EXPLAINABLE ARTIFICIAL INTELLIGENCE IN MEDICINE: SYMBOLIC REASONING, MACHINE LEARNING, AND HYBRID APPROACHES

Chara Theocharous

A Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of

Science at the

University of Cyprus

Recommended for Acceptance by the

Department of Computer Science

December, 2023

APPROVAL PAGE

Master of Science Thesis

EXPLAINABLE ARTIFICIAL INTELLIGENCE IN MEDICINE: SYMBOLIC REASONING, MACHINE LEARNING, AND HYBRID APPROACHES

Presented by

Chara Theocharous

Research Supervisor: Elpida Keravnou Papaeliou

Committee Member: Constantinos S. Pattichis

Committee Member: George Pallis

University of Cyprus

December, 2023

ACKNOWLEDGEMENTS

In this acknowledgement section, I extend my appreciation to all those who have been part of this challenging but fulfilling journey. Firstly, I would like to express my gratitude to my supervisor, professor Elpida Keravnou-Papaeliou, for the guidance provided, which has been instrumental in the completion of this thesis. I also appreciate her invaluable expertise and knowledge, generously shared during various courses throughout my studies. Special thanks go to my family and friends, whose constant encouragement fueled my perseverance during the completion of this dissertation.

ABSTRACT

As the usage of Artificial Intelligence (AI) is growing exponentially, it has been incorporated in medical diagnosis as well as other domains. Although Machine Learning (ML) models have been widespread adopted, many of them remain mostly black-boxes, meaning that their reasoning and/or their results are not understandable by the users. In addition, the appearance of some inaccurate or unfair results of these systems, in combination with legal regulations, led to the need of explainable AI. Moreover, there are separate disciplines of AI, each having their advantages and disadvantages. On the one hand, modern ML and Deep Learning are characterised by high performance, but also limited interpretability. On the other hand, early symbolic AI approaches seem more interpretable, but also more costly, as rules are created through human intervention. Modernizing symbolic reasoning by incorporating ML may help the improvement of explainability in AI outcomes in medicine.

Chara Theocharous – University of Cyprus, 2023

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Motivation

The usage of Artificial Intelligence (AI) is growing exponentially, with the global AI market size to be expected to rise 37% every year from 2023 to 2030. AI - and Machine Learning (ML) specifically - seems to has the most substantial impact in data-intensive and highly regulated sectors, including banking, financial services, insurance, and healthcare sectors. The use of AI in medical informatics is highly appreciated, as providing healthcare is overwhelming. The reasons include the incomplete medical knowledge, the need for clinicians to learn from experience, deal with new cases, and manually examine a huge volume of medical data which is time-consuming and relies on the prolonged attention of the doctor, and the communication problems between patients and clinicians.

Although ML models have been widespread adopted, many of them remain mostly black boxes, meaning that their reasoning and/or their results are not understandable by the users. Modern ML and especially Deep Learning (DL) models turn input features into predictions, typically involving millions of non-trivial operations.

In addition, the appearance of some inaccurate or unfair results of these systems, in combination with legal regulations, including the general data protection regulation (GDPR) but also similar regulations outside of Europe that highlight the need of trustworthy AI systems, led to the need of eXplainable AI (XAI). For example, a healthcare AI algorithm favored white patients over black patients with the same health burden, as the recommendations for necessary medical treatment to black patients was more rare [1]. This study highlights the potentially life-threatening consequences of biased healthcare AI applications, reminding the importance of explaining the algorithm's results.

Unfortunately, explainabilty is a complicated and multifaceted issue, which still needs some more effort to have a precise characterization from the terminological point of view [2]. Moreover, there are separate disciplines of AI, each having their advantages and disadvantages. On the one hand, modern ML and DL are characterised by high performance, but also limited interpretability. On the other hand, symbolic AI approaches seem more interpretable, but also more costly, as rules are created through human intervention. Combining symbolic reasoning with deep neural networks may help the improvement of explainability in AI outcomes, as proposed in European Commission's White Paper on Artificial Intelligence [3], too.

1.2 Related Work

Numerous XAI techniques have been proposed, which can be distinguished based on different aspects. The first division is between model-specific and model-agnostic techniques.

1. Model-specific XAI refers to XAI techniques tailored for working only on specific models.

2. Model-agnostic XAI refers to XAI techniques working across a variety of models and, often, types of models and domains.

The second division is between global-level and local-level techniques.

1. Global-level techniques aim at understanding the algorithm's behavior at a high/dataset/populational level, something that is typically done by researchers and designers of the algorithm.

2. Local-level techniques aim at understanding the algorithm's behavior at a low/subset/individual level, typically those being targeted by an algorithm

Thirdly, XAI can be categorized in pre-hoc (a.k.a. pre-model) XAI, ante-hoc XAI (a.k.a. XAI by design, or explainable modelling), and post-hoc (a.k.a. post-model) XAI, based on the step of the Machine Learning (ML) development life cycle that explainability is applied.

1. Pre-hoc XAI aim to perform an analysis of the data distribution in order to understand the data before building the model (pre-model XAI). This comprehension of the data may contribute to having a higher confidence with the decisions that the model provides.

2. Ante-hoc XAI approaches are related to systems that are inherently explainable (XAI by design), meaning that they incorporate interpretability directly into their structure, allowing one to directly understand their mechanisms in providing a

result such as a conclusion (e.g., a diagnosis) or a recommendation (e.g., a treatment option). The predictions of these models

are explained in terms of their input features. Decision trees, rule-based models, additive modes, sparse linear models, linear regression, fuzzy systems and Bayesian models are, for example, commonly considered to be implicitly explainable. These models can provide both global and local explanations; global explanations because of the transparent nature, and local by utilizing information of the model's parameters and structure (e.g. a path in a decision tree, a single rule, or the weight of a specific feature in a linear model).

3. Post-hoc approaches perform posterior analysis of the model predictions (post-model XAI). They try to provide some explanation to the results reached by models, such as deep neural networks, random forests, support vector machines, and many others.

Chapter 2

The need for explainability

In this section we will discuss about some questions around explainability and examine some issues that need to be addressed in order to attain explainability in AI.

2.1 Definition of explanation/explainability

Based on social sciences, a common definition is that explanation is an answer to a question. Tim Miller [2] proposes a model for explanatory questions, where these questions are divided into three classes; what-questions (such as 'What event happened?'), how-questions (such as 'How did this event happen?'), and why-questions (such as 'Why did this event happen?').

Combi et al. [4] define explainability for artificial intelligence in medicine as the intersection of interpretability, understandability, usability, and usefulness. *Interpretability* is the degree to which a user may infer the reasoning behind a choice and thus their ability to predict a system's results [5]. *Understandability* is the degree to which a user is able to determine how the system functions, leading to user confidence in the system's output. *Usability* is the ease with which a user can learn to use a system, including system's input preparation and results interpretation. *Usefulness* is referring to the practical value or applicability of a system. Although the definition

refers to medicine applications, these four characteristics are needed for any information system, statistical model, or software application, so the above definition can be extended for AI applications in general.

Explainability is a complex concept based on this definition, due to the intersectionality of the four characteristics explainability consists of (see Figure 2.1). Firstly, usability and usefulness are highly related, as if a system is not usable, it is unlikely to be useful. Secondly, usability is enhanced via understandability, as an AI application that is understandable provides user confidence that its outcome was correct, making the application more likely to be usable. Interpretability and usefulness are complementary as well, as a user of an AI application is more likely to find it useful, if the application's decision is interpretable.

