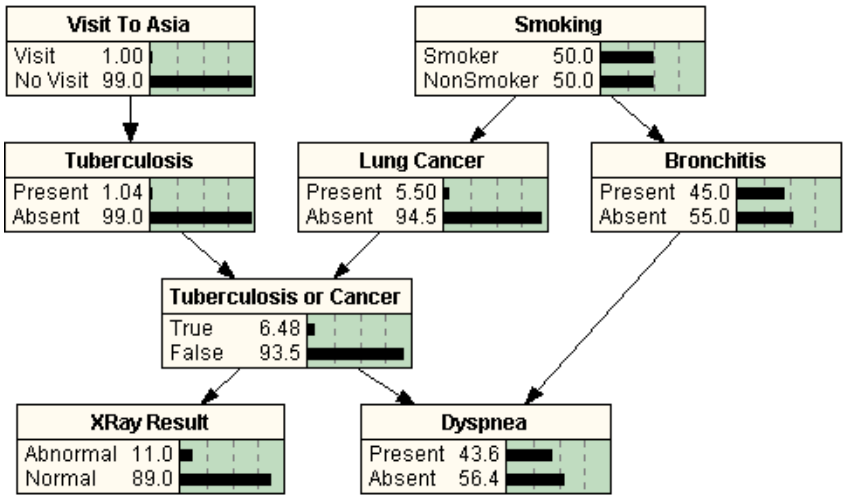



Figure 2.4. Classic example of Bayesian network [Norsys Software Corp.].

The best overview of the state of explainability of BNs, that will also be used as a framework for the methods developed in this thesis, is given by Lacave and Díez [2002] in the paper "A Review of Explanation Methods for Bayesian Networks"; here the authors identify various classification criteria for an explanation given by a BN:

- *description* vs. *comprehension*: the former consists in displaying the data set or providing further details regarding the output, the latter attempts to guide the user in understanding the model's conclusions.
- *micro-level* vs. *macro-level*: detailed description of how a single node in the network is affected vs. showing the main lines of reasoning.
- *linguistic* vs. *graphical*: "the most direct and intuitive way of showing the information embodied in a Bayesian network is to display the corresponding graph". When presenting probabilities, Henrion and Druzdzel [1990] strongly suggest that these be "linguistic probabilities" i.e., for the quantitative probabilities inherent in the model to be converted

to a qualitative equivalent. This is validated by research showing that linguistic expressions of probability are better understood than the equivalent numerical representation. Some of the many models surveyed in the paper use colours, shading and line thickness to represent the salience of links and nodes. The work carried out in this thesis will extensively use both linguistic and graphical explanations; the probabilities will also always be linguistic.

When compared with the taxonomies already presented in Section 2.2, it is obvious that these methods applied to a BN, would make it a *comprehensible system* or *post-hoc explainable* because the model would be emitting extra symbols (graphical, linguistic ...) geared towards explaining its outputs.

The authors also identify the three components of BNs that need to be explained: the *evidence*, the *model* and the *reasoning*. The first of these "consists of determining which values of the unobserved variables justify the available evidence" and is, in general, done by finding the solution to the most probable explanation problem (Definition 3.26). The explanation of the model is considered a *static* explanation (as opposed to *dynamic* ones, shortly be covered) and is simply the process of linguistically or graphically displaying the information already present in the data. The final element which needs explaining is what would most commonly be called an explanation in xAI circles, is the reasoning behind the model's outputs; a system may accomplish this by providing a justification for its outputs, for the results it did not give or via hypothetical reasoning. The last of these is maybe the most important, because it is paramount for any system, not just a BN, to be able to explain the reasons behind its outputs; returning to the medical setting, it was seen that physicians, in particular, are very reluctant to accept the advice of a machine if they cannot understand how it was obtained. A BN, unlike other ML systems, can also innately exhibit evidence for why it did not provide the output expected by the user and can also reason *counterfactually* i.e., provide alternative lines of reasoning. This will also be an important part of the work carried out in this thesis, as "counterfactuals"/"what-if analysis" will be used to help medical experts extract knowledge from data. These last two capabilities of Bayesian networks are particularly important, from an explainability perspective, in the light of the findings by Miller [2018] regarding the nature of explanations from a psychological perspective. The conclusions are that *explanations possess four primary characteristics*:

- Explanations are *contrastive*: that is people do not ask why an event happened but why another event did not happen instead. A Bayesian network, as noted in the previous paragraph, is capable of modelling counterfactuals which enables it to naturally give contrastive reasonings.

- Explanations are *selected*: people expect the explanation given to them to have been selected based on some criterion or cognitive bias; they do not expect a complete recount of all causes of an event. A BN has the ability to flexibly combine its constituent variables into an output and thus its explanations can be selected based on some criteria or be *partial*, for added simplicity; a BN's outputs needn't be *complete* i.e., constituted of all its parameters, unlike those of non-local models (for example, neural networks).

- To people, probabilities are not as important and not as well understood as *causal* relationships. BNs, as already mentioned, are closely related to *graphical causal models*, so their explanations have the possibility of being based on causal grounds. [Lipton, 2016], Rani et al. [2006]

- Explanations are *social*: that is they involve an *explainer* and an *explainee*. This is recognised in “Conversational Processes and Causal Explanation” by Hilton [1990], the most important work on the social aspects of conversation, who supports the view that every *explanation* is a *conversation*. A dialogue is an example of a *dynamical explanation*, in the framework set out by Lacave and Díez [2002]; these authors also recognise that an explanation “always means explaining something to somebody” and thus that “one of the key features of an effective explanation is the ability to address each user’s specific needs and expectations, which primarily depends on the knowledge he/she has”. So “In the case of a Bayesian network, the explanation generated for a user that is familiar with the concepts of prevalence, prior/posterior odds and likelihood ratios should be very different from the explanation generated for a user who has never heard about them”. Lacave and Díez [2002] again recognise that explainability is a graded notion but go further to note that practically all explainable BN systems have made the assumption of a *fixed user model* thus ignoring the possibility of users having varying levels of knowledge. However, some of the systems surveyed in the review make a step in that direction by incorporating an *importance threshold* mechanism that would let the user only display certain items; this enables these systems to display varying levels of detail without having defined a particular user model.

A *dialogical* explanation could probably make a BN an example of *explainable system*, in the framework developed by Doran et al. [2018] and *post-hoc explainable system* in that of Mittelstadt et al. [2019].

Bayesian networks, without any additional explainability methods, would most probably fall into the class of *interpretable systems*, as do many other ML models, in the taxonomy set forth by Doran et al. [2018]. Though, based on this review of the literature, it could be justifiably suggested that Bayesian networks are better equipped than other machine learning models to provide a meaningful explanation to humans. This could be claimed because it is quite widely believed that our brains, and thus our psychology, are near-optimal problem-solvers and as such approximate optimal Bayesian solutions. A standard view in the fields of psychology and neuroscience, as noted by Bowers and Davis [2012], is that our brain processes approximate the *rational player* as presented in the Dutch Book Argument (see Ch.7 of Anand et al. [2009]) and are thus Bayesian in nature. It is also worth noting that this view has recently been challenged, for example by Bowers and Davis [2012]. This said, even if our brains were not inherently Bayesian, the characteristics of Bayesian networks make them more capable than other ML models in being able to generate explanations tailored to our cognitive biases and psychology, as discussed when presenting the characteristics of an explanation as identified by Miller [2018].

## 2.6 “Explaining the Most Probable Explanation”

The paper “Explaining the Most Probable Explanation” by Butz et al. [2018] places itself in the literature concerned with the explainability of Bayesian networks. In particular, taking the classification proposed by Lacave and Díez [2002] presented in Section 2.5, it attempts to define a *linguistic explanation* of the *evidence* and of the *reasoning*. It differs from the previous attempt to define the explanation of the *evidence* given by Lacave and Díez [2002] and in other works, in that the paper is not concerned with finding the most probable assignment of variables that would explain the given evidence but, rather, the inverse problem. By starting with the evidence and finding a maximally probable configuration, the authors hope “to look at the

complete scenario to get an overview before deciding which variables should be focused on”; i.e., the goal appears to be to give the user an overview of the situation.

The initial claim of the paper is that BNs are still difficult to interpret for domain experts, even though these models provide a graphical structure to the *knowledge base*. The examples brought to justify the claim are that edges in the graph do not necessarily represent causal dependencies and that d-separation (Definition 3.22) may be confusing. The authors plan to address this claim by constructing a *dialogue* with the user and thus to continue in the long tradition of dialogical approaches to explaining BNs, many of which are presented in [Lacave and Díez, 2002].

The defining characteristic of their approach is that the domain expert is able to “argue” with the MPE and investigate alternative explanations. The complete methodology, executed over three steps, is shown in Figure 2.5. The first step is the construction of the “knowledge base”, which is nothing else than a probability tree representing a “chain of deduction” constructed following the strongest probabilistic dependencies between variables in the BN. Such a *knowledge base* is convenient because the document plan for the Natural Language Generation step is directly derived from it. One issue that is immediately apparent is that this greedy approach does not “generate the MPE solution” as the authors claim. This does not discredit the argumentative method as a whole, as *it is not necessary for the user to be arguing the MPE to derive a good explanation*; this ties into one of the main findings in the previous sections that many xAI researchers are only focusing on one half of the explanation. A good explanation is not given only by its formal properties but, most importantly, by how well it acts as an interface between the real *user* and the model. This is what Abdul et al. [2018] mean when they lament that “despite their mathematical rigour, these works *[referring to the existing explainability methods]* suffer from a lack of usability, practical interpretability and efficacy on real users”.

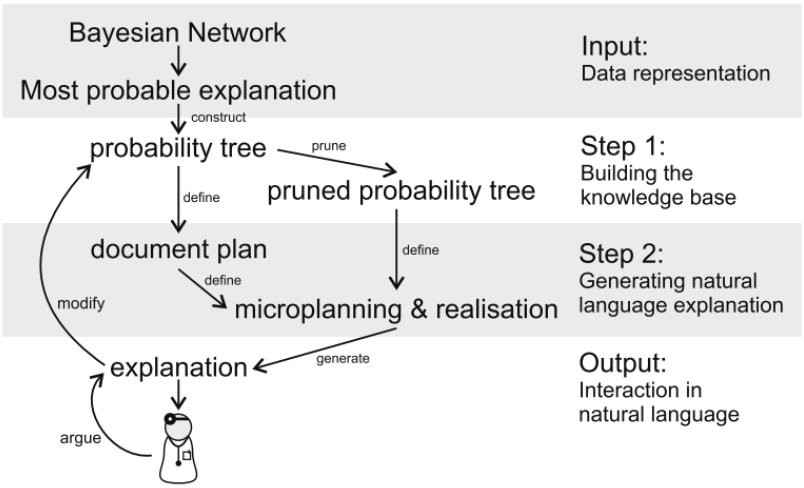



Figure 2.5. Overview of methodology followed by Butz et al. [2018].

The document plan for the argumentation follows the same chain of strongest dependencies constructed in the *knowledge base* until the expert disagrees; at that point, the user is presented with an alternative “MPE”. An example of how the document plan may look after interaction

with the user is shown in Figure 2.6. All the natural language phrasing is generated via boilerplates that take care of realising both the micro-planning phase and the generation of the text.

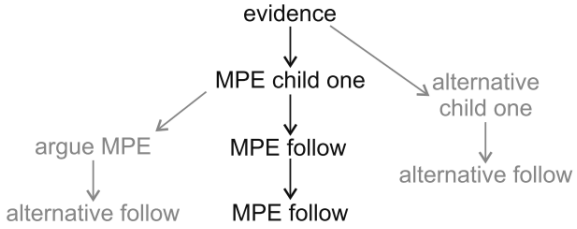



Figure 2.6. Document plan generated from the probability tree [Butz et al., 2018].

The authors recognise that such chains of deduction could become long and cognitively overloading in the case of larger BNs, as every variable in the tree is explained by all its ancestors. A solution they propose is that of *pruning* the probability tree by excluding d-separated nodes and those under a certain threshold of significance. They also adapt some methods from literature to perform *conflict analysis* i.e., only variables that contribute positively to the explanation are maintained in the document plan.

On the whole, Butz et al. [2018] offer a compelling explanation method for BNs by building on an established tradition of enabling explainability through dialogue. The work, though, takes some methodological missteps and also continues the "sin" of not validating its claims on real users, which is one of the primary gaps in the xAI field, as identified in the previous sections.

## 2.7 Summary

The findings outlined in this chapter refer to the concept of explainability of machine learning models, to its importance, to ways of evaluating it and to explainability in the specific case of Bayesian networks. The concept of explainability is central to the field of explainable AI, whose goal is to make machine learning systems "interpretable" or "explainable". *Explainability* can briefly be defined as the property of a system that is able to "explain or to present in understandable terms to a human" [Dosilovic et al., 2018] its outputs. Two main classes of explainable models have been identified by Mittelstadt et al. [2019]: *ante-hoc or transparent* and *post-hoc interpretable* ones; the former are inherently inspectable in their inner workings while the latter are made understandable by way of extra techniques. These two classes have also been refined by Doshi-Velez and Kim [2017] into a four-tier taxonomy consisting of: *opaque*, *interpretable*, *comprehensible* and *explainable systems*. An opaque system, also know as a "black-box model", is one whose inner workings are not inspectable from the outside; an interpretable system corresponds to an ante-hoc interpretable one; a comprehensible one emits extra information together with its output; an explainable system explicitly outputs a human-understandable line of reasoning aimed at clarifying its workings.

Explainability has become a central concept to the field of AI as a whole; as ML models take over more and more functions in our societies, the pressure for them to be able to explain their decisions is increased accordingly. The *General Data Protection Regulation* that became effective

in 2018 was viewed by many as increasing the societal pressure to make systems explainable; many may have been mistaken as regards the actual rules mandated by the regulation - that aren't really prescribing a broad "right to an explanation" [Edwards and Veale, 2018] - but nonetheless the feeling of urgency is sure to increase the focus of both researchers and laypeople. In general, explainability is framed as an issue of moral necessity as it easy to find a long series of situations where ML models displayed covert bias or what we would regard as bad moral judgement.

There are all manner of ways to measure the quality of an explaination and these can be classified into a three-layer taxonomy [Doshi-Velez and Kim, 2017] based on the assumption that the best type of evaluation is the one that most involves humans. The three classes are *functionally-grounded evaluation*, *human-grounded evaluation* and *application-grounded evaluation*, ordered from the one least involving real humans to the one where the presence of the human-in-the-loop is greatest. As the involvement of humans in evaluating models' explanations increases, so does the cost of such an experiment and its specificity, as the highest evaluation level necessarily entails the collaboration of domain-experts on specific tasks. A parallel taxonomy identifies: methods to explain black box models, methods to explain black box outcomes, methods to inspect black boxes and methods to design transparent boxes. An overarching notion that has been stressed is that *explainability is a graded notion* that depends on the knowledge and expertise of the particular user: different users, with different expertises and backgrounds, may rate the quality of a same explanation very differently.

Bayesian networks (BNs) have enjoyed widespread appeal in mission-critical domains like that of medicine and thus the drive to develop methods to explain their outputs has always been strong. BNs have three main elements that necessitate an explanation: the *evidence*, the *model* and the *reasoning*. [Lacave and Díez, 2002] Explaining the first "consists of determining which values of the unobserved variables justify the available evidence" and is done by solving the most probable explanation problem. For the second, a static explanation of the BN is achieved by displaying it graphically or linguistically. The last element is explained by showing the reasoning that brought the BN to give the outputs it did and can be achieved by providing a justification for its outputs, for the results it did not give or via hypothetical reasoning. The fact that BNs are able to naturally support counterfactual reasoning, combine single variables into composite outputs and model causality puts them at an advantage compared to other ML systems when generating an effective explanation for a user. This is because the capabilities of a BN enable it to generate explanations that are uniquely suited to our psychological biases and expectations of what an explanation should entail.

Some of the main gaps that were found during the review of the literature relate to how there is still a great confusion in the field of xAI regarding what an explanation really is and thus what would constitute a good instance of it. There is also a prevalent methodological confusion, as different authors use terms in incongruous ways, for example sometimes *intepretation* is taken to mean *explanation* while in other cases they refer to different concepts, for example in the taxonomy of interpretable systems proposed by Doshi-Velez and Kim [2017]. This naturally makes it difficult for the field to converge onto methods to evaluate such explanations and this is reflected in the barrage of methods present in the literature, each one focused only on a particular system or model. This confusion is exacerbated by a seeming lack of interest or awareness of xAI researchers for the sizeable corpus of work in psychology, philosophy, social sciences, neuroscience and human computer interaction that has already investigated the nature of explanations, what desiderata they may possess and which are most effective. In fact, the great majority of proposed approaches is only focused on proving formal explainabil-

ity and neglects the human side that is naturally present in any explanation; there has been little work carried out to validate approaches in real settings with real domain experts so many explainability methods are substantiated only at theoretical level. The underlying issue that has been seen running transversely across the various concepts investigated in this chapter is best summarised by the idea that "inmates are running the asylum", meaning that individual researchers are claiming that their models are interpretable referring only to their own personal views and biases and not to established literature and methods. It would be hard for them to do otherwise, as the field of xAI seems, at present, to be a collection of diverging strands without a comprehensive program able to help it converge onto its stated goal: to make machine learning systems understandable by their users and thus increase their social utility and acceptance.

# Chapter 3

# Mathematical Background

## 3.1 Introduction

This chapter will introduce and build up to a formal definition of Bayesian networks, a class of *probabilistic graphical models* used to represent systems under conditions of uncertainty.

The chapter is organised as follows:

- Section 3.2 introduces a series of basic concepts from probability theory focusing mainly on the basic concept of random variables and also establishing the notions of conditioning, independency and correlation.
- Section 3.3 presents information entropy and uses it to define entropy measures for random variables, *Kullback-Leibler divergence*, and distance measures for other objects, *Hamming and Jaccard distances*.
- Section 3.4 introduces the objects of graphs and polytrees and the central concept of *d-separation*.
- Section 3.5 uses the content of the previous sections to define the Bayesian network formalism and then gives an overview of structure learning algorithms and the notions of *conditional probability* and *maximum a posteriori queries*.

Not all the concepts introduced in this chapter are strictly needed for the description of the Bayesian network formalism, but all will be useful as a mathematical reference for the methods developed in later chapters of this thesis.

## 3.2 Probability Theory

We will be dealing with *standard probability* so random variables and probability measures will always be real-valued. We will also, in the work carried out in this thesis, only be considering the case of random variables that can assume a finite number of possible values/states. We will refer to these variables as *categorical* to indicate that there is no natural ordering among their states.

### 3.2.1 Random Variables

**Definition 3.1 (Event)** *Given $\mathscr{S}$ the space of all possible outcomes of interest, an event $\sigma$ is a subset of $\mathscr{S}$: $\sigma \subseteq \mathscr{S}$.*

*$\mathscr{F} \subseteq 2^{\mathscr{S}}$ is the set of all events that are under consideration.*

*Two events $\sigma$ and $\tau$ are called disjoint when $\sigma \cap \tau = \emptyset$.*

**Definition 3.2 (Random Variable)** *A random variable $X$ is a function $X : \mathscr{S} \to \mathscr{X} \subseteq \mathbb{R}$ that associates every outcome $s \in \mathscr{S}$ with a value.*

*Random variables* are a way of bringing to the fore the attributes of interest of events while dealing with them in a clean, mathematical way. The values that a random variable can take are a function of the events in sample space $\mathscr{S}$, with each of these having a value assigned by the random variable function.

**Definition 3.3 (Probability Measure)** *Given a sample space $\mathscr{S}$ and events $\mathscr{F}$, a discrete probability measure $\mathbb{P}$ is a function $\mathbb{P} : \mathscr{F} \to [0,1]$ that assigns a probability value to every event. In the discrete case all subsets of $\mathscr{S}$ can be treated as events thus $\mathscr{F}$ is the power set of $\mathscr{S}$. To be a valid probability measure, $\mathbb{P}$ must satisfy:*

- $\mathbb{P}(\mathscr{S}) = 1$;
- *If events $\sigma$ and $\tau$ are disjoint then $\mathbb{P}(\sigma \cup \tau) = \mathbb{P}(\sigma) + \mathbb{P}(\tau)$.*

Each event $\sigma \in \mathscr{F}$ must have a probability $\mathbb{P}(\sigma) \in [0,1]$ and the sum of all these must equal 1. An event with $\mathbb{P}(\sigma) = 0$ is deemed *impossible* while one with $\mathbb{P}(\sigma) = 1$ is *certain*.

**Definition 3.4 (Probability Mass Function)** *A probability mass function of a discrete random variable $X$ is a function $f_X : \mathscr{X} \to [0,1]$ defined, using a probability measure $\mathbb{P}$, as:*

$$f_X(x) = \mathbb{P}(\{s \in \mathscr{S} : X(s) = x\}),$$

*and thus assigns a probability to every value $x \in \mathscr{X}$ in the domain of $X$.*

The probability mass function returns the probability of a random variable $X$ taking on exactly its value $x$. This probability is the size of the subset of the event space $\mathscr{S}$ whose events $s$ are mapped to $x$ by the random variable function $X$.

Every random variable has a probability distribution induced by the cardinality of the subsets of its values; in the case of discrete one, such a distribution is *multinomial*.

Often, in the context of random variables the probability distribution $f_X$ is called the *marginal of $X$* and is usually contrasted with the notion of *joint probability distribution*.

**Definition 3.5 (Joint Probability Mass Function)** *The joint probability mass function of discrete random variables $X$ and $Y$ is a function $f_{XY} : \mathscr{X} \times \mathscr{Y} \to [0,1]$ defined as:*

$$f_{XY}(x,y) = \mathbb{P}(\{s \in \mathscr{S} : X(s) = x\} \cap \{s \in \mathscr{S} : Y(s) = y\}),$$

*and thus assigns a probability to every tuple $(x,y)$ with $x \in \mathscr{X}$, $y \in \mathscr{Y}$.*

In what follows, we will sometimes refer to the marginal probability $f_X$ of $X$ as $\mathbb{P}(X)$, to $f_X(x)$ as $\mathbb{P}(X = x)$, to joint probability $f_{XY}$ of $X$ and $Y$ as $\mathbb{P}(X, Y)$, and to $f_{XY}(x, y)$ as $\mathbb{P}(X = x, Y = y)$ as the only random variables we will be dealing with will be discrete. Notice that the notation $(E = e)$ is also often overloaded to signify an assignment of values to a set of random variables $E$; in this case what is meant is that every variable in the set $E = X_1, ..., X_k$ assumes a certain value from its own domain. We will denote sets in bold so $\boldsymbol{E} = \boldsymbol{e}$ stands to mean that every variable $E$ in the set of random variables $\boldsymbol{E}$ assumes a value $e$ from its own domain. Finally, recall that the set of values that $X$ can take is denoted by the cursive $\mathscr{X}$.

### 3.2.2 Probability interpretations

There are two main views through which to interpret the probability of an event: the *frequentist* and the *subjectivist/Bayesian* one.

The former views the probability of an event as the expected ratio of times it would occur over a great number of trials. That is, the probability of an event is seen as the *limiting frequency* of a repeatable event. So, for example, the probability of observing heads when tossing a coin is said to be 0.5 because over repeated throws heads was observed half the time.

The other view is the subjectivist or *Bayesian* one (from the 18th century mathematician Thomas Bayes) in which probabilities are instead viewed as the *subjective* degree of belief attributable to the manifestation of an event. In this interpretation, stating that a coin has probability of heads of 0.5 simply means that the person making the claim personally believes that the chances of seeing heads or tails are the same. This is useful in that it enables the characterisation of certain events that haven't yet come about or that are liable to happen only once or a small number of times (that is, they are not repeatable).

Philosophically, Bayesian inference assigns a probability to a hypothesis (a *prior*) while the frequentist method tests a raw hypothesis empirically before assigning it any probability. As Bayesian inference naturally embraces and deals with uncertainty, it is an enormously useful tool to model and reason about the real, stochastic world we live in.

From the Bayesian point of view, we would consider the probability of a state of a random variable as simply representing the subjective degree of belief we would have over a set of outcomes we believed could manifest themselves.

### 3.2.3 Conditional Probabilities

**Definition 3.6 (Conditional Probability)** *The conditional probability mass function of random variable $Y$ given $X = x$, $x \in \mathscr{X}$, $y \in \mathscr{Y}$ is:*

$$\mathbb{P}(Y = y \mid X = x) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(X = x)}.$$

*To be defined, it must be that $\mathbb{P}(X = x) > 0$.*

Definition 3.6 can easily be manipulated in order to obtain another basic result, called the *chain rule of conditional probabilities*:

$$\mathbb{P}(Y = y, X = x) = \mathbb{P}(Y = y \mid X = x)\mathbb{P}(X = x). \tag{3.1}$$

Equation 3.1 can be generalised to any number of variables:

$$\begin{aligned}\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = &\mathbb{P}(X_n = x_n \mid X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) \times \\ &\ldots \\ &\times \mathbb{P}(X_1 = x_1 \mid X_2 = x_2) \times \\ &\times \mathbb{P}(X_1 = x_1)\,.\end{aligned} \tag{3.2}$$

Intuitively, it means that we can decompose joint probabilities as products of conditional probabilities.

Another immediate, and crucial, consequence of Definition 3.6 is known as *Bayes' Theorem*, which lets us calculate the revised probability of an event given new knowledge regarding another event.

**Theorem 3.7 (Bayes' Theorem)** *Given random variables $X$, $Y$ and the events $X = x$, $Y = y$, $\mathbb{P}(Y = y) > 0$, it holds that:*

$$\mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(Y = y \mid X = x)\mathbb{P}(X = x)}{\mathbb{P}(Y = y)}\,.$$

Intuitively, this is a process of *belief revision* as the belief in event $X = x$ is revised by the new knowledge that $Y = y$.

### 3.2.4 Independence

**Definition 3.8 (Random Variables Independence)** *Two random variables $X$ and $Y$ with domains $\mathscr{X}$ and $\mathscr{Y}$ are independent when their joint probability mass $\mathbb{P}(X, Y)$ is equal to the product of their probability densities:*

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \times \mathbb{P}(Y = y) \quad \forall x \in \mathscr{X}, \forall y \in \mathscr{Y}\,.$$

In the real world it is hard or even impossible - if we consider Nature being based on chaos theory when viewed at a fine-enough level - to find two such perfectly non-interacting events. Thus, a more useful concept is that of *conditional independence* where two previously dependent events become independent when conditioned on a third one

**Definition 3.9 (Random Variables Conditional Independence)** *Two random variables $X$ and $Y$ with domains $\mathscr{X}$ and $\mathscr{Y}$ are conditionally independent on a third random variable $Z$ with domain $\mathscr{Z}$ when their probability densities conditioned on $Z$ are independent. That is, when the joint probability mass function conditioned on $Z$ is equal to the product of the conditional probability mass functions:*

$$\mathbb{P}(X = x, Y = y \mid Z = z) = \mathbb{P}(X = x \mid Z = z) \times \mathbb{P}(Y = y \mid Z = z) \quad \forall x \in \mathscr{X}, \forall y \in \mathscr{Y}, \forall z \in \mathscr{Z}\,.$$

Intuitively, this means that knowing any value of $Z$ makes the probability distributions of $X$ and $Y$ independent.

## 3.3 Information Theory

The birth of the field of *information theory* is usually traced back to the seminal paper "A Mathematical Theory of Communication" [Shannon et al., 1949] where the mathematical basis for the quantification of the amount of *information* transmissible over a noisy channel was set. In the authors' words, "the fundamental problem of communication is that of reproducing at one point, either exactly or approximately, a message selected at another point". The concepts of field are broad enough to have influenced practically every other scientific discipline and deep enough to have made the "digital age" possible, for example by enabling the creation of ever more complicated coding schemes for the compression, reconstruction and obfuscation of digital data.

### 3.3.1 Entropy

In classical mechanical statistics, *entropy* can be seen as a measure of the uncertainty, or randomness, of a physical system. This concept was reapplied by Shannon et al. [1949] to measure the amount of randomness, the converse of information, in a random variable.

**Definition 3.10 (Entropy)** *The entropy $H(X)$ is defined as the expected amount of information content carried by random variable X [Schneider, 2005]:*

$$H(X) = -\sum_{x \in \mathcal{X}} \mathbb{P}(X = x) \log_b \mathbb{P}(X = x).$$

The base $b$ of the logarithm defines the unit of measure. Shannon et al. [1949] used $b = 2$ as they were dealing with the transmission of digital, binary-coded data; in this case the unit of measure would be *bits*.

The entropy of a generic random variable's probability mass function is a *unimodal functional*[1] whose domain is the subset $[0, 1]$. Its maximum 1 is reached when applied to a uniform probability mass function while its minima appear in the presence of *degenerate probability mass distributions* i.e., those that are localised at a single value meaning that all values have probability zero, except one that is certain.

As always, we can define a *conditional version* of the notion:

**Definition 3.11 (Conditional Entropy)** *The conditional entropy $H(Y \mid X)$ is defined as:*

$$H(Y \mid X) = -\sum_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}}} \mathbb{P}(X = x, Y = y) \log_b \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(X = x)}.$$

The simplest example of how information entropy characterises a random variable $X$, is in imagining $X$ as modelling a coin and being tasked with predicting the probability of the outcome of a throw being *heads*. If the coin is fair ($\mathbb{P}(X = 0.5)$), we will not be any more "surprised" to see that the outcome is heads instead of tails: the entropy is maximum, as there is maximum uncertainty regarding the outcome. However, if the coin is not fair and tails is more probable than heads ($\mathbb{P}(X < 0.5)$), then we will be more surprised if we see that the outcome is heads: the entropy is sub-maximal because there is less uncertainty regarding the outcome, due to the fact that tails is more probable than heads. If one of the outcomes is impossible, for example if

[1] A function whose codomain is in the space of functions is called a functional.

the coin has two heads ($\mathbb{P}(X = 1)$), then the entropy of the coin is 0, as there is no uncertainty regarding the result of a toss. All these cases, and they converse, can easily be understood by looking at how the entropy $H(X)$ varies as a function of the probability of the outcomes in the graph in Figure 3.1.

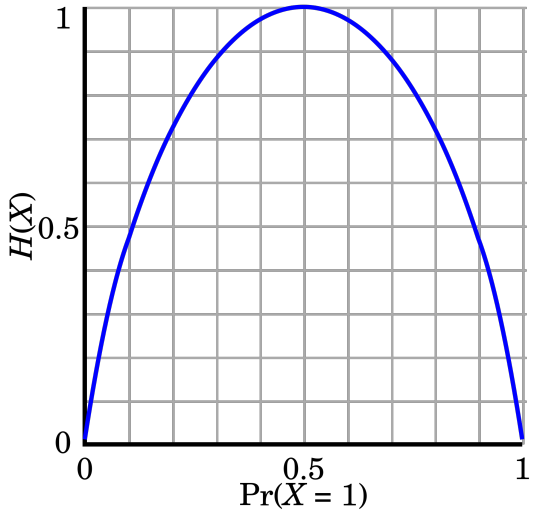



Figure 3.1. Entropy of the probability mass function over a Boolean variable as a function of the probability of the true state.

Plain entropy might be not the appropriate tool when trying to characterise random variables with codomains of different cardinality. Let us suppose that the objective is to find the variable with the least "entropic" distribution and imagine that their values have all been generated by the same process, say Gaussian. Simply calculating their entropies and ordering them according to this criterion will bias the selection process towards the variables with smallest cardinality. This is because we supposed them to be homoscedastic so there will naturally be less uncertainty when there are fewer possible outcomes. This can easily be understood by imagining the distributions to all be random uniform, as then each outcome of the smaller cardinality one will have a higher probability than those in the other ones (this topic is further analysed in Subsection 4.4.3).

To obviate this problem we need to *normalise* the entropy so that different-sized variables can be directly compared with each other. To achieve this, we can look at a measure of *normalised entropy*.

**Definition 3.12 (Normalised Entropy)** *The normalised entropy - or efficiency - of random variable $X$ with domain $\mathcal{X}$, given $n$ the number of values of $X$, is given by:*

$$H(X) = -\sum_{x \in \mathcal{X}} \frac{\mathbb{P}(X = x) \log_b \mathbb{P}(X = x)}{\log_b(n)}.$$

From Definition 3.12 it can be seen that $H(X)$ always takes values in the range $[0, 1]$; it is thus an unbiased measure that is comparable across variables of different cardinality. This ratio expresses the amount of entropy found in the distribution compared to the maximum possible

entropy for a variable of cardinality $n$, the case corresponding to the uniform distribution, as shown in Equation 3.3:

$$H\left(\underbrace{\frac{1}{n},\ldots,\frac{1}{n}}_{n}\right) = -\sum_{i=1}^{n}\frac{1}{n}\log_b\left(\frac{1}{n}\right) = -n\cdot\frac{1}{n}\log_b\left(\frac{1}{n}\right) = -\log_b\left(\frac{1}{n}\right) = \log_b(n). \tag{3.3}$$

### 3.3.2 Mutual Information

Another way, closely linked to entropy (Definition 3.10), of characterising the interrelatedness of two variables is through the concept of *mutual information* as defined in Cover and Thomas [2006]:

**Definition 3.13 (Mutual Information)** *The mutual information of two random variables $X$ and $Y$ with domains $\mathcal{X}$ and $\mathcal{Y}$ is given by:*

$$I(X,Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\mathbb{P}(X=x,Y=y)\log_b\left(\frac{\mathbb{P}(X=x,Y=y)}{\mathbb{P}(X=x)\mathbb{P}(Y=y)}\right).$$

NB: In Information Theory, the convention is that $0\log(0) = 0$.

