

Guide to explainability in Artificial Intelligence





© TIC Salut Social Foundation

This report has been elaborated by the Artificial Intelligence Area of the TIC Salut Social Foundation

Authors: Susanna Aussó, Didier Domínguez and Mariona Quintana.

Electronic edition: December 2022

This work is licensed under a Creative Commons Attribution - Non-Commercial - No Derivatives 4.0 license. Reproduction, distribution, and public communication is permitted as long as the authorship and publisher are acknowledged, and no commercial use is made. The transformation of this work to generate a new derivative work is not permitted.

Table of Contents

01/ Introduction	06		
02/ Explainable Artificial Intelligence (XAI)	10		
2.1 Interpretability v. Explainability	11		
2.2 Benefits of using explainability tools in AI	11		
2.3 Questions that XAI helps to answer	14		
2.4 Explainable AI cycle	14		
03/ Taxonomy of explainability tools	16		
3.1 XAI taxonomy models	17		
3.2 Intrinsic explainability and post-hoc explainability	19		
3.3 Global explainability and local explainability	19		
3.4 Transparent models and opaque models	20		
3.5 Model-agnostic techniques and model-dependent techniques	21		
3.6 Type of explainability	21		
04/ Explainability of algorithms based on digital medical imaging	22		
4.1 CAM (Class Activation Mapping)	23		
4.2 Grad-CAM (Gradient-weighted Class Activation Mapping)	24		
4.3 LRP (Layer-wise Relevance Propagation)	26		
4.4 LIME (Locally Interpretable Model-agnostic Explanations)	27		
4.5 SHAP (Shapley Additive Explanations)	28		
05/ Explainability of algorithms based on tabular data	29		
5.1 PDP (Partial Dependence Plot)	30		
5.2 ICE (Individual Conditional Expectation)	32		
5.3 Counterfactual Explanations	34		
5.4 LIME (Locally Interpretable Model-agnostic Explanations)	35		
5.5 Anchors	35		
5.6 SHAP (Shapley Additive Explanations)	36		
06/ Explainability of algorithms based on natural language processing	40		
6.1 Shapley Additive Explanations (SHAP)	42		
6.2 GbSA (Gradient-based Sensitivity Analysis)	43		
6.3 LRP (Layer-wise Relevance Propagation)	43		
6.4 LIME (Locally Interpretable Model-agnostic Explanations)	45		
07/ References	46		

/Index of Figures

Figure 1. Chart showing the generation of information in XAI and clinical experience.	15
Figure 2. Example 1 of a taxonomy map in XAI.	17
Figure 3. Example 2 of a taxonomy map in XAI.	18
Figure 4. Example 3 of a taxonomy map in XAI.	18
Figure 5. Example 4 of a taxonomy map in XAI.	19
Figure 6. Graphic representation of the results of the CAM method.	23
Figure 7. CAM method process.	24
Figure 8. Grad-CAM views for different models.	25
Figure 9. LRP-epsilon heatmap v. original annotation.	26
Figure 10. LIME explanation in medical imaging.	27
Figure 11. SHAP explanation in medical imaging.	28
Figure 12. Explainability PDP.	30
Figure 13. Representation of PDP explainability with 2 variables.	31
Figure 14. ICE explainability.	32
Figure 15. Centred ICE explainability.	33
Figure 16. Counterfactual explainability.	34
Figure 17. LIME explainability for tabular data.	35
Figure 18. SHAP variable importance graph.	36
Figure 19. SHAP variable importance graph for prediction with omics data.	36
Figure 20. SHAP force plot for two patients.	37
Figure 21. Graphic representation of SHAP.	38
Figure 22. Graphic representation of SHAP.	38
Figure 23. Graphic representation of SHAP.	39
Figure 24. Graphic representation of SHAP.	39
Figure 25. SHAP explainability in NLP.	42
Figure 26. Text with highlighted characters according to LRP values.	43
Figure 27. Uni-feature diagram.	44
Figure 28. Bi-feature and tri-feature diagrams.	44
Figure 29. LIME explainability with NLP.	45



1.

Introduction

The Programme for the Promotion and Development of Artificial Intelligence in the Catalan Health System aims to create an environment to aid the development and implementation of Artificial Intelligence (AI) solutions to optimise processes in the Catalan health system.

TIC Salut Social Foundation has created this guide to support those involved in the development of code for Artificial Intelligence algorithms applied to the field of health. This document focuses on the importance of explainability in these developments. It aims to list and classify the main existing techniques based on the type of results to be explained.


Advances in Artificial Intelligence (AI), consisting of creating systems that can reason like human beings and learn from experience, finding out how to solve problems in specific conditions, comparing information and carrying out logical tasks in all areas of society, are now a reality. The growing availability of electronic health records, digital medical imaging tests, omics data, and a long list of health-related datasets, has given AI vast potential to improve people's well-being [1].

Machine learning methods, and more specifically deep learning techniques, are used to create complex AI algorithms that can respond to the need to learn from the great diversity of health data sources [2]. However, the use of these kinds of techniques also requires understanding of the internal functioning of the algorithms created. The complexity of the artificial neural networks used means that the decision-making mechanisms are often unknown even to the people who develop the algorithms. We need to ask questions such as: Can we understand why the models give us a particular prediction? What areas do the algorithms focus on during the learning process? Are they automatic enough or do they need human involvement? All these aspects are collected in the document "Ethics Guidelines for Trustworthy AI" published by the European Commission [3].

Therefore, ensuring the explainability of AI algorithms is key to enabling widespread implementation of this type of tool in day-to-day clinical

practice. Health care professionals must be able to trust these AI solutions to support their work; and this must be built on principles such as transparency and high standards. Health is a challenging scientific domain, but also involves ethical and legal challenges, as the decisions taken have an immediate impact on people's well-being and life [1]. The trustworthiness of AI tools must thus be based on 3 components [3]:

- 1 **Legal AI:** compliance with all applicable laws and regulations.
- 2 **Ethical AI:** ensuring ethical principles and values.
- 3 **Robust AI:** from both a technical perspective (guaranteeing the robustness of the solutions), and a social point of view (taking account of the environment in which they operate)



Each of these concepts is necessary, but not sufficient, to achieve what we know as Trustworthy AI. Another aspect to highlight is the importance of human involvement in the development process for this type of tool. Although increasingly powerful AI techniques are being introduced and implemented to solve real-world problems, they are currently not fully autonomous systems in terms of decision-making. This is especially true for medical applications, where it is imperative for humans to be involved in the process. [4] This idea is captured in the human-in-the-loop (HITL) and human-on-the-loop (HOTL) models, which take advantage of the strengths of humans and machines to produce the best results. Humans can diagnose how and why AI methods fail and reveal their drawbacks. They are thus involved in a continuous feedback process [5]:

- Humans must provide quality data in order for the algorithm to learn in the most appropriate way. This stage would include processes such as data labelling, and bias control and mitigation, etc. The machine learning algorithm will learn to make decisions based on these data.
- Algorithms synthesise the model to infer what they have learned. This step can happen in various ways, which are often opaque to the developer. At this point it is important to introduce explainability tools so humans can interpret the algorithm's decision-making mechanism and analyse the results.
- People need to test and validate the model by qualifying its results, especially when the algorithm is not completely sure or is too sure of a wrong decision.



0101

01 0 1 00 011 0101

2.

Explainable Artificial Intelligence (XAI)

011

1 1

01 0

Explainable Artificial Intelligence (Explainable AI or XAI) allows the results of an AI algorithm to be understood by humans, as opposed to the “black box” concept, when it is not possible to know which mechanisms have been activated to produce a specific response or output to an input [6].

2.1.

Interpretability v. Explainability

Interpretability: An implicit ability of a system that allows it to be logical in the eyes of the people who look at it. Interpretability shows how well a machine learning model can associate a cause with an effect. This makes it possible to observe the cause-effect relationship, but it does not provide information on the parameters involved.

Explainability: Active ability of a system to execute actions that detail its internal workings. Explainability has to do with the capacity of a model’s parameters, often hidden in deep networks, to justify the results. The model can be explained in human terms, considering both the result and the entire internal decision-making process. Explainability thus provides information about the characteristics involved in a prediction.

2.2.

Benefits of using explainability tools in AI

The benefits of explainability must be analysed from multiple points of view, since several actors are involved throughout the lifecycle of an algorithm in which different implications will be observed. Thus, XAI is not a purely technological issue. It involves a series of medical, legal, ethical and social aspects [12].

2.2.1

AI developers

From a development perspective, explainability is useful to enable developers to check their AI models in terms of more than mere performance, so it can help detect when prediction performance is based on metadata rather than the data itself.

The complexity of artificial neural networks means that the developers of an algorithm with these characteristics are often ignorant of the internal mechanisms of the model when making a prediction. In these cases their task focuses on improving metrics by configuring parameters and hyperparameters. XAI makes it possible to improve knowledge of the algorithm's internal functioning. This deeper knowledge makes it possible to find out which characteristics are involved in the result and, therefore, aids improvement of the algorithm's performance. In short, explainability allows more accurate algorithms to be developed.

For example: In a solution to predict the prognosis of COVID-19 with chest X-rays, the classifying algorithm may focus on the visual difference between a conventional X-ray and a portable X-ray (which is usually performed on patients with reduced mobility such as those in the ICU). In this case, although the model's performance may be good, due to the correlation between the type of X-ray and the patient's condition, it would be advisable to revise the learning mechanisms.

This type of phenomenon is known as the Clever Hans Effect. Initially described in social science studies, it occurs when an experimenter unintentionally affects the individual being studied with involuntary signals, so the responses are conditioned by stimuli outside the area of study. The Clever Hans effect is currently being discussed in other areas, such as Automatic Learning [13].

2.2.2

Health professionals (AI users)

From a medical point of view, the application of Explainability tools in AI algorithms is key to achieving the necessary trust on the part of end users. They are unlikely to trust a 'black box' algorithm. AI algorithms are created to support decision-making by health professionals and trust is essential in this relationship.

The various possible types of explainability can provide information at different levels. They can also present the conclusions in various formats, adapting to the use case and experts' needs. A first-level (or global) explanation thus makes it possible to understand the general characteristics that a particular model takes into account, providing rankings of the importance of characteristics that explain which variables influence the predictions the most. In contrast, a second-level (local) explanation makes it possible to identify which characteristics are relevant for a particular patient using graphics, images or numerical values, depending on the needs. That is why it is important for developers of AI solutions to be advised by the health care professionals involved during the process of implementing explainability

methods. The format and types of the explanations should be agreed upon, reaching a compromise between what is technically possible and what is useful for end users.

Therefore, XAI aids analysis by end users, making it possible to quickly identify which characteristics have more weight in a prediction or which points of an image test have contributed to a particular diagnosis.

2.2.3

Patients

The fact that explainability tools improve the knowledge of algorithmic mechanisms for both AI developers and health care professionals, makes AI solutions more trustworthy. This increase in trustworthiness eventually translates into greater trust by people, who are ultimately the ones who use AI solutions. All in all, the process of developing an AI tool must be centred on the person, who has a right to know how decisions affecting them have been made.

2.2.4

Regulators

XAI brings a degree of transparency to AI, which builds confidence on the part of the regulators who set AI guidelines. Deciphering “black box” algorithms has become essential to ensure the trustworthiness of AI and for the application of AI in medicine.



2.3

Questions that XAI helps to answer

Through explainability we can address a set of open questions before executing an AI algorithm. These questions affect different areas:

Correctness: Are we sure that all, and only, the features of interest contributed to our algorithm's decisions?

Robustness: Are we sure the model is not susceptible to disturbances?

Bias: Are we aware of any specific biases in the data that unfairly penalise groups of individuals?

Improvement: In what specific way can the prediction model be improved?

Transferability: Specifically how can the prediction model from one application domain be applied to another application domain?

Human understanding: Can we explain the model's algorithmic machinery to an expert or even to a layperson?



2.4

Explainable AI cycle

As mentioned above, the process of developing an XAI solution must involve the expert professionals who will make use of the tool to ensure that the explanations given fit medical criteria and meet their needs in terms of level of understanding.