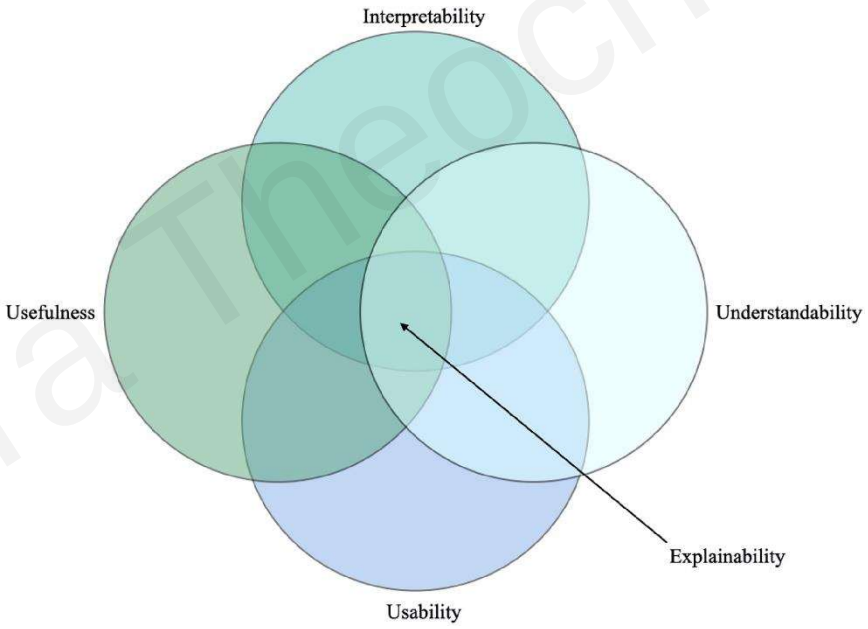

Figure 2.1: The Venn diagram of explainability. Explainability is defined as the intersection of usability, usefulness, interpretability, and understandability. [4]

A simpler definition, as it is stated in [6], defines Explainable AI (XAI) as AI/ML in which predictions/results are accompanied by explanations, presented in a way that humans can understand and formulate a line of reasoning that explains/justifies the decision-making process of the model.

2.2 Stakeholders and desiderata of explanation

The General Data Protection Regulation (GDPR), a European Union data privacy and protection law, has implications for the development and deployment of AI systems, as AI often relies on the collection and processing of large amounts of personal data. To ensure fair and transparent processing in automated decision-making, data subjects are according to GDPR entitled to relevant information about the reasoning behind the decision. Specifically, as written in Articles 13 (2) f), 14 (2) g) and 15 (1) h) of the regulation, the controller shall inform the data subject about the existence of automated decision-making, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject [7]. Explainability represents an important component in the framework proposed in the Ethics Guidelines for trustworthy AI too, published by the EU; "Whenever an AI system has a significant impact on people's lives, it should be possible to demand a suitable explanation of the AI system's decision-making process. Such explanation should be timely and adapted to the expertise of the stakeholder concerned (e.g. layperson, regulator or researcher)" [8]. Association for Computing Machinery US Public Policy Council (USACM) makes direct references to the need for explanation as well, in the Statement on Algorithmic Transparency and Accountability. Explanation is one of the seven principles for algorithmic transparency and accountability, and is stated that "Systems and institutions that use algorithmic decision-making are encouraged to produce explanations regarding both the procedures followed by the algorithm and the specific decisions that are made. This is particularly important in public policy contexts." [9].

The various classes of stakeholders in the context of XAI were discussed in the related bibliography. For instance, Preece et al. [10] distinguish between four stakeholder communities: developers, theorists, ethicists, and users. Arrieta et al. [11] distinct stakeholders into domain experts, users of the model affected by decisions, managers/executive board members, regulatory entities/agencies, data scientists, developers, and product owners. A similar categorization is done by Lange et al. [12], with the five classes of stakeholders being users, developers, affected parties, deployers, and regulators. However, each of these categories of stakeholders can be further clustered to finer categories, depending also in the field of use of the AI system. For instance, in the medical and healthcare field, among the possible users of the system are considered the clinicians, technicians, nurses, general practitioners, administrative staff, different kinds of students, and patients. The background knowledge of such stakeholders usually deeply differs and requires different user-centric solutions for a successful explanation.

Apart from the different type of stakeholders, the requirements of each stakeholder should be considered. The desiderata arising from the different classes of stakeholders are diverse. Lange et al. [12] propose a list of 29 possible desiderata holding by XAI stakeholders, including acceptance, accountability, confidence, education, fairness, legal compliance, morality/ethics, transparency, usability, and usefulness. The whole list of desiderata and their corresponding stakeholders is shown in Table 2.1. Such desiderata are not completely disjoint and may co-exist in a single XAI-system [12]. Hence, stakeholders in combination with their background and their desiderata is the main reason for the rising popularity of XAI, and guide explanation process.

Desideratum	Tentative description	Stakeholder
Acceptance	Improve acceptance of systems	Deployer, Regulator
Accountability	Provide appropriate means to determine who is accountable	Regulator
Accuracy	Assess and increase a system's predictive accuracy	Developer
Autonomy	Enable humans to retain their autonomy when interacting with a system	User
Confidence	Make humans confident when using a system	User
Controllability	Retain (complete) human control concerning a system	User
Debugability	Identify and fix errors and bugs	Developer
Education	Learn how to use a system and system's peculiarities	User
Effectiveness	Assess and increase a system's effectiveness; work effectively with a system	Developer, User
Efficiency	Assess and increase a system's efficiency; work efficiently with a system	Developer, User
Fairness	Assess and increase a system's (actual) fairness	Affected, Regulator
Informed Consent	Enable humans to give their informed consent concerning a system's decisions	Affected, Regulator
Legal Compliance	Assess and increase the legal compliance of a system	Deployer
Morality/Ethics	Assess and increase a system's compliance with moral and ethical standards	Affected, Regulator
Performance	Assess and increase the performance of a system	Developer
Privacy	Assess and increase a system's privacy practices	User
Responsibility	Provide appropriate means to let humans remain responsible or to increase perceived responsibility	Regulator
Robustness	Assess and increase a system's robustness (e.g., against adversarial manipulation)	Developer
Safety	Assess and increase a system's safety	Deployer, User
Satisfaction	Have satisfying systems	User
Science	Gain scientific insights from the system	User
Security	Assess and increase a system's security	All
Transferability	Make a system's learned model transferable to other contexts	Developer
Transparency	Have transparent systems	Regulator
Trust	Calibrate appropriate trust in the system	User, Deployer
Trustworthiness	Assess and increase the system's trustworthiness	Regulator
Usability	Have usable systems	User
Usefulness	Have useful systems	User
Verification	Be able to evaluate whether the system does what it is supposed to do	Developer

Table 2.1: An exemplary list of XAI desiderata and stakeholders holding these desiderata. [12]

2.3 Representation of explanation

Previous research has discussed different kinds of explanation. A first distinction, as proposed by Tjoa et al. [13], is between perceptive and mathematical structured explanations (see Figure 2.2). *Perceptive interpretability* refers to methods that generate items that are usually immediately interpretable, meaning that they are interpretations that can be humanly perceived through the visual highlighting of important input features with respect to a given output (saliency), through the observation of the stimulation of neurons (signal interpretability), or through the composition of logical statements or sentences (verbal interpretability) [13]. *Interpretability via mathematical structure*, refers to methods that, using mathematical models or data-oriented approaches, generate outputs that require cognitive processing before reaching the interpretability [13].

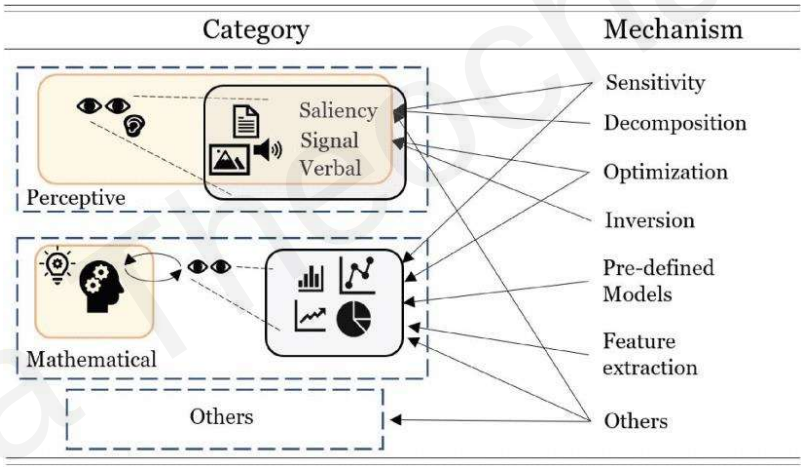

Figure 2.2: Overview of categorization of perceptive and mathematical interpretability. Orange box: interpretability interface to separate between interpretable information and the cognitive process required to understand them. Gray box: algorithm output/product that is proposed to provide interpretability. Black arrow: computing or comprehension process. The eyes and ear icons represent human senses interacting with items generated for interpretability. [13]

Another way to differentiate is between explanations generated by model-agnostic (or *post-hoc*) approaches and model-specific (or *anti-hoc*) approaches. While the first approaches attempt to provide explanatory information only by observing input/output associations, model-specific approaches consider also specific features of the model under explanation [12]. Model-agnostic approaches have the advantage of working regardless of the type of model, while model-specific approaches on the other hand are usually more efficient, accurate, and explanatory powerful, as the explanatory information's level of detail is higher with regard to individual phenomena [12].