$I(X,Y)$, intuitively, measures the amount of information that $X$ and $Y$ share and can also be seen as the degree to which one variable is informative of the other. If $X$ and $Y$ are independent then they share no mutual information and knowing something about one of the two gives no new information about the other. This can be immediately understood by rewriting Definition 3.13 as:

$$I(X,Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) \tag{3.4}$$

The mutual information I(X,Y) is the reduction in uncertainty of one of the variables given the knowledge of the other. If $X$ and $Y$ are perfectly correlated (correlation $\rho(X,Y)$, as defined by Sto [2006], is a measure of the degree to which two random variables are linearly dependent) then they both convey the same amount of information and $I(X,Y)$ is equal to the entropy i.e., $H(X) = H(Y)$.

### 3.3.3 Kullback-Leibler Divergence

The *Kullback-Leibler divergence* was first defined by Kingman and Kullback [2007] and is another measure for the difference between two random variables, which is also closely related to the concepts of entropy (Definition 3.10) and of mutual information (Definition 3.13).

**Definition 3.14 (Kullback-Leibler Divergence)** *The Kullback-Leibler divergence $D_{KL}(Y \parallel X)$ - also known as information gain - between two random variables $Y$ and $X$, with the same domain $\mathcal{Z}$, is given by:*

$$D_{KL}(Y \parallel X) = -\sum_{z\in\mathcal{Z}}\mathbb{P}(Y=z)\log_b\left(\frac{\mathbb{P}(X=z)}{\mathbb{P}(Y=z)}\right) = \sum_{z\in\mathcal{Z}}\mathbb{P}(Y=z)\log_b\left(\frac{\mathbb{P}(Y=z)}{\mathbb{P}(X=z)}\right).$$

As always, the base $b$ of the logarithm defines the unit of measure; KL divergence is defined for any $b$.

Intuitively, it can be seen as measuring the amount of information gained when revising one's beliefs from the distribution of $X$ to the distribution of $Y$. Unlike information gain, it is

not a *distance measure* as it is evidently asymmetric i.e., $D_{\mathrm{KL}}(Y \parallel X) \neq D_{\mathrm{KL}}(X \parallel Y)$; if there is an *information gain* when moving from $X$ to $Y$, there is obviously an *information loss* when revising one's beliefs from $Y$ to $X$.

The *mutual information* $I(X,Y)$ is related to *information gain* (Definition 3.13) by:

$$I(X,Y) = D_{\mathrm{KL}}(\mathbb{P}(X,Y) \parallel \mathbb{P}(X) \times \mathbb{P}(Y)) \tag{3.5}$$

That is, the mutual information of two random variables is equal to the difference in information content between the product of the marginals and the joint distribution. This can better be understood through the terms of information theory: the mutual information of $X$ and $Y$ is the extra number of symbols needed to discriminate between the distributions of $X$ and $Y$, when they are coded using their marginal distributions and not their joint. This is to be expected because the definition of mutual information is exactly the information shared by $X$ and $Y$ and all that is needed to discriminate between the marginals is this particular "set" of information. The equivalence shown in Equation 3.5 can be proved by simply substituting $X = \mathbb{P}(X = x) \times \mathbb{P}(Y)$ and $Y = \mathbb{P}(X,Y)$ into Definition 3.14; the result will be Definition 3.13.

### 3.3.4 Hamming Distance

The *Hamming Distance* is a widely-used distance measure that quantifies the similarity of strings.

**Definition 3.15 (Hamming Distance)** *The Hamming Distance $D_H(x,y)$ between two vectors $x$ and $y$ is given by:*

$$D_H(x,y) = \sum_i \Gamma(x_i, y_i), \tag{3.6}$$

*with $\Gamma(i,j)$ defined as:*

$$\Gamma(i,j) = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases} . \tag{3.7}$$

Given strings of characters of equal length, or more in general vectors over some field, their Hamming Distance is the number of positions where they differ. It can be seen as the number of substitutions needed to transform one into the other.

This is a valid distance measure because:

- it is *non-negative*: $D_H(x,y) \geq 0$;
- it fulfils the *identity of indiscernibles*: if $x = y$ then $D_H(x,y) = 0$;
- it respects the *triangle inequality*: $D_H(x,y) \leq D_H(x,z) + D_H(z,y)$.

For example, strings $x = 01234$ and $y = 15244$ have $D_H(x,y) = 2$, as they differ in two positions.

### 3.3.5 Jaccard Distance

The *Jaccard Distance* is a popular metric to measure the dissimilarity of sets.

**Definition 3.16 (Jaccard Similarity Coefficient)** *The Jaccard Similarity Coefficient $J(A,B)$ of two sets $A$ and $B$ is given by:*

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} .$$

**Definition 3.17 (Jaccard Distance)** *The Jaccard Distance $D_J(A,B)$ between two sets A and B is given by:*

$$D_J(A,B) = 1 - J(A,B)\,.$$

This is a valid distance measure because:

- it is *non-negative*: $D_J(A,B) \geq 0$;
- it fulfils the *identity of indiscernibles*: if $A = B$ then $D_J(A,B) = 0$;
- it respects the *yriangle inequality*: $D_J(A,B) \leq D_J(A,C) + D_J(C,B)$.

For example, the sets $A = \{0,2,3,4,1\}$ and $B = \{1,3,5\}$ have $D_J(A,B) = 1 - J(A,B) = 1 - \frac{2}{6} = \frac{4}{6}$, as their intersection is of cardinality 2 and their union of cardinality 6.

## 3.4 Graph Theory

Many problems in machine learning do not involve classification or prediction of single data points in isolation, but of sets of entities that may present a more, or less, complex relation with each other. Most real-world phenomena fit into the latter framework. Graphs are one of the most powerful tools for the modelling of this class of problems, as their structure naturally captures the wide variety of relations that may exist between entities. These range from the atomic structure of a molecule to a social network of friends. In all these examples graphs help in reasoning, visualising and making inferences and predictions.

### 3.4.1 Directed Graphs

**Definition 3.18 (Directed Graph)** *A directed graph is a tuple*

$$\mathcal{G} = (\mathcal{V}, \mathcal{E})\,,$$

*with $\mathcal{V} = \{v_1 \dots v_n\}$ the set of vertices/nodes and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ the set of edges.*

We will not be concerned with the subclass known as *undirected graphs* that are characterised by $\mathcal{E}$ being a *set* of unordered pairs; that is, of sets of the form $\{x, y\}$, with $x, y \in \mathcal{V}$.

The class of graphs which interest us at present are those where there can be at most a single directed edge between any pair of nodes in $\mathcal{V}$; that is, we are not considering *multigraphs*. We are also interested in enforcing that there be no *cycles* in the graph so there can be no subset of edges in $\mathcal{E}$ that when followed starting from vertex $v_i$ eventually ends up in $v_i$ again. A cycle is a *walk* - a sequence of edges which joins a sequence of vertices - of nodes of the form $v_i, v_j, \cdots, v_i$ i.e., a walk where only the first and last vertex are repeated. Thus we have also automatically excluded the special case of cycle called *self-loop*: an edge from a node to itself. The resulting graph possessing only directed edges and no cycles is commonly called a *directed acyclic graph*, or DAG for short.

**Definition 3.19 (Directed Acyclic Graph)** *A directed acyclic graph is a graph where every edge is directed and there are no cycles.*

In a DAG we may qualify nodes based on their "relationship status":

**children** the children of node $u$ are all nodes $k$ for which there is a *directed edge* from $u$ to $k$;

**parents** the parents of node $u$ are all nodes $k$ for which there is a *directed edge* from $k$ to $u$;

**descendants** the descendants of node $u$ are all nodes $k$ for which there is a *directed path* i.e., a walk where all vertices are distinct, from $u$ to $k$;

**ancestors** the ancestors of $u$ are all nodes for which there is a directed path from $k$ to $u$.

An example of a DAG, containing five nodes, is shown in Figure 3.2.

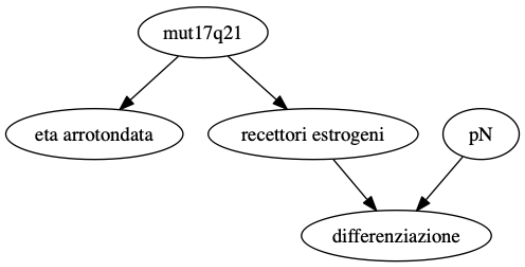



Figure 3.2. Example DAG representing a subset of the data set used in this thesis.

Polytrees and trees will also be defined because these are fundamental concepts for the work carried out in this thesis.

**Definition 3.20 (Tree)** *A tree is an undirected graph where there is one and only one walk between every node.*

**Definition 3.21 (Polytree)** *A polytree is a DAG whose underlying undirected graph is a tree. That is, if the directionality of edges is removed from the DAG, the resulting object is a tree.*

### 3.4.2 D-separation

*Dependence-separation* or *d-separation*, as the name entails, is a concept relating to the conditional dependence between variables and was first presented by Pearl and Dechter [1988]. We define the notation $u \to v$ to signify that there is a *trail* between $u$ and $v$ in the graph, with a trail $(u, \ldots, v)$ being a walk where all edges are distinct.

$u$ and $v$ and a node $z$ may be arranged in the graph in one of the following four configurations, called *v-structures* in this context:

- *chain*: $u \to z \to v$
- *chain*: $u \leftarrow z \leftarrow v$
- *fork*: $u \leftarrow z \to v$
- *collider*: $u \to z \leftarrow v$

We say that these v-structures are *closed* by the set $Z$ when:

- *chain*: $u \to z \to v$ and $z \in Z$
- *chain*: $u \leftarrow z \leftarrow v$ and $z \in Z$
- *fork*: $u \leftarrow z \to v$ and $z \in Z$
- *collider*: $u \to z \leftarrow v$ and $z \notin Z$ and no descendant $z'$ of $z$ i.e., $z'$ such that $z \to z'$ exists, is also in the set $Z$

We say that $u$ and $v$ are *d-separated* by $Z$ if every v-structure they appear in is closed by $Z$. Conversely, if there is at least one *open* v-structure then $u$ and $v$ are *d-connected*.

If $u$ and $v$ are d-connected, then knowing something about $u$ also tells us something new about $v$, and viceversa. An intuition for this can be given by interpreting the paths *causally*. In the case of a *chain*, $z$ is the cause of $v$ so knowing $z$ tells us everything we need to know about the value of $v$ (or of $u$, if the chain is reversed). In a *fork*, conditioning on the *common cause* $z$ has the same effect: $z$ is sufficient to know $u$ and $v$. This is also called the "Common Cause Principle" [Sober, 1988].

A good intuition for the behaviour of *colliders* was given by Pearl and Dechter [1988]: imagine that there are two independent causes for a car refusing to start ($z$): having no gas ($u$) and having a dead battery ($v$): $u \to z \leftarrow v$. Only knowing that the battery is charged gives no information about the car having fuel or not. But if we now know that the battery is charged after observing that the car won't start, we know for sure that it must be out of fuel. So knowing something about $u$ is informative about $v$, after conditioning on $z$.

**Definition 3.22 (D-Separation)** *Given vertices u and v and a set of vertices Z, then u and v are d-separated by Z if:*

- $Z \neq \emptyset$ *and u and v are never part of a collider;*
- $Z = \emptyset$ *and u and v are always part of a collider.*

The independencies between variables are encoded in the structure of the DAG so every probability distribution modelled by a BN that has the same connections between nodes, also necessarily has the same independencies regardless of the values of the variables.

A series of examples using the DAG presented in Figure 3.2 are shown in Figures 3.3, 3.4, 3.5. We can see how the network's topology and the nodes chosen to be in the observed set $Z$, define the resulting separations. In all cases $u =$"eta arrotondata" and $Y = V \setminus u \setminus Z$; we are asking for the set of all nodes in the DAG that are d-separated from $u$, given evidence $Z$. This can easily be answered by enumerating all the v-structures in the network and applying Definition 3.22. In the case shown in Figure 3.3 we see that the node "eta arrotondata" is separated from nodes "recettori estrogeni", "differenziazione" and "pN" given the observed evidence "mut17q21". The reason for this is because "eta arrotondata" $\leftarrow$ "mut17q21" $\to$ "recettori estrogeni" is a *fork* and thus the flow of information from the rest of the network is blocked. The way in which changing the conditioning set $Z$ also changes the independencies, can clearly be seen by comparing Figures 3.4 and 3.5.

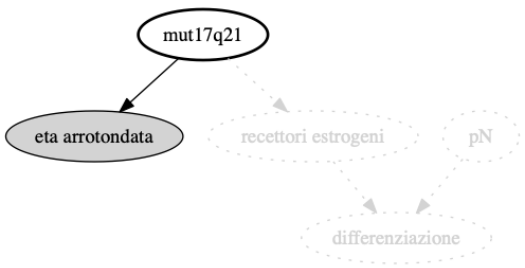



Figure 3.3. D-Separations in a subset of the provided data set (see Section 4.2).

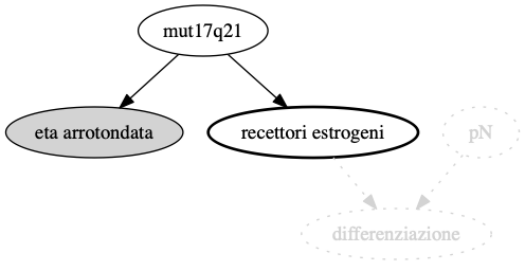



Figure 3.4. D-Separations in a subset of the provided data set (see Section 4.2).

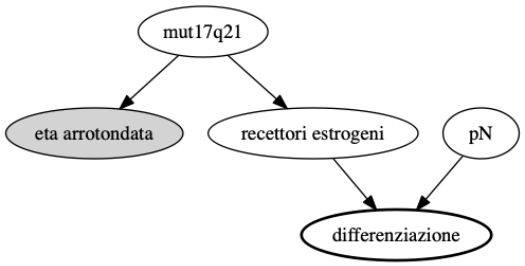



Figure 3.5. D-Separations in a subset of the provided data set (see Section 4.2).

## 3.5 Bayesian Networks

### 3.5.1 Bayesian Networks Definition

Given a variable $X$ and a set of variables $\boldsymbol{Y} = \{Y_1, ..., Y_n\}$, a *conditional probability table* (CPT) is a table whose columns are in one-to-one correspondence with one of all the possible combinations of values of the variables in $\boldsymbol{Y}$. Each column is a probability mass function over $X$, say $P(X \mid Y_1 = y_1, \ldots, Y_n = y_n)$, conditional on the tuple $(Y_1 = y_1, \ldots, Y_n = y_n)$. An example CPT can be seen in Table 3.4.

**Definition 3.23 (Bayesian Network)** *A Bayesian network (BN) is a probabilistic graphical model represented by a DAG $\mathcal{G}$ whose vertices are in one-to-one correspondence with a set of random variables $\boldsymbol{U}$ and the edges model the dependencies among these. The probability distribution of each variable $U \in \boldsymbol{U}$ is given by a CPT whose entries depend only on its parents in the graph structure.*

*The so-called Markov Condition states that every variable $U$ is independent of all nodes in the network, except from its descendents $Desc(U)$, given knowledge of its parents $Pa(U)$:*

$$\forall U \in \boldsymbol{U} : (U \perp \neg Desc(U) \mid Pa(U)).$$

The *Markov condition* (named after the 19th century mathematician Andrej Markov) imposes a d-separation on the DAG where each node/variable $U$ is d-separated from all others given the conditioning set $Pa(U) = \boldsymbol{Z}$.

A BN model is basically a way of representing an explicit joint distribution of random variables $\mathbb{P}(U_1 = u_1, \ldots, U_n = u_n)$ in a compact way. The compactness is achieved by leveraging the Markov condition i.e., the independencies that exist among the random variables, and the *chain rule* (Equation 3.1) to rewrite the joint as:

$$\mathbb{P}(U_1 = u_1, \ldots, U_n = u_n) = \prod_i \mathbb{P}(U_i = u_i \mid Pa(U_i)) \tag{3.8}$$

A BN gives the flexibility to drop the many weak dependencies that are bound to exist between variables, thus leading to an even simpler model. A full probability table for a joint distribution of random variables obscures the independencies and requires an exponential number of entries for the representation. A Bayesian network on the other hand can represent the same distribution using only a linear number of parameters. The power of BNs comes from the additional information encoded in their structure as first explicitly described in its entirety by Pearl and Dechter [1988], who defined the concept of dependence separation (Definition 3.22) and applied it to Bayesian networks. A classic example of BN has been shown in Figure 2.4.

One convenient characteristic of BNs is that they very naturally model the type of mixed causal and stochastic processes that are found in all of Nature. Imagine we want to represent the process modelled by joint distribution $\mathbb{P}(U, V)$; using the chain rule for conditional probabilities (Equation 3.1) we can write this as $\mathbb{P}(V \mid U) \times \mathbb{P}(U)$. A BN modelling this process would be composed of two nodes $U$ and $V$ with an edge from the former to the latter $U \rightarrow V$, $U$ is identified as the "parent" of $V$. Each of these two nodes would have its own probability table, with $\mathbb{P}(U)$ representing the *prior* distribution over $V$ and $\mathbb{P}(V \mid U)$ the *conditional probability distribution* of $V$ given $U$.

We can now see why these types of models are named *Bayesian* networks: the inference process is based in a predetermined *prior* distribution/belief and evolves through a parent $\rightarrow$ child relationship to constantly yield an updated *posterior* belief. The BN's DAG encodes a

generative sampling where each variable's value is determined stochastically by Nature, based on the value of its parents. This process is also highly compatible with our view of *causality* and this is one of the reason that could make BNs highly *interpretable*. The prior $\mathbb{P}(A)$ can be seen as the result of some stochastic process caused by a series of latent (unmodelled) variables while the posterior $\mathbb{P}(B \mid A)$ is stochastically, causally determined by $A$. As mentioned in the previous paragraphs, there are probably no truly "prior" distributions in the Universe, at the modelling scale we are usually interested in. Only on arriving on the quantum particle level may we find "pure" stochastic, uncaused processes created by quantum collapse.

A good example of how BNs are compatible with our notion of causality may be to imagine $A$ as the random variable modelling the predisposition to having a certain disease and $B$ the one indicating actually developing the symptoms for it. *First*, genetic and epigenetic factors such as the environment stochastically contributed to having the predisposition and *then* the development of the symptoms was stochastically determined by the degree of predisposition. Adding an extra time dimension certainly helps in understanding this class of probabilistic models.

If the example show in Figure 3.2 is taken as the underlying graph structure of a Bayesian network, each node now represents a random variable with an associated *conditional probability table* that defines its probability distribution based on that of its parents. The distributions for "eta arrotondata" and "mut17q21" in the Bayesian network in question are shown in Table 3.1 and 3.2. "Mut17q21" is a root node i.e., has no parents, in the DAG so its probability distribution is unconditional or *marginal*. "Eta arrotondata", on the other hand, is a child of "mut17q21" so the probability of its values is the results of a conditioning on those of its parent and is thus represented by a CPT. For example, "eta arrotondata" takes on value "<40" 44% of the time when "mut17q21" has value "mut", but only 4% of the time when "mut17q21" has value "unknown".

Table 3.1. "mut17q21" mass function

| | | |
|---|---|---|
| **mut17q21** | mut | 0.01 |
| | unknown | 0.99 |

Table 3.2. "eta arrotondata" CPT

| | | **mut17q21** | |
|---|---|---|---|
| | | mut | unknown |
| **eta arr.** | <40 | 0.42 | 0.04 |
| | 40-50 | 0.42 | 0.17 |
| | >50 | 0.15 | 0.78 |

Probabilistic graphical models such as Bayesian networks are often used to express expert knowledge about a particular domain and perform reasoning on that problem. Alternatively, the specification of the network can be automatically achieved from a sufficient amount of data about the variables under consideration for a particular reasoning task. In this thesis we focus on the case of Bayesian network learned from data, but the methods presented in Chapter 4 would also apply to a user-designed network, as would be the case in an *expert system*. As the Bayesian network formalism consists of both a qualitative element (the directed graph) and a

quantitative one (the conditional probability tables), in the following sections we will detail how these two components can be obtained automatically from data.

### 3.5.2 Learning Bayesian Networks Structure from Data

Learning a BN DAG from data is typically addressed as an optimisation task and is known as the *Bayesian network structure learning problem*. In many probabilistic models initialisation is fast but then fitting the data is slow (for example in *k-means*). For Bayesian networks the converse is true: fitting is fast as only sums of the counts in the data are needed, but learning the correct graph structure can take super-exponential time - more precisely, the space of Bayesian networks that have $|V|$ variables has size $2^{O(|V|^2)}$ [Berzan, 2012] - in the number of variables and this easily leads to an intractable problem.

Let us consider the specification of a BN over the variables $\boldsymbol{X} = (X_1, \ldots, X_n)$ and denote as $\boldsymbol{D}$ a data set of joint and complete observations of $\boldsymbol{X}$. A *score function* is a map $f$ giving a score to any possible DAG $\mathcal{G}$ whose nodes are in correspondence with $\boldsymbol{X}$ as a function of the data set $\boldsymbol{D}$. The resulting score $f(\mathcal{G}, \boldsymbol{D})$ is a measure of how well a BN with graph $\mathcal{G}$ fits the data set $\boldsymbol{D}$. The simplest approach consists in using the likelihood (the probability assigned by the BN to the data) as a score. Yet, additional terms that penalise complex models are added to prevent over-fitting. Given the score, the problem is basically a search over the set $\Gamma$ of all the possible DAGs with $|V|$ nodes i.e., $\mathcal{G}^* = \arg\max_{\mathcal{G} \in \Gamma} f(\mathcal{G}, \boldsymbol{D})$. Such a task is NP-hard but approximate search procedures can be defined to solve it efficiently.

The methods to solve this problem can be roughly categorised into three categories:

**Search and Score** This is the most naïve method as it does a brute force search over all the possible graphs' structure space - i.e., all DAGs with the same number of variables as the input data - and scores all these depending on some cost function.

There are many cost functions that have been proposed over the years; for example, a Bayesian cost function represents the probability of the DAG $\mathcal{G}$ given the data $\boldsymbol{D}$: $\mathbb{P}(\mathcal{G} \mid \boldsymbol{D})$ while an information theory one scores the fitness of a DAG by its ability to balance graph description length and data description length given the graph.

This process is super-exponential in the number of variables but through the use of dynamic programming and heuristic search algorithms it can become sub-exponential. Nonetheless, solving the exact problem is only feasible up to $\approx 30$ variables.

**Constraint Learning** Methods of this type calculate some measure of correlation to identify the presence and direction of edges between nodes and are much less used than the other ones presented. A typical test is to iterate over all triplets while testing for conditional independencies. Thanks to the d-separation properties outlined in Subsection 3.5, this test is able to identify the correct edges. The algorithm is quadratic in time and in the number of vertices.

**Approximations** Several heuristical approaches have been developed to be able to find good network structures in an efficient manner. Examples of these are:

- *Chow-Liu*, which builds a tree approximation of the probability distribution;
- *greedy hill-climbing*, which adds/removes/flips an edge at a time;

- *optimal reinsertion*, which iteratively calculates the optimal *Markov blanket* (the subset of all nodes that are sufficient to determine the value of another subset) of an ever-smaller subset of nodes.

### 3.5.3 Learning Bayesian Networks Parameters from Data

Once the DAG structure is given, learning the CPTs from the data $\boldsymbol{D}$ can be accomplished by one of two approaches:

**Frequentist** The frequentist approach treats the parameters $\theta$ to be learned as unknown but fixed and attempts to find a $\theta^*$ that maximises the likelihood function $\mathbb{P}(\boldsymbol{D} \mid \theta)$. Given symbol $j$ in the CPT of variable $i$ conditioned on the parents having value $k$ and $N_{ijk}$ the count of times that this combination of symbols appears in the data $\boldsymbol{D}$ then the Maximum Likelihood Estimator $\hat{\theta}_{ijk}^{MLE}$ for the entry $j,k$ in the CPT of $i$ is given by:

$$\hat{\theta}_{ijk}^{MLE} = \frac{N_{ijk}}{\sum_{j'} N_{ij'k}} \tag{3.9}$$

**Bayesian** The Bayesian method instead treats $\theta$ as a random variable with a prior probability $\mathbb{P}(\theta \mid \alpha)$, with $\alpha$ virtual pseudo-count, and uses Bayes' Rule (see Theorem 3.7) and a likelihood $\mathbb{P}(\boldsymbol{D} \mid \theta)$ to calculate the posterior probability $\mathbb{P}(\theta \mid \boldsymbol{D}, \alpha)$. Given symbol $j$ in the CPT of variable $i$ conditioned on the parents having value $k$ and $N_{ijk}$ the count of times that this combination of symbols appears in the data $\boldsymbol{D}$ then the Maximum a Posteriori Estimate $\hat{\theta}_{ijk}^{MAP}$ for the entry $j,k$ in the CPT of $i$ is given by:

$$\hat{\theta}_{ijk}^{MAP} = \frac{N_{ijk} + \alpha_{ijk}}{\sum_{j'} \left(N_{ij'k} + \alpha_{ij'k}\right)} \tag{3.10}$$

### 3.5.4 Bayesian Networks Updating

All the types of inference that will be presented are instances of *diagnostic reasoning*, also known as *abductive reasoning*. Abductive reasoning is a form of inference that starts from observed evidence and looks for the best/most simple *explanation* for them. Unlike *deductive reasoning*, abductive reasoning is not based on logical necessity, but on inferences based on observations; thus it cannot verify the conclusion with absolute certainty but only yield, at best, a highly probable explanation. The direction of inference is reversed in abduction - this is why it is sometimes called *retroduction* - as compared to deduction. It would be a logical fallacy known as "affirming the consequent" to state that any explanation were certain, because there may be multiple possible explanations for the same observation.

The following examples may help to clarify the difference between the two inference processes:

**deductive reasoning** given "Every man is mortal"$(a_1)$ and "Diogenes is a man"$(a_2)$ it necessarily follows that "Diogenes is mortal"$(b)$: $(a_1) \wedge (a_2) \implies (b)$.

**abductive reasoning** given that "Diogenes is mortal"$(b)$ it is very probable that "Diogenes is a man"$(a_2)$; this is not certain as it may also be that "Diogenes is a dog"$(a_3)$: $(b) \not\Rightarrow (a_2) \wedge (b) \not\Rightarrow (a_3)$.

Abductive reasoning can either be modelled as a conditional probability or a MAP query and is of fundamental importance in many important problems of machine learning, including medical diagnosis which is of particular interest for this thesis. Let us define three sets of variables of interest in the BN: $\boldsymbol{U_q}$ the *query variables*, $\boldsymbol{U_e}$ the *evidence variables* and $\boldsymbol{U_m}$ all remaining variables.

We can thus define the conditional probability query as the updated probability of $\boldsymbol{U_q}$ based on the observation of the values of $\boldsymbol{U_e}$.

**Definition 3.24 (Conditional Probability Query)** *The conditional probability query* $\mathbb{P}(\boldsymbol{U_q} \mid \boldsymbol{U_e} = \boldsymbol{e})$ *for variables* $\boldsymbol{U_q}$ *given evidence* $\boldsymbol{U_e} = \boldsymbol{e}$ *is:*

$$\mathbb{P}(\boldsymbol{U_q} \mid \boldsymbol{U_e} = \boldsymbol{e}) = \frac{\mathbb{P}(\boldsymbol{U_q}, \boldsymbol{U_e} = \boldsymbol{e})}{\mathbb{P}(\boldsymbol{U_e} = \boldsymbol{e})} = \frac{\sum_{\boldsymbol{U_m}} \mathbb{P}(\boldsymbol{U_q}, \boldsymbol{U_e} = \boldsymbol{e}, \boldsymbol{U_m})}{\sum_{\boldsymbol{U_m}, \boldsymbol{U_q}} \mathbb{P}(\boldsymbol{U_q}, \boldsymbol{U_e} = \boldsymbol{e}, \boldsymbol{U_m})} = \frac{\sum_{\boldsymbol{U_m}} \prod_i \mathbb{P}(U_i \mid Pa(U_i))}{\sum_{\boldsymbol{U_m}, \boldsymbol{U_q}} \prod_i \mathbb{P}(U_i \mid Pa(U_i))}.$$

The solution will be the *posterior* probability of $\boldsymbol{U_q}$.

Another common type of question we might ask a BN is the following: "given evidence $\boldsymbol{U_e}$ which is the most likely assignment to a subset of variables $\boldsymbol{U_q}$?". This is know as *Maximum a posteriori (MAP)* inference and is a much harder problem that a conditional probability query. The solution is found by solving an optimisation problem.

**Definition 3.25 (Maximum a Posteriori Query)** *Given sets* $\boldsymbol{U_q} \subseteq \boldsymbol{U}$ *and* $\boldsymbol{U_z} = \boldsymbol{U} \setminus \boldsymbol{U_e} \setminus \boldsymbol{U_q}$*, the MAP query for* $\boldsymbol{U_q}$*,* $MAP(\boldsymbol{U_q} \mid \boldsymbol{U_e} = \boldsymbol{e})$*, is:*

$$MAP(\boldsymbol{U_q} \mid \boldsymbol{U_e} = \boldsymbol{e}) = \arg\max_{\boldsymbol{q}} \sum_{\boldsymbol{z}} \mathbb{P}(\boldsymbol{U_q} = \boldsymbol{q}, \boldsymbol{U_z} = \boldsymbol{z}, \boldsymbol{U_e} = \boldsymbol{e}).$$

The solution will be the assignment of values $\boldsymbol{q}$ to all variables in the set $\boldsymbol{U_q}$ that maximises their joint probability.

An important thing to note is that the greedy assignment, where each variable picks its most likely value, can be very different from the most likely joint assignment of all variables. A simple example showing this is given by Koller et al. [2009, pag. 26]. Suppose a Bayesian network is composed of two nodes $U$ and $V$ with the former the parent of the latter: $U \to V$. Assume their CPDs are represented by the CPTs shown in Tables 3.3 and 3.4. The most probable value for $U$ is $U = u_1$ and this constrains $V$ to choose equivalently between $V = v_0$ or $V = v_1$. The probability of the assignment $(U = u_1, V = v_1)$ given by $(\arg\max_u \mathbb{P}(U = u), \arg\max_v \mathbb{P}(V = v)))$ is $0.6 \times 0.5 = 0.3$. On the other hand, the most likely joint assignment given by $\arg\max_{u,v} \mathbb{P}(U = u, V = v)$ is $(U = u_0, V = v_1)$ and has probability $0.4 \times 0.9 = 0.36$.

Table 3.3. "U" CPT

| | | |
|---|---|---|
| **A** | $u_0$ | 0.4 |
| | $u_1$ | 0.6 |

The MAP problem is hard to solve efficiently because it is part of the $NP^{PP}$-hard complexity class, as proved by Shimony [1994]. Attempting to solve it in a brute-force way would mean listing all the possible variable-value tuples and computing their joint probabilities; as the number of these combinations is exponential in the number of variables the problem is evidently intractable.

Table 3.4. "V" CPT

| | | **A** | |
|---|---|---|---|
| | | $u_0$ | $u_1$ |
| **B** | $v_0$ | 0.1 | 0.5 |
| | $v_1$ | 0.9 | 0.5 |

The problem remains intractable even in a Bayesian network. Such a model may possess a linear number of parameters but the underlying distribution is still implicitly exponential in size. Explicitly calculating the MAP defeats the very purpose of the BN: computational efficiency. For this reason, a host of approaches exist to optimise MAP: elimination algorithms, gradient methods, simulated annealing and other stochastic local searches, belief propagation, and integer linear programming.

A special case of MAP, which is also an easier problem to solve, is the *most probable explanation (MPE)*.

**Definition 3.26 (Most Probable Explanation Query)** *Given set $\boldsymbol{U_q} = \boldsymbol{U} \setminus \boldsymbol{U_e}$, the MPE query for $\boldsymbol{U_q}$, $MPE(\boldsymbol{U_q} \mid \boldsymbol{U_e} = \boldsymbol{e})$, is:*

$$MPE(\boldsymbol{U_q} \mid \boldsymbol{U_e} = \boldsymbol{e}) = \arg\max_{\boldsymbol{q}} \mathbb{P}(\boldsymbol{U_q} = \boldsymbol{q}, \boldsymbol{U_e} = \boldsymbol{e}).$$

The solution will be the assignment of values $\boldsymbol{q}$ to all variables in the set $\boldsymbol{U_q}$ that maximises their joint probability.

An intuition for why the MPE is easier to solve can be given by comparing Definitions 3.25 and 3.26; unlike MPE, MAP presents both a summation and a maximisation and as such is part conditional probability query, part MPE query. All algorithms for the computation of MAP obviously apply to MPE too, but efficient approximate algorithms exist for MPE that do not generalise to MAP such as Loopy Belief Propagation [Pearl and Dechter, 1988] and Stochastic Local Search [Kask and Dechter, 1999].

## 3.6 Summary

The chapter has introduced a number of concepts from probability, information and graph theory to be used as groundwork for the formal description of Bayesian networks, a widely used class of probabilistic graphical models.

The section dealing with probability theory opens by describing *random variables*, a mathematical construct that associates a value to every outcome in the set of possible events, which is used to bring to the fore the attributes of interest while dealing with them in a clean way. The probability of an event may be interpreted through the *frequentist* or the *Bayesian* lens; the former sees probability as simply the limit of the ratio between the number of times the event of interest occurred and the total number of trials; the latter views probabilities as the subjective degree of belief regarding the manifestation of the event. The main results presented are *Bayes' Theorem*, which states how to update a prior belief in the light of new knowledge, and the concept of *independence* between events, which will be central when introducing *d-separation*.