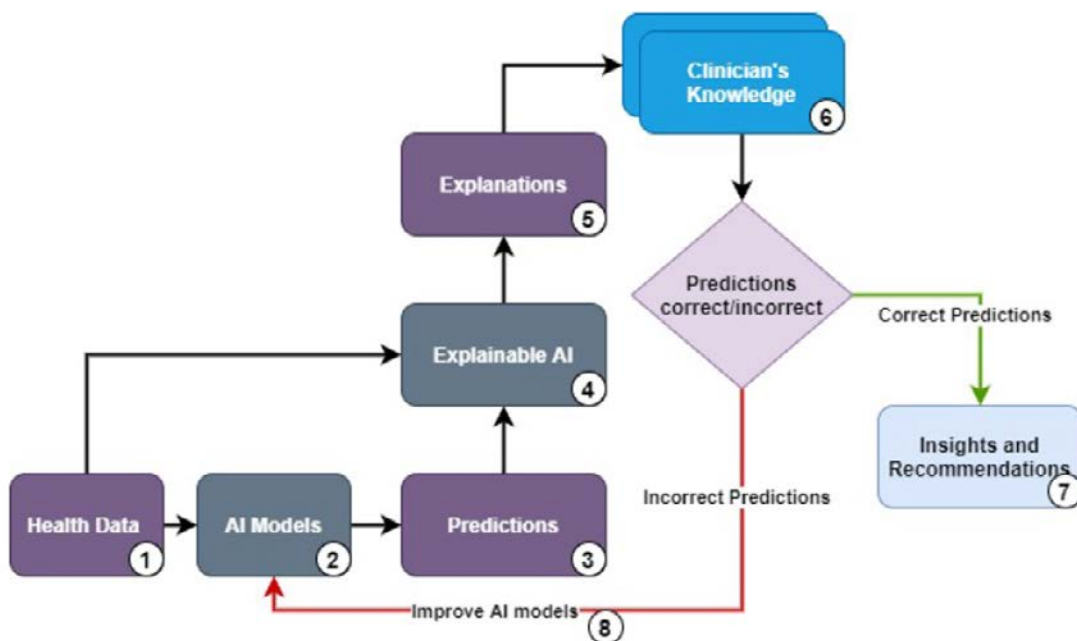


Figure 1. Chart showing the generation of information in XAI and clinical experience [11].

- Smart health apps are trained on a dataset (1) and use the resulting models (2) to make a prediction (3).
- The models obtained are used by XAI techniques (4) to generate explanations (5).
- These explanations, together with the predictions, are analysed with the help of the knowledge of health professionals who are experts in each case (6).
- If the explanations obtained do not satisfy the experts involved, despite the AI model's predictions being correct, it is necessary to review the actions that need to be taken to improve explainability. If it is detected that the model focuses on clearly erroneous features, the model must be reconsidered to analyse possible improvements (8). It may also be the case that the parameters of the explainability tool itself need to be modified, or the technique changed, because the results do not suit the model in question. This process will be repeated iteratively until a satisfactory result is achieved.
- If the explanations and predictions are validated by experts, who have verified that the algorithm focuses on parameters that make sense to them, and that this explanation meets the experts' needs, then the algorithm can be considered to correctly explain the findings (7).



3.

Taxonomy of explainability tools

3.1.

XAI taxonomy models

In general, there is a lack of consensus regarding the classification of techniques that follow XAI models. Various examples of taxonomy models are set out below.

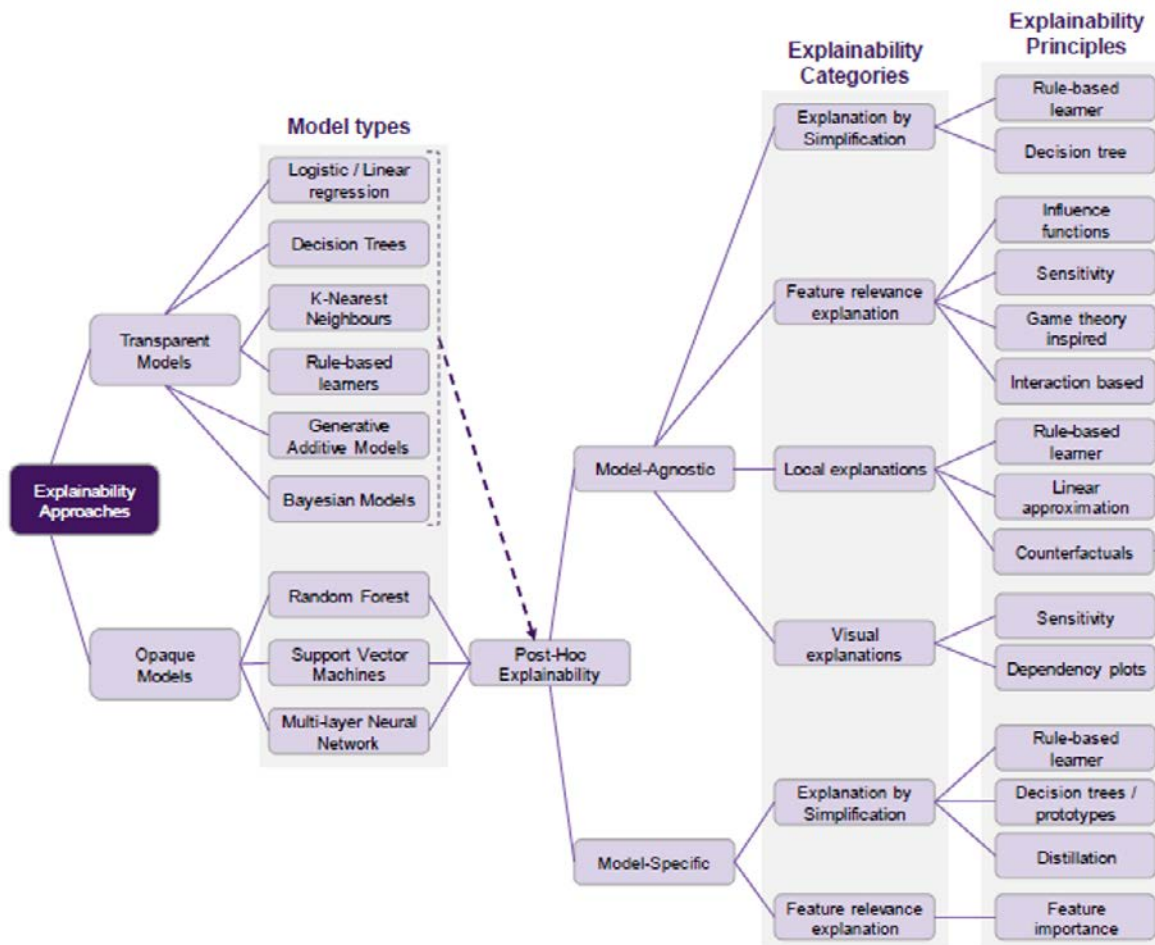


Figura 2. Example 1 of a taxonomy map in XAI [7].

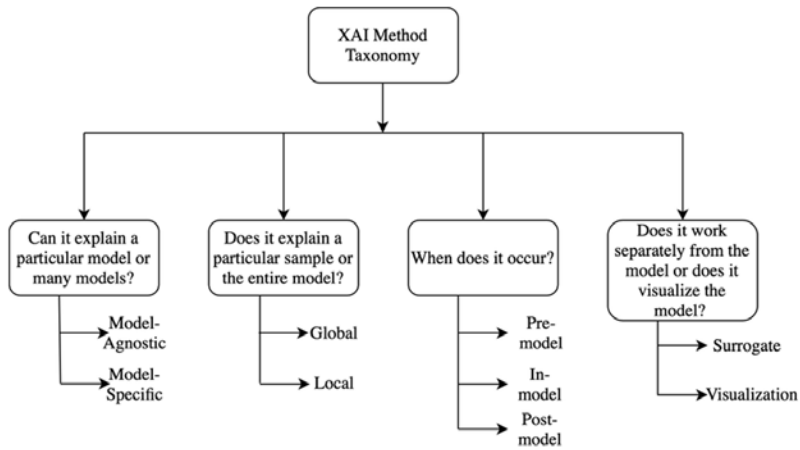


Figure 3. Example 2 of a taxonomy map in XAI [8].

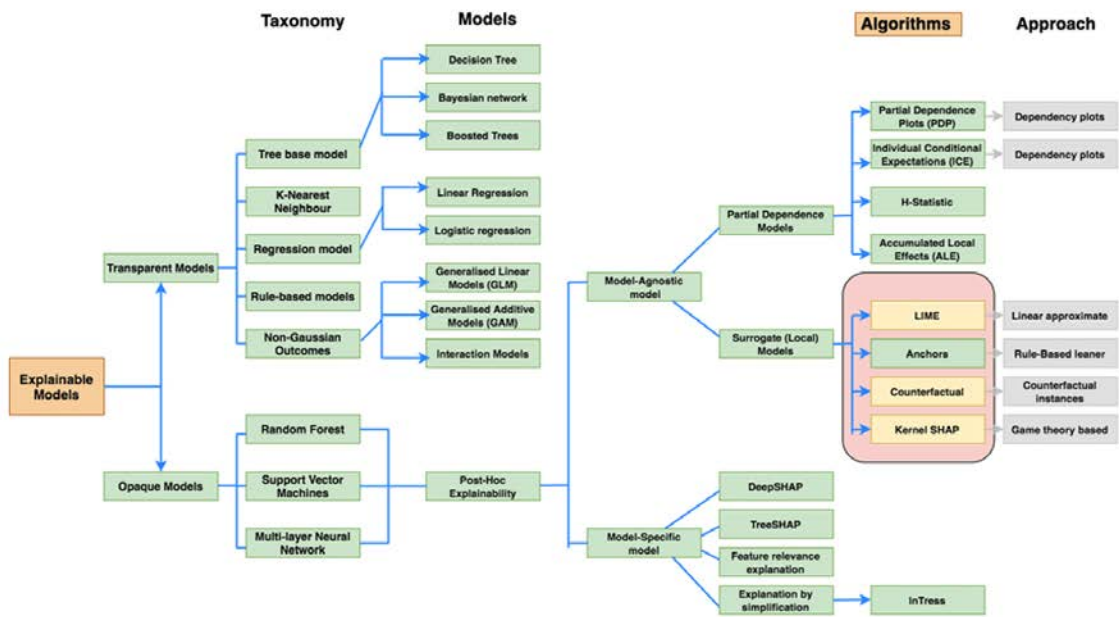


Figure 4. Example 3 of a taxonomy map in XAI [9].

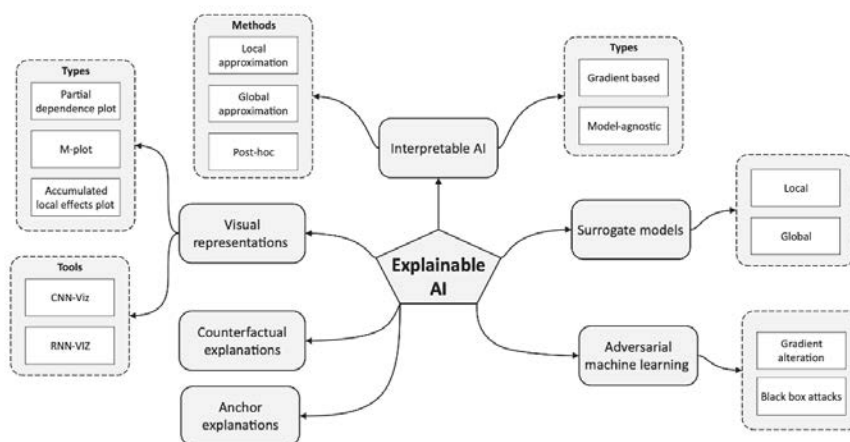


Figure 5. Example 4 of a taxonomy map in XAI [10].

As one can see, machine learning methods can be classified according to several criteria.

3.2.

Intrinsic Explainability and Post-hoc Explainability

We distinguish between whether explainability is obtained through the model's intrinsic nature or by applying methods that analyse the models after training (post hoc). **Intrinsic explainability** refers to machine learning that is considered interpretable because of its simple structure, for example short decision trees or sparse linear models. **Post hoc explainability** refers to the application of interpretation methods after the model is trained.

3.3.

Global Explainability and Local Explainability

XAI techniques can be applied **globally**, showing a general explanation of the model as a whole (importance of input features in the output prediction), or **locally**, focusing on a particular case study (a specific patient).

Global methods describe the broad behaviour of a machine learning model. They are useful methods when the modeller wants to interpret and analyse general mechanisms. Local methods allow us to better understand the predictions for each specific case.

3.4

Transparent Models and Opaque Models

AI-generated models can be transparent (e.g. a model obtained from a logistic regression) or opaque (e.g. a model obtained from a convolutional neural network). Transparency represents a human-level understanding of the model's inner workings. Three dimensions can be considered. **Simulability** is the first level of transparency and refers to the model's ability to be simulated by a human. Only models that are simple and compact fit into this category. At the second level is **Decomposability**, which is the ability to break a model down into parts (input, parameters and calculations) and explain those parts. The third level of transparency expresses the ability to understand the procedure the model goes through to generate its output. It is known as **Algorithmic Transparency** and must allow the model to be inspected with mathematical analysis.