A last aspect to consider for distinguishing explanations is their scope. *Local scope* explanations focus on single predictions/classifications of the supported system, offering visualized prototype outcome examples [12]. *Global scope* explanations are designed to explain the overall reasoning mechanism of the model, by approximating complex models with simpler ones that are inherently explainable [12].

2.4 Evaluation of explanation

While many approaches to XAI have been proposed, it is crucial to assess the quality of the explanation. In the survey conducted by Miller [2] a list of important characteristics of explanations in human-like interactions is provided. Major findings include:

1. Explanations are contrastive: People usually don't only ask why a certain prediction was made but rather why this prediction was made instead of another prediction.
2. Explanations are selected: People rarely expect any explanation covering all aspects of reasoning. Influenced by certain cognitive biases, they rather focus on one or two possible causes from a sometimes infinite number of causes to be the explanation.
3. Probabilities don't matter: People consider causal explanations more relevant than probabilities or statistical relationships in explanation, despite the fact that truth and likelihood are important in explanation.
4. Explanations are social: Explanations are considered as transfer of knowledge in the social interaction between the explainer and the explainee, involving their beliefs as well.

A list of properties is defined in [14] too, that can be used to evaluate the quality of explanation methods. Some of these properties are:

1. Accuracy: the ability that explanation of a given decision generalises to unseen instances.
2. Fidelity: how well the explanation approximates the prediction of a model.
3. Consistency: how similar the explanations are for similar instances, generated from different models.

4. Stability: how similar the explanations are for similar instances, generated from the same model.
5. Comprehensibility: readability and size of explanations.
6. Certainty: whether the explanations reflecting the certainty a model has about its predictions.
7. Degree of Importance: how well the explanation reflects the importance of returned items (e.g., features, rules).
8. Novelty: whether explanations would reflect the fact that the explained instance is not contained or well represented in the training set.
9. Representativeness: it describes how many parts of the model are covered by the model explanation.

However, the required characteristics of the explanations may get in conflict. According to Leilani et al. [15], explanations should be interpretable and complete, and answer 'why' questions, but the most accurate explanations are difficult for people to interpret, and conversely, the most interpretable descriptions frequently lack predictive power. So, there is a challenge for XAI to generate explanations that are both interpretable and complete.

Given these, in a review [6] on XAI applications in Medical Domain, explainability systems are evaluated based on:

1. accuracy vs interpretability,
2. quality human-friendly explanations based on some of Miller's criteria,
3. domain experts' active role in the design and evaluation of the system,

4. overall assessment and trustworthiness, including whether the performance is acceptable, and the explanation is helpful for the user.

Chapter 3

Artificial Intelligence and Explainability in Medicine

3.1 A chronology of AI in Medicine

The integration of AI systems in patient-care systems has started decades ago. In the late seventies and early eighties, medical expert systems emerged, aiming to support diagnostic decisions in specialized medical domains [16]. Figure 3.1. illustrates an expert system schema of early eighties. The answer of a medical expert system is derived by a knowledge base (KB) formed by medical experts, and an inference engine.

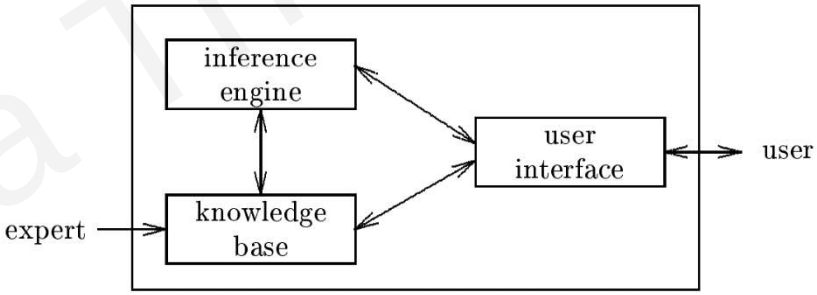

Figure 3.1: An expert system schema of early '80s. [16]

Rules were proposed as a prime formalism for expressing knowledge in symbolic way, because of their advantage of simplicity, uniformity, transparency, and ease of inference. However, acquiring knowledge directly from experts to form rules has high cost, and includes the risk of capturing the biases of one expert. Also, the integration of the rules can reveal inconsistencies and gaps, when there is lack of a global hierarchical organization of rules.

These limitations in combination with the increased availability of databases of example cases led to the need of learning the rules from such data, which seemed a more efficient, less biased, and more cost-effective. Hence, in the late eighties and early nineties early machine learning algorithms developed, with the C4.5 system [17] being the most popular. Although experts were not needed anymore to explicitly construct the rules, they are still actively involved in the development of the systems, as they provide the example cases, and validate the resulting rules [16].

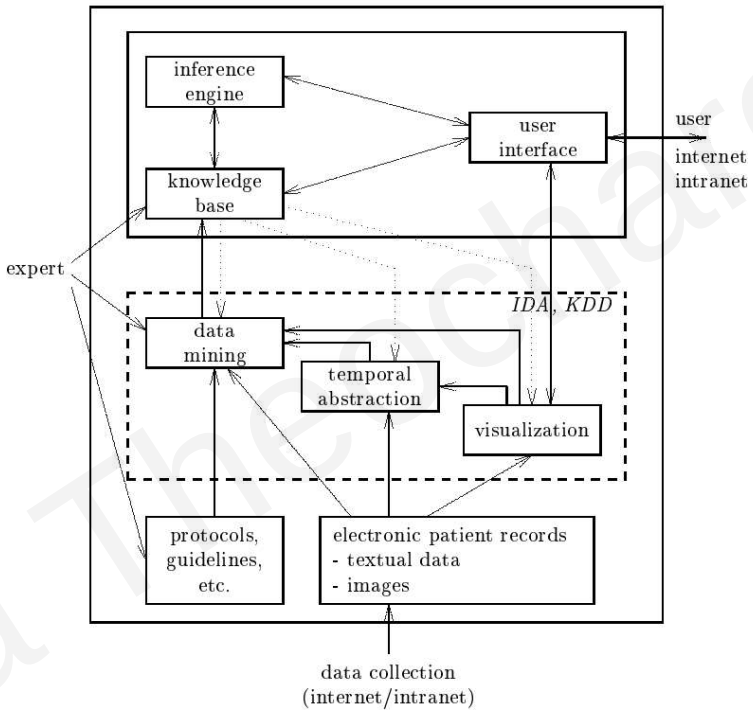

Figure 3.2: A decision support system schema of late '90s. [16]

Other Intelligent Data Analysis (IDA) methods were also developed in nineties to bridge the gap between the massive storage of data and the understanding of the data by discovering principles encoded within the data. IDA methods are categorized in Data Abstraction methods for intelligent interpretation of data, and Data Mining (DM) methods for the discovery of medical knowledge [16]. DM could be performed through symbolic classification methods (rule induction, regression trees, inductive logic programming, constructive induction, and case-based reasoning) and through sub-symbolic classification methods (instance-based learning, neural networks,

Bayesian classifier). Figure 3.2 illustrates a decision support system schema of late nineteenth-century that uses the method of IDA.

Case-based Reasoning (CBR) is a problem-solving methodology that reuses previously solved and memorized problem situations, called *cases*. CBR is suitable for any kind of problem, but it has found a fruitful application in health sciences, where cases are readily available in form of patients [18]. Given a new case, the process to determine a diagnosis or therapy for the patient consists of five stages (see Figure 3.3); interpretation, retrieve, reuse, revise, and retain, explained as follows:

1. Interpretation: The description of the new case is given, and the system interprets it in its knowledge representation language, through abstraction and in particular often time abstraction.
2. Retrieve: The most similar cases to the new case are found, searching the case base or a subset of cases in the base, by applying a similarity measure. The resulting set of cases, named *retrieved cases*, is ranked in descending order.
3. Reuse: The top case(s) from the list of retrieved cases is reused, after adaptation or interpretation depending on the task. This creates the *solved case* that constitutes the proposed solution to the new case.
4. Revise: The solved case is tested in the real-world environment, or evaluated by an expert, a simulation, or a known solution from a test set. After repairing the solved case, it becomes the *tested case* (or repaired case).
5. Retain: The tested case is added in the memory as a complete solution of the target case.