The second section introduces a few key concepts from information theory relating to entropy and distance measures. *Entropy* is a measure for the expected amount of information carried by a random variable, first introduced by Claude Shannon by drawing parallels to mechanical statistics. A convenient derived quantity is *normalised entropy* - also known as *efficiency*; this measure varies in $[0, 1]$ and thus enables random variables and probability distributions of different cardinalities to be compared. A second method, closely related to entropy, of measuring the interrelatedness of two random variables is that of *mutual information*; this measure quantifies the amount of information of one variable already contained in the other. Three popular distance measures are introduced: *Kullback-Leibler divergence, Hamming and Jaccard distances*; the first is closely connected to the concept of entropy and is another measure for the interrelatedness between two variables, the second measures the similarity of strings, based on the number of substitutions needed to transform one into the other, and the last quantifies the similarity of sets, given the size of their intersection over union.

The third section relates to graph theory and starts by defining the basic notion of *directed graph* and of *directed acyclic graph*, a special case of graph presenting no cycles between vertices. *Trees* and *polytrees* are briefly introduced and characterised, for future use. Finally, *d-separation*, first introduced by Judea Pearl, is discussed. D-separation is a concept relating to the conditional dependence between variables; sets of variables may become independent i.e., not influence each other, based on conditioning on a third set of evidence variables. The independence properties depend on the topology of the graph, specifically in how the variables of interest are connected to each other; they may be organised into *chains*, *forks* or *colliders*.

The final section of the chapter deals with introducing *Bayesian networks*, using many of the concepts laid out in the previous sections. A Bayesian network is a probabilistic graphical model represented by a DAG where each vertex corresponds to a random variable and the edges model the dependencies among these. The basic idea is to factorise a complete joint distribution of the constituent variables into a series of conditional probability distributions, one for each variable, that are assigned to the nodes in the DAG. The defining characteristic is that each variable's node values depend only on those of its parents. Such a representation efficiently represents a joint distribution and very naturally models the type of mixed causal and stochastic processes found in Nature. The DAG of a BN can either be given or learned directly from data; learning is a super-exponential problem and there are three main classes of algorithms that may be applied to solve it: *search and score*, *constraint learning* and *approximations*. Once a DAG has been learned, the problem moves to querying (*updating*) the BN; the main classes of queries are *conditional probability* and *maximum a posteriori queries*. The first class asks for the value of a set of variables given the observation of the values of others in the network. The second class, known as MAP queries, asks to find the most probable assignment of values to a subset of variables, given the observation of the values of another subset. This is, in general, a hard problem but efficient solutions exist for a special case known as the *most probable explanation*, where the set of query variables is the complementary subset to the evidence one.

# Chapter 4

# Methodology

## 4.1 Introduction

The inspiration for the work carried out in this thesis was the paper [Butz et al., 2018] reviewed in detail in Section 2.6. This work proposed a system that would learn a "knowledge base" i.e., a tree representing the chain of most probable deductions, starting from a Bayesian network modelling a medical data set, rooted in a set of initial evidence. This tree, deemed to represent the solution to the MPE query (Definition 3.26), could then be used to generate a dialogue in natural language with the medical expert, which the authors claimed could lead to the extraction of extra knowledge from the original data set. The driving hypothesis of the paper was that Bayesian networks and the solution to the MPE problem would be a powerful tool in helping medical experts gain insights into data.

The paper did not provide any indication that a such a system had ever been built and any validation of the method was left by the authors for future work. As discussed in Chapter 1 and 2, this lack of real-world validation has been seen to be the unfortunate norm in most papers published under the explainable AI moniker. Many works are content to only give a *functionally-grounded evaluation* (in the taxonomy of Doshi-Velez and Kim [2017]) for the methods they propose. The one by Butz et al. [2018] does not even present such an evaluation. As of the finalisation of this thesis in early September 2019, there has been no work carried out in substantiating the methods of Butz et al. [2018]. As introduced in Chapter 1 and discussed in detail in Section 2.3, there is an ever greater need for machine learning models and systems to be explainable, especially in mission-critical domains such as healthcare.

For these reasons the hope was that building a proof of concept system, whose logic was inspired by the method presented in the aforementioned paper, and validating it with real medical experts, could prove to be a positive endeavour. What is being proposed is to carry out an *application-grounded evaluation* of the developed system. The objective of this thesis is not only to provide an assessment of the paper and of Bayesian networks in general, but also to set a methodological precedent for the evaluation of a machine learning system with real medical experts on an actual tasks. This, as discussed in Chapter 1 and 2, is one of the main gaps existing today in the field of xAI. Thus, carrying out such an evaluation could be seen as introducing an element of novelty. It is also hoped that the proof of concept system may be of concrete use to the medical experts who are provided with it, in performing their daily work.

As already set out in Section 1.3, the work carried out in this thesis benefited from a high

degree of collaboration with a third party, the *Istituto Cantonale di Patologia*, based in Locarno (Switzerland).

The chapter is organised as follows:

- Section 4.2 opens with an introduction to the medical partner involved in the evaluation of the methods: *Istituto Cantonale di Patologia*. The institute is based in Locarno in the Swiss canton of Ticino and specialises in the histological analysis of tissue samples in support of cancer diagnosis.
- Section 4.3 gives an overview of the standard tools used in building the proof of concept system, which was then evaluated with the pathologists of the ICP.
  - Subsection 4.3.1 presents the main Python libraries employed in implementing the system.
  - Subsection 4.3.2 shows the algorithms that are part of the methods of this thesis but that make use of standard techniques; these include a classic algorithm for *d-separation* by Koller et al. [2009].
- Section 4.4 introduces the novel algorithms and methods that were developed specifically for the implementation of the prototype system and explains the rationale behind the selection method used to build the "knowledge base" (see Section 2.6) and presents an algorithm for the calculation of the *mutual information* (see Subsection 3.3.2) between pairs of variables in the BN.
  - Subsection 4.4.1 gives a detailed presentation of the methods developed, including the three variants of the *dialogue* - which are adaptations and developments to the method presented by Butz et al. [2018], an algorithm to generate alternative explanation branches when the domain expert disagrees with the system during a dialogue, a greedy algorithm to construct a "pseudo-MPE" branch from random evidence and a procedure to compare it to the true MPE solution.
  - Subsection 4.4.2 concentrates on the methods needed to interface effectively with the user.
- Section 4.5 presents the methodology used to validate the proof of concept tool, which implements all of the methods presented in the previous sections, from the point of view of its clinical relevance and its capacity to surface explainations to the user.

## 4.2 The Benchmark Data Set

This section focuses on presenting the data set that was used to learn the Bayesian network of the prototype system. The data set was provided by *Istituto Cantonale di Patologia* and was preprocessed before being used to learn the model.

### 4.2.1 The Medical Partner: Istituto Cantonale di Patologia

Istituto Cantonale di Patologia (ICP)[1] is an institute based in Locarno (Switzerland) specialised in the histological analysis of tissue samples received from private patients, clinics and hospitals, mainly in support of cancer diagnosis. Its laboratory of molecular pathology supports

[1] https://www4.ti.ch/dss/dsp/icp/istituto/

the diagnostic approach to neoplastic diseases through the application of biomolecular and cytogenetics techniques, which focus on understanding the impact of molecular alterations in carcinogenesis. One of the most powerful methods used is *fluorescence in situ hybridization* (FISH) (see Figure 4.1 for an example of the outputs of this technique) which is able to localise the presence or absence of specific DNA sequences in individual chromosomes. These tests are aimed at identifying the precise profile of the cancer cells and thus inform the clinician on the best treatment for the specific patient.

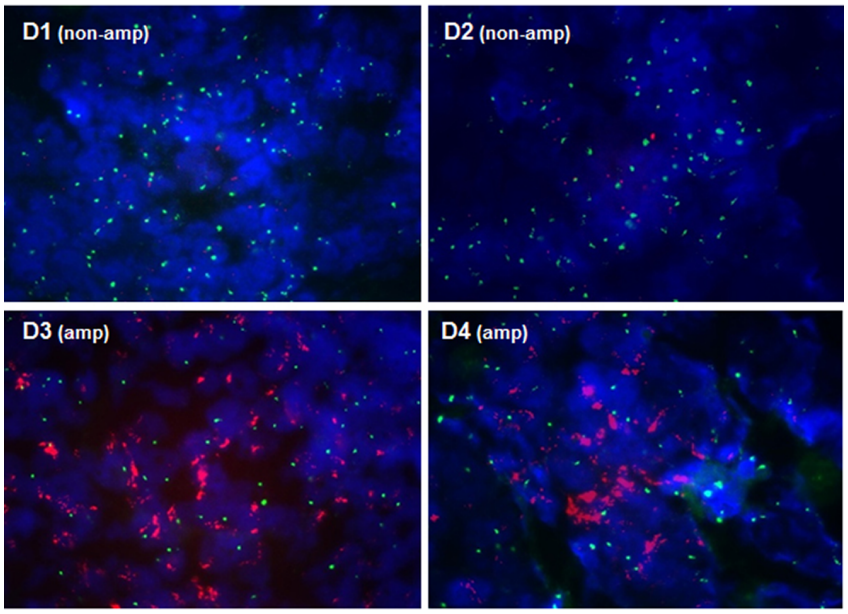



Figure 4.1. FISH analysis of HER2 gene expression in samples of breast tumour. The probe mix consists of a mixture of Texas Red-labelled DNA probe against HER2 gene (which is located on chromosome 17) and a fluorescein (green)-labelled probe targeted at the centromeric region of chromosome 17. The upper panels (D1 and D2) show normal expression - 2 green and 2 red signals per cell. The lower panels (D3 and D4) show HER2 amplification whereby there is a clear increase in the red signal [BioIVT].

Molecular diagnosis is the most recent approach in pathology. It mainly aims to support pathologists in the diagnosis of neoplastic diseases through the employment of molecular biology and cytogenetic techniques; it enables the deep understanding and monitoring of patients' profiles, as well as defining prognoses and ad-hoc therapies. In its totality, it allows for the development of an increasingly personalised medicine. This personalisation is also applied to research, for example by being able to understand pathological mechanisms and the consequent possibilities for intervention. However, it is still limited in its capabilities by the presence of a dishomogeneous information pool, especially in terms of knowledge extraction; for the pathologist, a non-uniform data set entails an increased cognitive load because of having to manage a high number of heterogeneous variables. These factors necessarily force her to resort to problem decomposition and reduction.

In addition to its clinical support activities, the ICP also carries out scientific research aimed at better understanding certain types of cancers at the translational level i.e., at the level of

protein synthesis. In the last ten years, the ICP has published more than 200 peer-reviewed papers and more than 100 works in non-peer reviewed journals and is active at a national and international level.

### 4.2.2 Motivation

The Istituto Cantonale di Patologia was already collaborating with the Dalle Molle Institute for Artificial Intelligence (IDSIA)[2] to investigate a series of specific issues, whose details are outside of the scope of this thesis. The prospect of the work to be carried out in this thesis was deemed of interest because it extended beyond the breadth of the existing collaboration. The institute had originally expressed interest in bringing machine learning into its workflow in order to both augment its profiling capabilities for patients and to be able to extract new knowledge from existing data. This was paired with an attention for more experimental research directions, as is the facilitation of human-machine interaction.

Given that the theoretical work carried out in this thesis is, at its core, an investigation into the explainability of Bayesian networks, collaboration with the Institute provided the precious opportunity to implement a proof of concept system (described in Chapters 4 and 5) based on a real medical data set and, most importantly, opened an opportunity for an *application-grounded evaluation* of it (see Section 2.4).

The first contact with the ICP was in January 2019, during a meeting with Vittoria Martin (PhD), molecular cytogenetist, and Luca Mazzucchelli (Dr. Med.), Director of the institute. Since then, the clinicians and researchers of the ICP have been able to validate the model software that has been developed from an explainable AI and clinical relevance point of view. That is to say, they have validated, to an extent that will be made clear in Chapter 5, the developed software both in its adherence to established medical literature and in its capacity to support clinical decision making and to surface clarifying explanations of the data set. This is a great opportunity because the lack of real-world validation of ML systems with actual domain experts is one of the prime gaps in the existing xAI literature (as discussed in depth in Chapter 2).

An example application for a clinician of the ICP, would be ability to "fill in the blanks" of a patient's profile, as it is not uncommon, for a variety of reasons, to have missing data. This may be because of degraded or insufficient tissue samples or because some test may not yet be part of the standard diagnostic procedure, even though their importance may already be suggested by clinical research. In other cases, patients may be missing a result because the specific test had not yet been invented, for example FISH was not available prior to 2010, so an *a posteriori* inference could be made possible. Another crucial use may be in understanding and effectively quantifying the relationship between clinical variables. It is not uncommon for some variables to have been observed but for their clinical relevance not to have yet been determined; learning their relationship with other variables could potentially not only help in defining their importance in tracing new patient profiles in terms of diagnosis, prognosis and support in decision making but also in placing these profiles in terms of pathological mechanisms. These are all examples highlighting the importance of the *inference* capabilities of machine learning and *uncertain reasoning* techniques, but the current work aims to principally address the interfacing of the human user with the software while carrying out these queries. Nonetheless, a system that is capable of interfacing in a meaningful manner with the clinician, is certainly in a bet-

[2] http://www.idsia.ch

ter position to enable all the important applications just discussed. It is hoped by the ICP that facilitating the process of knowledge-extraction may lead towards the confirmation of current scientific theories or may even be the first step towards the formulation of novel ones.

### 4.2.3 Provided Data Set

The provided data set was created by *Registro Tumori Ticino*[3] (Locarno, Switzerland) in order to highlight possible new relations between clinical, histopathological and molecular features, as well as to potentially discover novel biomarkers involved in the progression of the disease. It consists of the histopathological records, over 38 variables of interest, of *3218 breast cancer patients* who have been diagnosed between the years 2005 and 2014 within the Ticino canton of Switzerland. The data set had already been pre-processed by collaborators of IDSIA under supervision of the ICP, with 13 of the variables being dropped because not relevant. In particular, all variables relating to patients post-treatment were discarded as well as those recording the diagnosis date. The data set was also anonymised, for obvious privacy issues. Some of the variables were initially numerical (for example "FISH") but all were converted to categorical during this process.

In Table 4.1 is a description of the remaining variables, together with their clinical meaning. The distribution of the densities of the data set variables is shown in Table 4.2. The indications from Dr. Martin on how to further preprocess the data are shown in Table 4.3. Note that some variable names were simplified and that the conversion to coarser categories could aid in boosting the explainability of the data set by reducing the number of possible values of each variable.

[3] https://www4.ti.ch/dss/dsp/icp/registro-cantonale-dei-tumori/home/

Table 4.1. Data set variables

| **Variable** | Clinical meaning |
|---|---|
| **Codice globale** | Unique patient identifier |
| **mut17q21** | Mutation of chromosome 17 |
| **loss 17** | Loss of chromosome 17 |
| **età arrotondata** | The age of the patient at diagnosis |
| **Lateralità** | The affected breast |
| **Situ SUBGROUP MZ** | The primary site code of the tumour |
| **Morfologia SUBGROUP MZ** | The morphology classification of the tumour |
| **pT SUBGROUP MZ** | Primary tumour dimensions in the TNM classification for breast cancer |
| **pN SUBGROUP MZ** | Pathologic nodes involvement in the TNM classification for breast cancer |
| **M 8.2.96** | Distant metastasis in the TNM classification for breast cancer |
| **Differenziazione** | Tumour grade |
| **Recettori estrogeni percento 1.1.2003** | Expression of oestrogen receptors |
| **Recettori progestinici percento 1.1.2003** | Expression of progestin receptors |
| **c erbB 2 cod percento 1.1.2003** | ErbB2 marker expression |
| **Ki67 cod percento** | Tumoural proliferation index |
| **FISHRatio** | FISH analysis result |

Table 4.2. Data set distribution before pre-processing

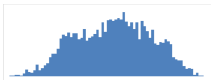
| Variable | Cardinality | Distribution |
|---|---|---|
| **mut17q21** | 2 |  |
| **loss 17** | 3 |  |
| **eta arrotondata** | 74 |  |
| **lateralità** | 3 |  |
| **situ** | 5 |  |
| **morfologia** | 5 |  |
| **pT** | 23 |  |
| **pN** | 6 |  |
| **M** | 3 |  |
| **differenziazione** | 5 |  |
| **recettori estrogeni** | 40 |  |
| **recettori progestinici** | 40 |  |
| **c erbB 2** | 4 |  |
| **Ki67** | 52 |  |
| **FISHRatio** | 5 |  |

Table 4.3. Data set preprocessing steps

| **Variable** | Action |
|---|---|
| **Codice globale** | Remove variable |
| **mut17q21** | Remove variable |
| **loss 17** | Remove variable |
| **eta arrotondata** | Bin into “$< 40$”, “$40 - 50$”, “$\geq 50$” |
| **lateralita** | Remove blanks and “sconosciuta” |
| **situ** | Remove blanks |
| **morfologia** | Remove blanks and “unuseful” if performance on classification is subpar |
| **pT** | Remove blanks and “unuseful” |
| **pN** | Remove blanks and bin into “0” and “$\neq 0$” |
| **M** | Remove blanks |
| **differenziazione** | Remove blanks and “Sconosciuto o non applicabile” |
| **recettori estrogeni** | Remove blanks and bin into “negativo” if $\leq 10$, “debolmente positivo” if $\leq 50$, “fortemente positivo” if $> 50$ |
| **recettori progestinici** | Remove blanks and bin into “negativo” if $\leq 10$, “debolmente positivo” if $\leq 50$, “fortemente positivo” if $> 50$ |
| **c erbB 2** | Remove blanks |
| **ki67** | Remove blanks and bin into “<14”, “14-20”, “20-30”, “>30” |
| **FISH** | Remove variable |

## 4.3 Methods

### 4.3.1 Libraries

The system developed in this thesis was coded in Python and as such made use of an array of standard and less familiar packages. The most significant for the development of the system are Pomegranate and Pgmpy, both implementing probabilistic graphical models.

#### Probabilistic Graphical Models Packages

*Pomegranate*[4] is an open-source probabilistic models package for Python. Its core philosophy is that every probabilistic model, from hidden Markov to Bayesian network, can be seen as a probability distribution and, as such, can be flexibly composed into hierarchical mixture models [Schreiber, 2017]. The package implements Bayesian networks as well as many other probabilistic models, but currently only supports discrete Bayesian networks, so the random variable of each node must have a categorical distribution. This is not an issue as the provided data set (see Section 4.2) was already composed of only categorical variables. Also, working with discrete entities should make explainability easier as the number of possible variable values at hand can be reduced at will; this should in turn reduce the cognitive load requested on the user's part.

Pomegranate was chosen among other packages because of its satisfactory implementation of Bayesian networks and its performance. The package is written in Cython and natively supports multi-core parallelism and out-of-core learning. Network *structure learning from data* is claimed to be particularly efficient, thanks to a novel *constraint learning* (see Subsection 3.5.2) method that implements prior knowledge into the graph selection process [Schreiber and Noble, 2017]. The claim made by the authors is that this innovative graph selection process should possess the speed of a heuristic approach, while yielding a far better quality estimate of the correct graph structure.

*Structure learning* from data is achieved using the `from_samples` method of the `BayesianNetwork` class, with the default algorithm being the novel one proposed in [Schreiber and Noble, 2017]. The *probability* of a sample is calculated using the `probability` function; the `predict_proba` function is used to return the probability of each variable in the model given some evidence. *Predictions* are run by passing an object a matrix with `None` as placeholders for missing values to the `predict` function. *Fitting* is done through the `fit` function which uses MLE estimates to update each node's distribution in the model based on the input data.

A `BayesianNetwork` object can also be displayed graphically by calling its `plot` function. The output is a `DOT` file that is generated using the PyGraphviz package[5], a Python interface to the famous Graphviz[6] graph visualisation software. An example of such an output is shown in Figure 4.2. For flexibility and compatibility reasons, a custom function was written to display a Pomegranate BN graphically.

*Pgmpy*[7] is, like Pomegranate, another recent probabilistic graphical model package for Python. Unlike Pomegranate, it natively implements various exact and approximate inference algorithms (see Subsection 3.5.4), like variable elimination, belief propagation and max-product linear programming.

[4] `https://Pomegranate.readthedocs.io/en/latest/`
[5] `https://pygraphviz.github.io`
[6] `https://www.graphviz.org`
[7] `http://pgmpy.org`

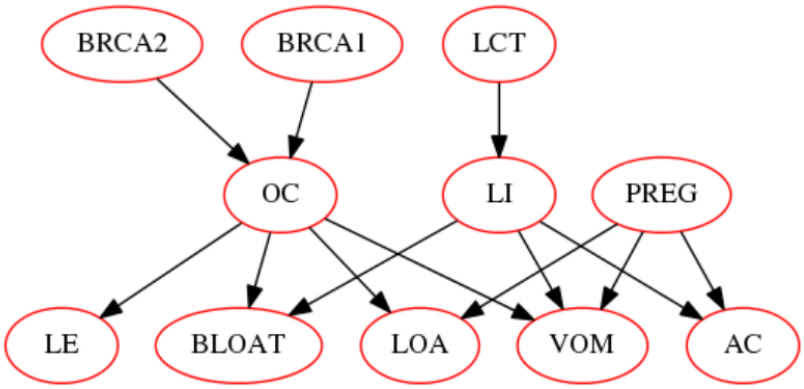



Figure 4.2. Example output of `plot` [Pomegranate tutorial].

The reason that two different probabilistic graphical model libraries were used, is because there is currently no Python package offering all the necessary functionality. Pomegranate implements a novel structure learning algorithm but is severely lacking in functionality in many other areas. Pgmpy, on the other hand, has a very good API as regards inference.

### DAOOPT

*DAOOPT*[8] is an open-source implementation of the sequential AND/OR branch-and-bound algorithm proposed by Marinescu and Dechter [2006]. Search-based algorithms traverse the model's space and are much more efficient in their use of memory, compared to inference-based algorithms such as variable elimination.

DAOOPT builds an AND/OR search space to generate an AND/OR graph that takes advantage of information encoded in the graphical model, namely its independencies. The DAOOPT implementation that was used is an exact solver for finding an MPE solution in Bayesian networks. The software is written in C++ and accessible through a command-line interface; the only required parameter is a `.uai` file representing a Markov random field or a Bayesian network but in most cases an optional `.uai.evid` file, containing the observed evidence, will also be given.

The `.uai` file format is a simple text file used to represent problem instances. Such a file is composed of:

- *preamble*: containing the type of the network (`MARKOV` or `BAYES`), the number and cardinality of variables and the cliques, that in the `BAYES` case are simply the variables appearing in each conditional probability table;
- *function tables*: containing the actual definition of the CPTs i.e., the values of each node given its parents or, in the case of root nodes, the marginal probabilities.

The `.uai.evid` is a very simple file containing the number of variables in the evidence set followed by the index of each variable and its observed value. In both formats the variables and their values are represented only by a numerical index, starting from 0, with the ordering

[8] `https://github.com/lotten/daoopt`

being defined in the preamble of the `.uai` and maintained consistent throughout both the `.uai` and `.uai.evid`.

Following, is the `.uai` representing the network shown in Figure 3.2, which has been the running example throughout the last chapters. Lines starting with `c` are interpreted as comments; these are misinterpreted by DAOOPT and are thus removed when running it, but are here shown for clarity and because they are part of the official `.uai` format. The file starts by stating that the model is a Bayesian network composed of 5 random variables; these will then be referenced by an ordinal index starting at 0. The first variable (index 0) is of cardinality 3, the second (index 1) is of cardinality 2 and so on. We can then see the definition of the cliques or more precisely, as the model is `BAYES` and not `MARKOV`, of the CPTs; there are 5 of these, each one associated to one of the five variables just stated. The first CPT involves 2 random variables: the first (0) and the second (1); the second CPT involves only one variable (1) and this tells us that variable 1 is a root node in the BN's DAG. The ordering is such that the child node is the last in the definition of each CPT's nodes so, for example, in the first CPT we find that variable 1 is the child of variable 0. Finally, we have a complete definition of the function tables/CPTs. The tables are printed so that each row corresponds to the conditional probability value of the child node and increasing rows correspond to increasing enumeration of the parents' states, in the order given when defining the variables involved in the CPTs. The first table corresponds to the CPT of "eta arrotondata", shown in Table 3.2. It contains 6 elements as it involves variables 0 ("mut17q21") and 1 ("eta arrotondata)" that are of cardinality 2 and 3, respectively. So, each row corresponds to the probability distribution of the three states of variable 1, given each of the two states of variable 0.

```
c
c Bayesian network exported from Pomegranate - Thomas Tiotto (2019)
c

BAYES
5
3 2 3 3 2

c
c Cliques
c

5
2 0 1
1 1
2 2 1
3 3 2 4
1 4

c
c CPTs
c

6
```

```
 0.42105263157894735 0.42105263157894735 0.15789473684210523
 0.043798177995795384 0.17063770147161877 0.7855641205325858

2
 0.006613296206056387 0.9933867037939436

6
 0.6842105263157895 0.0 0.3157894736842105
 0.1373510861948143 0.021723896285914507 0.8409250175192712

18
 0.004385964912280701 0.2412280701754386 0.7543859649122807
 0.022598870056497175 0.11864406779661016 0.8587570621468926
 0.10344827586206899 0.41379310344827586 0.4827586206896552
 0.2121212121212121 0.45454545454545453 0.3333333333333333
 0.14094488188976378 0.6362204724409449 0.22283464566929131
 0.289612676056338 0.5677816901408451 0.1426056338028169

2
 0.5315001740341107 0.4684998259658893
```

The following is an example of a randomly generated `.uai.evid` evidence file that simply states that the evidence set has cardinality 2 and contains variable 4 (in the ordering given in the `.uai`) in its state 1 and variable 3 in state 2.

```
 2
  4 1
  3 2
```

Both the `.uai` and the `.uai.evid` were generated by the custom functions presented in Subsection 4.3.2 under the "MPE" header. These are able to export a `Pomegranate` model and randomly generated evidence to the correct input format for DAOOPT.

### Standard Packages

*Pandas*[9] is an extremely widely-used open-source Python library providing data structures and methods to support data analysis. The package excels in the manipulation of tabular data in the form of `DataFrame`, which is the analogous to R's `data.frame`. A `DataFrame` can be seen as a "general 2D, size-mutable structure with potentially heterogeneously-typed columns". The syntax for slicing is very close to R's, as are many other functionalities; this is because one of Pandas' explicit goals is to offer all of CRAN's functionalities and to be easily approachable by anyone already knowing the other language.

Pandas was the default choice for this thesis' implementation because it is the *de facto* standard in data analysis applications when using Python. Its flexibility in reading Excel spreadsheets (the format of the provided data set presented in Section 4.2) and in then manipulating the data confirmed that this was a good choice. Note that the additional `xlrd` package is needed to read files in the Excel format.

[9] `https://pandas.pydata.org/about.html`

*Scikit-learn*[10] aims at providing a unified API for basic machine learning; it does not include advanced paradigms such as reinforcement learning or graphical models for structured learning. The latter omission was the reason that lead to select Pomegranate as the basis for the implementation of the prototype system. What is included are a stack supervised and unsupervised ML tools to prepare data sets, define machine learning models ranging from spectral analysis-based to ensemble methods to clustering and multiple evaluation and model selection utilities.

*NumPy*[11] is another *de facto* standard package when performing scientific computing with Python. Most scientific packages (including Pandas, Scikit-learn and TensorFlow) depend on NumPy for low-level operations; this is because NumPy provides a fast implementation of n-dimensional array objects together with powerful manipulation functions. In addition to this, NumPy implements linear algebra operations, Fourier transform and random number generation. The closest parallel to NumPy - as R was for Pandas, is MATLAB.

*NetworkX*[12] is another widely-used package; it is specialised in the creation and manipulation of graph-structured data. The main use for this package was in building the "knowledge base" structure that the dialogue with the expert is based on.

### 4.3.2 Algorithms

This section is concerned with presenting algorithms and methods that were adapted and used for this thesis, but that were not original work.

#### Model Construction

The data was given in `.xlsx` format and was imported using Panda's `read_excel` function that returned a `DataFrame` object. The imported data was then preprocessed by dropping unwanted records and binning the remaining ones following the instructions outlined in Table 4.3. The actual BN representation is learned at runtime by calling the `from_samples` method of Pomegranate's `BayesianNetwork` to solve the structure learning problem (defined in Subsection 3.5.2).

The binned data was codified into integer representations before being passed to Pomegranate's structure learning algorithm. Thus the network's state names are in natural language but the internal representation of the values of each random variable is an integer number. A dictionary object is used to translate one representation into the other when interacting with the user.

#### D-separation

A naïve implementation to check for d-separation between nodes $X$ and $Y$, according to Definition 3.22, would have a complexity in the order of the number of trails between $X$ and $Y$; this would lead to an exponential running time in the size of the graph's vertices set. Luckily, Koller et al. [2009] present a linear time algorithm to solve the problem, whose complete pseudocode is shown in Algorithm 1.

The `reachable` procedure, as defined in the book, takes as input the DAG representing the Bayesian network $\mathcal{G}$, a source variable $X$ and a set of observed variables $\boldsymbol{Z}$; on exit it returns

[10] `http://scikit-learn.github.io/stable`
[11] `http://numpy.org`
[12] `https://networkx.github.io`

**Algorithm 1** reachable procedure by Koller et al. [2009]

```
1: 𝒢 BN graph
2: X source variable
3: Z observations
4:
5: L = Z                                                    ▷ Phase 1
6: A = ∅
7: while L ≠ ∅ do
8:     Select some Y from L
9:     L = L \ {Y}
10:    if Y ∉ A then
11:        L = L ∪ Pa(Y)
12:    end if
13:    A = A ∪ {Y}
14: end while
15:
16: A = {(X, ↑)}                                            ▷ Phase 2
17: V = ∅
18: R = ∅
19: while L ≠ ∅ do
20:    Select some (Y, d) from L
21:    L = L \ {(Y, d)}
22:    if (Y, d) ∉ V then
23:        if Y ∉ Z then
24:            R = R ∪ {Y}
25:        end if
26:        V = V ∪ {(Y, d)}
27:        if d =↑ and y ∉ Z then
28:            for each Z ∈ Pa(Y) do
29:                L = L ∪ {(Z, ↑)}
30:            end for
31:            for each Z ∈ Ch(Y) do
32:                L = L ∪ {(Z, ↓)}
33:            end for
34:        else if d =↓ then
35:            if Y ∉ Z then
36:                for each Z ∈ Ch(Y) do
37:                    L = L ∪ {(Z, ↓)}
38:                end for
39:            end if
40:            if Y ∈ A then
41:                for each Z ∈ Pa(Y) do
42:                    L = L ∪ {(Z, ↑)}
43:                end for
44:            end if
45:        end if
46:    end if
47: end while
48: return R
```

the set of variables $R$ that are reachable from $X$. The procedure runs in two phases, traversing the graph twice: first bottom-up from leaves to roots, then vice-versa. During the first stage, the algorithm finds all nodes $\boldsymbol{A}$ that are ancestors of the evidence set $\boldsymbol{Z}$. During the second phase, the procedure distinguishes the direction it visits each node in order to determine if it is traversable or not. Any node $Y$ that is not in the evidence set is marked as reachable; if it is being visited in direction "up" ($(Y, \uparrow)$) it can be traversed as the v-structure (see Subsection 3.4.2) is a *chain*. All the parents of $Y$ are marked to be visited in the "up" direction (i.e., "from below") and the converse is done for $Y$'s children. If $Y$ is being visited "from above" ($(Y, \downarrow)$) its children are again added to be visited in the "down" direction, because $Y$ is traversable. Additionally, if $Y$ happened to be in the set $\boldsymbol{A}$, found in the first step, then $Y$'s parents are marked to be visited in the "up" direction because the *collider* is active and $Y$ can be traversed (a collider is open if and only if the central node or any of its descendants are observed).

The implementation in this thesis follows the pseudocode of the book very closely but the procedure `d-separated`, instead of finding all nodes $R$ that are d-connected to the input $X$, only tests if a given target $Y$ is d-separated from $X$ or not, as is shown in Algorithm 2. This gives some extra flexibility in how the function can be used. To find the set $S$ of all nodes d-separated from $X$, `d-separated` is iterated in order to test all nodes $V$ in the BN.

**Algorithm 2** d-separation algorithm

1: $separated\_list = \emptyset$
2: **for** target $Y \in V$ **do**
3: append $d-separated(X, Y, E)$ to $separated\_list$ ▷ will return true or false
4: **end for**

MPE

The solution to the most probable explanation problem (Definition 3.26) can be found by using DAOOPT (described in Subsection 4.3.1 under the "DAOOPT" header) as an external solver. The latest version of DAOOPT was downloaded from the official repository[13] and compiled into an executable. DAOOPT only offers a command line interface so some extra work is needed in order to integrate it with the Python-based application under development. The connection is provided by first writing to stable storage a `Pomegranate.uai` containing the model definition and a `Pomegranate.uai.uai.evid` with the chosen evidence. These files are then fed to DAOOPT by using Python's `subprocess` module, by running the following command in a background shell:

```
./daoopt -f Pomegranate.uai -e Pomegranate.uai.evid
```

The shell output is captured and also written to stable storage, in order for the solution to be parsed from it.