Transparent models are a set of models whose architecture satisfies at least one of the three levels of transparency. The following are examples of transparent models [7]:

- **Linear or logistic regression:** this refers to a class of models used to predict continuous/categorical objectives, respectively, under the assumption that this objective is a linear combination of the predictor variables.
- **Decision trees:** these contain a set of conditional control statements arranged hierarchically, where the intermediate nodes represent decisions and the nodes can be class labels (for classification problems) or continuous quantities (for regression problems). Decision trees are most commonly used when it is necessary to understand the application.
- **K-nearest neighbour algorithms:** these deal with classification problems by predicting the class of a new data point by inspecting the classes of its k-nearest neighbours (where the neighbourhood relationship is induced by a measure of distance between the data points). The majority class is then assigned to the instance in question.
- **Rule-based learning:** this builds on an intuitive foundation of producing rules to describe how the model generates the output.
- **Generalised Additive Models (GAMs):** a class of linear model in which the result is a linear combination of some functions of the input characteristics.
- **Bayesian networks:** probabilistic relationships between variables are explicitly represented by a directed graph. Because of the clear characterisation of the connection between the variables, they examine only probabilistic relationships. They have been used in a wide range of applications.

Opaque models hinder observation of their internal mechanisms. To understand these opaque models we can use several methods [7]:

- **Random Forest (RF):** this is seen as a way to improve the accuracy of decision trees, which often suffer from overfitting and consequently little generalisation. This method combines multiple trees to make the model smaller, leading to better generalisation of the resolution. An entire forest is more complex to explain than a decision tree, so it requires a post hoc explanation technique to be applied to gain understanding.
- **Support Vector Machine (SVM):** a class of deeply-rooted models with geometric approaches. Initially introduced for linear classification, they were later extended to the non-linear case, making them suitable for real-life applications. In SVMs we find the maximum margin between data points.

- **Artificial Neural Networks (ANN):** a class of models that has been widely used in a wide range of applications. Their mathematical/theoretical understanding has not been sufficiently developed, which makes them “black box” models. From a technical point of view, neural networks are made up of successive layers of nodes that connect the input features with the target function. Each node is an intermediate layer that collects and aggregates the outputs of the previous layer and then produces an output on its own by passing its aggregate value through a function (called an activation function). This process continues layer by layer until the output layer is reached. Therefore, the more layers the model has, the more difficult it is to interpret. Examples of this type of network are: convolutional neural networks, recurrent neural networks, graph neural networks, etc.

3.5

Model-agnostic techniques and model-dependent techniques

Another important classification consists of differentiating between techniques that depend on the AI model resulting from the training process and techniques that are model-agnostic.

Model-agnostic techniques: They must be flexible enough not to depend on the model’s intrinsic architecture. They can be useful for non-standardised architectures or customised models to which specific techniques are not suited. They may also be used in cases in which there are several models with different architectures for which homogeneous explainability is desired.

Model-dependent techniques: these aid the development of more efficient algorithms and more specific explanations, based on the features of the model itself. The main characteristic is that they are limited to specific architectures.

3.6

Type of Explainability

- **Visual explanation** seeks to generate visualisations that aid understanding of a model. They can be applied to both images and tabular data.
- **Explanation by feature relevance** seeks to explain a model’s decision by quantifying the influence of each input variable. They are very useful in models that use tabular data.
- **Explanations by example** extract representative instances from the training dataset to demonstrate how the model works.
- **Explanations by simplification** are techniques that approximate an opaque model by using a simpler one that is easier to interpret.



4.

**Explainability of
algorithms based
on Digital Medical
Imaging**

Medical imaging comprises the set of techniques and processes used to create images of the human body, or parts of it, for clinical purposes such as diagnosis, treatment and/or monitoring a disease. Some models have demonstrated extraordinary accuracy in image analysis tasks in recent years. One significant problem is that these deep learning models are “black box” algorithms, so they are intrinsically inexplicable.

Explanation methods attempt to show the reasoning in classification cases, preferably by building a degree of trust between the system, the health professional and the patient. Most image classification models use post hoc methodologies to analyse the features learned by the model. These techniques show the discriminative areas of the image. There are many different studies that apply post hoc explanatory methods to breast, prostate, lung, brain and liver cancer [14].

4.1.

CAM (Class Activation Mapping)

The CAM (Class Activation Mapping) method is one of the most popular for visual image explanation. It is capable of finding the features in an image that are responsible for classification in a neural network model.

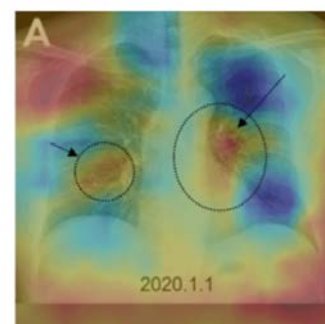
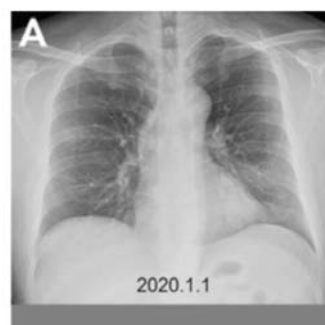


Figure 6. Graphic representation of the results of the CAM method [15].

The process is a network consisting of convolutional layers in which, just before the final output layer, global average pooling (GAP) is performed on the convolutional features that will be used for a connected layer that will produce the desired output. This simple connectivity structure allows us to identify the importance of image regions by giving a weight to the output layer of the convolutional features. Global average pooling produces the spatial average of each unit's feature map in the last convolutional layer. A weighted sum of these values is used to generate the final output. Similarly, a weighted sum of the feature maps of the last conventional layer is calculated to obtain the class activation maps [16].

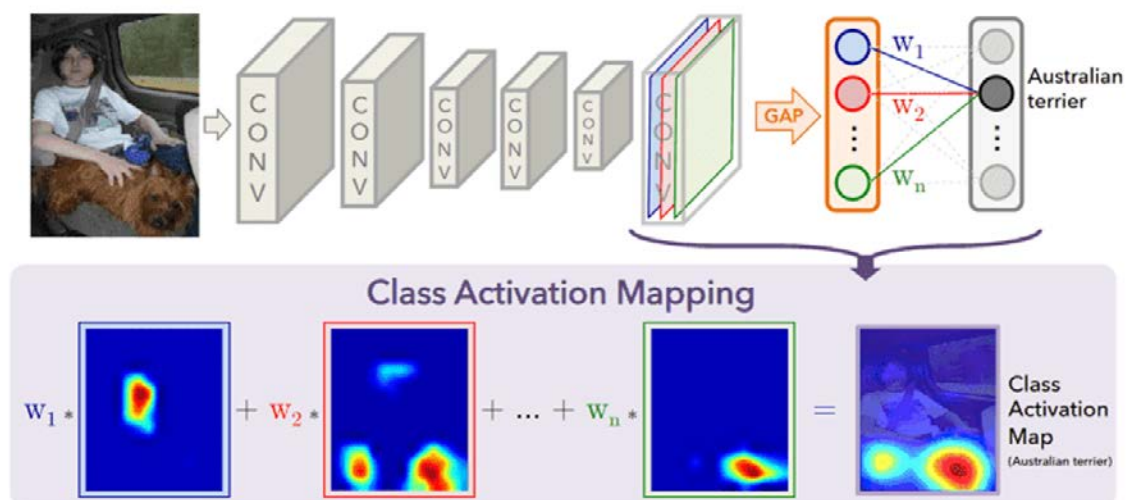


Figure 7. PCAM method process [17].

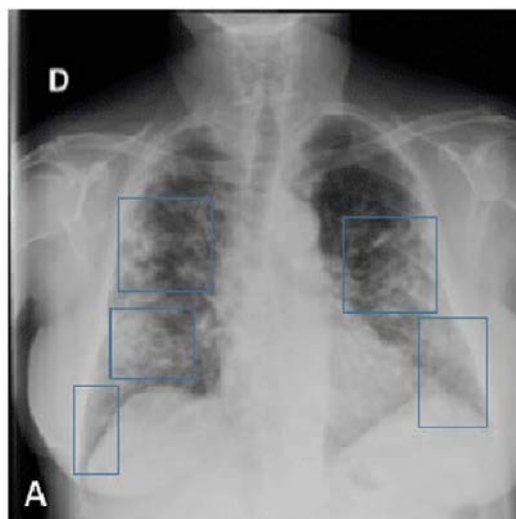
4.2.

Grad-CAM (Gradient-weighted Class Activation Mapping)

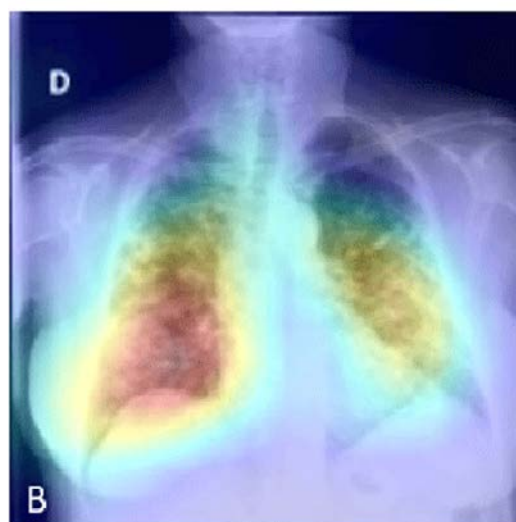
Grad-CAM (Gradient-weighted Class Activation Mapping) is an extension based on CAM, explained above, which uses the gradients for the target class that derives in the final convolutional layer. Grad-CAM produces a localisation map that highlights important pixels for image classification. Unlike CAM, this method

does not require any retraining and is broadly applicable to any architecture based on convolutional neural networks (CNN).

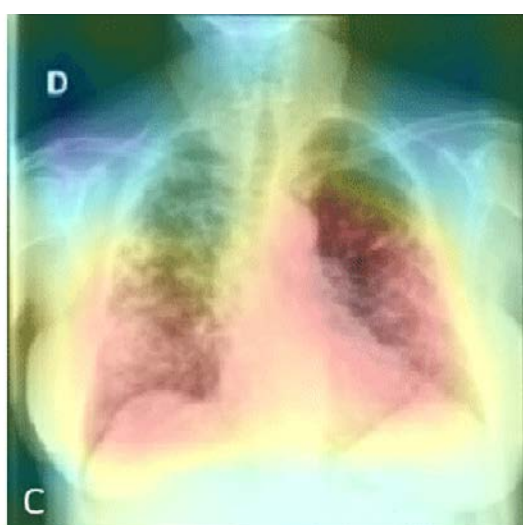
First, the class score gradient is calculated for the activation maps in the last convolutional layer. The gradients are returned after averaging them over the size of the activation map, and then the importance weights are calculated. The weighting factor shows the importance of the features for the class [16].



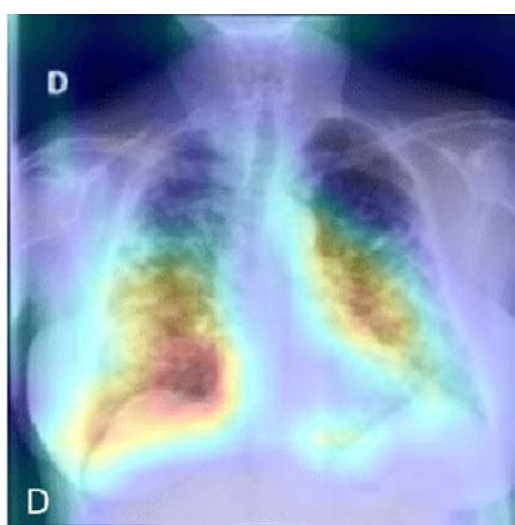
A



B



C



D

Figure 8. Grad-CAM views for different models.

4.3.

LRP (Layer-wise Relevance Propagation)

Layer-wise Relevance Propagation (LRP) is another visual explanation technique that displays a heatmap in the input space that shows the importance/relevance of each voxel that contributes to the final classification result. This method does not interact with network training, so it can be applied to pre-trained algorithms.

LRP uses network weights and neural activations to propagate the return output through the network to the input layer. There one can see which pixels actually contributed to the output.

The network is a classifier in which each entry corresponds to a different class. In the output layer, a neuron or class that we want to explain is chosen. For this neuron the relevance is equal to its activation, so the relevance of the other neurons in the output layer will be zero. It is said to be a conservative technique, which means that the magnitude of the output is preserved through the backpropagation process and is equal to the sum of the relevance map of the input layer [19].