The use of CBR in medicine includes some important challenges, such as the choice of an appropriate similarity measurement, the management of different treatment actions performed by

different physicians, and the auditing of old cases, as medical knowledge is continuously changing [16].

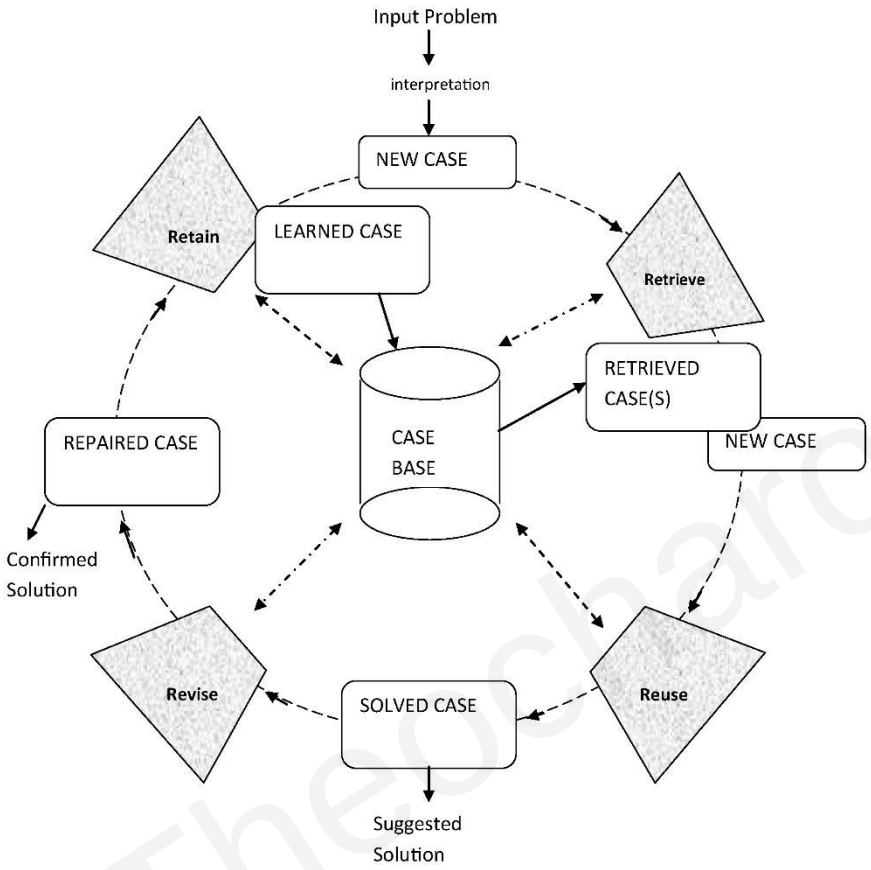

Figure 3.3: The classical CBR reasoning cycle. [18]

Nowadays, Machine Learning (ML) and especially Deep Learning (DL) take the lead from the rest of AI methods used in medicine, as well as in many other domains where AI is used. ML is the study of algorithms that learn from examples and experience instead of relying on hard-coded rules, and can make predictions on new data. DL is a subfield of ML focusing on learning data representations as successive layers of increasingly meaningful representations.

A Neural Network (NN) or Artificial Neural Network (ANN) is the general form of a DL model. Its name and structure are inspired by the human brain, because it is mimicking the way that biological neurons signal to one another. Figure 3.4 illustrates the structure of a NN. It consists of layers that usually have a state, encoded as weights, and has input data (x) and output data or targets (y). Between the input and the output, there can be one or more hidden layers of

connected nodes. The architecture of the model can be simple to complex, depending on the number of hidden layers, the number of nodes per layer and connections between nodes. A loss function is calculated as the difference between the predicted and the actual target, which is used as a feedback signal. Lastly, the network has an optimizer, which determines how the network's weights will be updated based on the loss function. As Deep Neural Network (DNN) can be characterized any NN with more than one hidden layers.

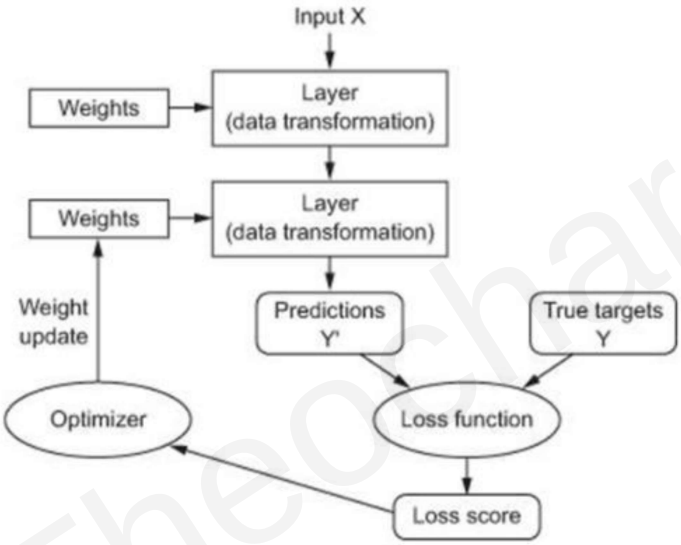

Figure 3.4: Structure of a Neural Network. [51]

Convolutional Neural Networks (CNNs) are NNs that are used for processing data that has a grid-like topology (e.g. time-series data, image data, video data). Hence, CNNs are very popular in the medical domain as pictures is an important source of medical data (radiographs, photographs, MRI, X-ray, and other images).

Medical data include also narratives recorded by clinicians, and other results reported as text. These textual data can be analyzed with Natural Language Processing (NLP) methods. NLP makes use of ML for building models that can learn patterns from data in the form of text. However, DL has gained significant prominence in NLP in recent years. Neural networks including Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs),

and more recently, Transformer models, such as BERT and GPT, have shown state-of-the-art performance in various NLP tasks.

3.2 XAI in Medicine

As previously stated, the provision of explanations for AI systems in general is significant. In a critical domain such as medicine, the significance of explaining the AI generated decision is even bigger. There is need for interpretable decision systems, that allow the understanding of their logic to facilitate debugging and the understanding of the recommended outcome.

3.2.1 Explainability in First Generation Expert Systems

Explainability in first generation expert systems (rule-based expert systems) is an easy task, as the explanation of a decision can be produced by revealing the chain of rules in the derived inference trees. An example of such system is MYCIN [19], which aims at diagnosis of blood infections. MYCIN applies deductive reasoning, implemented through backwards chaining, allowing in this way the justification of the system's recommendations. Moreover, TEIRESIAS system [20] is a KB system with syntactic and semantic knowledge on MYCIN's knowledge, acting as an intelligent debugger, and GUIDON system [21] as an intelligent tutor, aiming at tutoring medical students.

Despite their groundbreaking importance, issues quickly emerged with rule-based explanations, rendering them largely insufficient. These explanations are not meaningful as they are just rule 'playbacks'.

3.2.2 Explainability in Second Generation Expert Systems

Second-generation expert systems (deep knowledge-based expert systems), offered new, promising avenues towards more adequate symbolic explanations. For instance, NEOMYCIN [22] supports knowledge in the form of a causal model, aiming to act as an alternative means to solving problems, as well as to augment rule-based explanations with a more detailed and deep justification.

GUIDON2 [23] was also designed as a new version of GUIDON to provide more successful tutoring.

However, the explainability in these systems is still weak, as the rational basis of strategies is not explained. The explanations are not tailored to the user, and there is no ability for handling and justifying possible exceptions.

3.2.3 Explainability for Case-Based Reasoning

The hierarchical data organization in CBR knowledge base allows additional explanation mechanisms, named explainable case-based reasoning (XCBR) approaches. Schoenborn et al. [24] propose a taxonomy for CBR explanations (see Figure 3.5), divided into four categories: *definition*, *model agnostic*, *model-based*, and *visualization*.

Definition explanations are further divided based on their *goal*, which can be explained by using the questions on the kinds of explanations, including

- justification: why the decision has been reached
- transparency: how the decision has been reached
- relevance: which information were relevant for the decision making process
- contextual: further information on the current situation
- learning: teaching the user.