To exemplify the process, we return to the example used while presenting DAOOPT in the relevant Subsection in 4.3.1. Given the `.uai` representing the BN and the `.uai.evid` random evidence:

[13] `https://github.com/lotten/daoopt`

```
 2
  4 1
  3 2
```

DAOOPT would give the following output:

```
--- Starting search ---
[0] u 3 4 -1.3581 5 2 1 2 2 1
[0] Cache statistics: . . . .

--------- Search done ---------
Problem name:  Pomegranate
OR nodes:      3
AND nodes:     4
OR processed:  3
AND processed: 8
Leaf nodes:    2
Pruned nodes:  4
Deadend nodes: 1
Time elapsed:  0 seconds
Preprocessing: 0 seconds
-------------------------------
-1.3581 (0.0438433)

p 2 1 2
l 2 1 6
s -1.3581 5 2 1 2 2 1
```

The end of the final line is the one of interest, as it is the assignment of values to the variables that solves the MPE problem. The `5 2 1 2 2 1` string is to be interpreted as meaning:

- there are 5 variables in the solution
- the variable indexed by 0 (in the ordering given in the preable of the `.uai`) is assigned its third value (the ordering is inferred by the CPTs defined in the `.uai`) in the MPE solution
- variable 1 is assigned its second value
- variable 2 is assigned its third value
- variable 3 is assigned its third value
- variable 4 is assigned its second value

Variables 3 and 4 are constrained to assume the value they were specified with in the evidence `.uai.evid`; in this case the second (1) and third (2), respectively.

All the functionality relating to solving the MPE with DAOOPT is encapsulated by `daoopt_solver`, a function that given the input `.uai` files returns the MPE solution.

## 4.4 Novel Contributions

This section deals with all those algorithms and methods presenting a substantial element of novelty and which are at the core at the work carried out in this thesis.

### 4.4.1 Algorithms

An important part of the work in this thesis was developing the algorithms needed to actually implement the ideas sketched in the paper [Butz et al., 2018] (see Section 2.6). The basic method, which included the construction of the "knowledge base" through a constructive dialogue with the domain expert and the generation of counterfactual explanation branches, was studied, adapted to the current case, and expanded. The current work goes beyond the ideas presented in the paper, as it aims to expand and validate these based on the literature regarding the explainability of Bayesian networks (see Section 2.5). Thus the methods developed in this thesis aim to enable the actualisation of a mix of *graphical*, *textual* and *dialogical* interaction modes and support their subsequent validation in a concrete setting, by real medical domain experts.

#### Dialogues

The so-called "pseudo-MPE" algorithm is inherently wrapped up with the concept of *dialogue* and is central to the explanatory methods being developed in this thesis. The algorithm was formulated as a way of implementing the "MPE branch" of the "argumentative probability tree" hypothesised by Butz et al. [2018]. It was termed "pseudo-MPE" because a review of a literature made it clear that there are no guarantees of it returning the true MPE solution; this can be confirmed by consulting [Koller et al., 2009, pag. 26]. The procedure proposed in the paper chooses the most probable ("strongest dependence") item at each step and adds it to the "MPE probability tree" that is being built, as can be understood by looking at Figure 2.6 and [Butz et al., 2018, sec. 3.1 and 4.3]. The procedure being advocated is equivalent to each variable in the data set "choosing" its most likely value in the current setting and is thus analogous to:

> *[...]* the assignment where each variable individually picks its most likely value can be quite different from the most likely joint assignment to all variables simultaneously. This phenomenon can occur even in the simplest case, where we have no evidence.
>
> [Koller et al., 2009, pag. 26]

An example showing how this may be the case is presented in Subsection 3.5.4. This proves that the true MPE solution is not returned by such a procedure, thus the term "pseudo-MPE".

At a lower level of detail, the algorithm may be broken into:

- a dialogical part, which interfaces with the expert user through the use of natural language, menus and visualisations;
- the part responsible for constructing the "pseudo-MPE branch".

The former procedure was informed and shaped by the results obtained by the methods described in Subsection 4.5.2. At its core the latter part is a greedy procedure whose aim is to select the "best" next *(state,value)* tuple at each step, based on some measure of optimality and

on the variables already in the evidence set. In the actual implemented system the two parts are intertwined, given their close inter-dependence.

The `dialogue` procedure starts by asking the user to select a subset of variables and their corresponding values, to be added as initial evidence. This initial evidence is used to root the MPE branch. It should be noted that in the description given by Butz et al. [2018], the *argumentative probability tree* is a real tree (Definition 3.20) as each node is guaranteed to have at most one parent. The current system, on the other hand, constructs an *argumentative probability polytree* (Definition 3.21) because, as will better be described in Chapter 5, it was seen early on that the users much preferred to be able to start from a set of initial evidence and not be limited to a single one.

The algorithm then proceeds to call the `next_most_probable_states` subroutine tasked with returning an ordered list of *(state,value)* pairs. It does this by calculating the posterior distribution given evidence of all the states not already in the evidence, then calculating the efficiency (Definition 3.13) and the maximally probable symbol of each state's distribution. It then returns the *(state,value)* tuples ordered according to the normalised entropy of the states (see Subsection 3.3.1). The tuple whose state has minimum entropy is thus at the head of the list.

The *(state,value)* pair at the head of the list is proposed to the user who has the faculty to accept the system's proposal or refuse it. If the user accepts it, the proposed tuple is added to the evidence set and to the "pseudo-MPE" branch under construction. Thus, the evidence set's cardinality increases by one each time a user accepts a proposal. The updated evidence will be used to calculate the new list of *(state,value)* pairs at the following round.

If the expert chooses to refuse, then she is iteratively presented with the remaining *(state,value)* items in order of decreasing efficiency. This is detailed in Subsection 4.4.3. Once she accepts one of the explanations given by the system, the `generate_alternative_branch` subroutine is called to automatically generate a maximally probable "pseudo-MPE" branch, rooted in the newest *(state,value)* node of the MPE branch (the algorithm is described in Subsection 4.4.1 under the "Alternative Explanation Branches" header).

The proposal loop for alternative states runs until there are increasingly less probable elements in the list and exits with a partial solution if the user refuses all of them at a given step.

Three slightly different operational modes of the algorithm are implemented. This was done for research purposes, in order to understand which of the three, if any, the expert users would find the best from a standpoint of usability and explainability. Another question was to understand if a combination of their distinctive features would be preferred over any single one:

- *exhaustive*: in the basic dialogue mode, the set of variables under consideration monotonically decreases by one every time the user accepts a system's proposal and the dialogue terminates only when the user has accepted all variables at least once or refused all proposals at a given round. In the first case the user will have the "pseudo-MPE" solution while in the second she will be left with a partial assignment to some of the variables not present in the initial evidence. The pseudocode is shown in Algorithm 3.

- *d-separated*: in the second variant, the set of variables considered at each step is dynamic and depends on the separation properties of the underlying Bayesian network's DAG and on the evidence set constructed by the user's choices.

Differently from the first type of dialogue, an additional `evidence_d_separation` subroutine is called before `next_most_probable_states` to calculate the set of variables that are d-separated from the evidence set, up to that step of the dialogue.
`next_most_probable_states` is then executed but the variables that the previous function found to be d-separated from the evidence are removed from the returned list. This way, variables that can have no effect, given the current evidence, are not proposed.

As the d-separation operation is not monotonic, adding new nodes to the evidence set can both increment or diminish the number of nodes that will be proposed at each subsequent round. D-separation is not "monotonic" because, as can be understood by looking at the definitions of when a v-structure is *closed* (given in Subsection 3.4.2), a *collider* is only closed when its central node is *not* in the evidence set. Thus, if a new node is added to evidence it is liable to *open* some v-structure and this may make previous d-separated nodes become d-connected.

The user is shown an updated view of the independencies of the graph at each step; an example of such an output is shown in Figures 4.5 and 4.6. The pseudocode is shown in Algorithm 4.

- *thresholded*: the final variant of the algorithm prunes the set of variables using a different strategy from the one presented previously. In this case, the *(state,value)* pairs in the list returned by `next_most_probable_states` are dropped automatically based on their probability. Pairs whose probability is below a user-defined threshold or are "worse than random" (for ex. a *(state,value)* tuple will be discarded if $state$ is binary and the probability of $value$ is lower than 0.5) are removed and not proposed to the user. This thresholding strategy based on the probability of the tuples is paired with a threshold on the number of times that the expert can refuse a particular *(state,value)*.

  In the generic dialogue, tuples can be proposed multiple times, with an ever lower probability, if the user has previously refused them; in the thresholded scheme a *(state,value)* pair can only be proposed a maximum number of times before being permanently discarded.

The underlying Bayesian network representing the data set is learned and queried through the `Pomegranate` (see Subsection 4.3.1) API but the great majority of all the code is completely custom-written. This was necessary because `Pomegranate`, while having a powerful backend, was found to be severely lacking in the breadth and flexibility of its API. Many basic operations, such as the calculation of a joint distribution, were not available so the only way was to implement lower-level workarounds while still using `Pomegranate` for the most basic operations, for example for the calculation of a posterior distribution. In particular, `dialogue` is implemented with the only direct calls to the API being when learning the network and when calling `predict_proba`, that queries the `BayesianNetwork` object to calculate the posterior distribution of the states given the current evidence. D-separation, in the second variant of the algorithm, is calculated via the `evidence_d_separation` procedure that implements the pseudocode presented in Algorithm 2.

### Alternative Explanation Branches

The function to generate alternative branches to the main "pseudo-MPE" branch in the dialogue tree is called after the user refuses a *(state,value)* in the dialogue and accepts one of the

alternatives.

The motivation for this functionality is to present the user with a "what-if" analysis, thus helping to formulate a reply to the question: "had I accepted the *(state,value)* presented me by the system, what would the configuration of the remaining *(state,value)* pairs have been?". The question is answered by generating a maximally probable, alternative "pseudo-MPE" sub-branch that is rooted in the last node in the main "pseudo-MPE" branch.

The alternative branch is generated by what is essentially an automated version of `dialogue` that always accepts the first suggestion returned by `next_most_probable_states`, compatibly with the pruning strategy of the main `dialogue`. Given that `generate_alternative_branch` and the `dialogue` are essentially one and the same, the latter inherits the same pruning strategies as the former. That is, `generate_alternative_branch` called during the exhaustive dialogue will generate a maximally likely assignment (remembering that the resulting assignment is *not* the MPE solution) over all variables in $V \setminus E$ while when invoked from one of the other two variant of the dialogue algorithm, it will apply their same pruning strategies.

The implementation of the *pseudo-MPE polytree* is based on the NetworkX Python package (see Subsection 4.3.1). The creation of a chain of nodes is done by keeping a local pointer $alt\_node$ that refers to the last added node or set of nodes, if the node being added is the successor of multiple initial evidence.

An example of the output shown to the user at each step of the dialogue is shown in Figure 4.7 while the pseudocodes for the three variants are shown in Algorithms 6, 7 and 8. Note that $alternative\_evidence$ is local to this algorithm and is separate from the $evidence$ used in the main `dialogue` procedure.

### "Pseudo-MPE" from Initial Evidence

In order to compare the "pseudo-MPE" output with the true MPE solution, an algorithm was implemented that, starting from an initial set of evidence, generated the relative "pseudo-MPE" solution. The random initial evidence set is constructed by stochastically choosing a number in the interval $k = [1, |V|]$, with $V$ the set of vertices in the BN, and then randomly selecting $k$ of the random variables in $V$ to yield the set of variables $E$. The value of each variable is randomly chosen among the set of its values; as all variables are categorical and thus discrete, this is straightforward.

An optional threshold can be set by the user in order to discard *(state,value)* pairs whose probability is deemed too low.

The implementation is based on the NetworkX Python library (see Subsection 4.3.1) since what is being constructed is not a tree but a *polytree* (Definition 3.21), as nodes may have multiple parents. Note that $alternative\_evidence$ is considered separate from the main $evidence$ used in `dialogue`. The pseudocode is shown in Algorithm 9.

### MPE Algorithms Comparison

In order to compare the quality of the solution found using the simple "pseudo-MPE" heuristic, an experiment was set up to compare it to the exact MPE solution.

The user is asked to select the number of iterations over which to average the results and at each one of these a "pseudo-MPE" solution is generated using Algorithm 9. The same initial evidence is fed to DAOOPT using the `daoopt_solver` interfacing procedure (shown in Subsection 4.3.2) and to Pgmpy's `map_query` function.

At each iteration a new starting random evidence is generated; the solutions given by the "pseudo-MPE" algorithm, Pgmpy, and DAOOPT are scored against each other using Hamming (Definition 3.15) and Jaccard (Definition 3.16) distances on the vectors representing the returned assignments. The sequence of distances is then averaged over all the iterations to return the averaged distances between the three solutions to the MPE problem.

The implementation of the algorithm had to deal with the fact that DAOOPT did not converge to a single output for some instances of the input evidence; these instances have been simply discarded and substituted for one based on a different initial random evidence.

#### Pairwise Correlations

An interesting addition, in terms of both explainability and as a *novel contribution*, is the implementation of an algorithm to measure and graphically display the interrelatedness between pairs of variables. The Definition 3.13 of mutual information presented is extended to account for the current state of the model i.e., the set of observed variables.

The mutual information is displayed on the edges of the BN and scales them accordingly, in their thickness, as can be seen in Figure 5.3. To the best of our knowledge, there is no work that has applied this visual explanation approach to Bayesian networks.

The *conditional mutual information* between each pair of parent → child variables is calculated as:

$$I(X,Y|\boldsymbol{E}=\boldsymbol{e}) = \sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} \mathbb{P}(X=x,Y=y \mid \boldsymbol{E}=\boldsymbol{e}) \log_b\left(\frac{\mathbb{P}(X=x,Y=y \mid \boldsymbol{E}=\boldsymbol{e})}{\mathbb{P}(X=x \mid \boldsymbol{E}=\boldsymbol{e})\mathbb{P}(Y=y \mid \boldsymbol{E}=\boldsymbol{e})}\right) \tag{4.1}$$

This is not done if the parent or child variable, in the $X \rightarrow Y$ tuple under consideration, is in the evidence set; this because observing a variable in a BN conceptually corresponds to disconnecting it from its children.

The implementation takes advantage of Pgmpy's inference capabilities. To do this, a function to convert a Pomegranate-based BN to an equivalent Pgmpy-based one had to be written. The queries to the model are then done using the variable elimination algorithm, which is more than suitable for a small BN like the one learned from the provided data set (described in Section 4.2). The marginals $\mathbb{P}(X)$ and $\mathbb{P}(Y)$ for $X$ and $Y$ are calculated directly from the joint distribution $\mathbb{P}(X,Y)$, by marginalising it in turn over the two variables.

The function exits with $-1$ if the parent or child variable is in the evidence set. As mutual information only assumes values in range $[0,1]$, this signals to the calling function to treat this pair of variables differently. The pseudocode for the algorithm is shown in Algorithm 10.

### 4.4.2 Interfacing with the User

The driving design *desideratum* was for the system to be as transparent as possible to the final user. This intent was shared by the Istituto Cantonale di Patologia, whose members felt strongly about not wanting to have to deal with the inner workings and algorithms of the system but be able to concentrate only on the data and the task in hand.

The interactions with the system were designed to take advantage of the best practices identified in the literature review in Section 2.5. Thus, natural language was identified as the most spontaneous way to make the expert user aware of the knowledge in the data set and for her to interact with the system, given that Lacave and Díez [2002] recognise linguistic outputs as one of the most naturally comprehensible output formats for humans. Thus the main

interaction mode with the implemented system is in using natural language, both as input and as output. This is not to say that the authors are claiming that natural language is *the best way to explain the outputs of the system*, but only that it is the *most readily comprehensible output modality*.

Natural language denominators come "for free" when dealing with the node names and their values, regardless of their internal representation in the implemented model, as these are simply the column names and values found in the original input data set. A bit more care needs to be taken when quantifying probabilities to the user; these are in the form of "linguistic probabilities" as suggested by Henrion and Druzdzel [1990]. The format chosen should fit in as seamlessly as possible with the rest of the phrases generated and be able to conform to the user's prior expectations regarding probabilities. It should be both *informative* and *precise*. The chosen coding was taken from Butz et al. [2018] and is shown in Table 4.4. The mapping is used every time there is a need to translate a probability into natural language to be output to the user; for example when proposing tuples during a dialogue (see Subsection 4.4.1), an example of which can be seen in Figure 4.8.

In Figures 4.9, 4.10 and 4.11 we can see how the interface to the system's functions leans heavily on natural language elements. The design ideal was for the system to be as accessible and the least intimidating possible as possible, so that the user may easily be able to extract the maximum amount of information. Not only the terms but also the phrasing were deemed to be important: every interaction is broken down into what are believed to be manageable steps. For example, a user isn't asked to provide all the evidence at once but is asked for one at a time, when necessary. The idea is to reduce cognitive load on the expert's part so that she may be able to better concentrate on the task at hand, not on the tool she is using to achieve it.

Another important aspect is the translation of the information into a language that would make semantic sense to the user; that is, it should be part of the user's *ontology*, meaning the set of concepts part of her *worldview/forma mentis*. An example can be seen in Figure 4.9 where the notion of d-separation in the underlying BN's DAG is remapped to a higher-level concept that is part of the user's ontology: that of variables affecting each other's values or not. The underlying theme of the work carried out in this thesis is to aid in translating *information* present in a data set into *knowledge* for a medical expert user. This is done in various ways, for example through the use of dialogue, in order to bring the data to a form that is manageable and comprehensible.

Lacave and Díez [2002] identify that "the most direct and intuitive way of showing the information embodied in a Bayesian network is to display the corresponding graph". For this reason, visualisations are used extensively throughout all interaction modes, as can be seen in the multitude of figures throughout this chapter and the next.

A final element that was implemented is a coherent use of colours, even in the experimental command-line interface. This may seem secondary, but many of the BN-based systems surveyed by Lacave and Díez [2002] took extensive advantage of colours in their visualisations. In the generated phrases the source variables are always highlighted in *magenta*, the evidence in *green* and the results of the queries in *cyan*. It is hoped that this may guide the user's eye towards the important elements of what she had set out to achieve, thus anchoring her experience and letting the marginal elements fade into the background.

The grammars for the generation of the natural language are presented in Algorithms 11, 12, 13 and 14; these are defined in *extended Backus-Naur form* (EBNF).

One of the most important parts of this work is the validation of the system's ability to interact with real medical expert users, the results of which are presented in Chapter 5. This is done in order to address one of the main gaps identified in the current literature: the lack of

*application-grounded evaluations* (see Section 2.4).

### 4.4.3 Entropy-Based Selection

The criterion for the selection of the strongest dependencies while constructing the "knowledge base", as defined by Butz et al. [2018] and described in Section 2.6, was chosen not to be the probability, but the *normalised entropy* of each item. In the case of the methods currently under examination, such items are *(state,value)* pairs with the *value* being the most probable among the categorical *state*'s values. For example, the tuple for the state shown in Figure 4.3 would be *(X, b)*. Among all available *(state,value)* pairs, the one whose *state* has lowest normalised entropy is chosen as the strongest dependency to be added to the "knowledge base".

If states are of the same cardinality, then the one selected will be the one presenting a value with a probability higher than that of all other states in the remaining variables. Within each state, the most probable value (the one with highest probability) is the one chosen to be paired in the tuple *(state,value)*. The state selected will be the least entropic; this state will roughly correspond to the one with the highest "probability peak" on one of its values. Thus, in the case where variables have the same cardinality, selecting the most probable *(state,value)* pair based on the *values*' probabilities or on the normalised entropy of the state gives the same result. Using normalised entropy as the state selection criterion, thus factoring out the variable cardinality, makes this result applicable to states with a dishomogeneous number of values.

This normalised entropy-based selection method, that a review of the literature hasn't surfaced as having been used either within or outside the xAI community, is justified by the intuition that simply selecting based on the probability of the values results in a bias against higher cardinality variables. Only selecting based on the absolute probability of the states is also taking into consideration less information than a choice made by also looking at other characteristics of the random variable. Assuming that the underlying process generating the variables is similar, for example Gaussian, then a variable with higher cardinality will see its probability mass more spread out i.e., it will be less "dense". Thus it will be less probable for a high cardinality variable to present a state with a "probability peak" high enough to see it selected. An example of this phenomenon can be seen by comparing the distribution of the variable displayed in Figure 4.4 with that of Figure 4.3.

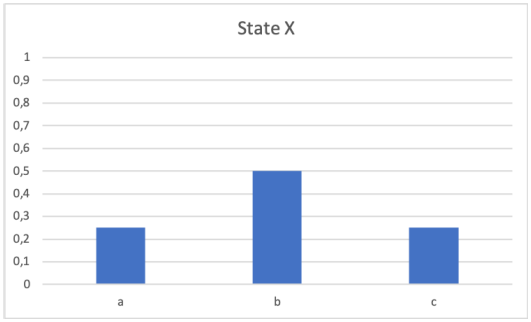



Figure 4.3. Distribution of state $X$ with possible values a, b and c.

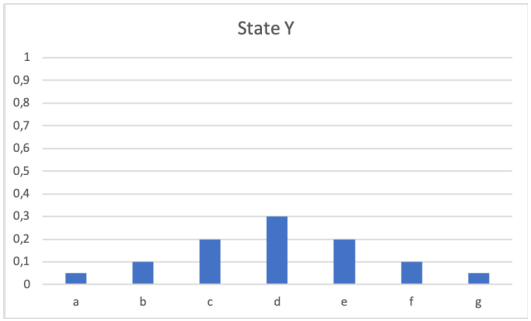



Figure 4.4. Distribution of state $Y$ with possible values a, b and c.

**Algorithm 3** Exhaustive dialogue algorithm

```
1:  evidence = user selected (state,value) tuples
2:  MPE_polytree = MPE Polytree rooted in evidence
3:  while True do
4:      mpe_states = next_most_probable_states(evidence)
5:      if mpe_states is not empty then
6:          next_state = head of mpe_states
7:          propose next_state to user                    ▷ the least entropic state
8:          if the user refuses next_state then
9:              for alternative_state in mpe_states \ next_state do
10:                 propose alternative_state to user     ▷ the next least entropic states
11:                 if the user accepts alternative_state then
12:                     call generate_alternative_branch() on MPE_polytree
13:                     add alternative_state to MPE_polytree
14:                     evidence = evidence ∪ alternative_state
15:                 else
16:                     continue
17:                 end if
18:             end for
19:         else
20:             add next_state to MPE_polytree
21:             evidence = evidence ∪ next_state
22:         end if
23:     else
24:         return
25:     end if
26: end while
```

**Algorithm 4** Independencies dialogue algorithm

```
1: evidence = user selected (state,value) tuples
2: MPE_polytree = MPE Polytree rooted in evidence
3: while True do
4:     separated = evidence_d_separation(evidence)      ▷ based on evidence of previous step
5:     mpe_states = next_most_probable_states(evidence)
6:     mpe_states = mpe_states \ separated
7:     if mpe_states is not empty then
8:         next_state = head of mpe_states
9:         propose next_state to user                    ▷ the least entropic state
10:        if the user refuses next_state then
11:            for alternative_state in mpe_states \ next_state do
12:                propose alternative_state to user     ▷ the next least entropic states
13:                if the user accepts alternative_state then
14:                    call generate_alternative_branch() on MPE_polytree
15:                    add alternative_state to MPE_polytree
16:                    evidence = evidence ∪ alternative_state
17:                else
18:                    continue                          ▷ go to next proposal
19:                end if
20:            end for
21:        else
22:            add next_state to MPE_polytree
23:            evidence = evidence ∪ next_state
24:        end if
25:    else
26:        return
27:    end if
28: end while
```

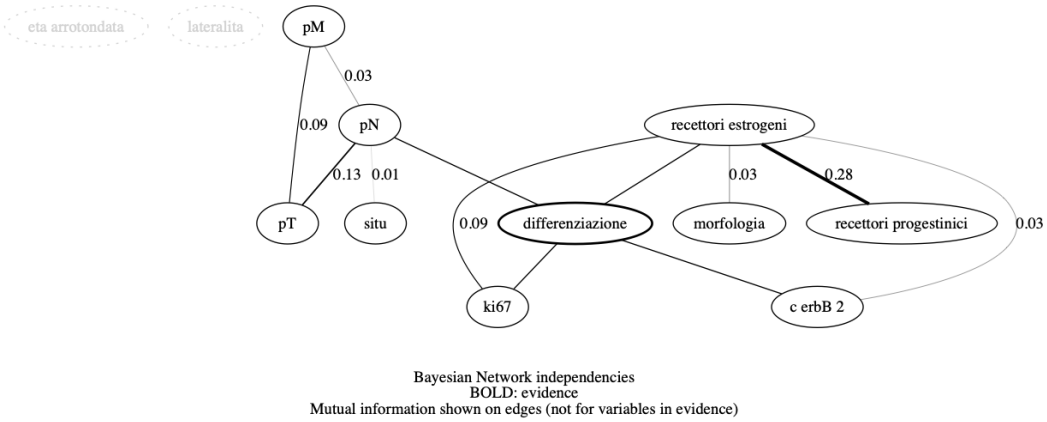



Figure 4.5. Example output during the first round of the d-separation-aware variant of `dialogue`. The variable "differenziazione" is the initial evidence.

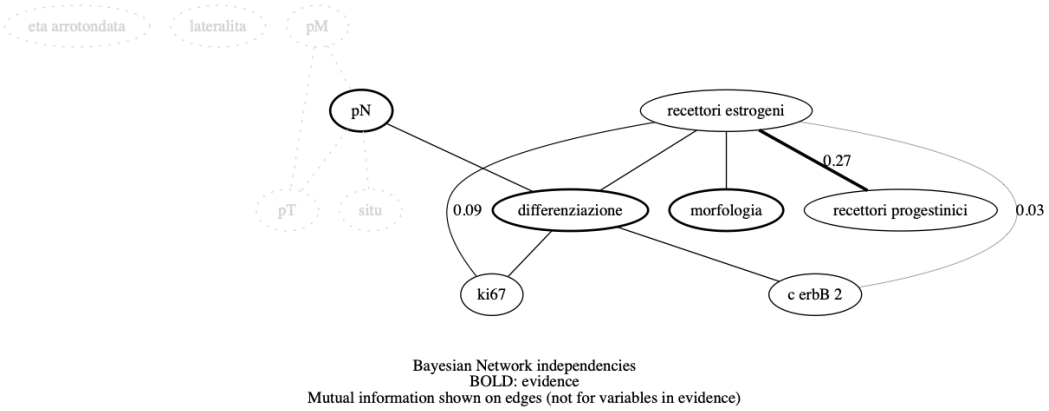



Figure 4.6. Example output during the third round of the d-separation-aware variant of `dialogue`. "pN" and "morfologia" are added to the evidence set and this makes a part of the network redundant.

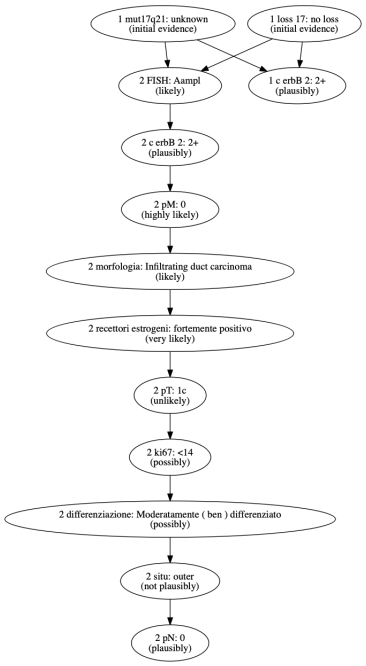



Figure 4.7. Example output during the d-separation-aware variant of `dialogue`. The tuple ("FISH","Aampl") was proposed but the expert refused it and accepted the alternative ("c erbB 2","2+"). The main "pseudo-MPE" branch has ID 1 while the "what-if" one has ID 2.

**Algorithm 5** Thresholded dialogue algorithm

```
user selected refuse_bound
refuse_thresholds = ∅
lower_thresholds = ∅
for v ∈ V do
    for value k of v do
        element [v, k] = 0 in refuse_thresholds
    end for
    element [v] = 1/|v| in lower_thresholds                ▷ refuse worse than random pairs
end for
evidence = user selected (state,value) tuples
MPE_polytree = MPE Polytree rooted in evidence
mpe_states = next_most_probable_states(evidence)
while True do
    for s ∈ mpe_states do
        if probability of s < lower_thresholds[s] ∧ refuse_thresholds[s] > refuse_bound
    then
            remove s from mpe_states
        end if
    end for
    if mpe_states is not empty then
        next_state = head of mpe_states
        propose next_state to user                          ▷ the least entropic state
        if the user refuses next_state then
            increment refuse_thresholds[next_state]
            for alternative_state in mpe_states \ next_state do
                propose alternative_state to user           ▷ the next least entropic states
                if the user accepts alternative_state then
                    call generate_alternative_branch() on MPE_polytree
                    add alternative_state to MPE_polytree
                    evidence = evidence ∪ alternative_state
                else
                    increment refuse_thresholds[alternative_state]
                    continue
                end if
            end for
        else
            add next_state to MPE_polytree
            evidence = evidence ∪ next_state
        end if
    else
        return
    end if
end while
```

**Algorithm 6** Exhaustive alternative explanation branch algorithm

```
1: alternative_evidence = evidence
2: alt_node = last node in the main MPE Polytree
3: branch_id = branch_id + 1
4: while True do
5:     mpe_states = next_most_probable_states(alternative_evidence)
6:     if mpe_states is empty then
7:         return
8:     else
9:         next_state = head of mpe_states
10:        create next_state node, tag it with branch_id and make it son of alt_node
11:        update alt_node node to be next_state node
12:        alternative_evidence = alternative_evidence ∪ next_state
13:    end if
14: end while
```

**Algorithm 7** Independencies alternative explanation branch algorithm

```
1: alternative_evidence = evidence
2: alt_node = last node in the main MPE Polytree
3: branch_id = branch_id + 1
4: while True do
5:     separated = evidence_d_separation(alternative_evidence)
6:     mpe_states = next_most_probable_states(alternative_evidence)
7:     mpe_states = mpe_states \ separated
8:     if mpe_states is empty then
9:         return
10:    else
11:        next_state = head of mpe_states
12:        create next_state node, tag it with branch_id and make it son of alt_node
13:        update alt_node node to be next_state node
14:        alternative_evidence = alternative_evidence ∪ next_state
15:    end if
16: end while
```

**Algorithm 8** Thresholded alternative explanation branch algorithm

```
1: alternative_evidence = evidence
2: alt_node = last node in the main MPE Polytree
3: branch_id = branch_id + 1
4: while True do
5:     for s ∈ mpe_states do
6:         if probability of s < lower_thresholds[s] ∧ refuse_thresholds[s] > refuse_bound
   then
7:             remove s from mpe_states
8:         end if
9:     end for
10:    mpe_states = next_most_probable_states(alternative_evidence)
11:    if mpe_states is empty then
12:        return
13:    else
14:        next_state = head of mpe_states
15:        create next_state node, tag it with branch_id and make it son of alt_node
16:        update alt_node node to be next_state node
17:        alternative_evidence = alternative_evidence ∪ next_state
18:    end if
19: end while
```

**Algorithm 9** pseudo-MPE from inital evidence algorithm

```
1: evidence defined by user
2: threshold defined by user
3: MPE_polytree = MPE Polytree rooted in evidence
4: last_node = evidence
5: alternative_evidence = evidence
6: while True do
7:     mpe_states = next_most_probable_states(alternative_evidence)
8:     if mpe_states is empty then
9:         return
10:    else
11:        next_state = head of mpe_states
12:        if probability of next_state ≤ threshold then
13:            return constructed MPE_polytree
14:        end if
15:        create next_state node and make it son of last_node
16:        update alt_node node to be next_state node
17:        alternative_evidence = alternative_evidence ∪ next_state
18:    end if
19: end while
```

**Algorithm 10** Mutual information algorithm

$X$ parent variable in the BN DAG
$Y$ child variable in the BN DAG
$E$ set of current evidence in the BN
**if** $X \in E$ **then**
    return -1
**end if**
$joint = \mathbb{P}(X, Y \mid \boldsymbol{E} = \boldsymbol{e})$
$Y\_marginal =$ marginalise $joint$ over $X$
$X\_marginal =$ marginalise $joint$ over $Y$
$mutual\_information = 0$
**for** y in $Y\_marginal$ **do**
    **for** x in $X\_marginal$ **do**
        $j =$entry in $joint$ corresponding to $y$ and $x$
        **if** $j$ is 0 **then**
            $mutual\_information += 0$
        **else**
            $mutual\_information = j * \log(\frac{j}{y*x})$
        **end if**
    **end for**
**end for**
return $mutual\_information$

**Algorithm 11** Grammar generating dialogue output

Next_state := 'var 1' | ... | 'var n'
Value := 'val 1' | ... | 'val k'
Probability := 'highly unlikely' | 'very unlikely' | 'unlikely' | 'not plausibly' | 'plausibly' | 'possibly' | 'likely' | 'very likely' | 'highly likely' | 'certain'
Output := 'The next most probable state is ' Probability Next_state 'Enter to accept or n to refuse: ' (Still_to_explain ',')* | 'Is state ' Next_state ' with value ' Value ' more correct?'
Still_to_explain := 'var 1' | ... | 'var n'

**Algorithm 12** Grammar generating independencies query output

Source := #user input#
Evidence := 'var 1' | ... | 'var n'
Separated := 'var 1' | ... | 'var n'
Output := 'Given source variable' Source 'and given evidence: ' (Evidence ',')* 'the following variables have no effect: ' (Evidence ',')*

**Algorithm 13** Grammar generating conditional probability query output

Source := #user input#
Evidence := 'var 1' | ... | 'var n'
Output := 'Given target variable' Source 'and observed evidence: ' (Evidence 'with value: ' Value)* 'then the predicted values for ' Source 'are: ' (Evidence 'with value: ' Value)*

**Algorithm 14** Grammar generating MPE query output

1: Source := #user input#
2: Evidence := 'var 1' | ... | 'var n'
3: Value := 'val 1' | ... | 'val k'
4: Output := 'Given observed evidence:' (Evidence 'with value: ' Value)* 'the most probable configuration of the other variables is: ' (Evidence 'with value: ' Value)*

Table 4.4. Probability quantifiers in natural language

| Probability range | Natural language quantifier |
|---|---|
| (0, 0.2) | "highly unlikely" |
| (0.2, 0.3) | "very unlikely" |
| (0.3, 0.4) | "unlikely" |
| (0.4, 0.5) | "not plausibly" |
| (0.5, 0.6) | "plausibly" |
| (0.6, 0.7) | "possibly" |
| (0.7, 0.8) | "likely" |
| (0.8, 0.9) | "very likely" |
| (0.9, 1) | "highly likely" |
| (1) | "certain" |

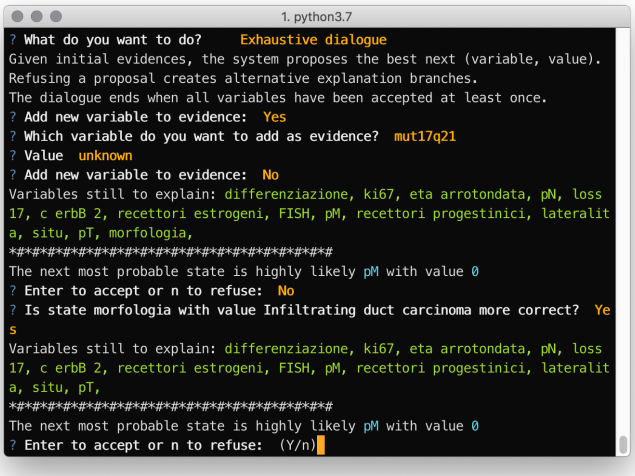



Figure 4.8. Interface while executing a dialogue.