In LRP-epsilon a small epsilon is added to propagate the relevance with numerical stability.

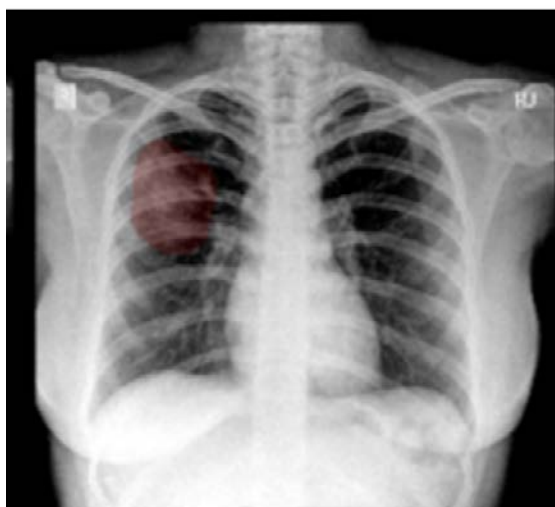
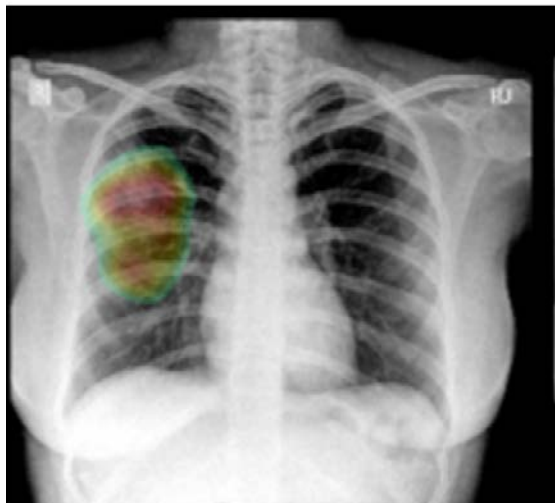


Figure 9. LRP-epsilon heatmap v. original annotation [18].

4.4.

LIME (Locally Interpretable Model-agnostic Explanations)

LIME (Local Interpretable Model-agnostic Explanations) are explanations that highlight the most relevant features for the output. This is a local type of explanation, so it does not attempt to explain all the decisions a network can make across all possible inputs. Instead, it only considers the factors it uses to determine their classification in an individual prediction [20].

This technique generates several samples that are similar to the input image by turning some of the image's superpixels on and off. The weight of each artificial image to measure its importance is calculated to explain the most important features [16].

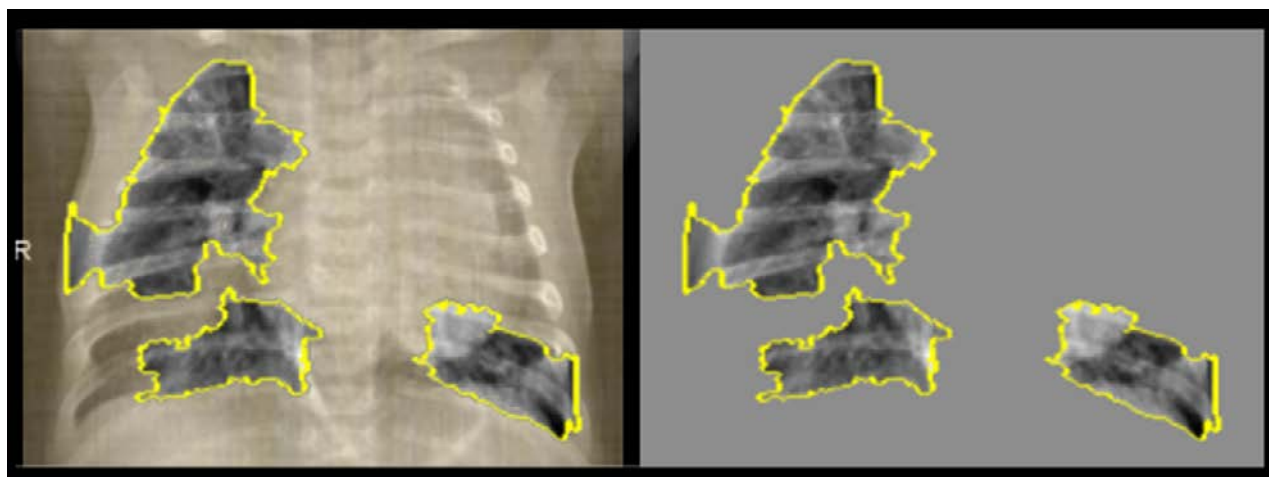


Figure 10. LIME explanation in medical imaging [18].

4.5.

SHAP (Shapley Additive Explanations)

SHAP (Shapley Additive Explanations) is a method for explaining individual predictions based on the theoretically optimal values of the Shapley game. The goal is to explain the prediction of an instance by calculating the contribution of each feature.

Shapley values arise from a context in which n players participate collectively and obtain a reward p that is intended to be distributed equitably to each of the players according to their individual contribution. In an ML model each player corresponds to a feature and the reward is the prediction [21].

Image classification tasks can be explained by the scores of each pixel in a predicted image, which shows how much it positively or negatively contributes to the probability.

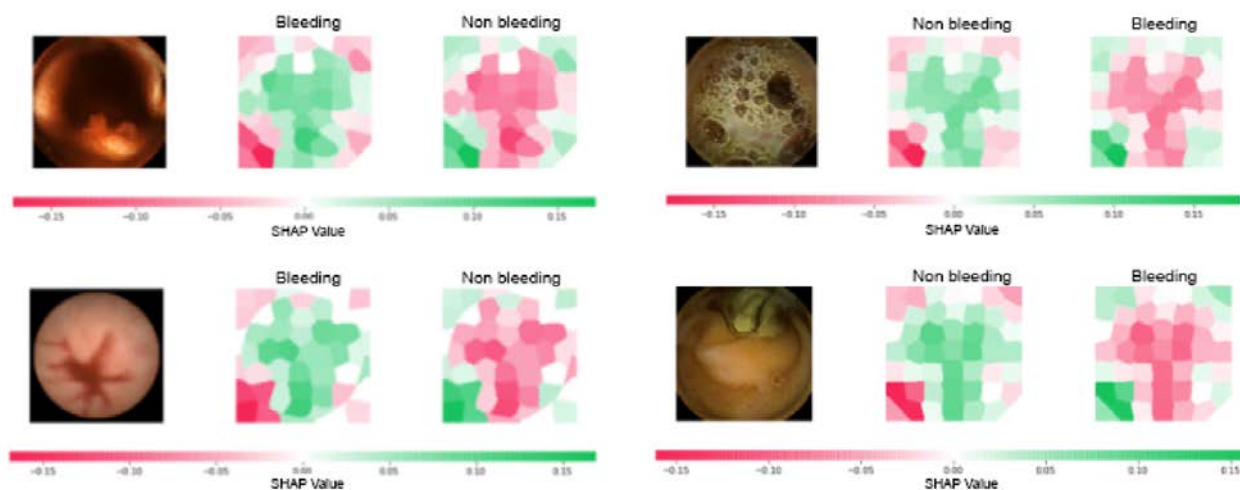
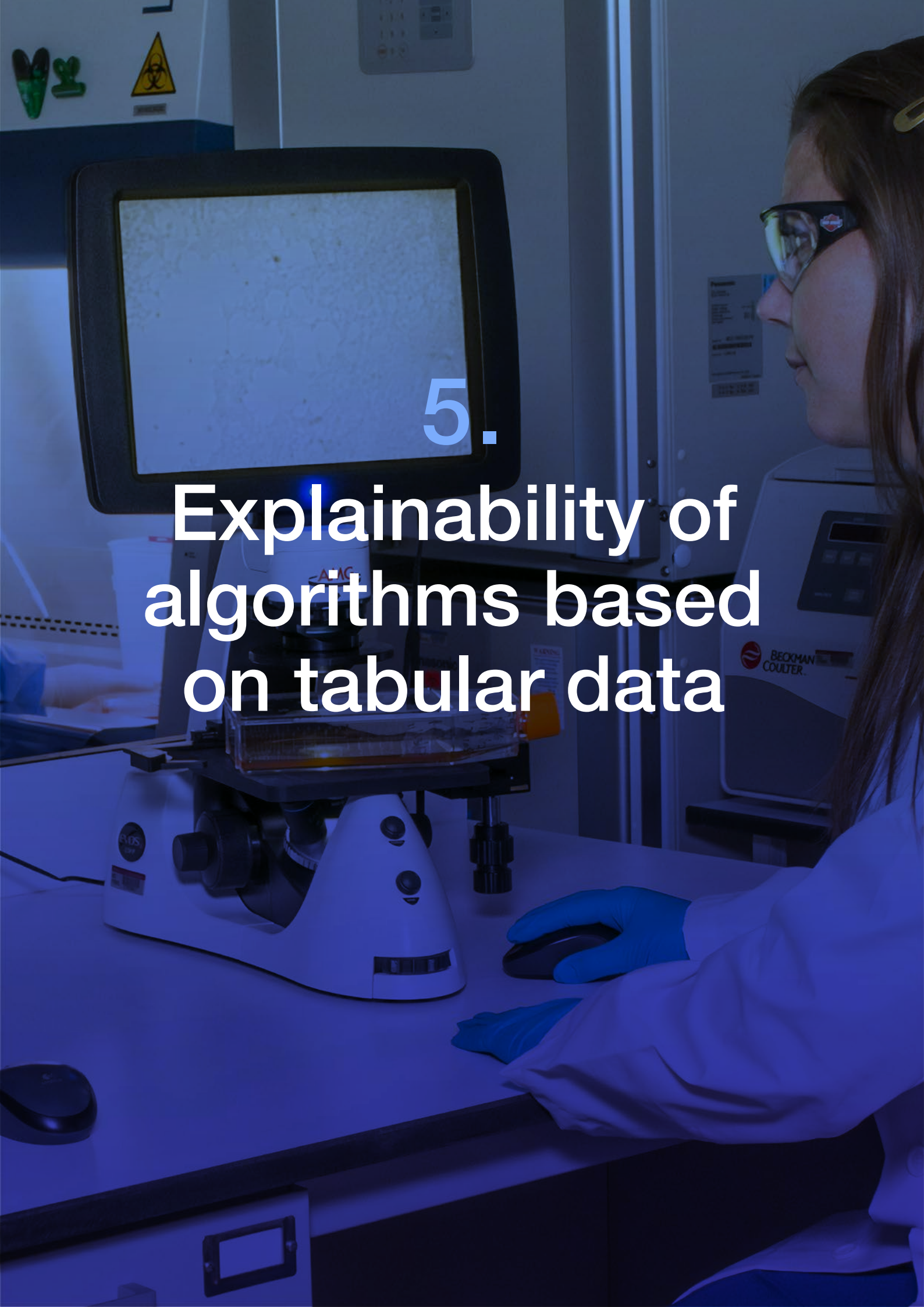


Figure 11. SHAP explanation in medical imaging [22].



5.

Explainability of algorithms based on tabular data

In the field of health, a large volume of tabular data from multiple sources and formats is used every day, from variables taken from direct measurements or tests on patients (blood tests, vital signs, omics data, etc.) to population registers or hospital management data.

5.1

PDP (Partial Dependence Plot)

The PDP (Partial Dependence Plot) shows the marginal effect that one or two features have on the predicted result of a machine learning model. In practice, feature set S normally contains only one feature or a maximum of two, since one feature produces 2D plots and two features produce 3D plots [23].