The model-agnostic nearest neighbor explanations approaches are categorized into:

- probabilistic approach: use machine learning methods to define the similarity measure and the prediction error of retrieving wrong cases
- counterfactuals: convert nearest neighbors into nearest unlike neighbors, used as contrastive explanations

- explanations using examples: do not need domain knowledge, as the explanation is formed by the similar cases

Graph structures are used for explanations:

- patterns: combines textual explanation and graph structure of the case (based on its attributes), the query, and their similarity
- workflows: the workflow event trace is presented and compared to other, similar workflows, and the parts of the workflow contributing to the similarity calculation can be displayed
- routing: navigation routes stored as cases with additional context information

Model-based explanations are based on the application domain, hence they are more precise and targeted. There are:

- probabilistic and analogical approaches, requiring knowledge-intensive models, Bayesian Networks, or ontologies
- recommendation systems, that tend to focus on the domain model, or the user's behavior

Visualization explanations are divided into textual and graphical. Textual approaches include:

- explanation patterns and templates: explaining by identifying categories and extending cases
- free text explanations: reusing past cases and using both justifications and counterfactuals
- interactive approach: justify recommendations that have missing attributes based on user's preferences

Graphical explanations include graph structures, plots and boxes (scatter plots, rainbow boxes with elements representing cases).

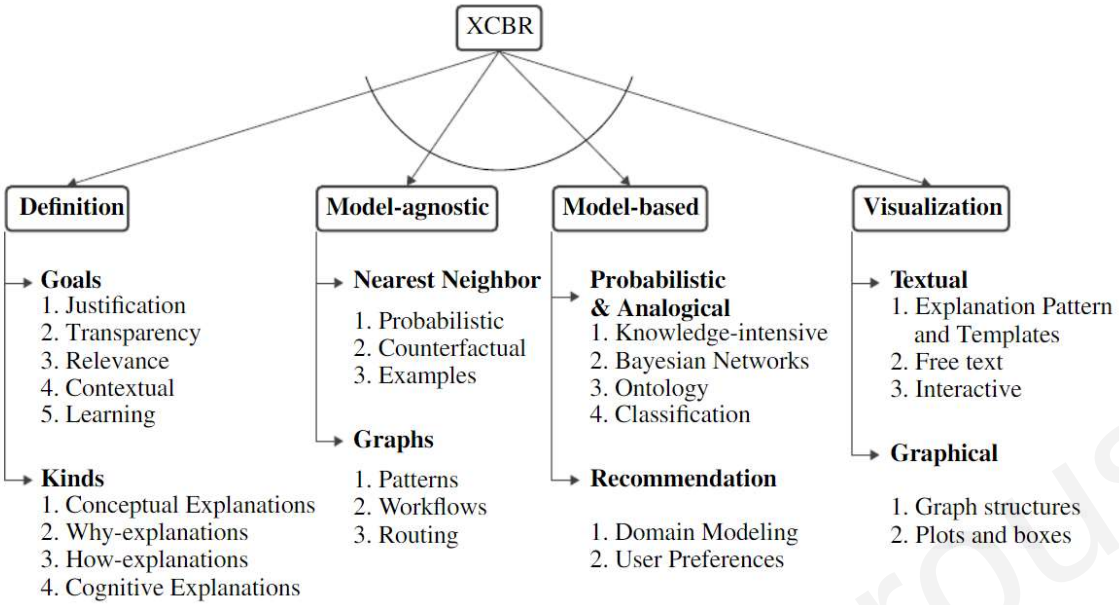

Figure 3.5: Taxonomy of explanation in XCBR. [24]

Lamy et al. [25] propose a CBR method for breast cancer that uses graphic visualization for explainability. It can be executed automatically as an algorithm and presented visually in a user interface as well, for providing visual explanations or for visual reasoning. It combines two visual interfaces, a quantitative and a qualitative (see Figure 3.6), enabling the user to easily classify a query through explainable visual reasoning.

The quantitative is visualized by a scatter plot based on multidimensional scaling in polar coordinates, preserving distances involving the query. While q represents the query, similar cases are scattered around q, where smaller distance to q represents more similarity. Different shapes and colors are used to represent cases of different classes. The qualitative is visualized using rainbow boxes with elements representing cases, that are ordered in columns based on their similarity to q. The boxes including cases' attributes and their values have width that spans over the cases with this attribute-value pair, and height proportional to the global similarity of the given attribute. In addition, boxes' colors are uniquely associated with each solution class when there are not conflicting classes among the cases which correspond to this box. Otherwise, the colors are mixed proportionally to the occurrence of the attribute in a given solution class.

After it was tested in three public data sets, the qualitative method showed accuracy similar to k-nearest neighbors algorithm, but better explainability. For a real data set which is related to breast cancer, the visual approach was found interesting by medical experts, as it explains why cases are similar through the visualization of similar patient characteristics.

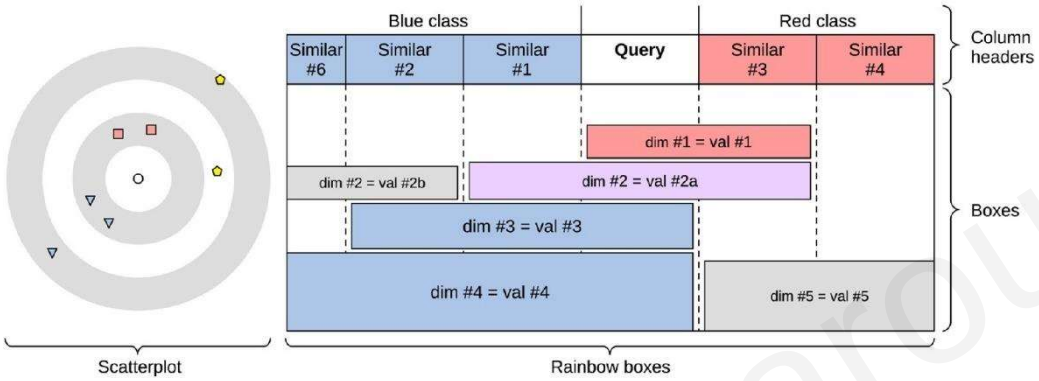

*Figure 3.6: Quantitative (scatter plot) and qualitative (rainbow boxes) visualizations fo*r a CBR result. *[25]*

3.2.4 Explainability in modern Machine/Deep Learning models

To make modern machine/deep learning approaches explainable, post-hoc explainability methods are utilized. Figure 3.7 depicts an overview of the explainability process. These methods

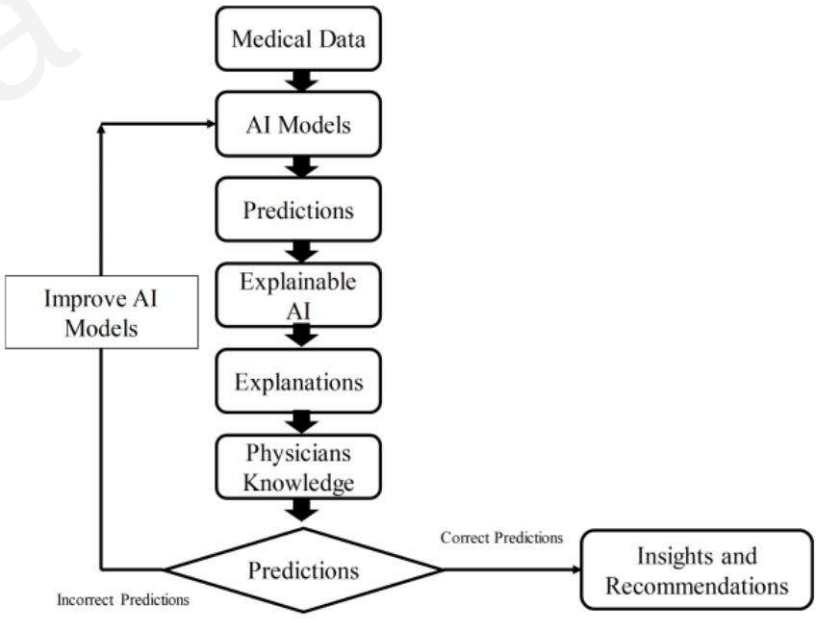

Figure 3.7: A step by step overview of a post-hoc XAI system. [52]

can be distinguished between *model agnostic* methods and *explaining DL methods* [6]. Model agnostic techniques try to provide an understanding of the knowledge of a trained black-box model, and can be further categorized as:

1. Example based techniques: select particular instances of the data set to explain the predictions of a model (e.g. explainable case-based reasoning techniques). These techniques may be insufficient because there are fewer tools for obtaining influential cases and prototypes [52].