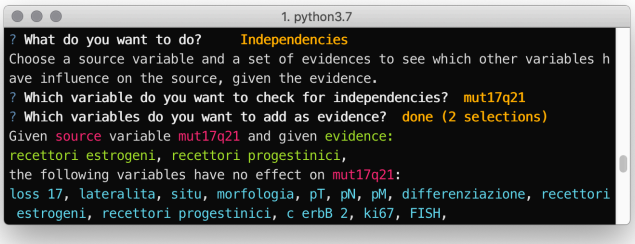



Figure 4.9. Interface while executing a query on the d-separations.

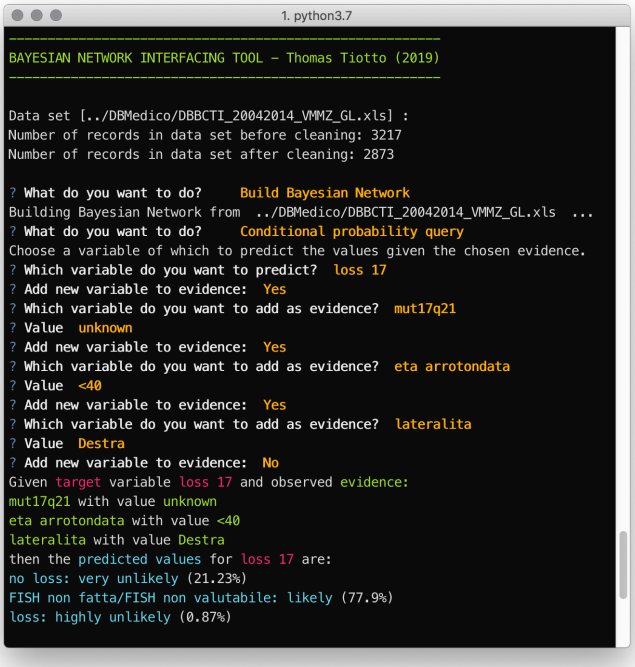



Figure 4.10. Interface while executing a conditional probability query.

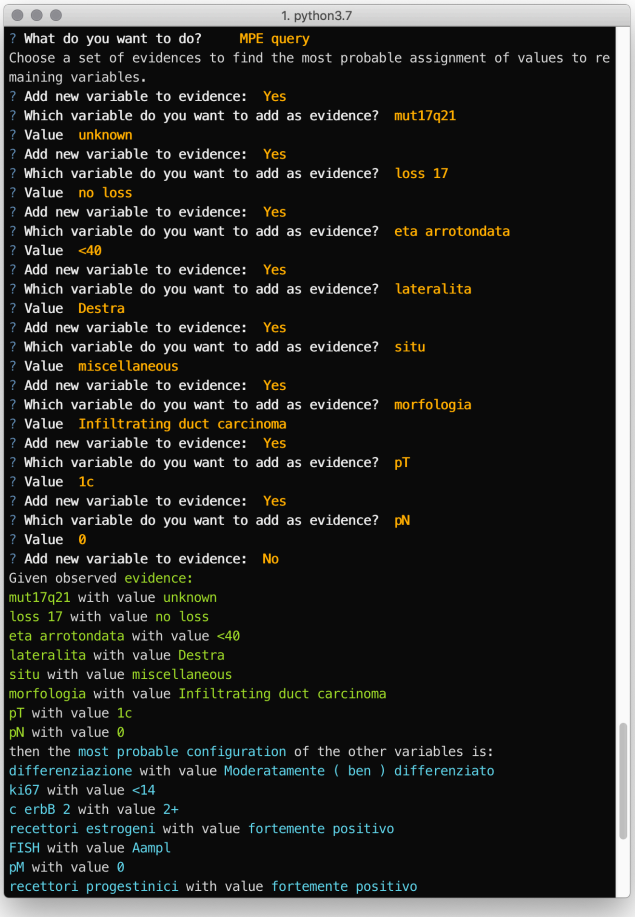



Figure 4.11. Interface while executing an MPE query.

## 4.5 Validation Methodology

Having direct access to expert pathologists has not only helped in guiding research into the theoretical explainability properties of the system but also enabled their *application-grounded evaluation* (see Section 2.4). There are two main validation points of view to be addressed: the clinical (Subsection 4.5.1) and the explainability (Subsection 4.5.2), with the results of the latter depending on part on those of the former.

### 4.5.1 Clinical Validation

A validation of the methods carried out in this thesis in their adherence to established clinical literature is of paramount importance. A failure on the Bayesian network's part in capturing the true relationships between the variables would hamper it in being able to give any meaningful representation of them. For the experts to even start to trust the system or to be able to make sense of its outputs, it is vital that there be as little cognitive dissonance between their basic beliefs and expectations and those that they see represented in the system.

For this reason, the initial validation phase with the ICP concentrated on the clinical aspect. The methodology chosen to clinically validate the system was for the ICP to formulate a series of natural language queries; each one of these questions was annotated with the queried variable and its value, together with the values of any evidence variables. The experts included the expected reply to the queries together with its likelihood, based on the latest medical literature and their personal, knowledge-based expertise. These questions can be abstracted as:

"Given that the value of $var_1$ is $a_1$ and ... and the value of $var_n$ is $a_n$, what is the probability that $var_{n+1}$ takes value $a_{n+1}$?".

The natural language questions formulated by the ICP can be classified along two axes:

- based on their intended purpose: *validation* vs. *research*. The former questions' replies are known from established clinical literature and are the queries that will actually be used to validate the system from a clinical point of view. The latter are queries that don't have a definite clinical answer but that are nonetheless extremely interesting in helping to understand the types of questions a domain expert may want to ask the system.
- based on the way they may be answered: by a *conditional probability query* (Definition 3.6), a *d-separation query* (Definition 3.22) or an *MPE query* (Definition 3.26).

The complete series of thirty questions has been organised according to the second criterion. Appendixes B.1 and B.2 present fourteen questions that can be answered by conditional probability queries. Appendix B.3 shows a series of eight natural language questions that can be answered by running a d-separation query. Appendix B.4 presents five questions that could be answered by a conditional probability query but also, at a higher level, by a d-separation query. This is because what is being asked, is basically wether changing the value of the evidence variable has an influence on that of the target variable. This could be answered by running multiple conditional probability queries and comparing the resulting target variable values or, more simply, by checking if the target and evidence variables are d-connected or not. The first method would give a finer grained answer as it would also *quantify* the magnitude of the effect of one variable on the other; checking for d-separation would only give a *qualitative* answer, which may nonetheless be sufficient. Finally, Appendix B.5 shows three questions that are naturally mapped onto a query of the MPE type.

Most importantly at this stage, all questions can be implemented on the proof of concept system and consequently this shows a good coverage on the tool's part of the use cases that can be imagined by a domain expert. If the system can, in principle, answer every question imagined by the expert then this is an indication that it conforms to her *worldview* and thus could be well positioned to interact fruitfully with her.

The questions marked as *validation* will be posed to the system, in autonomy, by the ICP's representatives, who will then compare the outputs with the result they would have expected, based on established medical literature and their expertise. The columns containing the experts' expected results and their comments have been omitted from the natural language questions shown in Appendix B and included directly in the discussion of the results in Subsection 5.3.2, alongside the system's outputs. If the system's outputs conform to the experts' preconceived ideas in a high number of cases (as confirmed by the experts themselves) then the system can be said to have been *clinically validated*. This is important because the enabling condition for the user to trust the predictions made by the software is that these shouldn't be in strong discordance with her existing beliefs. Not having a strong *cognitive dissonance* is a *necessary* - but not sufficient - condition to enable trust and therefore explainability.

### 4.5.2 Explainability Validation

In general, there is strong resistance to novelty in the field of medicine, both for ethical reasons and because of the need for clinicians to be conservative in attending to established best practices in the field. Any tool that is too onerous in terms of time and cognitive load is liable to remain underutilised. In this field, *a tool must therefore only be the means by which a question is answered*, not itself become a question; the methods developed in this thesis aim to conform to this objective, barring the experimental nature of the software and the consequent lack of refinement of its interface. The need for a comprehensible and efficient tool is especially present because the goal of a pathologist is to arrive at a diagnosis, containing the elements useful to define prognosis and therapeutical approach, in the briefest time possible. The main reasons are ethical, since for a patient waiting for a report is extenuating, and clinical, because a timely diagnosis is the first factor at the base of life expectancy. Obviously, the highest possible accuracy is always strived for. The clinical field and that of biomedicine are forced to embrace uncertainty, as this is an integral part of their practice. Consequently, any tool able to support in comprehension and decision-making is automatically useful, once it has been clinically validated; in other words, even though a specific system may not be decisive or applicable to all reviewed cases, it will nonetheless be taken into account.

Thus, a system validated in terms of its adherence to clinical literature could then also meaningfully be validated from an explainability point of view. The main question to be addressed is its capacity to relate to the expert user. Is the system able to engender the user's trust? In doing so, is she able to extract more knowledge from existing data when using the system than not? Especially in cases where there may be a dearth of data, can the expert maximise the benefit from the available information? Does the user subjectively feel that the system may positively impact her work? These are all hard questions to answer, as there is a very high degree of subjectivity involved. Thus to attempt to answer them, the chosen methods were borrowed from the social sciences.

In an earlier stage, the experts were introduced to the system in prototype form and instructed on the use cases it offered. This process would enable the collection of feedback on the functionalities of the system and help in shaping its subsequent design.

The finalised system was, in a later phase (early August 2019), provided to the experts at the ICP for use in their daily work. To quantify the performance of the system, as perceived by its users in a real setting over an extended period of time, a follow-up was done after three weeks by way of an "explainability evaluation questionnaire", designed to test the gaps identified in Chapter 2. The full questionnaire can be found in Appendix C.

The "explainability evaluation questionnaire" presents five sections:

- *confidence*: aimed at assessing wether the use of the system incremented the confidence the clinician felt in making her decisions;
- *features*: to understand in more detail which interaction modes were perceived as most useful and the subjective reasons for this. Of particular interest is the understanding of the perceived quality of the dialogical interaction modes and of the "pseudo-MPE" query;
- *time*: questions focusing on the the temporal element, mainly the time needed to understand various explanations offered by the system. This element is often overlooked in the relevant xAI literature (see Section 2.5);
- *tool*: general questions regarding the use of tool and if any important use-case was felt to be missing;
- *clinical*: investigating if the tool was clinically relevant in day-to-day work. Unlike the clinical validation presented in Subsection 4.5.1, these questions investigate *a posteriori* the use of the tool and as such should provide a broader evaluation of its clinical relevance;
- *satisfaction*: simple question asking to rate the general satisfaction with the proof of concept system.

As discussed throughout Chapter 2 and summarised in Section 2.7, one of the main gaps in the field of explainable AI is the absence of real-world validation of the - supposedly - explainable models. The objective of the questionnaire is to act as an *application-grounded evaluation*, in the taxonomy proposed by Doshi-Velez and Kim [2017] and presented in Section 2.4, and thus provide what is considered the gold standard for the evaluation of a machine learning system. Also included, since it is almost always neglected in literature, is a focus on the *temporal element* of the explanations that was noted as important by Gilpin et al. [2018]. Of particular interest is evaluating the Bayesian network - underlying the tool's capabilities - in its capacity to surface cogent explanations for the target user; the questionnaire inflects the questions in order to identify which particular characteristics of the system and BN were perceived by the user as the most useful in order to gain an understanding of the underlying data set. As noted in Section 2.5, by acknowledging the psychological characteristics of an explanation identified by Miller [2018], explanations have various essential characteristics that seem to also be inherent in BNs; the questionnaire thus seeks to understand if these are actually present and perceived as useful, in the sense of enabling explainability, by the domain experts.

The questionnaire is not the only source of the results relating to the *application-grounded evaluation* of the developed system; similarly to [Stumpf et al., 2009] in their "think-aloud experiment", many results and details throughout Chapter 5 will be the outcome of observing and listening to the expert users while they were engaging with the system. We refer to these as "informal explainability evaluation results" contrasting them with the "formal explainability evaluation results" that will be the outcomes of the questionnaire.

## 4.6 Summary

This chapter has presented a series of methods whose aim is to enable the creation of a proof of concept software tool inspired by the paper "Explaining the Most Probable Explanation" [Butz et al., 2018] and to enable the validation of this system from the point of view of its explainability. This evaluation with expert pathologists at the Istituto Cantonale di Patologia, whose results will be presented in Chapter 5, aims at validating the methods proposed by Butz et al. [2018] together with newly proposed ones, BNs' explainability in general and to act as a methodological framework for future work. This is important because, as discussed in Chapter 2, the lack of evaluation of explainability is one of the main gaps present in xAI literature.

The chapter opens by presenting the data set that was supplied by the ICP and how this was integrated into the system and used to learn the structure and conditional probability tables of the Bayesian network, which was implemented using the *Pomegranate* Python package. A series of algorithms - based on standard techniques - useful to learn the BN from the data, to calculate the *d-separation* between sets of nodes in the BN and to use the external solver DAOOPT to calculate the optimal solution to the MPE problem have been presented.

After these, a set of novel algorithms constituting the core of this thesis are proposed: three variants of the *dialogue* - which in its basic form is the realisation of the interaction method proposed by Butz et al. [2018] - together with a procedure to generate alternative explanation branches to the "knowledge base" when - and if - the expert user dissents with the system on a proposal. The remaining two algorithms are connected to evaluating the solutions found by the "pseudo-MPE" algorithm as compared to the true MPE.

Then, the rationale relating to the user interface has been presented together with the boilerplates in *extended Backus-Naur form*, which are used to generate the natural language outputs of the system. Regarding interfacing with the user, the method for calculating and displaying pairwise *mutual information* between the variables in the BN is also introduced.

The final topic discussed is the evaluation methodology, which will be used to assess the proof of concept system. This evaluation will consider both the tool's adherence to clinical literature i.e., its capacity to give outputs coherent with the experts' beliefs, and its explanatory powers i.e., its capability to explain its outputs to the users of the ICP and to support them in their daily work.

# Chapter 5

# Results

## 5.1 Introduction

At a high level, the overall objective of this thesis has been to investigate the *explainability* of a medical AI system. To this end, a proof of concept system based on a Bayesian network model was developed whose methods were inspired by the paper [Butz et al., 2018] The result is a tool geared towards addressing a key gap in existing xAI research: the lack of application-grounded validations of explainability of ML systems, as discussed in Sections 1.3 and 4.1. These are: to substantiate the claims made in [Butz et al., 2018] regarding the explainability of their method, to extend this method to better explore the space of BN explainability and thus investigate the effectiveness of Bayesian networks in providing explanations to expert users. An additional objective is also to set a methodological precedent for an *application-grounded evaluation* of an AI system in the medical domain. The methods, algorithms and tools underlying the developed system were presented in Subsections 4.3.2 and 4.4.1 while the design rationale was explained in Subsection 4.4.2.

This chapter will present the experimental results obtained through the implementation of the proof of concept system, informing on both the clinical significance of the results (as described in Subsection 4.5.1) as well as on the tool's capacity to explain the data set and meaningfully interact with the expert medical user (see Subsection 4.5.2). A successful clinical evaluation of the system, undertaken in collaboration with the medical professionals of the Istituto Cantonale di Patologia (see Section 4.2), is of paramount importance as this is one of the prerequisites for the effectiveness of the system and the user perception of its trustworthiness. Intuitively, there is little hope for a meaningful interaction between man and machine if the former distrusts or does not believe in the latter. The evaluation of the explainability of the system - and, as a consequence, of both Bayesian networks at large and of the method proposed in [Butz et al., 2018] - has also been carried out together with the ICP, by using methods akin to those of the social sciences. This approach was necessitated by to the nature of the research, as an *application-grounded evaluation* (see Section 2.4) process involves humans to a high degree. As identified in Chapter 2, it is important for the field of explainable AI to borrow methods from outside computer science, statistics and mathematics in order to meaningfully validate the explainability of ML models.

The tool was compared to the concepts identified in Lacave and Díez [2002] (see Section 2.5), as this review was deemed to be the most complete present in the literature. These con-

cepts can easily be seen as defining a *taxonomy* for the classification of explainability approaches to Bayesian networks. As such, the concepts of the taxonomy will be the frame of reference that the software's functionalities will be compared against. The system was developed in order to explain all three of the elements that the authors identified as needing clarification in a BN: the *evidence*, the *model* and the *reasoning*. The explanation for the *reasoning* and the *evidence* are at the core of the methods of this thesis, as the "dialogues" (see Subsection 4.4.1) are essentially a way of approximating the MPE problem, which the authors identify as the means to explain the *evidence*. The *model* is explained by means of both *graphical* and *linguistic* descriptions. The description of the *reasoning* is also achieved by the "dialogues" and by the "pseudo-MPE" method of interaction, as their aim is to make the "line of thought" of the system clear to the user. Every interaction mode is also compared with the characteristics that Miller [2018] identified as inherent to an explanation, from a psychological perspective; that is, explanations should be: *contrastive*, *selected*, *causal* and *social*.

The chapter is organised as follows:

- Section 5.2 presents the developed proof of concept system in detail from a user-centric point of view. Each subsection is dedicated to describing a user interaction mode and integrated throughout these are explainability and clinical significance remarks that were observed first-hand or collected informally from the ICP; that is, these results were not collected through the use of the questionnaire described in Subsection 4.5.2. Every user feature is described in terms of the taxonomy presented by [Lacave and Díez, 2002] and the characteristics identified by Miller [2018].
- Section 5.3 reports the formal results of the clinical and explainability validation of the developed system, based on the methods presented in Section 4.5.
  - Subsection 5.3.2 presents the results of the clinical validation of the system by outlining the results to the questions described in Subsection 4.5.1 and by developing a discussion on the significance of the results.
  - Subsection 5.3.3 exhibits and discusses the results of the explainability validation of the developed system by means of analysing the answers given to the questionnaire presented in Subsection 4.5.2.
- Section 5.4 shows the results of the comparison between the "pseudo-MPE" (see 4.4.1 under the "pseudo-MPE from Initial Evidence" header) and the true MPE (Definition 3.26) solution, from a quantitative point of view. The qualitative evaluation by the medical experts has been included in the questionnaire (see Appendix C).
- Section 5.5 discusses the issues that presented themselves during the implementation of the methods of this thesis, together with the solutions and workarounds found to address them.

## 5.2 Implemented Tool

### 5.2.1 Overview

The system referenced in Section 4.4 and especially in Subsection 4.4.2, is a proof of concept terminal-based software tool that was developed in order to test the hypotheses referenced in

the Introduction to this chapter and laid out in Sections 1.3 and 4.1. The whole software tool has been made freely available for research purposes[1]. The major goal in the creation of such a tool is to have a working software that could be given to a number of clinicians at the ICP (see Subsection 4.2.1) in order to carry out the research program defined in Section 1.3, which is a response to the gaps identified in Chapter 2. This being only a prototype, the implementation was carried out using Python, as this was the language that enabled the best focus on rapid development, due to its familiarity and to its vast array of available libraries.

Despite never having been intended to be production software, particular care was taken in the design of the interfacing methods, as described in Subsection 4.4.2, in line with the spirit of this work that is to study human-machine interaction.

In the following section, the various interaction modes that were developed are presented using screenshots. The basic methods underlying the software tool have already been discussed at length in Section 4.4 so the current examination will focus on the user interface and how these methods have been incorporated into the system. Where relevant, the information and descriptions given in Section 4.4 will be integrated.

Figure 5.1 shows the initial screen presented during use. The user can input the path to the data set to use or can accept the hardcoded one, which in this case is the one described in Section 4.2. Next, the number of entries before and after preprocessing are shown; the data set in question sees its number of valid records go from 3217 to 2873, after the rules summarised in Table 4.3 have been applied. The "Inspect data set" and "ML" options are only for testing purposes; the former surfaces a pair of options to visualise the distribution of the data set's variables' values and their normalised entropies, the latter runs a series of tests that will not be discussed.

The user-oriented section of the software is the one accessed by selecting "Build Bayesian Network"; this is where all the methods discussed in Chapter 4 are to be found and will be the object of the present evaluation. Selecting this option automatically uses the Pomegranate package to construct a Bayesian network model using the previously selected data set. The user is then shown the main menu of the application, as can be seen in Figure 5.2.

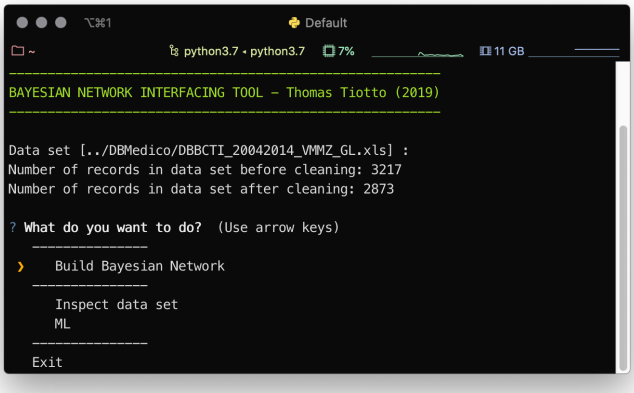



Figure 5.1. Initial screen in the developed tool.

[1] https://github.com/Tioz90/Bayesian-Networks-Explainability-Tool

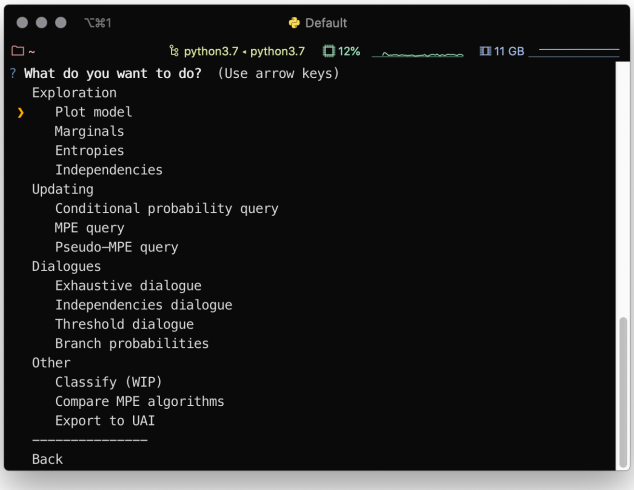



Figure 5.2. Main interaction menu.

### 5.2.2 Plot Model

The "Plot Model" interaction mode would be an example of a *static*, *graphical* explanation in the framework defined by Lacave and Díez [2002], aimed at *explaining the model*. Compared to the characteristics of an explanation identified by Miller [2018], this might be erroneously regarded as a *causal* explanation. Yet, it is important to remember that the directed graphs underlying a Bayesian network are not necessarily describing causal relations, but only probabilistic dependencies.

The "Plot model" interaction mode gives the expert an overview of the variables present in the system and their relationships by displaying the underlying BN's DAG, with the directionality of edges removed for the reasons explained in 5.2.8. Apart from the DAG, mutual information (Definition 3.13) between every pair of connected variables is shown on the edges in order to help the expert gauge the strength of the connection.

The users at the ICP considered this interaction modality a good solution to immediately visualise all the features of the data set at a high level together with their relationships; i.e., it gave the user a sense of the *context* of the data set at hand. The scaling of the thickness of an edge in a manner proportional to the mutual information of the variables it connects was also considered useful in helping to appreciate the varying strength of the correlations between clinical variables.

The output for the data set presented in Section 4.2 in shown in Figure 5.3.

### 5.2.3 Independencies

The "Independencies" interaction mode would be an example of a *static*, *linguistic* and *graphical* explanation in the framework defined by Lacave and Díez [2002] aimed at *explaining the model*. Compared to the characteristics of an explanation identified by Miller [2018], it could be seen as possessing the *selected* and *causal* elements.

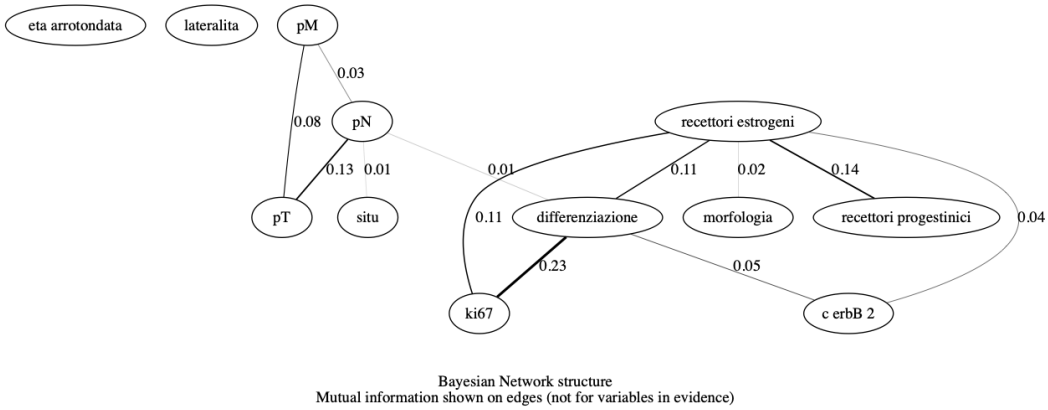



Figure 5.3. Plot model output.

The "Independencies" interaction mode gives the expert the possibility of verifying which d-separations (Definition 3.22) exist in the constructed Bayesian network's DAG (Definition 3.19) . The concept of d-separation is here reworded into a higher-level notion of "choosing a source variable and a set of evidence to see which other variables have influence on the source, given the evidence". This recasting was deemed necessary because the clinicians at the ICP initially had difficulty in conceptualising at the level of graph theory, probably due to the the misinterpretation of the directionality of the edges in the graphs of a Bayesian network. Graphs and trees are quite widely-used in clinical practice; however, the presence of an edge is commonly interpreted not as a correlation, but often as an indication of causality. Thus, the concept of d-separation could be misinterpreted because of this consolidated viewpoint. For this same reason, after having chosen first the source variable and then the observed set of evidence variables, the user is presented with an output both in graph (Figure 5.5) and in natural language (Figure 5.4) form. Having both output modalities present was seen to reduce the confusion that users trained in the medical sciences felt for such an unfamiliar concept.

The new visualisation of d-separation (see Figure 5.10), introduced on the basis the the ICP's suggestions, was confirmed by the clinicians to be very intuitive, especially when compared to the initial design (see Figure 5.11).

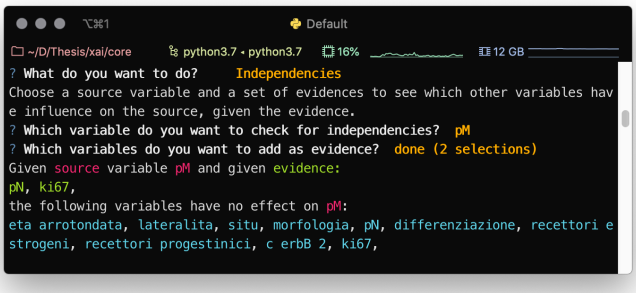



Figure 5.4. Independencies query natural language output.

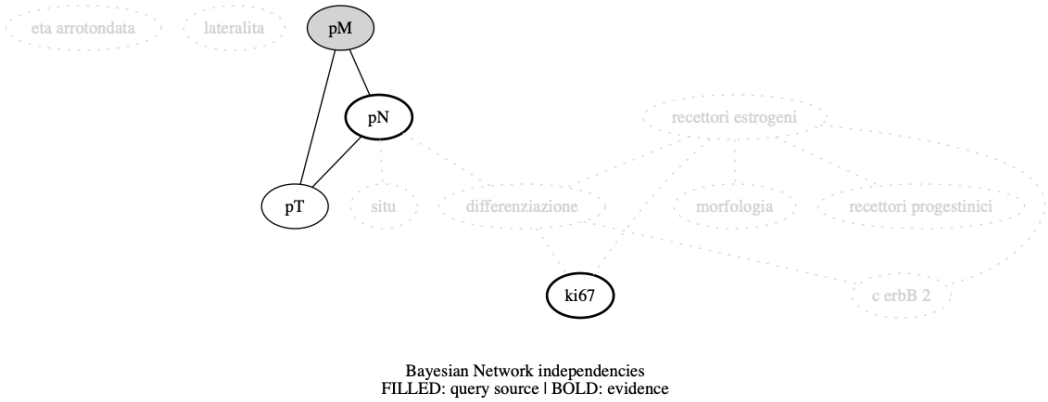



Figure 5.5. Independencies query graph output.

### 5.2.4 Conditional Probability Query

The "Conditional Probability Query" would be an example of a *static*, *linguistic* explanation in the framework defined by Lacave and Díez [2002] mainly aimed at *explaining the evidence*. Compared to the characteristics of an explanation identified by Miller [2018], it could be seen as possessing the *selected* element.

Conditional probability queries (Definition 3.6) were seen to be instinctively understood by the clinicians at the ICP. Indeed, many of the natural language questions that they defined to clinically validate the system (see Subection 4.5.1) could be framed as and answered by instances of this type of query.

As can be seen in Figure 5.6, the user is asked for a target variable (in magenta) of which to observe the conditioned values and for a set of variables, together with their observed values (in green). The output, in natural language, includes all elements of the query together with the colour-coding described in Subsection 4.4.2. The answer to the question (in cyan), shows the probability of each of the states of the target variable quantified in natural language i.e., as linguistic probabilities, using the coding defined in Table 4.4, and as raw probabilities, shown as percentages.

In addition, the use of colours was appreciated by the users at the ICP because they felt that it helped them to orient themselves among the different elements of the query and also to remember how they had posed it.

### 5.2.5 MPE Query

The "MPE Query" would also be an example of a *static*, *linguistic* explanation in the framework defined by Lacave and Díez [2002] mainly aimed at *explaining the evidence*. Compared to the characteristics of an explanation identified by Miller [2018], it could be seen as possessing the *selected* element.

Queries of the MPE type (Definition 3.26) were not initially understood until a bridge to concepts familiar to clinical practitioners had been established. When presented at an abstract, mathematical level, the experts of the ICP were not sure of the utility of such a query class. With some work, it was understood that an MPE query could be linked to a concept familiar to any

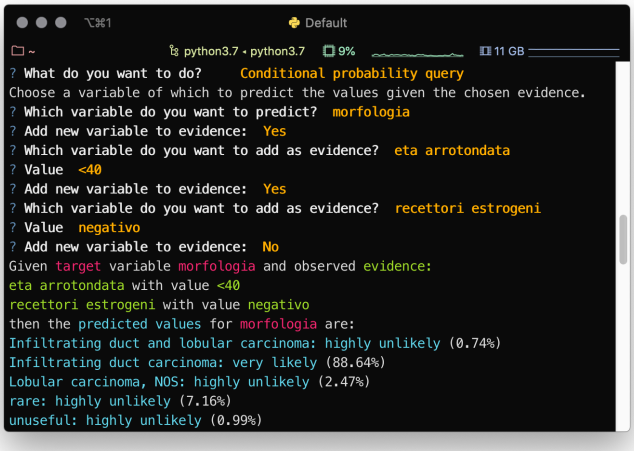



Figure 5.6. Conditional probability query output.

clinician: that of "a maximally likely patient profile". That is, given a set of known parameters it is of interest for the clinician to find which is the most likely assignment to the others. As each record in the data set represents a patient's clinical profile, this is equivalent to finding the most probable patient given a set of know values.

Another way than an MPE query makes clinical sense, is in the crucial task of predicting missing values for a patient. This is not an unlikely case, as discussed in Subsection 4.2.2, because there is more than one reason that patients may be missing one or more entries in their clinical profiles. Executing an MPE query with the known patient's values will yield the most probable assignments to the missing ones and is thus equivalent to a prediction task. The clinical significance of such an interaction mode can also be inferred from the fact that a number of the natural language questions, that were spontaneously defined to validate the system (see Section 4.5), were seen to map onto instances of this type of query.