For example, you can see the effect that variables such as age and number of years on hormonal contraceptives have on a cancer prediction:

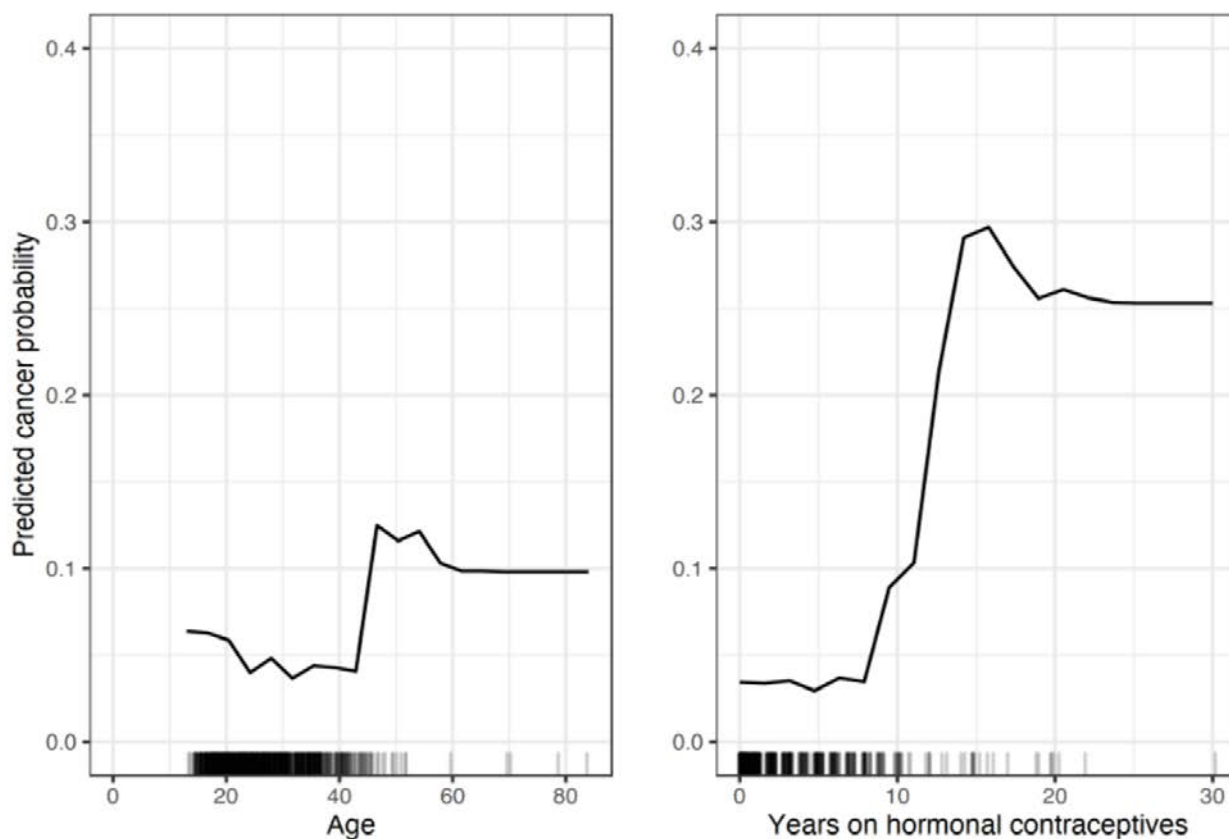


Figure 12. Explainability PDP [23].

We can also visualise the partial dependence of two features at the same time:

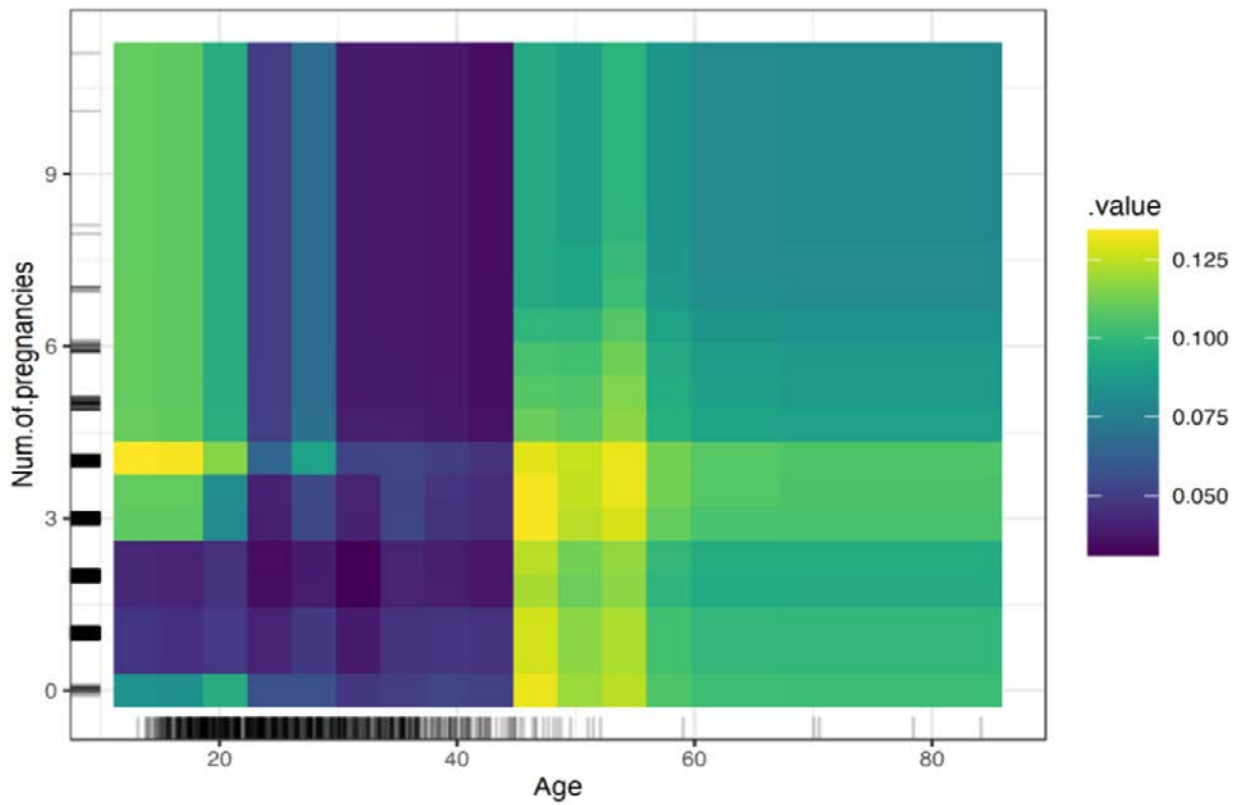


Figure 13. Representation of PDP explainability with 2 variables [23].

5.2

ICE (Individual Conditional Expectation)

Graphs generated by ICE (Individual Conditional Expectation) show how the prediction of the instance changes when a feature changes. The Partial Dependence plot for the average effect of a feature is a global method because it does not focus on specific cases, but instead on a global average. The equivalent of a PDP for individual data instances is called an Individual Conditional Expectation (ICE) plot [24].

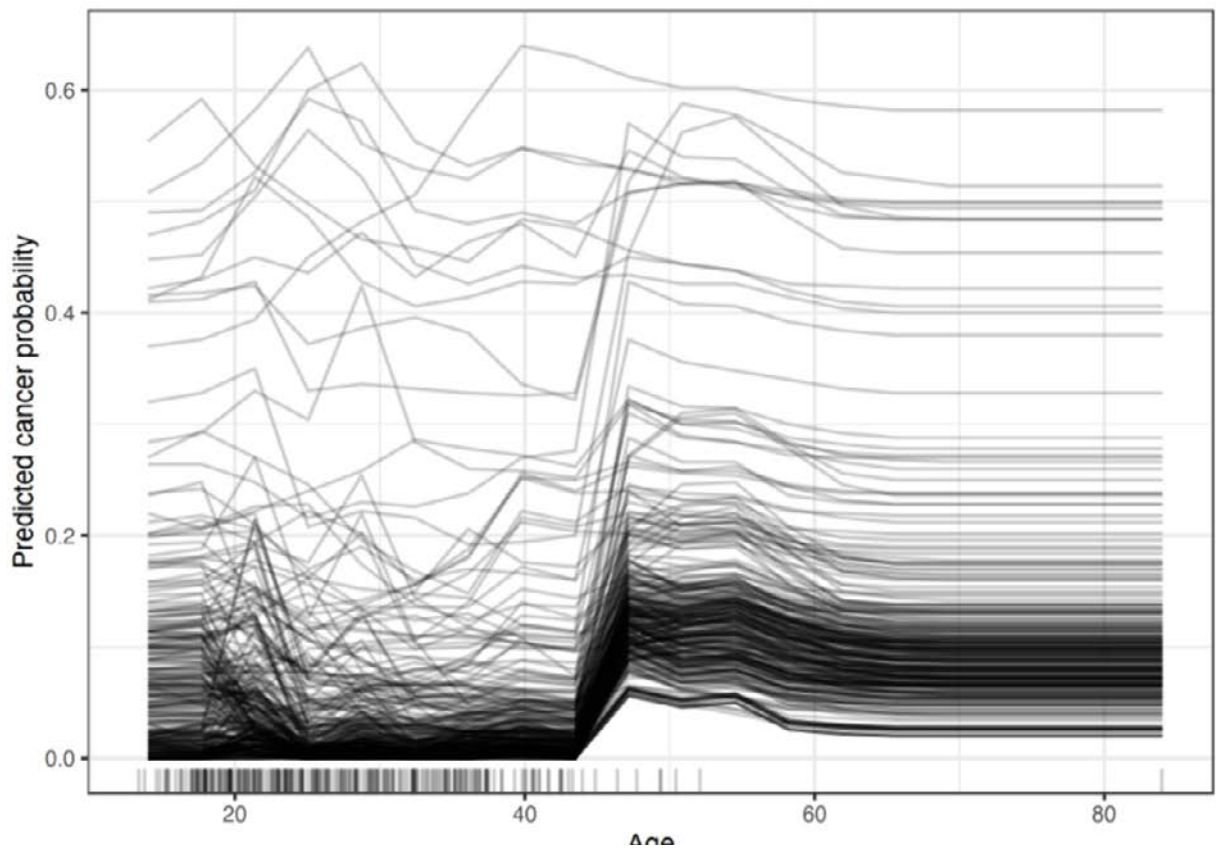


Figure 14. ICE explainability [24].

In the graph above, each line represents a patient.

5.2.1

C-ICE (Centred ICE)

One problem with ICE charts is that it can sometimes be difficult to tell if ICE curves differ between individuals because they start with different predictions. A simple solution is to centre the curves at a particular point and show only the difference up to that point. The resulting graph is known as a centred ICE (c-ICE) graph [24].

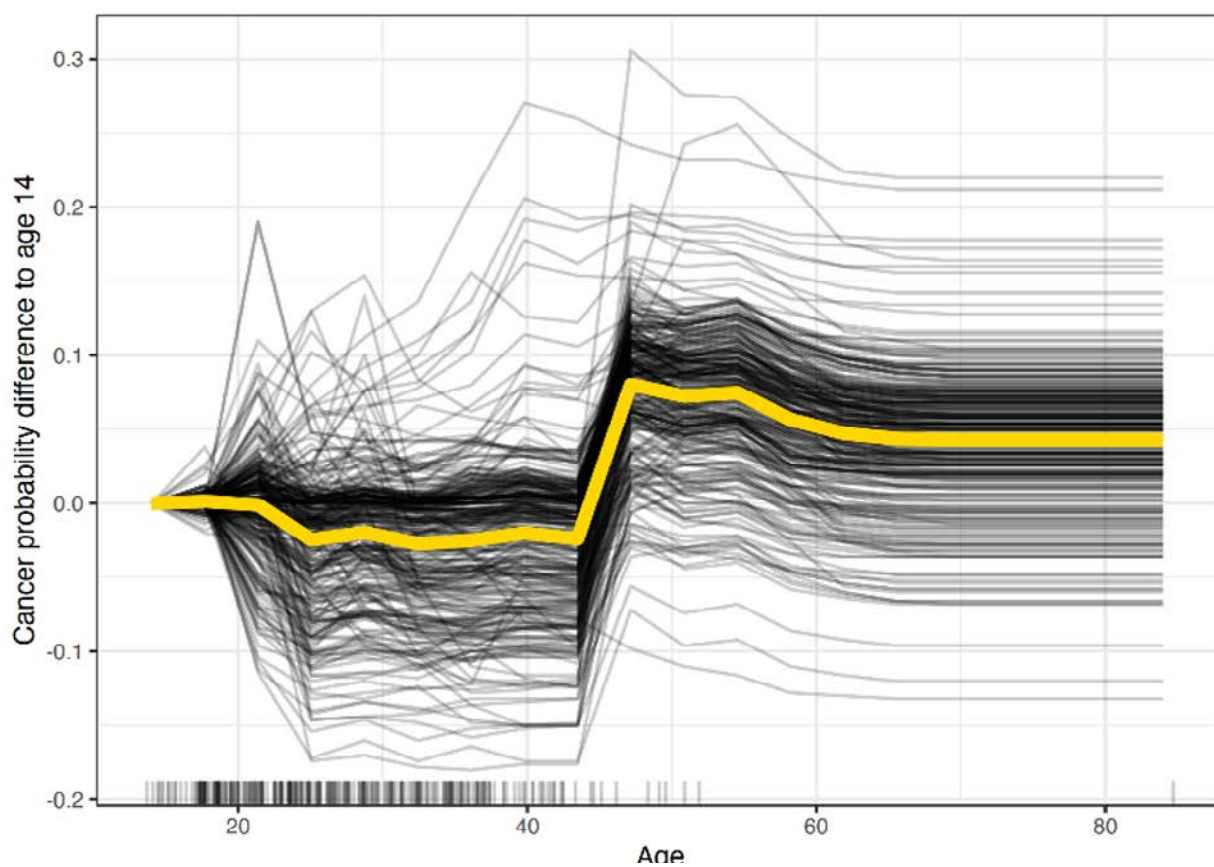


Figure 15. Centred ICE explainability [24].