2. Feature relevance (or importance or contribution) techniques: assign a score to each feature that indicates its relative importance to the prediction function of the model.

   One example of these techniques is *perturbation feature importance*, where the premise is that important features are detected by studying the effects of their perturbation on the model's prediction. A feature is considered important if the prediction error is increased when changing its values. Crabbe et al. [26] propose dynamic masks (Dynamask) for time series data, which produces instance-wise importance scores for each feature at each time step by fitting a perturbation mask to the input sequence. This mask is learned by backpropagating the error, which we get by comparing the original output and the perturbed output of the model (see Figure 3.8).

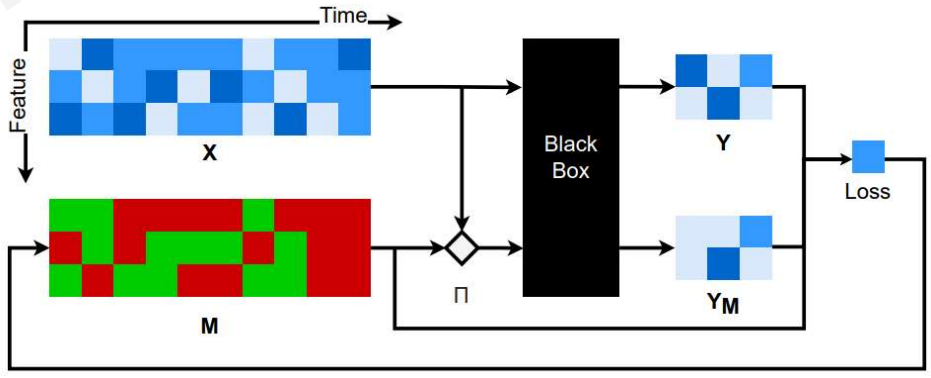

Figure 3.8: Diagram for Dynamask. Time series input matrix X is fed to a black-box to produce a prediction Y. To give a saliency score for each component of X, stored in a mask M (of same shape as the input X), the mask produces a perturbed version of X via a

perturbation operator Π. The perturbed X is fed to the black-box to produce a perturbed prediction Y(M). The perturbed prediction is compared to the original prediction and the error is backpropagated to adapt the saliency scores contained in the mask. [26]

Partial dependence plots (PDPs) show the dependence between one or two input features and the target variable, marginalizing all other features.

SHapley Additive exPlanations (SHAP) [27] is a game theoretic approach to explain the output of any machine learning model. Shapley values are used to assign each feature an importance value representing its contribution to the model's output by averaging the marginal contributions of each feature across all potential permutations of features. SHAP can be used for both local and global explanations. In a global SHAP explanation, the importance of each indicator that has a positive or negative impact on the target variable are determined. For local interpretability, SHAP values are determined for each instance, increasing transparency and aiding in explaining case prediction.

3. Model induction techniques: observe the behavior of the black-box trained model to infer a second model that can be used to explain the predictions of the first.

   One technique is the creation of *surrogate models*. They are usually trained on the inputs and predictions of the opaque model to build an interpretable model. There are two categories of surrogate models: global, which approximate the predictions of the black-box model as a whole, and local that explain individual predictions of a black-box model, e.g. LIME (Local Interpretable Model-Agnostic Explanations) [28]. The idea behind LIME is that, even if the overall decision boundary of a function is complex, in the neighborhood of a single decision the boundary is simple (see Figure 3.9). So, a single decision can be explained by learning an interpretable model locally around the given instance.

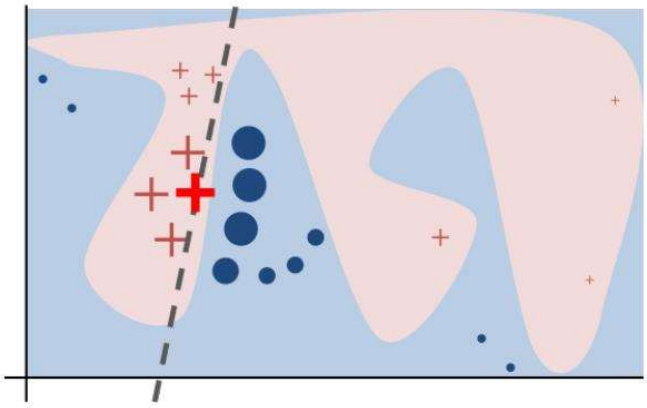

Figure 3.9: Toy example to present intuition for LIME. The black-box model's complex decision function f, that cannot be approximated well by a linear model, is represented by the blue/pink background. The bold red cross is the instance being explained. LIME samples instances, gets predictions using f, and weighs them by the proximity to the instance being explained (represented here by size). The learned explanation is the dashed line, which is locally (but not globally) faithful. [28]

Rule extraction is another model induction technique. It reveals the hidden knowledge of opaque models such as deep neural networks (DNNs), support vector machines (SVMs), and ensemble models. For instance, ANCHORS [29] explains the behavior of complex models with high-precision rules called anchors, representing local, sufficient conditions for predictions. A rule "anchors" a prediction locally if changes to the rest of the feature values do not affect the prediction.

In deep learning (DL), a model/deep neural network (DNN) learns representations from raw data, discovers patterns, and creates mappings from representations to output (prediction). The representations captured by the neurons in the intermediate layers of DNNs are usually not human interpretable, especially in the deeper layers of the network. Thus, many explanation techniques focus on understanding the learned representations. Montavon et al. [30] distinguish between DNN-interpretation and DNN-explanation, where the former is defined as the mapping of an abstract concept (e.g. a predicted class) into a domain that humans can make sense of, and the

latter as the collection of features of the interpretable domain, that have contributed for a given example to produce a decision (e.g. classification or regression). Based on [6] and other proposed taxonomies, techniques for explaining DL methods can be generally classified in non-attribution and attribution-based or visualization (similar or identical methods are classified as visualization or attribution methods in the literature).

1. Visualization or Attribution methods: According to Ras et al. [31], visualization methods explain by highlighting characteristics of an input that strongly influence the output, through a scientific visualization. Attribution-based methods perform post-model analysis on a neural network, aiming to decide the contribution of an input feature to the target neuron (usually the output neuron of the correct class in a classification problem) and visualize the network [32]. These methods can be further classified as backpropagation-based and perturbation-based.

   Backpropagation-based methods visualize features of relevance based on the volume of gradient passed through the layers of the network during its training. The attribution of the input features is computed with a single forward and backward pass through the network [6]. Activation Maximization [33], Gradients [32], Saliency Maps [34], LRP (Layer-Wise Relevance Propagation) [35], CAM and Grad-CAM [36], Deep SHAP [37] and DeepLIFT [38] are some examples of backpropagation-based algorithms.

   Perturbation-based methods are considered the simplest way to examine the effect of changing the input features on the output of the neural network. They visualize feature relevance by comparing network output between the original input and a modified copy of it [32]. Some examples are the occlusion sensitivity method [31] and shapley value sampling [39].

2. Non-attribution methods: develop a specific methodology to provide explainability and validate it on a given problem, instead of performing a separate analysis using pre-existing

methods. These include post-model methods, that may require specific changes to the model structure, and intrinsic methods that improve the interpretability of internal representations within methods that are part of the deep learning architecture [6]. These methods include attention-based mechanisms [40], concept vectors [41], joint training techniques [31], and other extended architectures.

Kary Framling proposed an alternative explanation approach for NNs results, named *Contextual Importance and Utility (CIU)* [42]. The CIU method consists of two important evaluation methods: Contextual Importance (CI), which approximates the overall importance of a feature in the current context, and Contextual Utility (CU), which provides an estimation of how favorable or unfavorable the current feature value is for a given output class. Both CI and CU for an input are continuous numerical values, that are easily computed using Monte-Carlo simulation or other methods. To produce explanations, CI and CU can further be transformed into symbolic values, by defining value ranges for 'good', 'bad', 'important', 'little important', etc., or by using fuzzy sets.

In another review on explainable AI for healthcare [52], XAI methods are divided into example-based, attribution-based, and model-based explanations, and their interpretability (characterized by simplicity and clarity) and fidelity (characterized by soundness and completeness) are evaluated for both local and global scope of explanation (see Table 3.1).