At a technical level, the MPE calculation is executed using Pgmpy's `map_query` function.

The output, shown in Figure 5.7, presents, in colour-coded natural language, the input evidence (in green) and most probable assignments to the remaining variables (in cyan).

### 5.2.6 Pseudo-MPE Query

The "Pseudo-MPE Query" would be an example of a *static*, *linguistic* and *graphical* explanation in the framework defined by Lacave and Díez [2002] mainly aimed at *explaining the evidence* but also the *reasoning*. Compared to the characteristics of an explanation identified by Miller [2018], it could be seen as possessing the *selected* and *causal* elements, while remembering that the implications resulting from a BN are not necessarily causal.

The "pseudo-MPE query" interaction mode is aimed at generating a "maximally probable" assignment using the methods described in Subsection 4.4.1 under the "pseudo-MPE from Initial Evidence" header. The hypothesis is that this should be a valid explainability tool, as it is not only a *linguistic* but also a *graphical* explanation, with the latter element being identified by

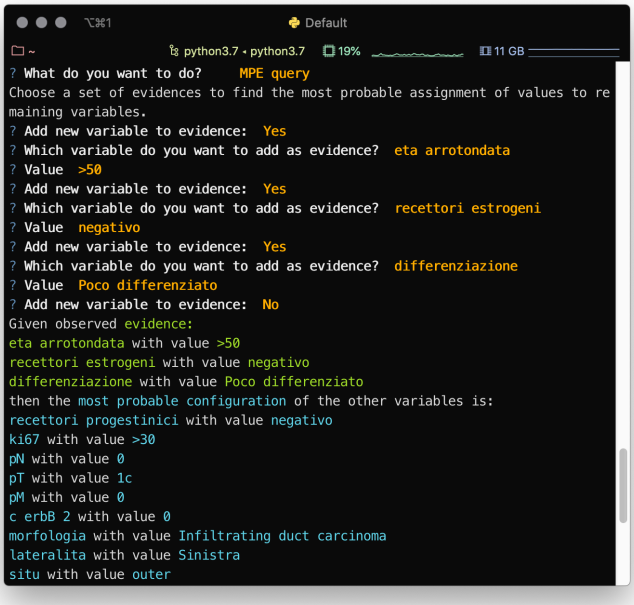



Figure 5.7. MPE query output.

Lacave and Díez [2002] as one of the most effective ways of giving a satisfactory explanation in a BN.[2]

The user is first asked for the probability threshold under which to discard the *(state,value)* pairs whose probability is deemed too low. Then, after being asked for the initial observed evidence, the expert is presented with the constructed polytree (Definition 3.21); an output example can be seen in Figure 5.8. This polytree will have the initial evidence, that the expert specified, as roots and a single chain of *(state,value)* pairs, each one quantified with its probability (in natural language) given all of its ancestors.

A doubt, that presented itself quite early during the ICP's evaluation, concerned the quantification of probabilities in the chain. The pathologists were unsure of why *(state,value)* pairs appeared before others that had been reported as more probable. For example, in Figure 5.8, *("morfologia","Infiltrating duct carcinoma")* that is considered *likely* appears before *("recettori estrogeni","fortemente positivo")* which is considered "very likely". The pathologist's intuition brought her to expect this deduction chain to be monotonically decreasing in probability from the initial evidence (that, as such, is certain).

What turned out to be the point of confusion, was that it was unclear that the probability of every node added to the chain depends on all its ancestors. In the specific example, *("recettori estrogeni","fortemente positivo")*'s evidence set also contains *("morfologia","Infiltrating duct carcinoma")*. There is thus, mathematically, no reason for the chain to be monotonically decreasing in probability because adding new evidence is liable to boost the likeliness of some unobserved variables. Returning to the example, the marginal probability $\mathbb{P}$(("recettori estrogeni","fortemente positivo"))

[2] If the cardinality of the set of variables to explain is one, i.e., $|E| = |V| - 1$, with $E$ the evidence set and $V$ the set of variables in the BN, then the "pseudo-MPE" and true MPE assignments will be identical.

may very well have been less probable than “very likely”, maybe it was only “likely” or even “unlikely”, but this says nothing about the posterior probability $\mathbb{P}$((“recettori estrogeni”,“fortemente positivo”) | (“morfologia”,“Infiltrating duct carcinoma”)), which is what the polytree displays.

The unclearness of the chain of inferences is certainly not a point to underestimate, as the hope was for the “pseudo-MPE” output to be able to clarify the underlying reasoning process of the models and thus help in guiding the expert’s through process. If this reasoning process itself were unclear, this could hardly lead to a good explanation; thus an effort will have to be made to explain the underlying assumptions better, while also being mindful when evaluating if this output mode genuinely presents the characteristics of a good explanation.

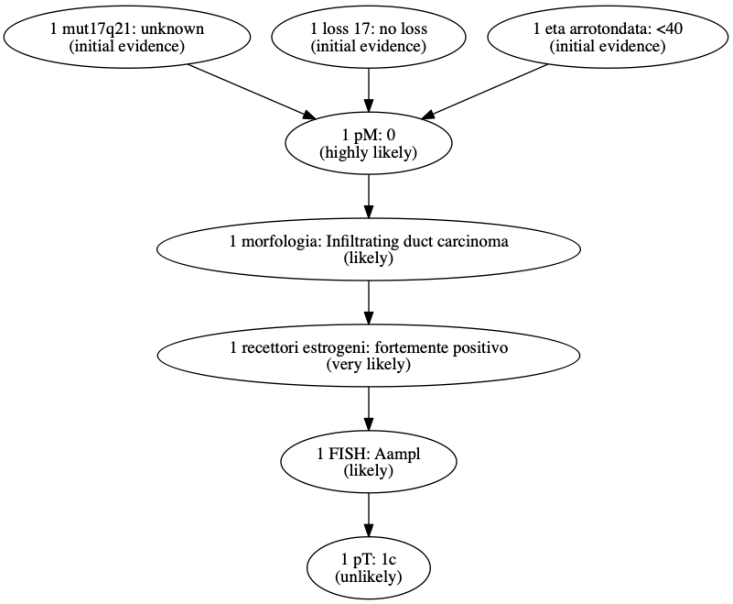



Figure 5.8. Pseudo-MPE query output with threshold 0.5.

### 5.2.7 Exhaustive Dialogue

The three dialogue variants would be examples of *dynamic*, *contrastive*, *linguistic* and *graphical* explanations in the framework defined by Lacave and Díez [2002] aimed at *explaining the evidence* but also, and most importantly, to *explaining the reasoning*. Compared to the characteristics of an explanation identified by Miller [2018], these could be seen as possessing all the necessary elements: *contrastive*, *selected*, *causal* and *social*.

The three “dialogues” are the most experimental interaction modes and thus also the most alien to a user. None of the natural language questions defined by the ICP in the form described in Subsection 4.5.1 could be directly mapped onto such a dialogical process. On the other hand, the dialogue aims to build an *expert-driven MPE approximation* and could thus be regarded as essentially answering the same question as the “pseudo-MPE” and “MPE” queries (Subsection 5.2.6). The research hypothesis is wether this could be a better explainability tool, as it is not only a *linguistic* and *graphical* explanation but also a *dynamic dialogue* that Hilton [1990] and Lacave and Díez [2002] identify as a key ingredient in having an effective explanation. Another

important fact is that the dialogues offer a *counterfactual* branch when the expert dissents with the model; Miller [2018] singles out being *contrastive* as one of the defining characteristics of an effective explanation, as this feature closely aligns with our expectations of what an explanation should entail.

Because of the novel nature of such a knowledge-extraction process, three different versions were implemented with each one adding a different set of behaviours to the "exhaustive" version described in this subsection. This helped in exploring the space of possibilities and aided in understanding which features were preferred by the clinicians of the ICP, both as a means for knowledge-extraction from the data set and from a comprehensibility point of view. It should here be noted that comprehensibility of the outputs is a *necessary* but *not sufficient condition* to be able to gain knowledge from data. Both variants to the basic dialogue - the independencies-aware and the thresholded one - aim to prune the space of variables proposed to the user in order to reduce her cognitive load. This is in keeping with the insight by Miller [2018] that explanations are *selected*, meaning that we humans expect that the explaining factors be picked based on some criterion.

The "exhaustive dialogue", as described in much more detail in Subsection 4.4.1 under the "Dialogues" header, is so named because it ends only when the expert user has reviewed all the variables present in the data set. It starts by asking the clinician for a set of initial evidence and from thereon after iteratively proposes the *(state,value)* pair with the least entropic *state*, based on the accumulated evidence (the rationale behind this is explained in Subsection 4.4.3). An example of such an ongoing interaction is shown in Figure 5.9.

An issue that was highlighted early on was that the experts had great trouble in building the *knowledge base* from a single evidence; this was the driving motive that pushed the representation of the "pseudo-MPE" branch beyond a simple *tree* (Definition 3.20) - as in [Butz et al., 2018] - but towards a *polytree* (Definition 3.21). This way the expert is able to inject the query with as much domain knowledge as she feels comfortable with. It was an unrealistic assumption to expect a domain expert to bear the cognitive load of selecting a *single best initial evidence*; this would be a hard task to do in its own right but it is made even more difficult by the fact that the subsequent dialogue *depends* on the initial evidence. To effectively select one best evidence, the expert should also have been able to *predict* how the dialogue would have evolved from that initial point onwards. The dialogue is an *exploratory tool* that the user utilises with the objective of extracting knowledge from the data set; expecting the user to already know the outcome of her choices would mean that she already had the domain knowledge necessary to predict the consequences of those same choices; a clear instance of *circular reasoning*. This was confirmed by the ICP: having multiple initial evidence helped the users because it reduced the number of tuples proposed by the system and therefore the quantity of choices the users were tasked to deal with.

The general feeling being echoed by the users at the ICP was that the dialogue was the hardest interaction mode to understand and to utilise. The way they used the dialogues was by nearly always replying "yes" to its proposals, because they were mostly interested in seeing what the machine would propose. Nonetheless, the users understood the high potential of this method especially when applied with the objective of conducting research, but they reported they would need to "trust" the interaction mode before feeling comfortable with using it in such a manner. They felt that probably having more time to experiment with this kind of interaction might improve their confidence felt in using it, that was lower than that perceived for the other interaction methods.

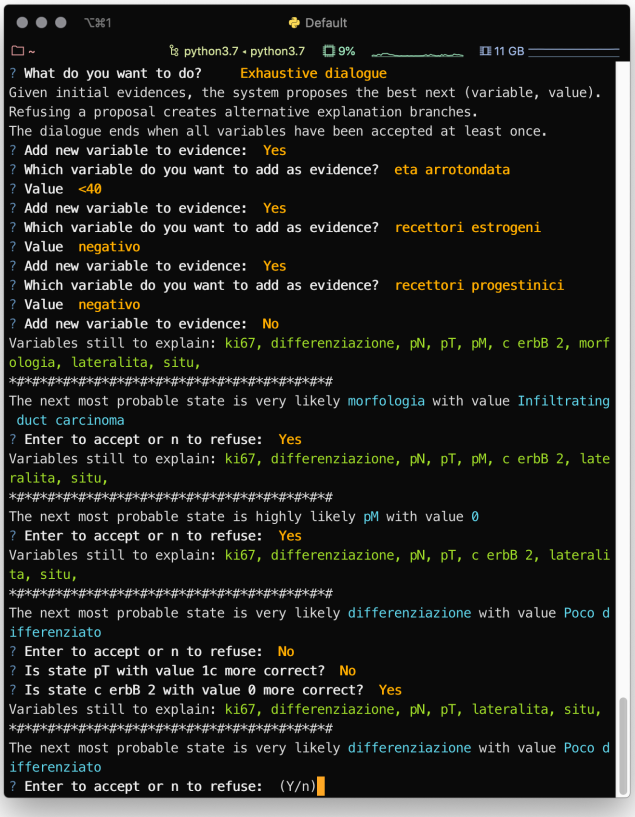



Figure 5.9. Ongoing Exhaustive Dialogue.

### 5.2.8 Independencies Dialogue

The first variant to the "exhaustive dialogue" takes the approach of excluding variables based on their d-separation properties (Definition 3.22) in the underlying DAG (Definition 3.19). Thus the cardinality of the set of variables proposed to the user varies in a non-linear way, depending on the topology of the graph and the order of insertions into the evidence set. In the "exhaustive dialogue", presented under the previous header, the relationship between the set of variables still to explain at step $t$, $W_t = V \setminus E_t$, with $V$ all the variables and $E_t$ those already added to evidence, obeys the recurrence relation:

$$\begin{aligned} W_0 &:= V, \\ E_0 &:= \emptyset, \\ |W_{t+1}| &= |W_t| - 1, \\ |E_{t+1}| &= |E_t| + 1. \end{aligned} \tag{5.1}$$

That is, at each step $t$ of the "exhaustive dialogue", one variable moves from the set still to explain $W$ to the explained one $E$ i.e., after any iteration step, the number of instantiated variables increases by one unit, while the number of variables to explain also decreases by one. In the "independencies dialogue", this relationship depends on the set of variables $Z$ that are d-separated from those already in $E$. The relationship between the cardinalities is modelled by an operator $\zeta$ that is unique to the DAG of the BN (or to any *i-equivalent*[3] one):

$$\begin{aligned} W_0 &:= V, \\ E_0 &:= \emptyset, \\ |W_{t+1}| &= \zeta(|E_t|), \\ |E_{t+1}| &= |E_t| + 1. \end{aligned} \tag{5.2}$$

As d-separation is not monotonic (adding a variable to $E$ may open new paths and d-connect new variables), the cardinality of the set $W$ may vary, from the point of view of the user, in an unpredictable manner. To attempt to offset this effect, during the dialogue the user is supported by an updated view of the independencies in the graph (an example during the dialogue is shown in Figure 5.10).

Before receiving feedback from the ICP, the visualisation of the independencies was the one shown in Figure 5.11. The most striking difference was the use of colour-coding to identify the role and the separation of variables with pink identifying the query variables, blue the evidence, red the separated variables and green the connected ones. As already noted in Subsection 5.2.3, the concept of d-separation turned out to be quite unfamiliar to the clinicians of the ICP so the first priority was to represent the concept visually in the clearest way possible. This was achieved, and confirmed in its efficacy by the pathologists, by fading the separated variables and marking those in evidence in bold, as can be seen in Figure 5.10. The fading of separated variables was felt to successfully reinforce the concept of these not influencing the remaining ones and its directness was especially appreciated.

The use of directed arcs to represent the DAG raised another critical issue that hadn't been foreseen. In the visual representation of a Bayesian network, an arc between two variables represents a correlation between their values while the direction identifies the *parent* and the *child* in the relationship; for example the graphical representation $X \to Y$ means that $X$ is the

[3] I-equivalence identifies classes of graphs that present the same d-separation properties.

parent of $Y$. This is a defining characteristic of such a model, because the fundamental idea of a BN is to factorise the joint distribution such that each variable's values depend only on that of its parents; the concept of conditional probability table is explained in Section 3.5 and some more examples can be seen in Subsection 4.3.2 under the "MPE" header. Nonetheless, the pathologists explained that the DAG representing a BN is very similar to diagrams used during clinical research, with the crucial difference that in those a directed arrow represents *causation* and not *correlation*. In these diagrams a correlative relationship would have usually been represented by an undirected edge. For this reason the DAG representation of the BN was *disoriented* in all visualisations. The ICP confirmed that this new formulation was more closely aligned with the intuition that could be expected by a clinician.

The third element of difference, is the addition of the mutual information coefficient (Definition 3.13) on the arcs connecting each couple of variables; the coefficient also scales the width of its associated edge, giving further visual feedback to the user. This functionality was a direct request from the ICP's representatives since after inspecting the initial DAG visualisation they felt the need for a feature that would increase their understanding of the relationships between the variables. D-separation is binary while mutual information can give the practitioner a much wider (theoretically infinite) range of information. For example looking at Figure 5.10 it is quite easy to see that, while "morfologia" and "recettori estrogeni" are d-connected, the amount to which they influence each other's values is small compared to other connected variables. Some arcs are missing the mutual information coefficient because one of the two variables is in the evidence set, e.g., those belonging to "recettori estrogeni".

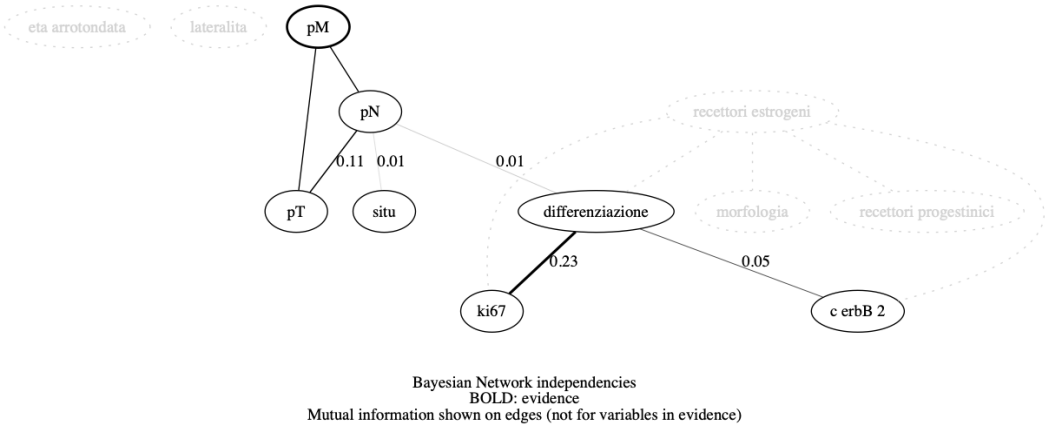



Figure 5.10. Ongoing Independencies Dialogue.

### 5.2.9 Thresholded Dialogue

The final "dialogue" variant adopts a different strategy for pruning; namely one based on the probability of the proposed tuples and on the number of times they have been refused by the expert. This implements a suggestion found in Lacave and Díez [2002] that explanations should be graded on the user and not on a *fixed user model*; one of the ways this has been addressed in literature is by the introduction of thresholds to filter unwanted information.

The cardinality of the set $W$ of states to explain decreases linearly, similarly to the "exhaustive dialogue", but potentially with a slope coefficient $\alpha \leq -1$, as many states may be

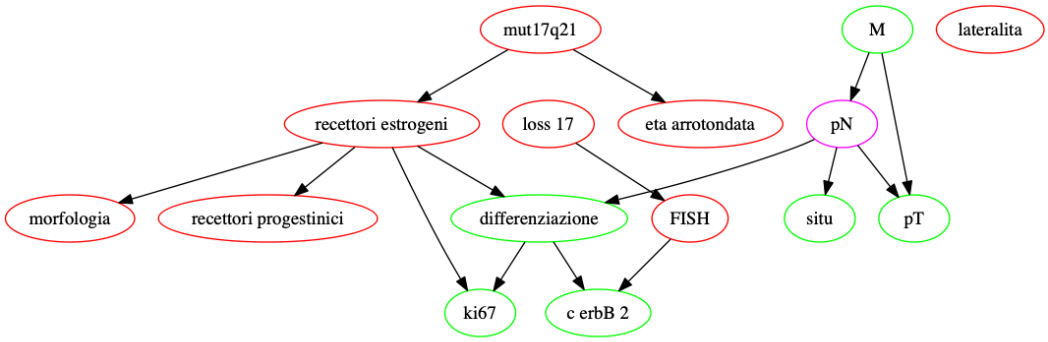



Figure 5.11. Previous visualisation during Independencies Dialogue.

infra-threshold i.e., too improbable to be considered. Unlike the independencies dialogue, the cardinality of $W$ cannot increase:

$$\begin{aligned} W_0 &:= V, \\ E_0 &:= \emptyset, \\ |W_{t+1}| &= \alpha |W_t|, \\ |E_{t+1}| &= |E_t| + 1. \end{aligned} \tag{5.3}$$

The default values for the threshold and the maximum number of times a *(state, value)* tuple could be proposed were decided together with the ICP and set to:

- *threshold*: 0.4, a *(state, value)* tuple is ignored if the probability of *value* is less that 0.4;
- *refusal limit*: 2, a *(state, value)* is ignored if it has already been refused twice.

## 5.3 Validation Results

This section will present the clinical and explainability results of the developed system, based on the methods outlined in Section 4.5.

### 5.3.1 Domain Experts' Initial Expectations for an Explanation

At the beginning of this project, it was not easy for the ICP's experts to understand what AI was exactly and what form an interactive tool in this domain could take. For these reasons, they considered the idea of receiving an output in natural language highly intriguing, since this was the output modality they had the most experience working with.

Because of the novelty of the approach developed in this thesis as applied to the field of medicine the experts were very receptive to receiving many different forms of explanation, for example textual, graph-based and tabular. Nonetheless, in their mind, the preferred one would still be a natural language output, as they imagined it as being simpler, clearer and more compact than any other output modality. Thus, their ideal explanation would be a natural

language output corroborated by the values in the data, presenting a summary of their inputs to the system (e.g., the used query and evidence variables) and understandable in terms of probability.

The ICP representatives felt that their preference for *linguistic* explanations over any other form might have stemmed from the fact that *nearly all medical literature is highly linguistic* in the way it communicates content; tables are little used and graphs are often ignored because there is little standardisation in the way they are presented across the subfields of the medical domain. Clinicians thus prefer to focus on reading the textual, conversational description that accompanies the results presented and this habit may have shaped their expectations of what form an explanation should take.

### 5.3.2 Clinical Validation

#### Natural Language Questions Results

The natural language questions marked as *validation*, presented in Subsection 4.5.1 have been discussed and validated by the clinicians at the ICP. The questions tagged as *research* necessarily have an *unknown* "Expected result" but, as noted in Subsection 4.5.1, these queries are nonetheless extremely interesting in order to understand how the user might want to relate to the system. The system's outputs to the natural language questions are shown together with the results they would have expected based on established medical literature and their professional expertise. The ICP's representatives had the faculty to *agree* or *disagree* with the software's verdict and, where they felt it necessary, were able to leave notes, which have also been reported in this section.

Tables 5.2, 5.3, 5.4 and 5.5 present the clinical validation results for the natural language questions in Appendixes B.1, B.2, B.3, B.4 and B.5. Table 5.1 shows the experts' assessments, aggregated by summing the number of times each evaluation appeared over the thirty questions. Comments followed by "to further explore" were aggregated together; the assessment of question number 4 "agree, to further explore" was also counted in the "agree" category. Note that questions 12 and 13 are compound i.e., they were run multiple times by changing only the *evidence values* for the same *evidence variable*.

Table 5.1 strongly supports the idea that the developed system, and by extension the underlying Bayesian network, is able to capture the clinical relevance of the variables in the data set, in the current application domain. Twenty-four query answers out of a total of thirty-five found the ICP experts as either in "full agreeance" of in "agreeance" with the system's predictions, one was deemed "acceptable" while none were refuted outright.

Three queries (numbers 23, 25 and 27 in Appendix B.4) were not executed by the domain experts; this is because they had framed them as conditional probability queries but, in actual fact, these were questions that could easily have been answered by d-separation queries. For example, question number 23 reads:

> In young patients, does a negative expression of the progestinic receptors influence the lymph nodes' state?

Framing this as a conditional probability query would mean having to obtain the value of "lymph nodes' state" ("pN" in the benchmark data set) for all the combinations of values of "age" and "progestinic receptors" ("eta arrotondata" and "recettori progestinici" in the data set) in order to see if a change in the latter produced a variation in the former. Instead, a d-separation query can

directly answer if the nodes of the variables representing “lymph nodes’ state” are d-separated from those for “age” and “progestinic receptors”. The response given by the d-separation query is less informative than that of a conditional probability query in the case that the nodes are not d-separated, but nonetheless answers the question as it was posed. This misunderstanding of the use of the tool is certainly something to take note of, but the instances of confusion are also limited to only a particular phrasing of questions among all those posed. As such, it could probably be addressed by adjusting how the information is presented to the user and not by radically changing the interaction mode.

The answers to the remaining nine questions were not excluded *a priori* by the ICP’s representatives but were deemed interesting enough “to further explore”.

Table 5.1. Aggregation of the experts’ evaluations of the answers given by the software tool.

| Expert comment | Counts |
|---|---|
| fully agree | 4 |
| agree | 20 |
| acceptable | 1 |
| to further explore | 9 |
| query not executed | 3 |
| **Total** | **35** |

More discussion of Natural Language Questions Results

Regarding the use of separations in a clinical setting, it was observed that the approach under investigation has had particular success in evidencing the interesting contextualisation of the variable “eta arrotondata”, which represents the age of the patient at first diagnosis of the tumour. Current medical understanding assumes that tumours could have some differences in their clinical morphological status depending on the age of onset. The novel knowledge that the age at diagnosis does not influence and does not depend on tumour morphology, dimensions, nodes, differentiation, hormonal status and proliferation opens the door to new clinical and therapeutical approaches. Hence, extracting information in terms of the relationship between variables could help not only in understanding the function of these variables in patient profiling but also in helping in the potential definition of novel, optimised guidelines relating to the best practices in the presence of evidence. The ability to define the “informational flux” of the explanation depending on the evidence could allow for more robust patient profiling, especially in the presence of partial knowledge or reduced resources (e.g., budget and time). For example, Figure 5.5 shows the graph for the query on “pM” and how many other clinical variables are not relevant once the result of “ki67” and “pN” are known.

In current practice, it is quite common to have a single, standardised, certified procedure for the management of the diagnosis and treatment, independent of the evidence that is already present (or missing). Typically, it is a simple decision tree with a “yes/no” progression that is independent on the pieces of evidence already known. Information about d-separation could help to introduce a novel concept of data managing “tuned” to the specific case.

It is worth highlighting that the available molecular biomarkers undergo a process of continuous updating thanks to ever more accurate and accessible high throughput screening. Thus, new variables can be integrated for specific patients, in the presence or absence of evidence, and this can progressively aid the onset of personalised medicine. At the same time, features that are already well established in clinical practice could find new interpretations, allowing the creation of innovative hypotheses and thus of an evolution in clinical practice. For example, despite it being conceivable that the independency of the variable "lateralita" (that have, generally, long been annotated for all the patients) from the other features has a biological meaning; to date there is not a well established validation of this hypothesis.

In clinical practice, considering d-separation could help in defining alternatives that could enable the optimisation of time, cost and diagnosis. For example, it is quite common for a sudden cyto-histo-molecular analysis to have to be performed before surgery, in order for the doctors to decide on the best way to proceed. In this case, the ICP prioritises the analyses of the sample of the patient who is to undergo surgery. It is possible that the biological sample could be degraded or not in sufficient quantity to enable completion of all the necessary assays. In this context, the importance of being able to obtain the information of interest in a quick but accurate way is self-evident. The type of analysis that is to be carried out should be prioritised in order to be able to formulate a diagnosis and d-separation could lead to the exclusion of certain tests that may turn out to be redundant, given some already known exam results or patient characteristics.

At the same time, in the case of degraded material, the possibility of inferring the missing value using MPE or conditional probability queries, in agreement with the opinion of the expert pathologist, could offer further basis and support for the clinician.

### 5.3.3 Explainability Validation

The "explainability evaluation questionnaire" introduced in Subsection 4.5.2 - visible in its entirety in Appendix C - was submitted to the ICP in late August, around three weeks after giving the institute members access to the proof of concept system developed as part of the methods of this thesis (see Sections 4.3 and 5.2). The clinicians at the ICP (see 4.2.1) took ownership of the survey and replied to the questions it posed frankly and to the best of their knowledge.

In the following, the answers to the "explainability evaluation questionnaire" are presented, one section at a time together with the unedited answers that the ICP representatives gave where requested. A presentation of the various interaction modes available in the tool is found in Section 5.2.

#### Confidence

The first section, Confidence, deals with the "confidence" brought about by the system in the expert user. From the answers, it can be deduced that the developed system did indeed help in clinical decision-making.

The "MPE query" interaction mode was highly-valued because of its capability to "fill in the blanks" in a patient's profile; given a series of known values for a patient, the most probable assignment to the other variables is immediately found and thus complete a profile is obtained. As discussed in Subsection 4.2.2, this was one of the initial hopes that the ICP had and the tool seems to have fulfilled it. This seems to validate the claim by Lacave and Díez [2002] (see

Section 2.5) that the solution to the MPE problem (Definition 3.26) is the way to explain the *evidence*.

"D-separation queries" were instead valued because of their ability to give a high-level overview of the data set and of the relationship between variables. This enabled the prioritisation of certain clinical tests over others, as having observed the value of a variable representing the outcome of a given analysis may render others redundant. This information is contained in the BN's DAG topology and could turn out to be a powerful tool in clinical practice. The claim "the most direct and intuitive way of showing the information embodied in a Bayesian network is to display the corresponding graph" [Lacave and Díez, 2002] seems to have been somewhat confirmed by the expert users' evaluation since the "plot" and "d-separation" modes are inherently *graphical* in nature. Though, this may not actually be the case because d-separation was also implemented with a *linguistic* output and so this may have been the characteristic that triggered the appreciation.

[NB: A *cohort* - in clinical setting - is a group of individuals who share a common trait.]

## Confidence

1. Did the tool increase the confidence in diagnosis when diagnostic screening results were missing for a patient? Why?
   ⊠ Yes O No
   *Validation of the tool by queries on well-known interactions between some clinical features helped in considering reliable the proposed variables for missing data in patients' profile.*

2. Did the tool help in characterising a particular patient's profile?
   O Not at all O Somewhat ⊠ Absolutely
   *MPE gives at once the full profile of missing variables, for example in patients affected by triple negative breast cancer (see details below).*

3. Did the tool help in your confidence of understanding the cohort characteristics? How?
   O Not at all O Somewhat ⊠ Absolutely
   *Plot and d-separation are able, by a quick visualization, to give the general idea about the presence or the absence of a relationship between variables, thus giving at once the general idea about the characteristics of the entire cohort.*

4. Did the tool improve your confidence in your clinical decision-making? How?
   O Not at all ⊠ Somewhat O Absolutely
   *Complete integration of the tool in the clinical decision-making workflow requires further time; at the moment it has been used for validation of corroborate data and for exploration of new hypothesis.*

5. Did having the tool at your disposal improve your confidence when making time-constrained decisions? How? (for example, did it improve confidence in prioritising some tests over others?)
   O Not at all ⊠ Somewhat O Absolutely
   *Knowing 'independencies' between variables could help in prioritizing some tests over*

> *others, for example in case of poor tumor material we can decide to investigate only one specific related marker rather than more unrelated markers.*

Features

The Features section of the questionnaire was designed to probe the various interaction modes in more detail, so as to understand which characteristics were perceived as the most useful.

The most conspicuous result is that *natural language* was perceived as a more useful output modality than *graphically* displaying the results; this is a step towards confirming the hypothesis just laid out when discussing Confidence, that the "d-separation query" was principally appreciated not because of its *graphical* nature but because of also being *linguistic*. This is also supported by the preference for the "MPE query" over the "pseudo-MPE" one; the former is a purely *linguistic* explanation while the latter is nearly completely *graphical* (compare Figures 5.7 with 5.8). The quantitative comparison between the solutions, presented in Section 5.4, shows that there is very little difference between the two answers and this further lends credence to the *output modality* of the explanations being the discriminant factor. However, in the light of Subsection 5.3.1, the preference for a *linguistic* explanation over a *graphical* one could be explained by the experts' *preconceived notions* carrying over until the end of the testing phase. It cannot be excluded that given a longer hands-on period with the system and thus the prospect of acquiring more familiarity with the alternative output modalities, their belief of preferring a *linguistic* over a *graphical* explanation may have been reversed.

Lacave and Díez [2002] contrasted the two output modalities but seemed to lean towards stating that the former were more important than the latter when explaining a BN; the current evaluation does not seem to support such a claim, barring all the reservations that have just been set forth. Naturally, this does not disprove the importance of a graphical explanation and it may be the case that different graphical outputs might have been more efficacious in communicating the model to the user. For example, the "pseudo-MPE" output, as noted in Section 5.2, had been deemed confusing already in the "informal evaluation". Nonetheless, the feeling is that a linguistic output could be easier to design and tailor to the specific application and could be made as explicit and deep as desired by increasing its verbosity. A graphical output, on the other hand, may risk being more easily misinterpreted because of the greater preexisting knowledge needed to decipher it. A graphical explanation could certainly be more intuitive than a textual one, but it may be much more difficult to properly design it as such.

The second main result is that regarding the "dialogue", which is probably the central method developed in this thesis. The *dialogical* output (an example of a *dynamic* explanation in the considered classification [Lacave and Díez, 2002]) was appreciated because of its ability to offer "what-if" cases; but the perception, based both on this "formal" and on the "informal" assessment, is that this interaction mode is too cumbersome for the average expert medical user. This could be because of the added cognitive load needed to keep track of a long dialogue together with the proposed counterfactual alternatives. The novel alternative dialogue modes proposed, d-separation-aware and thresholded, were rated higher than the plain exhaustive one; as these different dialogues aim to remove unnecessary proposals (as explained in Sections 4.4 and 5.2) this seems to go in the direction of confirming the *cognitive overload hypothesis*. Nonetheless, the experts recognise the potential of such an interaction mode and seem to feel that they could appreciate it more if they were given more time to test the system.

The display of the connection strength between variables in the network, a feature devel-

oped based on suggestions from the ICP, was not deemed essential to the comprehensibility of the outputs, where it was used.