5.3

Counterfactual Explanations

This type of explainability describes a cause-effect situation in the form: “If X had not occurred, Y would not have occurred.” In explainable machine learning, counterfactual explanations can be used to explain predictions of individual instances. The “event” is the predicted result of an instance, the “causes” are the specific values of features of this instance that were input into the model and “caused” a certain prediction [25].

We are interested in scenarios in which the prediction changes in a relevant way, such as a shift in the predicted class or in which the prediction reaches a certain threshold (for example, the probability of cancer reaches 10%). A counterfactual explanation of a prediction describes the smallest change in feature values that changes the prediction for a predefined output.

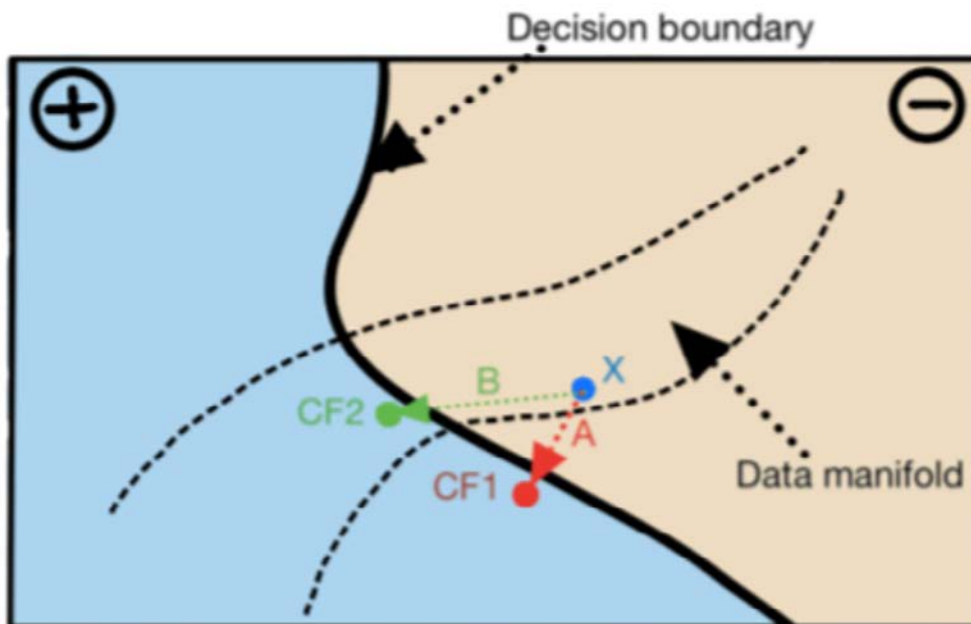


Figure 16. Counterfactual explainability [25].

The graph shows two possible paths for a point X (in blue), originally classified in the negative class, to cross the decision boundary. The endpoints of both paths, CF1 and CF2, are shown in red and green respectively.

With this type of method it is possible to find out which variables can be influenced in order for a prediction to change from a potentially ‘negative’ state to a ‘positive’ one, for example.

5.4

LIME (Locally Interpretable Model-agnostic Explanations)

We can use LIME for a classifier model with images, as seen above, but also with tabular or text data. In the case of tabular data, LIME provides a type of explanatory graph representing the importance of each of the variables and the direction of their contribution to the result (positive or negative).

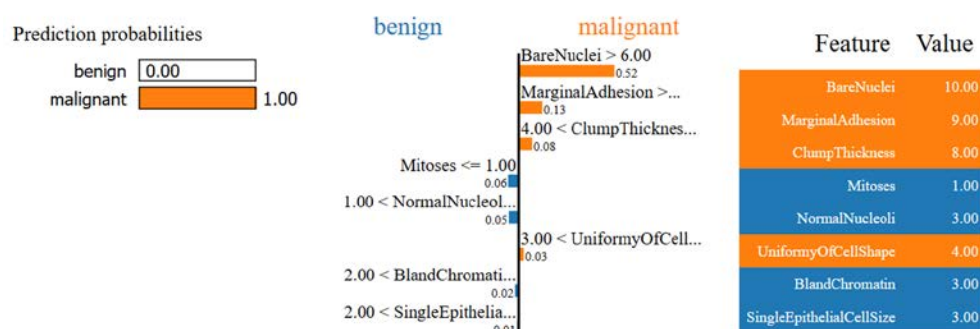


Figure 17. LIME explainability for tabular data.

5.5

Anchors

This type of explanation is used for individual predictions in any black box classification model by searching for a decision rule that sufficiently anchors the prediction. A rule anchors a prediction if changes in other feature values do not affect the prediction.

This approach deploys a perturbation-based strategy to generate local explanations for predictions of “black box” machine learning models. The resulting explanations are expressed as easy-to-understand *IF-THEN rules* [26].

This method provides a result explanation such as the one below:

```

IF PSA < 2.5 ng/ml
  AND Age < 50
THEN PREDICT Cancer = false
WITH PRECISION 97 %
AND COVERAGE 15 %
  
```

5.6

SHAP (Shapley Additive Explanations)

As explained above, Shapley's explanations show the global importance of each feature. We thus average the absolute Shapley values and order them in descending order according to their importance in the final prediction [27].

Let's look at the example of the contribution of each feature to the prediction of uterine cancer:

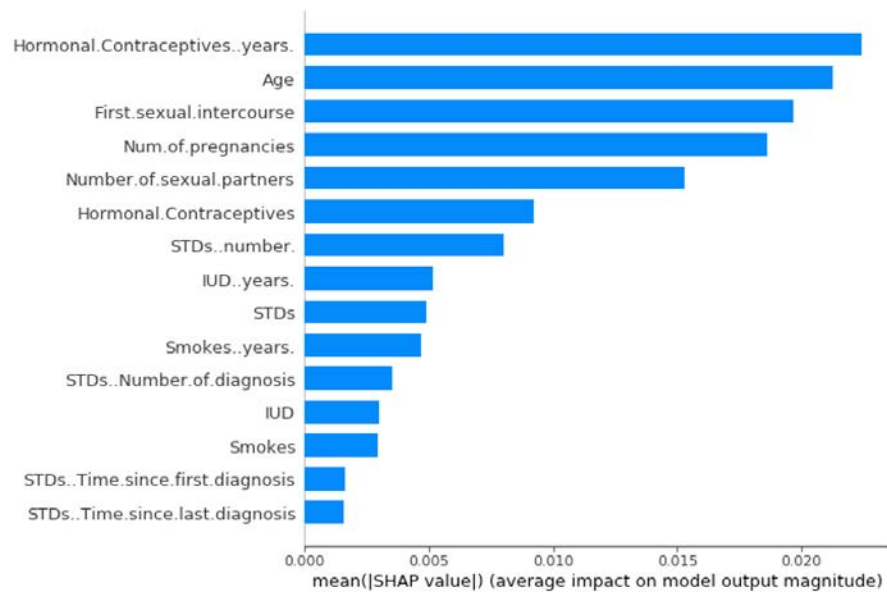


Figure 18. SHAP variable importance graph [27].

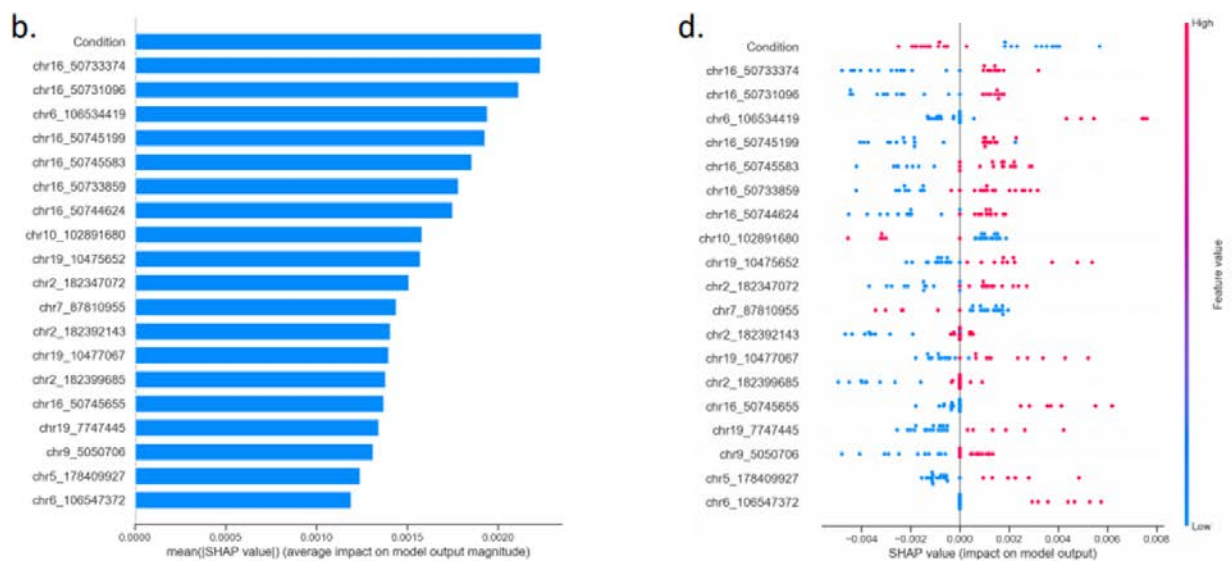


Figure 19. SHAP variable importance graph for prediction with omics data.

Shapley values can also be visualised as “forces”. Each feature value is a force that increases or decreases the prediction. This prediction starts from the baseline that corresponds to the average of all predictions. In this plot each Shapley value is an arrow that pushes to increase (positive value) or decrease (negative value) the prediction. These forces balance each other in the actual prediction of the instance [27].

The following chart (Figure 20) shows two SHAP explanation force plots for two patients from a uterine cancer dataset:

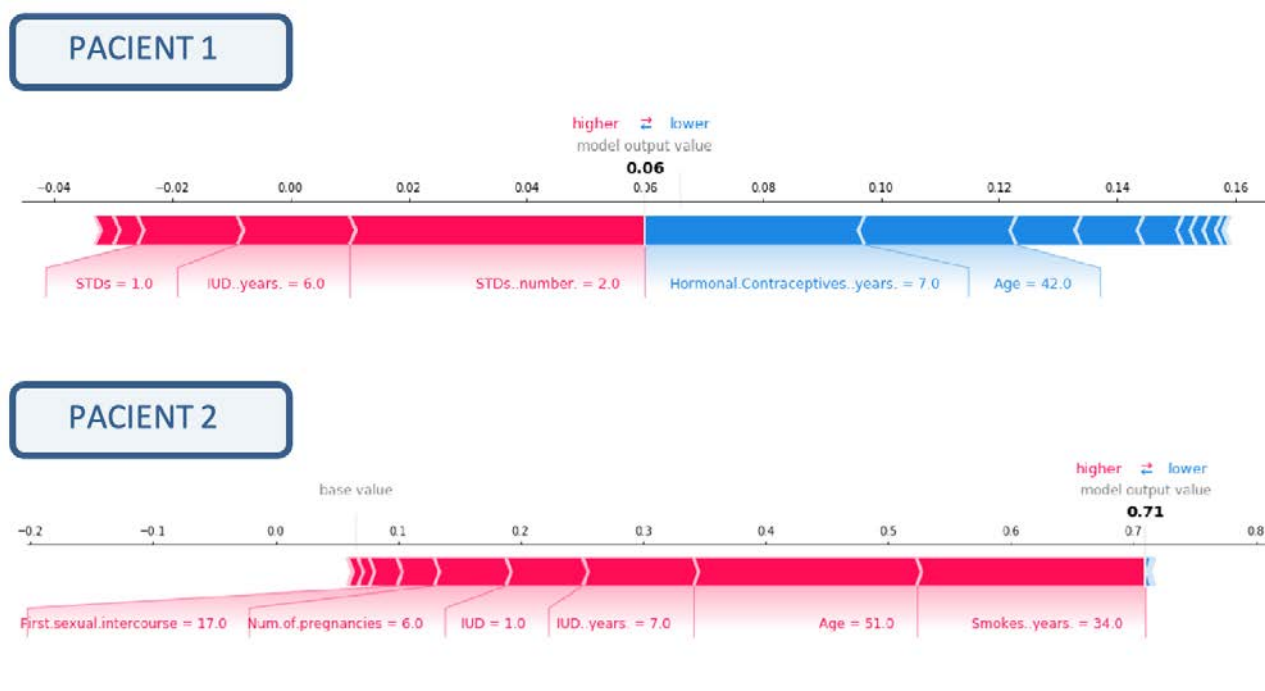


Figure 20. SHAP force plot for two patients [29].