Method	Explanation type	Scope	Fidelity		Interpretability	
			Soundness	Completeness	Parsimony	Clarity
Post-hoc explanation	Attribution	Local	General quantitative metric	Unavailable matrices	Satisfied if an instance or feature is human-intelligible.	General quantitative metric
	Attribution	Global	General quantitative metric	Unavailable matrices	Satisfied if an instance or feature is human-intelligible.	Satisfied
	Model	Local	General quantitative metric	Satisfied	General quantitative metric	Unavailable matrices
	Model	Global	General quantitative metric	Satisfied	General quantitative metric	If the model is incapable of providing various rationales, the model is satisfied.
	Example	Local	Unavailable matrices	Unavailable matrices	Satisfied if an instance or feature is human-intelligible.	Unavailable matrices
	Example	Global	Unavailable matrices	Unavailable matrices	Satisfied if an instance or feature is human-intelligible.	Satisfied

Table 3.1: Quality evaluation for post-hoc explainability methods. [52]

Knapic et al. [43] approach XAI for human decision-support in medical image analysis, by utilizing three explainability methods: LIME, SHAP, and CIU. Their aim was to improve understandability of the decisions provided by a convolutional neural network (CNN) that predicts lesions on in-vivo gastral images (assigns them as bleeding or normal/non-bleeding) and increase health professional trust. A few of the LIME's explanations are shown in Figure 3.10. The explanation is provided by drawing yellow lines around significant features of the photos that affected the black-box model's judgment. The area that positively contributes to the bleeding class is marked in the case of a bleeding image, and the area contributing to the non-bleeding class is marked in the case of a non-bleeding image. Figure 3.11 depicts SHAP explanations provided for some examples, each illustrating contributions to both bleeding and non-bleeding classes. Important features of the images are marked in green color to represent a support for the class, and in red color to represent contradiction to the class. The CIU explanations, shown in Figure 3.12, are similar to those of LIME, marking the important image area that contributes to the given class, either bleeding or non-bleeding. User studies were performed with users from different non-medical backgrounds to evaluate the comprehensibility of the given explanations. The reason that normal people were invited is that explanations should be made as simple as possible in order to make them easily understandable even to non-medical people. The results showed that CIU explanations were more transparent and understandable to users. Between LIME and SHAP users that were provided SHAP explanations reported higher satisfaction with the explanations. According to the time needed to generate the explanations, the best performing method was CIU, followed by SHAP, and then LIME.

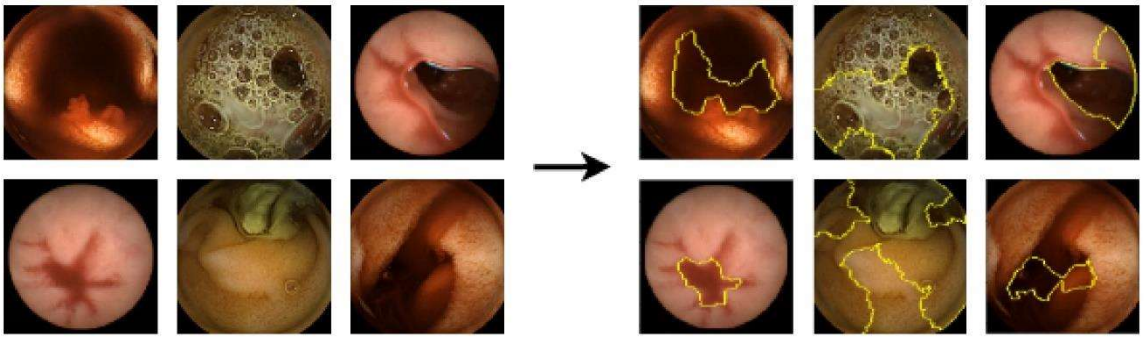

Figure 3.10: A few examples of LIME explanations for lesion prediction. [43]

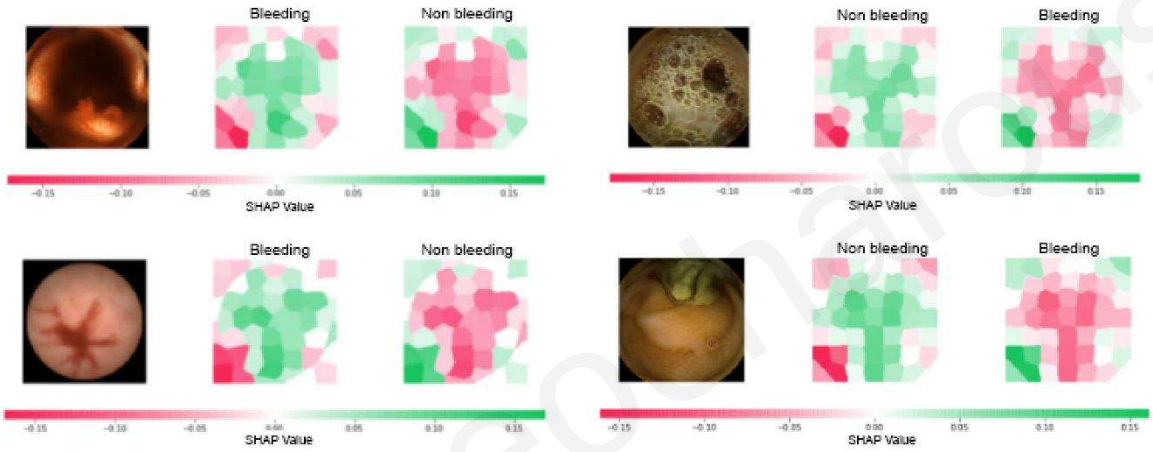

Figure 3.11: A few examples of SHAP explanations for lesion prediction. [43]

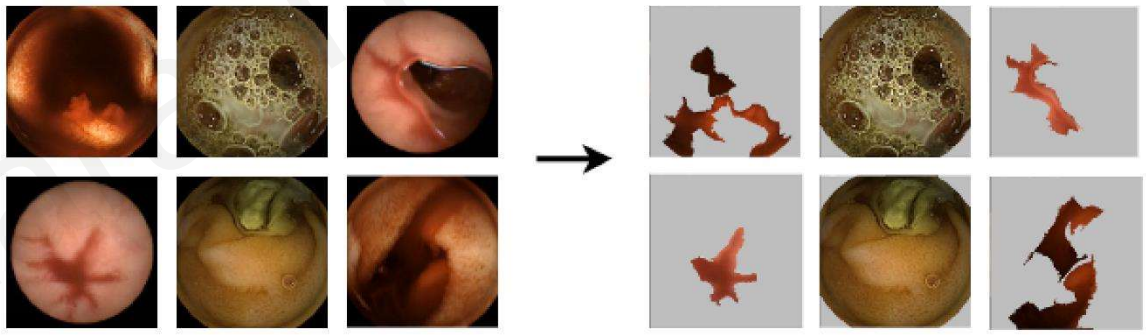

Figure 3.12: A few examples of CIU explanations for lesion prediction. [43]

3.2.5 Hybrid methods

Hybrid methods aim at integrating machine learning with symbolic reasoning to build an explainable AI model.

Prentzas et al. [44] propose applying argumentation on top of Machine Learning (ML) to build an explainable AI system for stroke prediction. Specifically, the ML model used is Random Forest

[45], and after it is fitted on the training data, IF-THEN rules are extracted from the trained model using *inTrees* [46]. Then an argumentation theory is developed in Gorgias framework [47], using the created rules. The resulting application supports queries to find which options (e.g. stroke, asymptomatic) are supported by the argumentation theory for a given set of input (patient data). The arguments (rules) which support the options are also provided, forming in this way the explanation of the predicted result.

Ren et al. [48] propose MulNet framework (see Figure 3.13) to predict pneumonia, combining an opaque and an explainable model. The former is a CNN that predicts pneumonia from chest X-rays as input, and the latter is a Bayesian Network with medical knowledge (see Figure 3.14). The final classification is performed by the Bayesian Network, that takes as input the CNN result and the presence or absence of the seven symptoms/indicators for pneumonia diagnosis (cough, hemoptysis, chest pain, fever, dyspnea, wet rales, dry rales). These seven features are extracted from textual reports of community acquired pneumonia cases. The experiment shows that this hybrid approach improves the performance by using multi-source data and providing explanations for the diagnosis results.