## Features

6. Given the modes of interaction with the system labelled as “dialogues”, do you think you would have had more difficulty in interpreting the data without the these modalities?
   O No ☒ Maybe O Yes
   *The other modes cover the vast majority of our queries; dialogue could be useful in exploring new hypothesis and evaluating different scenarios such as ‘what if is not...’ at the same time.*

7. Was natural language useful during the interaction? Why?
   O Not at all O Somewhat ☒ Absolutely
   *Natural language allowed 1) the better selection of the proper interaction mode depending on the question to answer; 2) useful recapitulation of evidence, variables and features selected during analysis; 3) easy comprehension of the output.*

8. Which type of “dialogue” did you feel was most useful? Why?
   O Exhaustive ☒ Separations ☒ Thresholded O A combination of the previous O All O None
   *In our opinion, exhaustive dialogue resulted in a time consuming and redundant process; separation and threshold offered a much more focused and efficient output.*

9. Did you feel that the dialogue helped you in cases of uncertainty? If yes, how? If no, why?
   O No O Somewhat ☒ Yes
   *In case of uncertainty dialogue shows the counterpart hypothesis in an uncomplicated way.*

10. Did you feel that the “dialogue” helped your clinical decision-making? If yes, how? If no, why?
    O No ☒ Somewhat O Yes
    *We understand the potential of this mode of interaction but we requires further practice to apply it in clinical decision-making process.*

11. Did the generation of “counterfactual branches” help in your understanding of the data? Why?
    O No ☒ Somewhat O Yes
    *Visualization helps in giving a immediate and direct approach to the output.*

12. Given the interaction mode labelled “pseudo-MPE query”, how would you rate the solutions it proposed from a point of view of their understandability? (1 poor, 5 good)
    O 1 O 2 O 3 O 4 ☒ 5

13. How would you rate the "pseudo-MPE" solutions from a point of view of their clinical usefulness?
O 1 O 2 O 3 O 4 ☒ 5

14. Do you feel that the interaction mode labelled as "MPE query" gave better solutions than that labelled "pseudo-MPE query"? Why?
O No O Maybe ☒ Yes
*MPE query output is much more easy to understand, probably due to the natural language.*

15. Did you find the "pseudo-MPE" or "MPE" interaction mode the most useful? Why?
O "pseudo-MPE" ☒ MPE O Both O None
*Gives at once the most probable profile of many different variables in natural language.*

16. How important was the highlighting of the independencies between variables?
O 1 O 2 O 3 ☒ 4 O 5

17. Do you think you would have had more difficulty in interpreting the data without the correlation strength displayed?
O No ☒ Maybe O Yes

18. Do you think you would have had more difficulty in interpreting the data without visualisations?
O No ☒ Maybe O Yes

19. Do you think you would have had more difficulty in interpreting the data without natural language output?
O No O Maybe ☒ Yes

Time

The Time section focuses on the *temporal* element of an explanation. This, as touched upon in Section 2.4, is still uncommon in current xAI literature and its inclusion has been advocated by Gilpin et al. [2018].

What is immediately noticeable is that the responses in this section align with - and confirm - both the ones in the other sections of the questionnaire and also the results of the informal validation; i.e., there is an inverse relation between the subjective rating of the quality of an explanatory mode and the time needed to understand it.

*Conditional probability queries'* outputs were the easiest for the experts to relate to; this was to be expected based on the number of validation questions (see Subsections 4.5.1 and 5.3.2) answerable by such a class of queries. As previously noted when discussing the clinical validation of the system, the fact that questions can be answered by the software means that it is probably well positioned to align with the expert's worldview; similarly, a high number of natural language questions answerable with a certain query class, likely denotes a tendency for the expert to think in terms compatible with that query type.

Accordingly, the "MPE" and "pseudo-MPE" queries' outputs were ranked slightly lower based on the time it took to understand them, but still higher than the "dialogues". This is to be expected based on the findings of the rest of the questionnaire, which seem to point to the

"dialogues" being the least easily accessible interaction modes for the users.

## Time

20. How would you rate the time it took to understand the dialogues' outputs? Which of the three was best? (1 bad, 5 good)
O 1 O 2 ☒ 3 O 4 O 5

21. How would you rate the time it took to understand the conditional probability query's outputs
O 1 O 2 O 3 O 4 ☒ 5

22. How would you rate the time it took to understand the MPE and "pseudo-MPE" query's outputs?
O 1 O 2 O 3 ☒ 4 O 5

23. Did natural language help in reducing the time needed to understand the outputs?
O No O Somewhat ☒ Yes

24. Did visualisations help in reducing the time needed to understand the outputs?
O No O Somewhat ☒ Yes

Tool

The Tool module of the questionnaire directly asks the users for their opinion regarding the developed system as a whole.

The findings support those of the other sections of the questionnaire, namely that the "conditional probability", "pseudo-MPE" and "MPE" queries were the explanatory modes perceived as most useful by the ICP. Somewhat surprisingly the "dialogues" were indicated among the most used interaction modes even though they had been ranked as the explanations that were hardest to relate to. The fact that they were used for longer, could precisely be a consequence of them being harder to utilise; this may be because the users would be spending more time on them to understand the outputs.

The other finding in this section is that the tool seems to have fulfilled the needs of the ICP and this is a good confirmation of the explanatory powers of the system itself. Considering the resistance to change inside the field of medicine, somewhat confirmed by the experts of the ICP who had expressly stated they did not want to be dealing with understanding a software tool but only to focus on their work, it would then be difficult to imagine them being satisfied with the developed system if its explanatory powers had been considered insufficient.

## Tool

25. Which interaction modes did you feel could be the most useful? Why?
O Plot model O Independencies ☒ Conditional Probability Query ☒ "pseudo-MPE" and MPE O Dialogues
*These two Modes are really user friendly and help in solving the majority of the queries.*

26. Which interaction modes did you use the most? Why?

O Plot model O Independencies O Conditional Probability Query ☒ "pseudo-MPE" and MPE ☒ Dialogues
*Cause we are mainly interested in patients profiling.*

27. How did you use the tool in your day-to-day work?
*For validation and research purpose.*

28. Is the tool missing any functionality that would address your needs? If yes, which ones?
☒ No O Yes

29. Did you have any difficulties in understanding which functionalities to use to address your needs? If yes, when?
☒ No O Yes

30. Did you have any difficulties in understanding the functionalities during usage? If yes, when?
☒ No O Yes

31. If you answered Yes to the previous question, how do you think this could be addressed?

32. Could you suggest any functionalities you would like to be implemented?
*Maybe could be useful the possibility to import other 'similar' data set and also have the possibility to save the textual and visual outputs. We would also like to be able to save queries and dialogues half-way and maybe have some ready-made query masks.*

Clinical

The Clinical part was included to understand if after personally using the system over a period of time, the experts at the ICP still felt that the software had the clinical significance confirmed in the first part of the evaluation (see Subsections 4.5.1 and 5.3.2). This was because the *clinical validation questions* in natural language were formulated before they had had the opportunity to interact meaningfully with the system. It was expected that a follow-up at the end of the testing period would be able to capture a clearer view of the clinical noteworthiness of the developed system.

The open question format of the questionnaire is undoubtedly an excellent way to elicit and understand the clinicians' judgement and, from reviewing the answers, it seems that the clinical worth of the system is still highly regarded by the pathologists of the ICP.

## Clinical

33. Did the tool help in recovering missing features of patients thus supporting diagnostic profile creation and decision making? If yes, which is/are the feature/s that benefited the most?
O No ☒ Yes
*The tool has the great powerful to recover missing features and scale their values. One of the features that benefited the most of this process is the 'lymph node involvement'*

*represented by the variable pN. Knowing this status is crucial for therapeutic decision in patients affected by breast cancer because it indicates to treat or not to treat with chemotherapy (i.e.,depending if it is positive or negative, respectively). In daily practice this datum come from an invasive exam that is performed as a secondary step in the clinical patients workup, therefore knowing a priori pN status on the basis of only data coming from the first standard diagnosis of the tumors (i.e.,morphology, dimension, differentiation) has a really high clinical impact.*

34. Did any of the tool's predictions have clinical confirmation later on? If yes, how?
O No ☒ Yes
*As example the tool predicted that in lobular carcinoma cerb expression is absent, confirming an established evidence in breast cancer literature.*

35. Did the tool help in highlighting new relationships between variables?
O No ☒ Yes
*The tool showed no relationship between patients' age at diagnosis and the other clinical morphological features (i.e.,morphology, TNM, grade, hormonal status), a new observation with a really high clinical impact that should be further explored.*

36. Did the tool help in highlighting new patient subgroups?
O No ☒ Yes
*The tool helped in better understanding features of patients with triple negative breast cancer showing that they carry a profile of aggressive tumors in terms of proliferation index (ki67) and grade (poorly differentiated), but without nodes involvement (pN=0) and no metastasis at diagnosis (pM=0).*

Satisfaction

The final question in the questionnaire aimed to be as general as possible to elicit the users' feedback on the whole system and their experience of using it.

The experts confirmed - and this is also corroborated by their other answers - that despite the fact that they lacked a complete understanding of the inner workings of the tool and that their technical understanding of the methods was far from clear, their initial expectations (see Subsec. 5.3.1) had been completely fulfilled. That is, they confirmed that the tool was entirely satisfactory in being able to provide them with understandable outputs they could use efficiently in their daily clinical work.

## Satisfaction

37. What is your general satisfaction with the tool? For what reasons?
O Completely dissatisfied O Somewhat dissatisfied O Neutral O Somewhat satisfied
☒ Completely satisfied

Table 5.2. Results for questions in Appendixes B.1 and B.2

| # | Expected result | System result | Expert comment |
|---|---|---|---|
| **1** | yes, with high probability | Plausibly high | agree |
| **2** | yes, with high probability | Highly likely low | agree |
| **3** | yes, with high probability | Highly likely low | agree |
| **4** | yes, with high probability | Plausibly low | agree, to further explore |
| **5** | yes, with high probability | Plausibly negative | fully agree |
| **6** | yes, with high probability | Possibily involved | fully agree |
| **7** | yes, with high probability | Highly likely low | agree |
| **8** | yes, with high probability | Highly likely low | agree |
| **9** | yes, with high probability | Plausibly high | agree |
| **10** | yes | Highly likely low | agree |
| **11** | yes | Plausibly low | agree |
| **12** | unknown | Very likely not very differentiated | agree |
| | unknown | Not plausibly quite well differentiated | acceptable |
| | unknown | Possibly quite well differentiated | agree |
| **13** | unknown | Plausibly negative | agree |
| | unknown | Possibly strongly positive | agree |
| | unknown | Highly likely strongly positive | agree |
| | unknown | Very likely strongly positive | agree |
| **14** | unknown | Plausibly positive | Agree, it's plausible that nodes are positive |

Table 5.3. Results for questions in Appendix B.3

| # | Expected result | System result | Expert comment |
|---|---|---|---|
| **15** | unknown | age, laterality, morphology and hormonal status don't influence nodes | new result, to further explore |
| **16** | unknown | age and laterality don't influence proliferation index | agree |
| **17** | unknown | age and laterality don't influence cerb | agree |
| **18** | unknown | age, laterality, TNM and situ don't influence oestrogen expression | new result, to further explore |
| **19** | unknown | age and laterality don't influence tumour grade | agree |
| **20** | unknown | age, laterality, morphology and hormonal receptors don't influence the presence of metastases at diagnosis | new result, to further explore |
| **21** | unknown | age, laterality, morphology and hormonal receptors don't influence tumour dimensions at diagnosis | new result, to further explore |
| **22** | unknown | no clinical morphological features influences the age at diagnosis | new result, to further explore |

Table 5.4. Results for questions in Appendix B.4

| # | Expected result | System result | Expert comment |
|---|---|---|---|
| **23** | unknown | - | query not executed |
| **24** | unknown | no influence | new result, to further explore |
| **25** | unknown | - | query not executed |
| **26** | unknown | plausibly negative | agree, to further explore |
| **27** | In young patients, does a negative expression of the progestinic receptors influence the expression of the c-ERBB2 marker? | - | query not executed |

Table 5.5. Results for questions in Appendix B.5

| # | Expected result | System result | Expert comment |
|---|---|---|---|
| **28** | unknown | • eta: >50<br>• lateralita: sinistra<br>• situ: outer<br>• morfologia: infiltrating duct carcinoma<br>• ki67: >30<br>• pT: 1c<br>• differenziazione: poco differenziato<br>• pN: 0<br>• pM: 0 | interesting, to further explore |
| **29** | unknown | • eta: >50<br>• lateralita: sinistra<br>• situ: outer<br>• morfologia: infiltrating duct carcinoma<br>• recettori estrogeni: negativo<br>• recettori progestinici: negativo<br>• c erbB 2: 0<br>• pT: 1c<br>• differenziazione: poco differenziato<br>• pN: 0<br>• pM: 0 | fully agree |
| **30** | unknown | • eta: >50<br>• lateralita: sinistra<br>• situ: outer<br>• morfologia: infiltrating duct carcinoma<br>• recettori estrogeni: fortemente positivo<br>• recettori progestinici: fortemente positivo<br>• c erbB 2: 0<br>• pT: 2<br>• differenziazione: moderatamente (ben) differenziato<br>• ki67: <14<br>• pM: 0 | fully agree |

## 5.4 Pseudo-MPE Evaluation

This is not a user-facing feature *per se*, but a way of running a test to compare the outputs of the "pseudo-MPE" algorithm with the exact solution (this is described in detail in Subsection 4.4.1 under the "MPE Algorithms Comparison" header).

The Hamming and Jaccard distances between the MPE calculated with Pgmpy's `map_query` and with DAOOPT should have been zero, as both use exact methods to solve the MPE problem. It was seen that this was not the case and this lead to the discovery of the problems described in Subsection 5.5.2. As is explained in Section 5.5, the benchmark against which to compare the "pseudo-MPE" was thus chosen to be Pgmpy's `map_query` function.

The tests were run over 1000 iterations; that is, 1000 initial random evidence $\boldsymbol{U_e}$ were generated and the "pseudo-MPE" solution was found using Algorithm 9 and was compared with the true MPE solution. Given that there are twelve variables in the BN, the maximum distance possible for both Hamming and Jaccard distances is 12. Thus, the average distances shown in Table 5.6, denote a good overlap between the ground-truth most probable explanation and the approximate configuration obtained with the "pseudo-MPE" algorithm.

Table 5.6. Distance of "pseudo-MPE" from true MPE solution

| Distance measure | Average distance |
|---|---|
| Hamming | 0.062 |
| Jaccard | 0.048 |

As discussed by Gámez et al. [2013], the MPE problem is *NP-hard* [Kwisthout, 2011] and remains such even in restricted network topologies, networks with restricted probabilities and for constant-bound approximations. MAP, as anticipated in Subsection 3.5.4 is a harder problem than MPE and is in fact $NP^{PP}$-complete and, unlike MPE, remains NP-complete when restricted to polytrees. The proposed "pseudo-MPE" algorithm, whose pseudocode is shown in Algorithm 9, reduces the MAP problem to the updating problem. Updating involves a polynomial number (namely quadratic with respect to the number of variables) of calls of standard updating in a BN and thus has a complexity $O(kV^2)$, in the number of nodes $V$ with $k$ a parameter depending on the algorithm used for the calculation of posterior probabilities.

In order to achieve the explanation of a node, we compute the updated probabilities for all the nodes, at every step the number of available *(state,value)* pairs decreases by 1 so one fewer updating call is done and the resulting complexity is quadratic $O(V^2)$. It is simple to see that there are at most only $V$ *(state,value)* pairs because every *state* is paired at each step with only one of its *values* - the most probable. The single most efficient pair is chosen at each step and then the corresponding *state* is no longer selectable.

The extra term $k$ depends on the algorithm used to compute the posterior probabilities, as at each step these are recalculated based on the updated evidence. The Pomegranate-based implementation does this inference utilising the `predict_proba` built-in function, based on *loopy belief propagation*. Loopy belief propagation is an *approximate inference algorithm* that estimates a solution to the exact belief propagation algorithm in *quadratic time*; it was introduced by Pearl [1982].

## 5.5 Issues

### 5.5.1 Zero Probabilities in Learned CPTs

It was noticed that some of the conditional probability tables that Pomegranate learned from the data set (i.e., by solving the structure learning problem defined in Subsection 3.5.2) presented entries with 0 value.

The assignment of the extreme probabilities 1 and 0, while perfectly coherent in the frequentist approach, is not in line with the Bayesian one. This is because of the different conception of probabilities between the two approaches, as discussed in Subsection 3.2.2. A frequentist practitioner would happily assign probability zero to an event not present in the data set while a Bayesian would refrain in doing so, as having a zero or one prior belief makes every posterior, calculated using Bayes' Rule (Definition 3.7), also zero or one. The necessity of avoiding the assignment of prior probability beliefs equal to 0 or 1, has been named "Cromwell's rule" by Jackman [2009].

A useful distinction to make that could help in deciding when to accept extreme probabilities or not, is between *a priori* and *a posteriori* propositions; the former are those whose truth value is not empirical but can, in general, be deduced based on logical necessity alone; the latter are those whose truth value is based on experience, for example any statement regarding the physical world. An example of *a priori* proposition could be a tautology, such as "every wife is married"; an example of *a posteriori* proposition could be an assertion regarding the state of the world, for example "yesterday it rained". It would be an epistemological error to believe in the absolute truth or falsity of any proposition that is not necessary and thus one should refrain from assigning absolute belief or disbelief to their truth value. This is because, apart from issues regarding the *ontological determinism of Reality* and of the *contingency of experience*, it is also against the intended use of statistics; statistical methods should only be applied in cases of uncertainty and not in those where a *deterministic mechanism* is implied. If a proposition's truth value can be known without resorting to the senses, then its value does not depend on the state of things in the sensible world. Thus, we can assign absolute confidence in the truth value of a priori statements.

To solve the issue in the context of this work, a simple post-learning correction was applied; specifically a small positive constant was added to every zero-valued entry in the CPTs in Pomegranate's model. The methodologically correct approach would have been to apply *Laplace smoothing*, a standard technique used to *smooth* categorical data. The method would have entailed adding a *pseudocount* $\alpha$ to every empirical probability estimated from the data set. Given $x_i$ the count of occurrences of event $i$ in a set of $N$ events, the un-smoothed empirical probability is:

$$p_i = \frac{x_i}{N}, \tag{5.4}$$

while the smoothed one would be given by:

$$p_i^* = \frac{x_i + \alpha}{N + \alpha d}, \tag{5.5}$$

with $d$ the number of possible categories.

The value of $\alpha$ should be chosen to reflect any prior knowledge regarding the events; in the case there were none, a non-informative prior should be chosen, as stated by the *principle of indifference* (in absence of any evidence one should distribute his beliefs uniformly). The simplest possible non-informed approach is to increment every event's count in the data set

by one, including the ones not appearing. Thus the relative frequency between events will be maintained but there will be no event $i : x_i = 0 = p_i$.

Unfortunately implementing this approach, while recognisably the most methodologically sound way of proceeding, would have been very time-consuming and outside the main focus of this thesis. The chosen strategy of adding a small positive constant $\epsilon$ to each empirical probability $p_i$, while not strictly correct, is extremely unlikely to change the learned CPTs. This cannot be ruled out *a priori* and would require *sensitivity analysis* to be decided, but we are working with the prior belief that this is very unlikely to happen.

The developed algorithm, termed "epsilon smoothing", is based on adding and subtracting "probability atoms" $\epsilon$ with the objective of removing zeros in the CPTs and maintaining the normalisation, so that the result will still be a valid probability distribution (based on Definition 3.3). We work with probability atoms $\epsilon$ so as not to incur in numerical imprecisions in the implementation, as only additions and subtractions need to be used. Every zero-valued element in a CPT column has a quantity $\epsilon \times \#!0$ added to it, with $\#!0$ the number of elements in the distribution that are not zero, and every non-zero entry has $\#0$, the number of zero elements, atoms subtracted from it. The end result is obviously still a correctly normalised probability distribution because given a probability distribution $P$, with $n$ elements $p_i$ of which $\#0$ are zero, and the distribution $P^*$ resulting from to the described procedure:

$$\sum_{i=0}^{n} p_i = 1 \tag{5.6}$$

$$\wedge \quad 1 - \{\#0 \times [(n - \#0) \times \epsilon]\} + \{(n - \#0) \times [\#0 \times \epsilon]\} = 1 \tag{5.7}$$

$$\implies \quad \sum_{i=0}^{n} p_i^* = 1 \tag{5.8}$$

The pseudocode is shown in Algorithm 15.

**Algorithm 15** Epsilon Smoothing algorithm pseudocode

```
1: ϵ = smallest positive constant
2: for s in model CPTs do
3:     for c in s's columns do            ▷ distributions of values are organised column-wise
4:         num_zeros = number of zero-valued entries in c
5:         num_non_zeros = |c| - num_zeros
6:         for v in c do
7:             if v equal to 0 then
8:                 v+ = ϵ × num_non_zeros
9:             else
10:                v− = ϵ × num_zeros
11:            end if
12:        end for
13:    end for
14: end for
```

If the procedure were applied to the CPT in Table 5.10, it would yield the one shown in Table 5.7.

Table 5.7. "recettori estrogeni" CPT

| | | **mut17q21** | |
|---|---|---|---|
| | | 0 | 1 |
| **rec. estr.** | 0 | $0.68-\epsilon$ | 0.13 |
| | 1 | $0.0+2\epsilon$ | 0.02 |
| | 2 | $0.31-\epsilon$ | 0.84 |

### 5.5.2 MPE Calculation

The main issue encountered during the implementation of the system described in Section 5.2, was that of correctly calculating the MPE. Initially, an attempt was made to write a custom function `export_model_to_uai` to generate a UAI file (described in 4.3.1) directly from the Pomegranate Bayesian network model. This UAI file was fed to DAOOPT to generate the MPE solution as recounted in Subsection 4.3.2 under the "MPE Algorithms Comparison" header. When compared with the MPE solution generated directly using Pgmpy's `map_query` function, it was seen that these disagreed in almost all cases. This shouldn't have been the case as both were generated using exact methods: *variable elimination* in Pgmpy's case and *AND/OR branch-and-bound* for DAOOPT.

While investigating the cause for this divergence, an undocumented feature of Pgmpy was discovered: a `UAIWriter` class that should have converted the Pgmpy-based model (which was converted in turn from the Pomegranate-based model, as outlined in Subsection 4.3.2 at the "Pairwise Correlations" header) to the correct UAI file representing it. When this alternative UAI file was used as input for DAOOPT, the resulting MPE not only diverged from that calculated based on `export_model_to_uai`, as was to be expected, but also from that calculated directly with Pgmpy using `map_query`, which was surprising.

The data flow for the MPE calculation is shown in Figure 5.12; initially a Pomegranate-based BN is learned from the data set by using the built-in `from_samples` method, then a custom `convert_to_pgmpy` function converts the Pomegranate-based model to a Pgmpy-based one. The custom function `export_to_uai` and Pgmpy's built-in `UAIWriter` class are used to generate the `.uai` and `.uai.evid` files that are the input for DAOOPT to generate the MPE solutions. The `map_query` method that is part of Pgmpy's API is also used to generate an MPE assignment.

The conversion from the Pomegranate-based to the Pgmpy-based model was thoroughly tested using conditional probability and independencies queries, so the issue is most likely to be found elsewhere. As the DAOOPT MPE solution generated starting from the UAI differs from the one calculated directly with Pgmpy, there must be a bug either in Pgmpy's UAI exporter or in its inference method. It is unclear if Pgmpy is still presenting issues in its inference methods[4] but a series of tests on simple networks, where the MPE calculations were carried out manually, seemed to confirm that `map_query` was returning the correct MPE solution.

For example, in the small BN whose structure is shown in Figure 5.13 and the CPDs of the nodes in Tables 5.8, 5.9, 5.10 and 5.11, Pgmpy's `map_query` and DAOOPT returned different

[4] `https://github.com/pgmpy/pgmpy/issues/856`

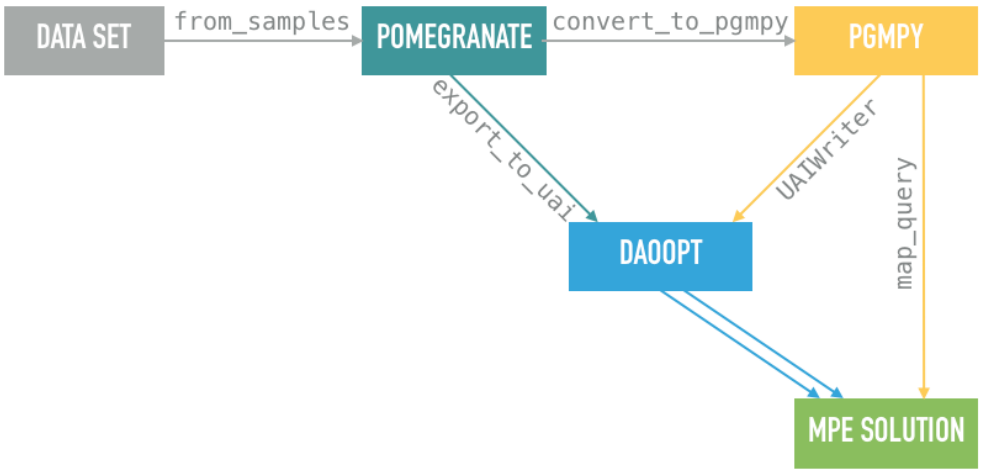



Figure 5.12. MPE calculation flow.

solutions to the following MPE query:

$$\text{MPE}(\text{“differenziazione”} = x, \text{“mut17q21”} = y, \text{“recettori estrogeni”} = z \mid \text{“eta arrotondata”} = 0) \tag{5.9}$$

`map_query` returned the assignment:

$$(\text{“differenziazione”} = 1, \text{“mut17q21”} = 1, \text{“recettori estrogeni”} = 2) \tag{5.10}$$

while DAOOPT on the UAI exported with Pgmpy's returned:

$$(\text{“differenziazione”} = 0, \text{“mut17q21”} = 1, \text{“recettori estrogeni”} = 2) \tag{5.11}$$

In such a small network it is easy to verify that the probability of Equation 5.10 is: $0.99 \times 0.84 \times 0.60 = 0.50$ and that it is the MPE solution. The fact that the probability of Equation 5.11 is $0.99 \times 0.84 \times 0.21 = 0.17$, makes it obviously incorrect.

Table 5.8. “mut17q21” distribution

| | | |
|---|---|---|
| **mut17q21** | 0 | 0.01 |
| | 1 | 0.99 |

### 5.5.3 Late Removal of Clinical Variables

As discussed in Section 4.2, it was decided to drop certain variables from the initial data set. Among these, “mut17q21”, “loss 17” and “FISHRatio” were initially included in the post-processed data set and became part of the Bayesian network. Thus, in the initial phases of development and validation the network topology was the one shown in Figure 5.14, which can be compared with the current one shown in Figure 5.3.

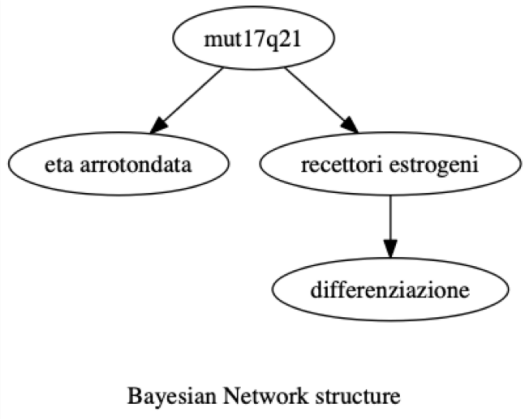



Figure 5.13. Independencies query natural language output.

Table 5.9. “eta arrotondata” CPT

| | | **mut17q21** | |
|---|---|---|---|
| | | 0 | 1 |
| | 0 | 0.42 | 0.04 |
| **eta arr.** | 1 | 0.42 | 0.17 |
| | 2 | 0.15 | 0.78 |

Table 5.10. “recettori estrogeni” CPT

| | | **mut17q21** | |
|---|---|---|---|
| | | 0 | 1 |
| | 0 | 0.68 | 0.13 |
| **rec. estr.** | 1 | 0.0 | 0.02 |
| | 2 | 0.31 | 0.84 |

Table 5.11. “differenziazione” CPT

| | | **recettori estr.** | | |
|---|---|---|---|---|
| | | 0 | 1 | 2 |
| | 0 | 0.012 | 0.16 | 0.21 |
| **diff.** | 1 | 0.18 | 0.43 | 0.60 |
| | 2 | 0.80 | 0.40 | 0.18 |

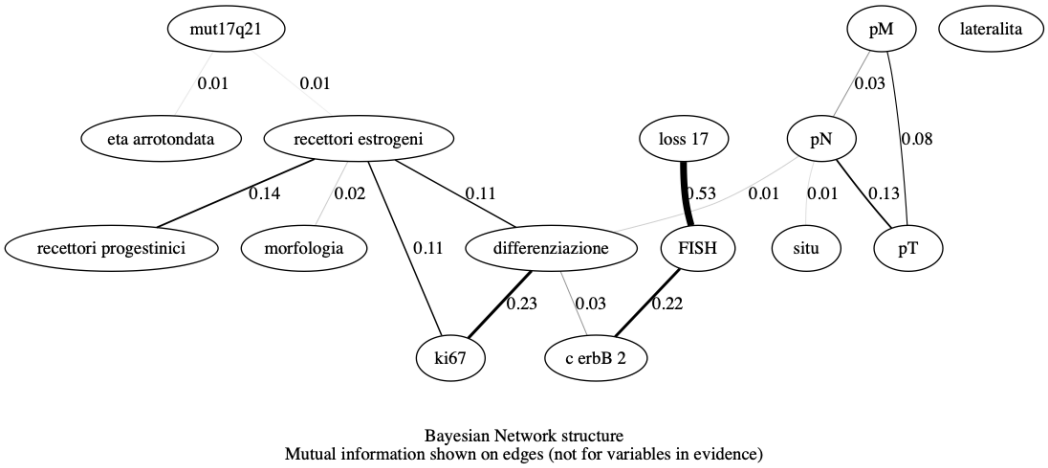



Figure 5.14. Bayesian network topology before the removal of "mut17q21", "loss 17" and "FISH".

At a later phase a decision was made, in agreement with the ICP, to remove these three variables from the data set during the preprocessing phase; this was due to their being extremely skewed in their values, as can be seen by inspecting the "Distribution" column in Table 4.2.

The reason for their skewness was briefly mentioned in Subsection 4.2.2 and can be traced to the fact that these variables are all connected to the technique of *fluorescence in situ hybridisation*. This can be understood by analysing the "Clinical meaning" column in Table 4.1 and knowing that FISH enables the analysis of specific DNA sequences on chromosomes and in particular of chromosome 17, because of its significance in breast cancer [Zhang and Yu, 2011]. FISH was a technique that was not available prior to 2010 and thus nearly 70% of the patients in the data set had a value of "NCO" for "FISHRatio", meaning this test had not been carried out on them. Even worse, "mut17q21" presented more than 99% "unknown" values and "loss 17" had 78% of "FISH non fatta/FISH non valutabile", meaning that the great majority of cases presented values with no real clinical meaning.

It can be seen in Figure 5.14 how strong the association between "loss 17" and "FISH" was - *a posteriori*, a clear case of spurious correlation - and how feebly "mut17q21" was connected to the rest of the network. Having these variables was introducing a very large amount of bias that confounded the resulting model. An example of this effect was experienced in the early stages of validation; a previous series of validation questions had been prepared by the ICP and included queries in a form similar to:

> In the general population, if *[...]* is *[...]*, then is it more/less probable that *[...]* is *[...]*?

and

> In young patients, if *[...]* is *[...]*, then is it more/less probable that *[...]* is *[...]*?

Queries of these forms, when compared against each other, reliably returned identical answers, indicating that the age of the patient ("eta arrotondata" in the benchmark data set) had little or no influence of the values of other variables. Inspection of Figure 5.3 will show how, after

the removal of "mut17q21", "loss 17" and "FISHRatio", "eta arrotondata" actually becomes disconnected from the rest of the network. The removal of the spurious "bridge" created by "mut17q21" now explicitly shows and confirms that patient age can have no influence on other variables' values, as had already been noticed before removal.

The ICP confirmed that removing these three variables helped its users in better understanding the differences between a correlation in the variables representing the data set and a correlation in the actual clinical meaning behind the data. That is, they felt that it reduced the *confounding factors* thus allowing a better appreciation of the explainability methods used.

## 5.6 Summary

This chapter has dealt with the outcomes of the work introduced in Chapter 4: the developed proof of concept software system which was used to validate the initial hypotheses regarding the explainability of Bayesian networks.

The implemented tool has been described and presented in detail by looking at every user interaction mode separately and at how these relate to the framework proposed by Lacave and Díez [2002] and the characteristics of an explanation advanced by Miller [2018]. Where present, the outcomes of an "informal evaluation" of the system have been included; these are the result of observing and interacting personally with the ICP's experts. The interaction modes analysed are:

- "plot model";
- "independencies/d-separation query";
- "conditional probability query";
- "MPE query";
- "pseudo-MPE query";
- three variants of "dialogue".

The first validation results that have been reported are those regarding the *clinical validation* of the system as the outcome of these underpins the validity of all others. The complete series of results to the natural language questions presented in Subsection 4.5.1 has been presented, together with the experts' comments. The results, when summarised, confirm that the ICP considers the system's outputs clinically valid. An extra discussion to better contextualise the clinical relevance of the system has also been included.