The first patient has a risk of 0.06. The variables that increase the risk, in red, are offset by effects that make it decrease, in blue. The second patient has a higher risk of 0.71. Variables that increase the risk predominate.

SHAP provides multiple graphic formats, as listed below:

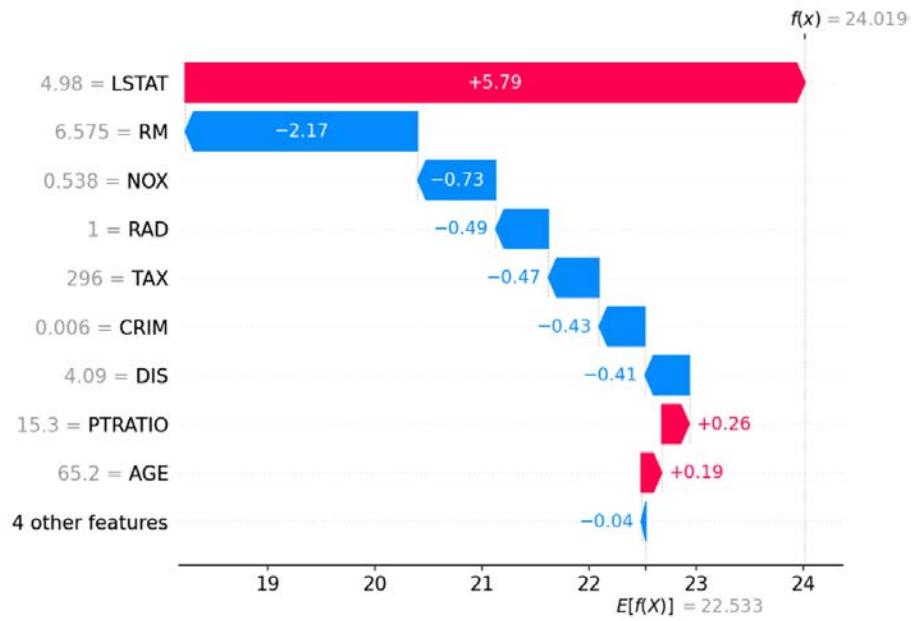


Figure 21. Graphic representation of SHAP.

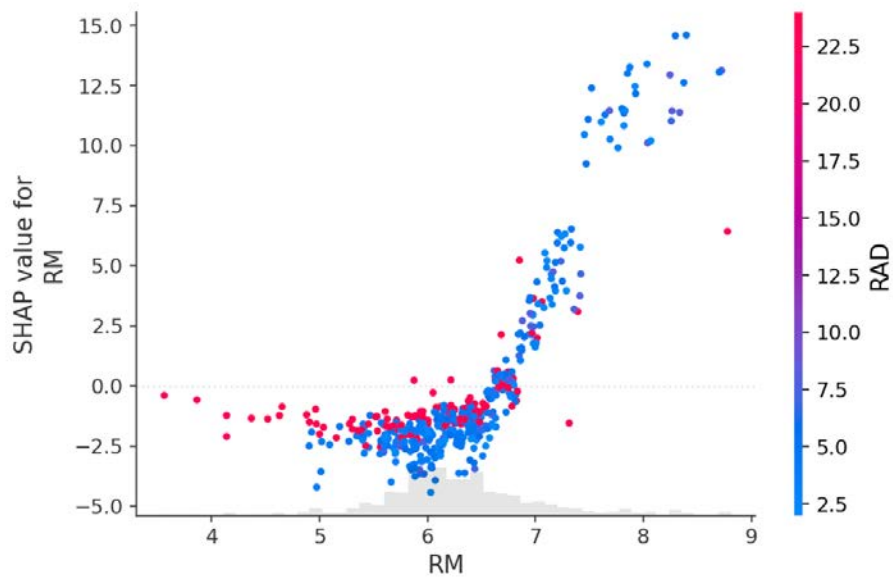


Figure 22. Graphic representation of SHAP.



Figure 23. Graphic representation of SHAP.

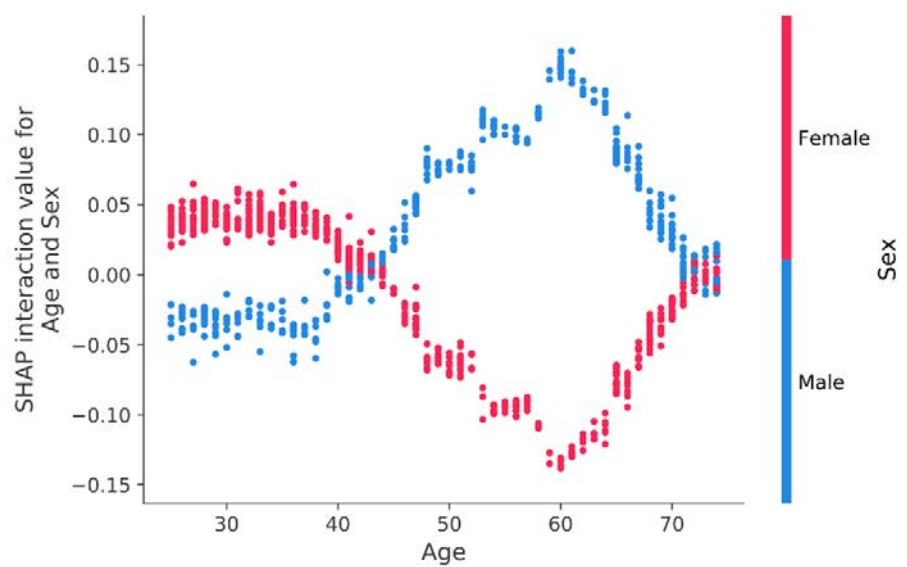


Figure 24. Graphic representation of SHAP.



6.

Explainability of algorithms based on Natural Language Processing

Natural Language Processing (NLP) allows, for example, the extraction of structured information from free-text (unstructured) reports with diagnostic, treatment or monitoring data [28] [29].

Neural networks in NLP are trained in an end-to-end manner on input-output pairs. Since linguistic features are not explicitly encoded, it is unsure what information is captured in neural networks. The answer to this depends on three elements [30]:

1. The methods used to analyse the network, such as classification or clustering.
2. The type of linguistic information that we assume the network captures, such as sentence length, parts of speech, or concepts.
3. The part of the neural network being investigated, such as weights, activations or embeddings.

6.1.

Shapley Additive Explanations (SHAP)

The SHAP technique can also be used to explain Natural Language Processing (NLP). The goal in this case is to see if the information is true or false. This model has been pre-trained with a manually-labelled dataset. The model's explanation is used to explain the output of these by assigning to each feature an importance value based on the prediction [31].

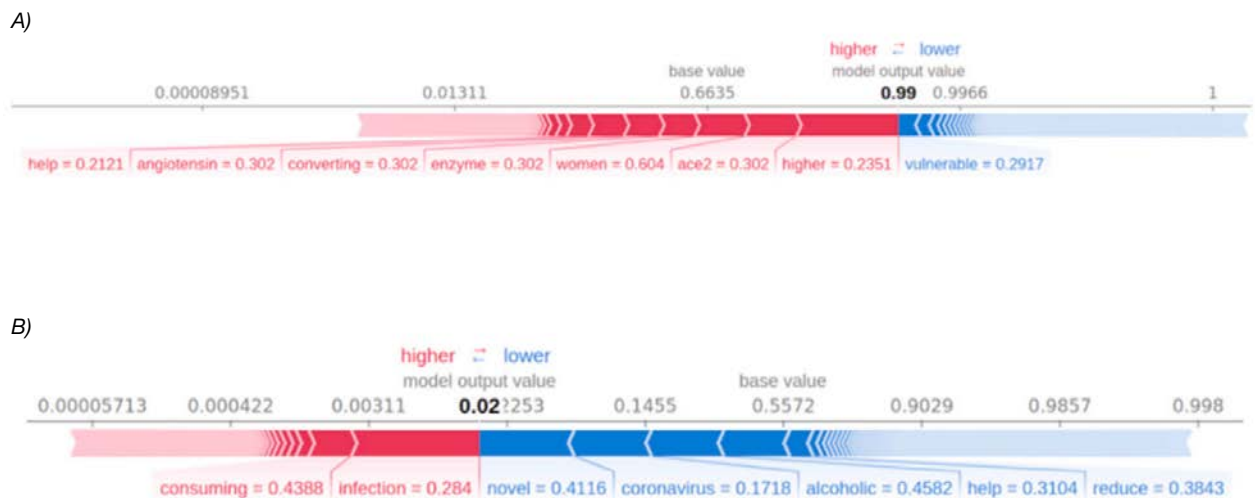


Figure 25. a. SHAP explainability in NLP. (a) Example of a true claim: “Men have higher concentrations of angiotensin-converting enzyme 2 (ACE2) in their blood than women, which may help to explain why men are more vulnerable to COVID-19 than women”. / (b) Example of a false claim: “Consuming alcoholic beverages may help reduce the risk of infection by the novel coronavirus”. Explained by SHAP [31].

Factors that push the predicted truth probability are shown in red while those that push it to be false are shown in blue. The first example (Figure 25, a) represents a true claim with a probability of 0.99. The words that contributed to producing the given prediction were *help*, *angiotensin*, *converting*, *enzyme*, *women*, *ace2* and *higher*. The second example (Figure 25, b) represents a false claim with a truth probability of 0.02. The words that contributed to the given prediction were *novel*, *coronavirus*, *alcoholic*, *help* and *reduce*.

6.2.

GbSA (Gradient-based Sensitivity Analysis)

A very simple way to relate the inputs to their outputs is to calculate the partial derivative of the output with respect to each input feature. Sensitivity analysis can be applied directly or indirectly to textual data. A vector of size D will be obtained with the sensitivity of each output where the norm of the squared gradient will have to be decomposed for the prediction function. A drawback is that this technique does not necessarily apply saliency to the feature, but may apply noise.

6.3.

LRP (Layer-wise Relevance Propagation)

LRP is used to decompose a text classifier's decision function and uses the relevance scores to provide the highlighted text.

Below are some records in which the highlighted characters help us clearly see why the model has predicted that this is a negative analysis [32].

Getting worse not better 1 30 appointment for a diagnostic and charge the AC got the car back after 6 00pm Sent a poor 17 year old kid to pick us up as their courtesy driver once it was finally ready to go
Unfortunately there is nothing special about this place My husband got the french dip and myself the mushroom panini Mine was rather disappointing the mushrooms were minced so tiny and the flavor was semi reminiscent of canned cream of mushroom soup on a sandwich I hate leaving bad reviews but it wouldn t help anyone if i lied sorry
Over priced and mediocre food
The nasty youngster working at the Wetzell s Pretzel counter ruined it man She was all pissed at me because she misheard my order and I bothered her to give me the right kind of pretzel Lamé Grow up little girl Rude
Duh what a wasteland of crappy products Gift card forced me to pop by in disguise

Figure 25, b. Text with highlighted characters according to LRP values [32].

Figure 27 shows the influence graph of the characters that contribute to a negative analysis (left) and to a positive analysis (right). The horizontal axis shows the impact on the model.