The evaluation of the system's explainability hinges on an "explainability evaluation questionnaire", a method borrowed from the social sciences, which was submitted to the ICP around three weeks after they had started using the software. The questionnaire aims to disentangle the explainability of the single interaction modes and thus trace them back to more general explainability concepts for BNs. The results seem to confirm that the system was indeed explainable, based on the warm reception it received from seasoned medical professionals, who had explicitly stated their aversion to having to deal with the inner workings of a ML tool. The dialogical mode of interaction was appreciated in its novelty and potential to conduct clinical research, but was ultimately deemed more onerous than the other implemented approaches. A

novel result was that *linguistic* explanations were consistently rated more highly than *graphical* ones, contradicting findings in [Lacave and Díez, 2002].

An evaluation has then been discussed for a non-user-facing feature comparing the quality of the solutions given by the "pseudo-MPE" algorithm with the true MPE.

The final section of the chapter has assessed the main issues encountered during the development and implementation of the methods in this thesis. These include zero-valued entries in the conditional probability tables learned by Pomegranate and an outline of the algorithm used to solve this problem, complications in the calculation of the true MPE solution using DAOOPT and Pgmpy, and the effect that a late-stage removal of three variables from the data set had on the resulting model.

# Chapter 6

# Conclusions

## 6.1 Discussion

The purpose of this thesis is to investigate the *explanatory powers* of Bayesian networks.

The motivation for undertaking this research is connected to the recent surge in the use and integration of AI into the fabric of our societies (see Section 1.1). It has thus become imperative for these systems to be *explainable*; that is, for its users to be able to understand the *reasoning* behind the machine's outputs (see Section 2.2). This need is even more pressing in mission-critical domains such as that of *medicine*.

The basis for the research developed in this thesis was the paper "Explaining the Most Probable Explanation" [Butz et al., 2018] (see Section 2.6). This foundational work, while proposing a seemingly appealing method to enable the understanding of a medical data set, failed - as many other xAI works do - to provide any validation for its claims. Thus a proof of concept system was developed[1] with the purpose of validating the claims made by the paper and, more in general, to investigate the ability of Bayesian networks to provide meaningful explanations to their users. The benchmark against which the system's outputs have been compared have been the *explainability framework for BNs* offered by Lacave and Díez [2002] and also the *psychological characteristics inherent to an explanation* identified by Miller [2018] (see Section 2.5). One of the main gaps in the xAI literature has been the absence of substantial validation of the models being proposed by researchers; therefore one of the main objectives of this thesis was to provide a methodological framework for the evaluation of machine learning systems with real domain experts i.e., an *application-grounded evaluation* methodology [Doshi-Velez and Kim, 2017] (see Section 2.4).

The prototype system was created by the implementation of standard techniques (see Subsection 4.3.2) and the development of novel ones (see Subsection 4.4.1). The underlying Bayesian network was learned through use of a real medical data set (see Section 4.2), which was provided by a medical partner with a high degree of involvement at every step of this research process (see Subsection 4.2.1). Expert pathologists aided by informing of the desiderata of the developed system and particularly in the crucial stage of validating it from a clinical relevance point of view and as regards its ability to interact meaningfully with them. That is, they both informed the design and evaluated its *explainability*.

[1] `https://github.com/Tioz90/Bayesian-Networks-Explainability-Tool`

The clinical relevance was evaluated by asking the medical experts to define a series of natural language clinical questions, which were then mapped and executed on the system's user interaction modes (see Subsection 4.5.1 and Appendix B). The ensuing results from these queries were compared by the medical experts with those that they would have expected, based on medical literature and their personal expertise (see Subsection 5.3.2). In this respect, it has been shown that the tool, and the underlying BN, were able to capture and respond in a significant manner to nearly all these questions thus validating the software and ML model from a clinical relevance point of view. This first validation step was important in order to establish a solid basis for the users to trust the system; if the system had been incapable of implementing the questions asked by the users or of offering them answers conforming to their expectations, it would have then been very difficult for it to then provide any meaningful explanation to the medical experts (see Section 2.3). This is because the users would not have trusted its outputs and, as discussed throughout Chapter 2, an explanation becomes such by virtue of a dialogue between an *explainer* (the machine) and an *explainee* (the user). If the two actors involved in an explanation are not able, or willing, to interface in a certain way, an explanation simply never comes into being. The manner in which humans and machines should interact (e.g., in terms of outputs, trust) in order for the former to explain something meaningful to the latter and the ways to elicit this interchange, are the focus of the explainable AI field.

The evaluation of the explanatory powers of the tool was carried out by both an *informal evaluation* (see Section 5.2), consisting in observing the experts using the tool and recording their impressions and issues, and a *formal one* (see Subsection 4.5.2 and 5.3.3), involving an "explainability evaluation questionnaire" (see Appendix C) geared towards probing the explainability of the system compared to the concepts given by Lacave and Díez [2002] and Miller [2018]. Both evaluations confirmed that the *dialogical* explanation mode proposed in Butz et al. [2018] was the least effective means to offer the experts an explanation. Thus, the claims made in the paper cannot presently be substantiated by this thesis; however, such results do not disprove the explainability of Bayesian networks as a whole. Another result, confirmed by both evaluations, was that the experts were very biased towards preferring the *linguistic* explanation modality over any other; this seems to disprove the statement "the most direct and intuitive way of showing the information embodied in a Bayesian network is to display the corresponding graph" [Lacave and Díez, 2002]. But, this result is also a step in vindicating the initial claim made in "Explaining the Most Probable Explanation" [Butz et al., 2018] that BNs are hard to interpret for medical domain experts, even though they provide a graphical representation of their knowledge base. These authors attempted to solve this issue through the use of *dialogue* but, as discussed, the conclusions of this thesis can only confirm that this is still an open problem. If the explainability of Bayesian networks is to be approximated in the satisfaction of the users of the system, then the work carried out in this thesis certainly seems to be a step in the direction of confirming the explanatory powers of such graphical models. The medical experts confirmed that they were able to interact meaningfully with the system and expressed the desire to continue using it on newer data sets.

The developed proof of concept software tool presented a mix of *static* and *dynamic* explanations together with *contrastive*, *linguistic* and *graphical* output modalities. Even though the "dialogues" - which are instances of a *dynamic*, *contrastive*, *linguistic*, and *graphical* explanation - were not easy for the users to use meaningfully, the other explanatory modes - consisting of "MPE", "pseudo-MPE", and "conditional probability" queries - seemed to completely satisfy the experts. These other explanatory modes are notable examples of the kind of *uncertain reasoning* that can be performed with Bayesian network inference, provided that suitable algorithms

for those computational tasks are provided. They also make good use of one of the core characteristics of BNs, namely that their outputs can be *selected*. This selectivity means that an explanation may contain only the information necessary to be efficiently conveyed to the user; neural networks, for example, are not capable of such an output. The results of this thesis thus seem to indicate that simple *selection* of the outputs may be more important to medical users that the other characteristics - present in the "dialogues" and to a certain extent in "pseudo-MPE queries" - of explanations identified by Miller [2018]: *contrastiveness*, *causality* and *sociality*.

Regarding the objective of laying out an *evaluation methodology groundwork* for future *application-based evaluations* of machine learning models in the medical domain, we believe that the work achieved by this thesis - barring the limitations recognised in Section 6.2 - was a worthwhile undertaking. The evaluation methodology served to develop results which did not align with the established literature and these thus merit further consideration; for example, the fact that the medical experts preferred *linguistic* explanations over any other. The *time* needed to understand an explanation was also included as part of this evaluation and this presents an element of novelty because, as noted by Gilpin et al. [2018], the temporal component is usually disregarded in xAI literature. Nonetheless, many interesting results specific to the medical domain (reported throughout Chapter 5 where relevant) were brought to light through the application of the *research methodology* and the prototype system, which was also a result of it, was warmly received by its expert users. This supports the soundness of this research methodology in its capacity to accurately characterise the domain of interest and to inform the building of effective *explanatory tools* within it.

## 6.2 Future Work

When considering possible future work, one needs to distinguish between tasks whose scope is to *address limitations of the current methods* and those related to *expansion of the current work* and *novel applications for it*.

### 6.2.1 Addressing Limitations of Current Work

#### Methodology

Based on the "formal" (see Subsection 5.3.3) and "informal" (see Section 5.2) feedback received by the medical partner, it appears that the "dialogue" interaction modes (see Subsection 5.2.7) should be modified, as the experts were not easily able to understand their workings, even after having perceived their potential. The experts believed that with extra time they would have been able to use this interaction mode productively, but this is a symptom of a failure on the part of the software as a system designed to be explainable should definitively require as little effort from the user as possible. The ICP has nonetheless confirmed its intention to continue using the software, focusing in particular on the dialogue modes of interaction, as they feel they have potential as research tools. The evaluation methodology borrowed techniques from the social sciences but could undoubtedly be improved by experts in this domain, who will certainly be better versed in the methodological details compared to the author of this thesis, whose academic background is firmly in computer science.

#### Additional Evaluations

A more extended evaluation period could certainly be recommended as it would also enable the assessment of the effect of *novelty* of certain interaction modes and help in factoring out the *experts' preconceptions* regarding what an explanation should look like (see Subsection 5.3.1).

The "pseudo-MPE" query also presented elements of ambiguity for the experts, as they were confused on the non-monotonicity of the probability of the elements in the deduction chain (see Subsection 5.2.6). This is unquestionably a point to investigate further by implementing and evaluating alternative *output modalities*, as well as potentially *linguistic* ones as these were proved to be preferred by the clinicians over all others. The objective would be to identify the *point of attrition* and discriminate if it were to be found in the actual underlying method or only in the way its outputs were displayed. In the same vein, the reasons for the experts' misunderstanding in how to implement questions 23, 25, 27 on the system (see Section 5.3) should also be investigated more thoroughly.

#### System

The system developed in this thesis is bound by the limitations inherent to any proof of concept software, namely a general lack of polish and of somewhat lacking usability. The methods themselves are studied to be able to surface explanations in the clearest way possible, but substituting the console-based frontend for a GUI - local or web-based - would beyond doubt lead to a marked improvement in the user experience. Also related to the experimental nature of the software is the fact that it was not built from the start with a coherent end-goal but was extended in a non-organic manner as new methods were selected for exploration or novel ones were developed. As a result the implementation presents some fragmentation and is rich in "workarounds". The choice of Python as the implementation language, while definitively advantageous for rapid development thanks to the comprehensive set of data science and machine learning libraries available, brought with it some clear limitations. A Python application, while portable across different systems, does not provide a *native* experience on any platform; the current project also relied heavily on the *Anaconda* package manager[2] so any user wishing to use the tool on their machine would need to deal with a potentially intricate setup process. A better alternative to a full rewrite in a compiled language would be a web GUI to a Python backend[3] that would enable portability without requiring to completely change the implementation. Nevertheless, a rewrite of the application that dropped many redundancies, unused code, and non-user-focused features, is a necessary part of future work.

### 6.2.2 Extensions and Novel Applications

The second class of future work concerns the expansion of the current techniques. During this research it was not possible to utilise the software tool DAOOPT (see Subsection 4.3.1 under the "DAOOPT" header) for exact MPE inferences and it was consequently not used as a benchmark for the "pseudo-MPE" algorithm (see Subsection 5.5.2). There is, however, a roadmap to follow up in more detail on the evaluation of the "pseudo-MPE" algorithm in a separate paper co-authored with IDSIA researchers.

[2] `https://www.anaconda.com`

[3] For example by using `https://github.com/epeios-q37/atlas-python`.

One of the issues encountered in this thesis was the late removal of three variables due to their lack of clinical significance, as described in Subsection 5.5.3. The ICP is in possession of a newer data set, homogeneous to the *benchmark data set* used throughout this thesis (see Section 4.2), where the values of two of the three variables ("FISH" and "loss17") are defined and thus have clinical noteworthiness. The third variable ("mut17q21") is still missing too high a number of values to be relevant but these could be predicted using the BN or other discriminative ML techniques. The estimation of the values of this last variable is bound to be accurate, as there is a real *causal* dependency between it and the other two; this is because "mut17q21", the mutation on chromosome 17, is identified through the technique of *fluorescent in situ hybridisation* and is thus also tightly coupled with the results of "loss17". These are all novel clinical variables and they are thus open to being the subject of many research questions; the medical partner has confirmed its interest in using the tool developed in this thesis to pursue such investigations and this could, potentially, lead to scientific publications.

Connected to the previous point, the expert users suggested developing a "workflow" for the importing of new data sets, thus extending the clinical capabilities of the tool. The experts also felt that the ability to save the textual and visual outputs of previous queries could undoubtedly be useful, together with the capacity to "snapshot" the state of a "dialogue" in order to resume it from a certain point onwards. Ready-made query masks that could reduce the time needed to execute similar questions were also a request.

The ICP has also confirmed that it will be investigating the clinical relevance, through literature reviews and experiments, of some of the results obtained while using the tool; in particular, these would be those related to:

- the lack of correlation between common clinical pathological features (i.e., morphology, TNM, grade, hormonal status) with age at diagnosis;
- the correlation between positive progesterone expression and low tumour proliferation index (ki67);
- the correlation between negative oestrogen receptor expression and high tumour proliferation index (ki67).

Finally, some interesting research avenues were not explored, for example a work by Kyrimi and Marsh [2016] who introduced a method that can be seen as the "inverse" of that proposed by Butz et al. [2018]. Instead of looking for the outcome best explained by the given evidence, this other paper proposes techniques to find the evidence that best explains the chosen target variables. Adapting this "inverse" method could potentially extend the *explanatory powers* of the system, for example by enabling clinicians to understand which variables best justify a series of observed features in a patient. It should not be too difficult an undertaking, as the current Python implementation is highly modular, by virtue of being based on standard open-source data structure libraries.

# Appendix A

# Acronyms

This first annex contains a list of acronyms used throughout this thesis, paired with the corresponding phrases of which they are the abbreviation. Throughout the text, the acronym is usually paired with the phrase it refers to when first used.

- **AI**: Artificial intelligence
- **API**: Application programming interface
- **BN**: Bayesian network
- **CPD**: Conditional probability distribution
- **CPT**: Conditional probability table
- **DAG**: Directed acyclic graph
- **EBNF**: Extended Backus-Naur form
- **FISH**: Fluorescence in situ hybridisation
- **GDPR**: General Data Protection Regulation
- **GUI**: Graphical user interface
- **ICP**: Istituto Cantonale di Patologia
- **IDSIA**: Dalle Molle Institute for Artificial Intelligence
- **MAP**: Maximum a posteriori
- **ML**: Machine learning
- **MLE**: Maximum likelihood estimation
- **MPE**: Most probable explanation
- **xAI**: Explainable artificial intelligence

# Appendix B

# Natural Language Questions

This annex contains the natural language questions formulated by the medical partner - Istituto Cantonale di Patologia - with the objective of clinically validating the proof of concept system developed as part of the methods of this thesis. These questions are referenced in the relevant Subsection 4.5.1.

Table B.1. Natural language questions answerable by conditional probability queries

| # | Natural language question | Type | Target variable | Target value | Evidence variable | Evidence value |
|---|---|---|---|---|---|---|
| **1** | At diagnosis, if estrogen receptors are negative, is tumor proliferative index high? | validation | ki67 | >30% | estrogeni | 0-10% |
| **2** | At diagnosis, if estrogen receptors are negative, is the risk of metastases low? | validation | pM sub | pM=0 | estrogeni | 0-10% |
| **3** | If estrogen receptors are negative and tumor proliferative index is high at diagnosis, is the risk of metastases low? | validation | pM sub | pM=0 | estrogeni<br>ki67 | 0-10%<br>>30% |
| **4** | If the diagnosis of mammary carcinoma happened at a young age, is tumour proliferative index high? | validation/research | ki67 | >30% | eta arrotondata | <40 |
| **5** | If the histologic diagnosis is of lobular carcinoma, is the expression of the c-erbB2 marker absent? | validation | cerb | 0 & 1 | morfologia | Lobular carcinoma, NOS |
| **6** | If the tumour is large, is lymph node involvement more probable? | validation | pN sub | pN!=0 | pT sub | pT>=2 |
| **7** | If the tumour is large and lymph nodes are involved, is the risk of metastases low at diagnosis? | validation | pM sub | pM=0 | pT sub<br>pN sub | pT>=2<br>pN!=0 |

Table B.2. Natural language questions answerable by conditional probability queries

| # | Natural language question | Type | Target variable | Target value | Evidence variable | Evidence value |
|---|---|---|---|---|---|---|
| **8** | If the tumour is of high grade at diagnosis, is the risk of metastases low? | validation | pM sub | pM=0 | differenziazione | poco differenziato |
| **9** | In young patients, if estrogen receptors are negative, is tumor proliferative index high? | validation | ki67 | >30% | estrogeni<br>eta arrotondata | 0-10%<br><40 |
| **10** | In young patients, if estrogen receptors are negative, is the risk of metastases low? | validation | pM sub | pM=0 | estrogeni<br>eta arrotondata | 0-10%<br><40 |
| **11** | In young patients, if estrogen receptors are negative and tumor proliferative index is high at diagnosis, is the risk of metastases low? | validation | pM sub | pM=0 | estrogeni<br>ki67<br>eta arrotondata | 0-10%<br>>30%<br><40 |
| **12** | How does the tumoural grade change if I know the oestrogen expression? | research | differenziazione | ben differenziato<br>moderatamente differenziato<br>poco differenziato | estrogeni | negativi (0-10%)<br>debolmente positivo (10-50%)<br>fortemente positivo (>50%) |
| **13** | How does the oestrogen expression change if I know the proliferation index? | research | estrogeni | negativi (0-10%)<br>debolmente positivi (10-50%)<br>fortemente positivi (>50%) | ki67 | negativo (0-14%)<br>14-20%<br>20-30%<br>positivi (>30%) |
| **14** | Does the negative expression of progestinic receptors influence lymph nodes' state? | validation | pN sub | pN!=0 | progestinici | 0-10% |

Table B.3. Natural language questions answerable by d-separation queries

| # | Natural language question | Type | Target variable | Target value | Evidence variable | Evidence value |
|---|---|---|---|---|---|---|
| **14** | Which clinical-pathological variables influence the lymph nodes' state at diagnosis? | research | pN | - | - | - |
| **15** | Which clinical-pathological variables influence tumoural proliferation index? | research | ki67 | - | - | - |
| **16** | Which clinical-pathological variables influence the expression of the c-ERBB2 marker? | research | cerb | | - | - |
| **17** | Which clinical-pathological variables influence the oestrogen expression? | research | recettori estrogeni | | - | - |
| **18** | Which clinical-pathological variables influence the tumoural grade? | research | differenziazione | | - | - |
| **19** | Which clinical-pathological variables influence the presence of metastases at diagnosis? | research | pM | | - | - |
| **20** | Which clinical-pathological variables influence the tumoural dimension? | research | pT | | - | - |
| **21** | Which clinical-pathological variables influence the age of the tumour onset? | research | eta | - | - | - |

Table B.4. Natural language questions answerable by conditional probability queries or, at a higher level, by d-separation queries

| # | Natural language question | Type | Target variable | Target value | Evidence variable | Evidence value |
|---|---|---|---|---|---|---|
| **22** | In young patients, does a negative expression of the progestinic receptors influence the lymph nodes' state? | research | pN sub | pN!=0 | progestinici<br>eta | 0-10%<br><40 |
| **23** | Does a negative expression of progestinic receptors influence the tumoural proliferation index? | research | ki67 | >30% | progestinici | 0-10% |
| **24** | In young patients, does a negative expression of the progestinic receptors influence the tumoural proliferation index? | research | ki67 | >30% | progestinici<br>eta | 0-10%<br><40 |
| **25** | Does a negative expression of progestinic receptors influence the expression of the c-ERBB2 marker? | research | cerb | 0 & 1<br>2<br>3 | progestinici | 0-10% |
| **26** | In young patients, does a negative expression of the progestinic receptors influence the expression of the c-ERBB2 marker? | research | cerb | 0 & 1<br>2<br>3 | progestinici<br>eta | 0-10%<br><40 |

Table B.5. Natural language questions answerable by MPE queries

| # | Natural language question | Type | Target variable | Target value | Evidence variable | Evidence value |
|---|---|---|---|---|---|---|
| **27** | How are tumours characterised by a triple negative profile from the point of view of the other clinical-pathological variables? | research | - | - | cerb<br>recettori estrogeni<br>recettori progestinici | 0<br>negativo<br>negativo |
| **28** | How are tumours characterised by high ki67 from the point of view of the other clinical-pathological variables? | research | - | - | ki67 | >30 |
| **29** | How are tumours characterised by nodes involvement from the point of view of the other clinical-pathological variables? | research | - | - | pN | !0 |

# Appendix C

# Questionnaire

This second addition contains the "explainability evaluation questionnaire" that was prepared in order to "formally" test the *explanatory powers* of the prototype system developed during this thesis. The questionnaire is references in the relevant Subsection 4.5.2.

## Explainability evaluation questionnaire

### Confidence

1. Did the tool increase the confidence in diagnosis when diagnostic screening results were missing for a patient? Why?
   O Yes O No
2. Did the tool help in characterising a particular patient's profile?
   O Not at all O Somewhat O Absolutely
3. Did the tool help in your confidence of understanding the cohort characteristics? How?
   O Not at all O Somewhat O Absolutely
4. Did the tool improve your confidence in your clinical decision-making? How?
   O Not at all O Somewhat O Absolutely
5. Did having the tool at your disposal improve your confidence when making time-constrained decisions? How? (for example, did it improve confidence in prioritising some tests over others?)
   O Not at all O Somewhat O Absolutely

### Features

6. Given the modes of interaction with the system labelled as "dialogues", do you think you would have had more difficulty in interpreting the data without the these modalities?
   O No O Maybe O Yes

7. Was natural language useful during the interaction? Why?
   O No O Maybe O Yes
8. Which type of "dialogue" did you feel was most useful? Why?
   O Exhaustive O Separations O Thresholded O A combination of the previous O All O None
9. Did you feel that the dialogue helped you in cases of uncertainty? If yes, how? If no, why?
   O No O Somewhat O Yes
10. Did you feel that the "dialogue" helped your clinical decision-making? If yes, how? If no, why?
    O No O Somewhat O Yes
11. Did the generation of "counterfactual branches" help in your understanding of the data? Why?
    O No O Somewhat O Yes
12. Given the interaction mode labelled "pseudo-MPE query", how would you rate the solutions it proposed from a point of view of their understandability? (1 poor, 5 good)
    O 1 O 2 O 3 O 4 O 5
13. How would you rate the "pseudo-MPE" solutions from a point of view of their clinical usefulness?
    O 1 O 2 O 3 O 4 O 5
14. Do you feel that the interaction mode labelled as "MPE query" gave better solutions than that labelled "pseudo-MPE query"? Why?
    O No O Maybe O Yes
15. Did you find the "pseudo-MPE" or "MPE" interaction mode the most useful? Why?
    O "pseudo-MPE" O MPE O Both O None
16. How important was the highlighting of the independencies between variables?
    O 1 O 2 O 3 O 4 O 5
17. Do you think you would have had more difficulty in interpreting the data without the correlation strength displayed?
    O No O Maybe O Yes
18. Do you think you would have had more difficulty in interpreting the data without visualisations?
    O No O Maybe O Yes
19. Do you think you would have had more difficulty in interpreting the data without natural language output?
    O No O Maybe O Yes

## Time

20. How would you rate the time it took to understand the dialogues' outputs? Which of the three was best? (1 bad, 5 good)
O 1 O 2 O 3 O 4 O 5

21. How would you rate the time it took to understand the conditional probability query's outputs
O 1 O 2 O 3 O 4 O 5

22. How would you rate the time it took to understand the MPE and "pseudo-MPE" query's outputs?
O 1 O 2 O 3 O 4 O 5

23. Did natural language help in reducing the time needed to understand the outputs?
O No O Somewhat O Yes

24. Did visualisations help in reducing the time needed to understand the outputs?
O No O Somewhat O Yes

## Tool

25. Which interaction modes did you feel could be the most useful? Why?
O Plot model O Independencies O Conditional Probability Query O "pseudo-MPE" and MPE O Dialogues

26. Which interaction modes did you use the most? Why?
O Plot model O Independencies O Conditional Probability Query O "pseudo-MPE" and MPE O Dialogues

27. How did you use the tool in your day-to-day work?

28. Is the tool missing any functionality that would address your needs? If yes, which ones?
O No O Yes

29. Did you have any difficulties in understanding which functionalities to use to address your needs? If yes, when?
O No O Yes

30. Did you have any difficulties in understanding the functionalities during usage? If yes, when?
O No O Yes

31. If you answered Yes to the previous question, how do you think this could be addressed?

32. Could you suggest any functionalities you would like to be implemented?

## Clinical

33. Did the tool help in recovering missing features of patients thus supporting diagnostic profile creation and decision making? If yes, which is/are the feature/s that benefited the most?
O No O Yes

34. Did any of the tool's predictions have clinical confirmation later on? If yes, how?
O No O Yes

35. Did the tool help in highlighting new relationships between variables?
O No O Yes

36. Did the tool help in highlighting new patient subgroups?
O No O Yes

## Satisfaction

37. What is your general satisfaction with the tool? For what reasons?
O Completely dissatisfied O Somewhat dissatisfied O Neutral O Somewhat satisfied O Completely satisfied

# Bibliography

*Statistics for Research*, volume 3. 2006. ISBN 047126735X. doi: 10.2307/3324586.

Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, page 582. ACM, 2018.

Paul Anand, Prasanta Pattanaik, and Clemens Puppe. *The Handbook of Rational and Social Choice*. Oxford University Press, 2009.

Constantin Berzan. An Exploration of Structure Learning in Bayesian Networks. *Tufts University Senior Honors Thesis*, 2012.

Or Biran and Kathleen McKeown. Human-centric Justification of Machine Learning Predictions. *IJCAI International Joint Conference on artificial intelligence*, pages 1461–1467, 2017. ISSN 10450823.

Nick Bostrom and Elizer Yudkowsky. The Ethics of Artificial Intelligence. *Cambridge University Press*, 39(1):56–57, 2011. ISSN 13573039. doi: 10.1016/j.mpmed.2010.10.008.

Jeffrey S. Bowers and Colin J. Davis. Bayesian Just-so Stories in Psychology and Neuroscience. *Psychological Bulletin*, 138(3):389–414, 2012. ISSN 00332909. doi: 10.1037/a0026450.

Raphaela Butz, Arjen Hommersom, and Marko van Eekelen. Explaining the Most Probable Explanation. In *Lecture Notes in computer science (including subseries Lecture Notes in artificial intelligence and Lecture Notes in Bioinformatics)*, volume 11142 LNAI, pages 50–63. 2018. ISBN 9783030004606. doi: 10.1007/978-3-030-00461-3_4.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics Derived Automatically from Language Corpora Contain Human-like Biases. *Science*, 356(6334):183–186, 2017.

Thomas Cover and Joy Thomas. *Elements of Information Theory*. 2006.

Derek Doran, Sarah Schulz, and Tarek R. Besold. What does Explainable AI Really Mean? A new Conceptualization of Perspectives. *CEUR Workshop Proceedings*, 2071, 2018. ISSN 16130073. doi: 10.2307/1541581.

Finale Doshi-Velez and Been Kim. Towards a Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608*, 2017.

Filip Karlo Dosilovic, Mario Brcic, and Nikica Hlupic. Explainable Artificial Intelligence: a Survey. *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2018 - Proceedings*, pages 210–215, 2018. ISSN 0277-0008. doi: 10.23919/MIPRO.2018.8400040.

Lilian Edwards and Michael Veale. Enslaving the Algorithm: From a 'Right to an Explanation' to a Right to Better Decisions? *IEEE Security & Privacy*, 16(3):46–54, 2018.

José A Gámez, Serafín Moral, and Antonio Salmerón Cerdan. *Advances in Bayesian Networks*, volume 146. Springer, 2013.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining Explanations: An Overview of Interpretability of Machine Learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

Anthony Gitter and Casey Greene.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A Survey of Methods for Explaining Black Box Models. *ACM computing surveys (CSUR)*, 51(5):93, 2018.

David Gunning. Explainable Artificial Intelligence (xAI). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2, 2017.

Max Henrion and Marek J Druzdzel. Qualititative Propagation and Scenario-based Scheme for Exploiting Probabilistic Reasoning. In *Proceedings of the Sixth Annual Conference on Uncertainty in artificial intelligence*, pages 17–32. Elsevier Science Inc., 1990.

Denis J. Hilton. Conversational Processes and Causal Explanation. *Psychological Bulletin*, 107 (1):65–81, 1990. ISSN 00332909. doi: 10.1037/0033-2909.107.1.65.

Simon Jackman. *Bayesian Analysis for the Social Sciences*. 2009. ISBN 9780470011546.

Kalev Kask and Rina Dechter. Stochastic Local Search for Bayesian Networks. *in Proceedings Seventh International Workshop on artificial intelligence and Statistics*, 1999.

J. F. C. Kingman and Solomon Kullback. *Information Theory and Statistics*, volume 54. 2007. ISBN 0844656259. doi: 10.2307/3613211.

Daphne Koller, Nir Friedman, Sašo Džeroski, Charles Sutton, Andrew McCallum, Avi Pfeffer, Pieter Abbeel, Ming-Fai Wong, David Heckerman, Chris Meek, et al. *Probabilistic Graphical Models Principles and Techniques*, pages 74–76. MIT press, 2009.

Johan Kwisthout. Most Probable Explanations in Bayesian Networks: Complexity and Tractability. *International Journal of Approximate Reasoning*, 52(9):1452–1469, 2011.

Evangelia Kyrimi and William Marsh. A Progressive Explanation of Inference in 'Hybrid' Bayesian networks for Supporting Clinical Decision Making. In *JMLR: Workshop and Conference Proceedings vol 52, 275-286, 2016*, 2016.

Carmen Lacave and Francisco J Díez. A Review of Explanation Methods for Bayesian Networks. *The Knowledge Engineering Review*, 17(2):107–127, 2002.

Shu Hsien Liao. Expert System Methodologies and Applications-A Decade Review from 1995 to 2004. *Expert Systems with Applications*, 28(1):93–103, 2005. ISSN 09574174. doi: 10.1016/j.eswa.2004.08.003.

Zachary C. Lipton. The Mythos of Model Interpretability. (Whi), 2016. ISSN 00010782. doi: 10.1145/3233231. URL http://arxiv.org/abs/1606.03490.

P Lucas. Bayesian Networks in Medicine: a Model-based Approach to Medical Decision Making. *Proceedings of the EUNITE workshop on Intelligent Systems in patient Care*, pages 73–97, 2001.

Radu Marinescu and Rina Dechter. AND/OR Search Spaces for Graphical Models. *artificial intelligence*, (171):73–106, 2006. ISSN 00043702. doi: 10.1016/j.artint.2006.11.003.

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

Tim Miller, Piers Howe, and Liz Sonenberg. Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. 2017. URL http://arxiv.org/abs/1712.00547.

Brent Mittelstadt, Chris Russell, and Sandra Wachter. Explaining Explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 279–288. ACM, 2019.

James Moor. The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years. *Ai Magazine*, 27(4):87–87, 2006.

Judea Pearl. *Reverend Bayes on Inference Engines: a Distributed Hierarchical Approach*. Cognitive Systems Laboratory, School of Engineering and Applied Science, 1982.

Judea Pearl and Rina Dechter. Identifying Independencies in Causal Graphs with Feedback. *Intelligence*, (1):43, 1988.

Alun Preece. Asking 'Why' in AI: Explainability of Intelligent Systems and Perspectives and Challenges. *Intelligent Systems in Accounting, Finance and Management*, 25(2):63–72, 2018. ISSN 21600074. doi: 10.1002/isaf.1422.

Pramila Rani, Changchun Liu, Nilanjan Sarkar, and Eric Vanman. An Empirical Study of Machine Learning Techniques for Affect Recognition in Human-Robot Interaction. *Pattern Analysis and Applications*, 9(1):58–69, 2006.

Melvin Rittel, Horst, Webber. Dilemmas in a General Theory of Planning. *Policy*, 2:155, 1973.

Thomas D. Schneider. Molecular Information Theory Primer. (301), 2005. URL http://users.fred.net/tds/lab/paper/primer/primer.pdf.

Jacob Schreiber. Pomegranate: Fast and Flexible Probabilistic Modeling in Python. 18:1–6, 2017. URL http://arxiv.org/abs/1711.00137.

Jacob Schreiber and William Noble. Finding the Optimal Bayesian Network Given a Constraint Graph. 2017. doi: 10.7287/peerj-cs.122v0.2/reviews/2. URL https://peerj.com/articles/cs-122/.

Klaus Schwab. *The Fourth Industrial Revolution*. Currency, 2017.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: from Theory to Algorithms*. Cambridge university press, 2014.

C. E. Shannon, W. Weaver, and R. E. Blahut. The Mathematical Theory of Communication. *Urbana: University of Illinois press*, 117(April 1928):379–423, 1949. ISSN 07246811. doi: 10.2307/3611062.

Solomon Eyal Shimony. Finding MAPs for Belief Networks is NP-hard. *artificial intelligence*, 68 (2):399–410, 1994. ISSN 00043702. doi: 10.1016/0004-3702(94)90072-8.

Elliott Sober. The Principle of the Common Cause. In *Probability and causality*, pages 211–228. Springer, 1988.

Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. Interacting Meaningfully with Machine Learning Systems: Three Experiments. *International Journal of Human-Computer Studies*, 67(8): 639–662, 2009.

Sjoerd T Timmer, John-Jules Ch Meyer, Henry Prakken, Silja Renooij, and Bart Verheij. Explaining Bayesian Networks using Argumentation. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 83–92. Springer, 2015.

Wei Zhang and Yingyan Yu. The Important Molecular Markers on Chromosome 17 and their Clinical Impact in Breast Cancer. *International journal of molecular sciences*, 12(9):5672–5683, 2011.