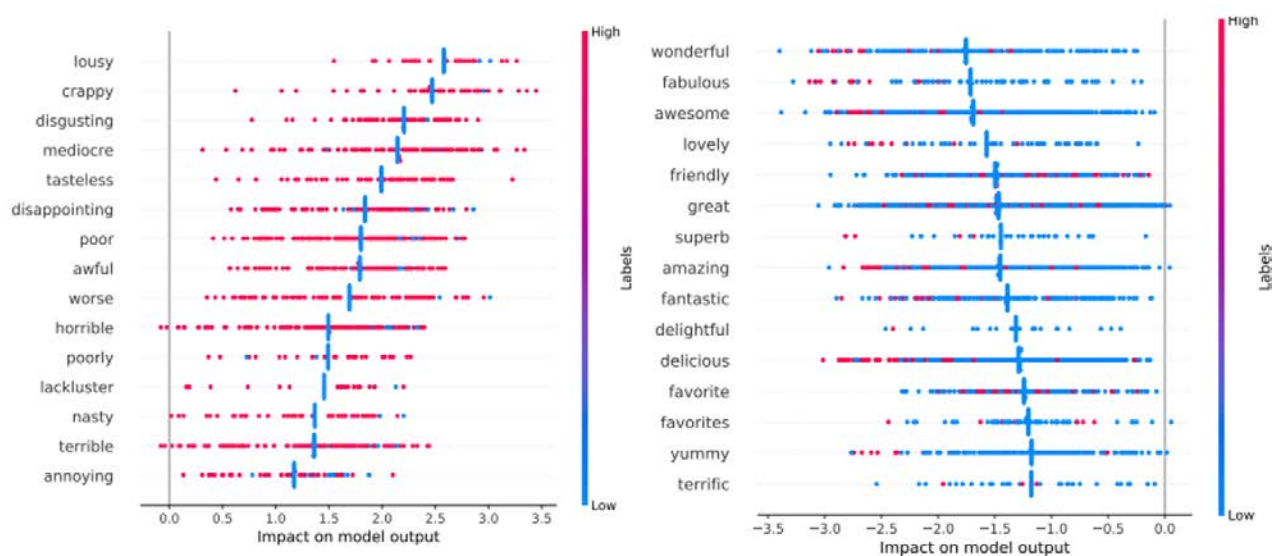


Figure 27. Uni-feature diagram [32].

Figure 28 shows the influence of the double/triple word on the prediction with LRP values for a negative analysis. When the two are compared, one can observe that the frequency of characters is lower than that of individual salience. These frequencies can be seen in the number of dots in each row. The vertical bar in the middle of each row is the average contribution of the feature regardless of whether it is uni-, bi- or tri-feature. Due to their low frequency, double and triple features have a very low impact on the analysis.

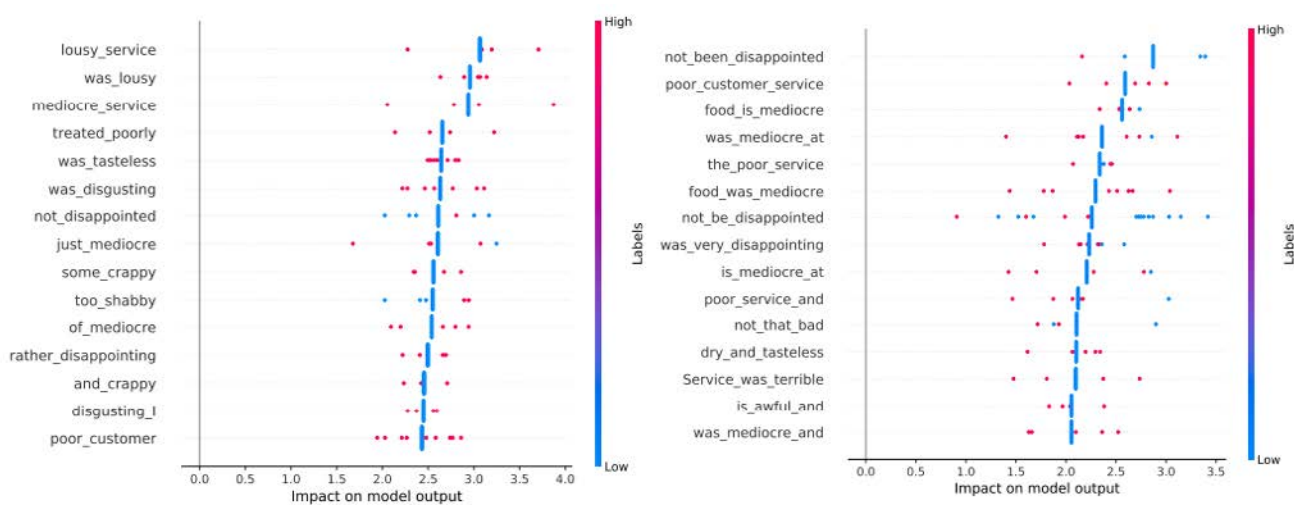


Figure 28. Bi-feature and tri-feature diagrams [32].

6.4.

LIME (Locally Interpretable Model-agnostic Explanations)

Model-agnostic explanations can also be used to understand the output of an NLP system in human terms.

With this technique it was considered that the words virus and protein contributed to defining this model as virology [33].



Figure 29. LIME explainability with NLP [33].



7.

References

- [1] MARKUS, AF, KORS, JA, & RIJNBEEK, PR (2021). The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 113 , 103655. <https://doi.org/10.1016/j.jbi.2020.103655>
- [2] MAADI, M., KHORSHIDI, HA, & AICKELIN, U. (2021). A Review on Human-AI Interaction in Machine Learning and Insights for Medical Applications. *Public Health*, 18, 2121. <https://doi.org/10.3390/ijerph18042121>
- [3] High-level expert group on Artificial Intelligence set up by The European Commission. (2019) Ethics Guidelines for Trustworthy AI. Retrieved on 13 December 2021 from <https://www.aepd.es/sites/default/files/2019-12/ai-et-hics-guidelines.pdf>
- [4] PATEL, BN, ROSENBERG, L., WILLCOX, G., BALTAXE, D., LYONS, M., IRVIN, J. & ET AL. (2019). Human-machine partnership with artificial intelligence for chest radiograph diagnosis. *Npj Digital Medicine* . <https://doi.org/10.1038/s41746-019-0189-7>
- [5] APPEN. Human in the Loop - Human in the Loop Machine Learning. (2019). Retrieved on 13 December 2021 from <https://appen.com/blog/human-in-the-loop/>
- [6] VILONE, G., & LONGO, L. (2020). *Explainable Artificial Intelligence: A Systematic Review*. arXiv. <https://doi.org/10.48550/arXiv.2006.00093>
- [7] BELLE, V., & PAPANTONIS, I. (2020). Principles and Practice of Explainable Machine Learning. <https://doi.org/10.3389/fdata.2021.688969>
- [8] SINGH, A., SENGUPTA, S., & LAKSHMINARAYANAN, V. (2020). Explainable deep learning models in medical image analysis . *Journal of Imaging*, 6(6) , 52. <https://doi.org/10.3390/jimaging6060052>
- [9] CHOU, Y., MOREIRA, C., BRUZA, P., OUYANG, C., & JORGE, J. (2022). Counterfactuals and Causability in Explainable Artificial Intelligence: Theory, Algorithms, and Applications. *Information Fusion*, 81 , 59-83. <https://doi.org/10.1016/j.inffus.2021.11.003>
- [10] NASSAR, M., SALAH, K., UR REHMAN, MH, & SVETINOVIC, D. (2020). Blockchain for explainable and trustworthy artificial intelligence. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery*, 10(1). <https://doi.org/10.1002/widm.1340>
- [11] PAWAR, U., O'SHEA, D., REA, S., & O'REILLY, R. (2020). Explainable AI in Health care. *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment, Cyber SA 2020* . <https://doi.org/10.1109/CYBERSA49311.2020.9139655>
- [12] AMANN, J., BLASIMME, A., VAYENA, E., FREY, D., & MADAI, VI (2020). Explainability for artificial intelligence in health care: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making* , 20 (1), 1–9. <https://doi.org/10.1186/S12911-020-01332-6>
- [13] LAPUSCHKIN, S., WÄLDCHEN, S., BINDER, A. ET AL. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nat Commun* 10, 1096. <https://doi.org/10.1038/s41467-019-08987-4>

- [14] GULUM, MA, TROMBLEY, CM, KANTARDZIC, M., & MARTINEZ, I. (2021). A Review of Explainable Deep Learning Cancer Detection Models in Medical Imaging. *Applied Sciences*, 11(10) , 4573. <https://doi.org/10.3390/app11104573>
- [15] DIALAMEH, M., HAMZEH, A., RAHMANI, H., RADMARD, AR, & DIALAMEH, S. (2020). *Screening COVID-19 Based on CT/CXR Images & Building a Publicly Available CT-scan Dataset of COVID-19*. EuropePMC preprint. Retrieved on 13 December 2021 from https://www.researchgate.net/figure/Plotting-the-results-of-Class-Activation-Mapping-CAM-The-CAM-highlights-the_fig4_347966202
- [16] PAPAISTRATIS, I. (2021). *Explainable AI (XAI): A survey of recent methods, applications and frameworks*. *theaisummer.com*. Retrieved on 13 December 2021 from <https://theaisummer.com/xai/>
- [17] ZHOU, B., KHOSLA, A., LAPEDRIZA, A., OLIVA, A., & TORRALBA, A. (2016). Learning Deep Features for Discriminative Localisation,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* , 2921-2929, <https://doi.org/10.1109/CVPR.2016.319>
- [18] JAIN, S., & GHOSH, R. (2017). *Visualizing Deep Learning Networks - Part II* . Retrieved on 13 December 2021 from <https://blog.qure.ai/notes/deep-learning-visualisation-gradient-based-methods>
- [19] LINDWURM, E. (2019). *InDepth: Layer-Wise Relevance Propagation*. Towards Data Science. Retrieved on 13 December 2021 from <https://towardsdatascience.com/indepth-layer-wise-relevance-propagation-340f95deb1ea>
- [20] RIBEIRO, M. (2016). *LIME - Locally Interpretable Model-Agnostic Explanations*. homes.cs.washington.edu. Retrieved on 13 December 2021 from <https://homes.cs.washington.edu/~marcotcr/blog/lime/>
- [21] LOPEZ, F. (2021). *SHAP: Shapley Additive Explanations*. Towards Data Science. Retrieved on 13 December 2021 from <https://towardsdatascience.com/shap-shapley-additive-explanations-5a2a271ed9c3>.
- [22] *Multi-class ResNet50 on ImageNet (TensorFlow) — SHAP latest documentation* . (n.d.). Retrieved on 13 December 2021 from https://shap.readthedocs.io/en/stable/example_notebooks/image_examples/image_classification/Multi-class%20ResNet50%20on%20ImageNet%20%28TensorFlow%29-checkpoint.html
- [23] MOLNAR, C. 8.1 *Partial Dependence Plot (PDP) | Interpretable Machine Learning*. (n.d.). Retrieved on 13 December 2021 from <https://christophm.github.io/interpretable-ml-book/pdp.html>
- [24] MOLNAR, C. 9.1 *Individual Conditional Expectation (ICE) | Interpretable Machine Learning* . (n.d.). Retrieved on 14 December 2021 from <https://christophm.github.io/interpretable-ml-book/ice.html>
- [25] MOLNAR, C. 9.3 *Counterfactual Explanations Interpretable Machine Learning*. (n.d.). Retrieved on 14 December 2021 from <https://christophm.github.io/interpretable-ml-book/counterfactual.html>

- [26] MOLNAR, C. 9.4 *Scoped Rules (Anchors)* | *Interpretable Machine Learning*. (n.d.). Retrieved on 14 December 2021 from <https://christophm.github.io/interpretable-ml-book/anchors.html>
- [27] MOLNAR, C. 9.6 *SHAP (SHapley Additive Explanations)* | *Interpretable Machine Learning*. (n.d.). Retrieved on 14 December 2021 from <https://christophm.github.io/interpretable-ml-book/shap.html>
- [28] DANILEVSKY, M., QIAN, K., AHARONOV, R., KATSIKIS, Y., & SEN, P. (2020). *A Survey of the State of Explainable AI for Natural Language Processing*. arXiv. <https://doi.org/10.48550/arXiv.2010.00711>
- [29] LI, Y. *Explainability for Natural Language Processing*. (2020). Retrieved on 14 December 2021 from <https://www2.slideshare.net/Yunyaoli/explainability-for-natural-language-processing>
- [30] VOLPATO, R. *NLP meets XAI: Top 5 Trends in NLP Explainability*. (2019). Retrieved on 14 December 2021 from <https://volpato.io/articles/1907-nlp-xai.html>
- [31] AYOUB, J., YANG, X., & ZHOU, F. (2021). Combat COVID-19 infodemic using explainable natural language processing models. *Information Processing & Management*, 58 (4), 102569. <https://doi.org/10.1016/j.ipm.2021.102569>
- [32] GHOLIZADEH, S., & ZHOU, N. (2021). *Model Explainability in Deep Learning Based Natural Language Processing*. arXiv. <https://doi.org/10.48550/arXiv.2106.07410>
- [33] GODAVARTHI, D., & A, M. (2021). Classification of Covid-related articles using machine learning. *Materials Today: Proceedings* <https://doi.org/10.1016/j.matpr.2021.01.480>