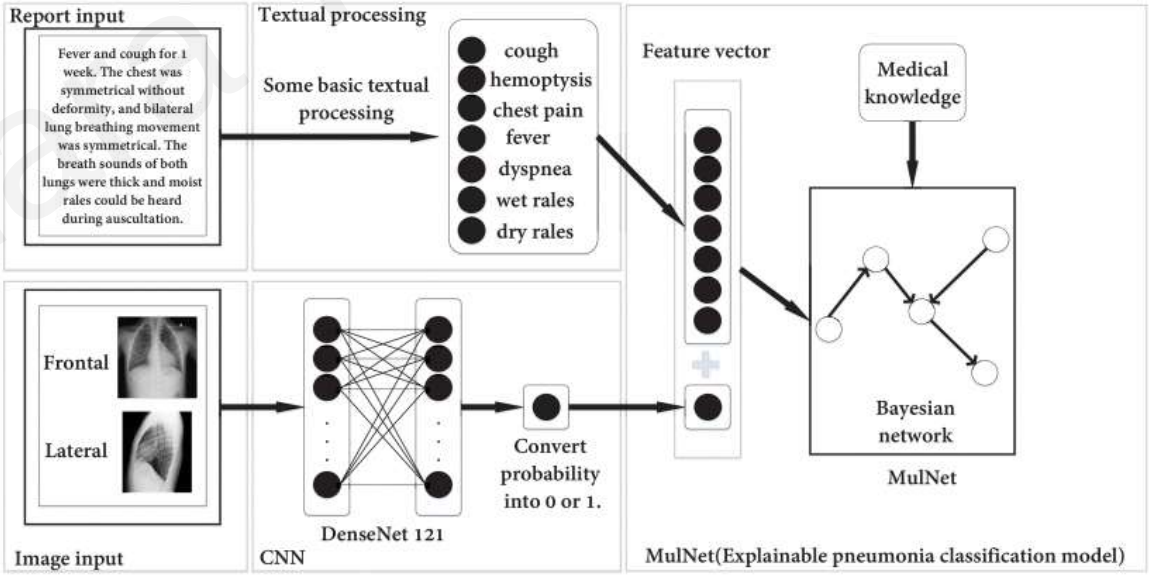

Figure 3.13: An overview of the MulNet framework. [48]

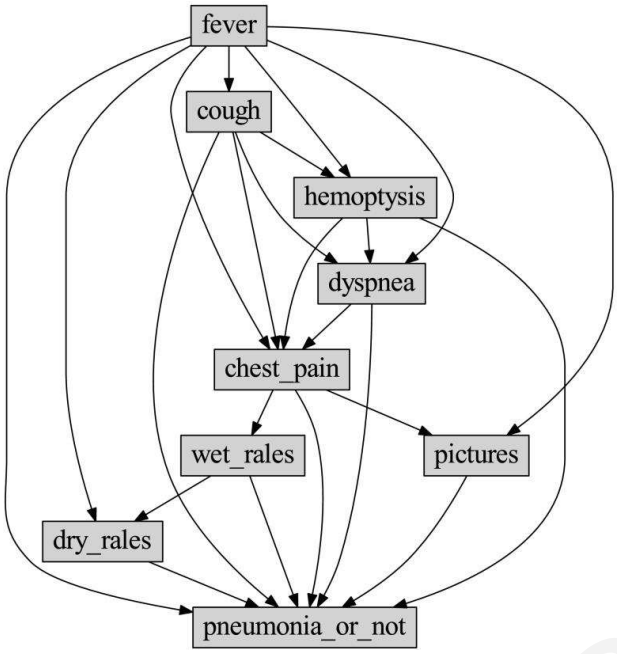

Figure 3.14: Bayesian network structure. [48]

Chapter 4

Conclusions

4.1 Remaining Challenges

The analysis of participants' feedback statements regarding the explanations provided by systems in different user studies [49, 43] have shown that users would like:

- more precise explanations
- explanations to be provided in various forms (both visual-based and linguistic explanations)
- to be able to interact with the explainable method in order to gain more in-depth information.

Applying XAI methods in the medical field include various challenges.

- ML models must be embedded in systems that are usable and acceptable to clinicians, motivated by their needs and the healthcare processes and environmental constraints
- As accuracy comes at the cost of interpretability and transparency, there is need for achieving a balance between them
- There are privacy issues of sharing personal medical data through multiple devices

- Intelligent exploitation of medical data because of their huge volume
- Data quality (i.e. incomplete/inaccurate/non-structured data, without documentation, understanding clinical narrative text)
- Data annotation is time-consuming and expensive process
- Difficulty of aggregation of medical data and use them for machine learning due to various data forms
- Maintenance of data used in the AI system, given the rapid developments in medical domain

4.2 Comparison of XAI methods in Medicine

As noted in previous chapters, the use of symbolic reasoning such as rule-based systems as an XAI method in medicine is a solid solution, as it provides explainability by design. Nevertheless, this approach has drawbacks, including the high maintenance cost and low performance.

Historically, there has been a trade-off between interpretability and the performance of the prediction models. Interpretable models, like decision trees, often perform less well on prediction tasks compared to less interpretable or opaque models like deep learning and machine learning (ML).

ML methods have the advantage of learning from the training data, and the ability to generalize, which may sometimes have negative impact in case of over-generalization. Moreover, the most popular explainability methods for ML approximate the black-box system with a local surrogate linear model, providing in this way only local fidelity, and losing faithfulness to the original model. However, there are other explainability methods that generate the explanation based on the contributing features of the data set.

Case-based Reasoning (CBR) provide the ability of adjusting on the most specific instance in the training set, in contrast with most other AI problem solvers such as Bayesian networks, artificial neuron networks, and decision trees, that have a known tendency to overgeneralize [50]. In addition, decisions made by experts often rely on their experience, yet most ML approaches cannot provide explanations based on specific experiences, as they don't retain them. In contrast, explainable case-based reasoning (XCBR) approaches can provide such explanations, because CBR is a realistic model that imitates human behavior, as humans also think about similar situations when encountering novel situations, which may be solved by adapting a retrieved case's solution. However, maintaining domain knowledge and appropriately populating its knowledge containers are essential for a working CBR system.

Taking everything into account, it seems like the best solution to make AI in medicine explainable is to combine the approaches above, in order to combine their advantages and eliminate their disadvantages. A medical system that provides predictions based on both traditional symbolic reasoning and modern machine learning approaches would increase the trust of the user (both the patient and the physician), apart from providing explainability. Hence, more study is needed to incorporate knowledge, rules, and learning in an efficient way for the medical domain.

Bibliography

[1] T. Simonite, “A Health Care Algorithm Offered Less Care to Black Patients,” *Wired*, Oct. 2019.

[2] T. Miller, “Explanation in artificial intelligence: Insights from the Social Sciences,” 2018. [Online]. Available: http://dx.doi.org/10.1016/j.artint.2018.07.007

[3] “White Paper on Artificial Intelligence - European Commission,” Feb. 2020. [Online]. Available: https://commission.europa.eu/system/files/2020-02/commission-whitepaper-artificial-intelligence-feb2020_en.pdf

[4] C. Combi, B. Amico, R. Bellazzi, A. Holzinger, J. H. Moore, M. Zitnik, and J. H. Holmes, “A manifesto on explainability for artificial intelligence in medicine,” *Artificial Intelligence in Medicine*, vol. 133, p. 102423, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365722001750

[5] M. Reyes, R. Meier, S. Pereira, C. A. Silva, F.-M. Dahlweid, H. von TenggKobligk, R. M. Summers, and R. Wiest, “On the interpretability of Artificial Intelligence in Radiology: Challenges and opportunities,” May 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7259808/

[6] N. Prentzas, A. Kakas, and C. Pattichis, “Explainable AI applications in the Medical Domain: a systematic review,” Oct. 2022.

[7] “General Data Protection Regulation (GDPR).” [Online]. Available: https://gdpr-info.eu/

[8] “Ethics Guidelines for Trustworthy AI.” [Online]. Available: https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf

[9] “Statement on algorithmic transparency and Accountability,” Jan. 2017. [Online]. Available: https://www.acm.org/binaries/content/assets/public-policy/2017usacm_statement_algorithms.pdf

[10] A. Preece, D. Harborne, D. Braines, R. Tomsett, and S. Chakraborty, “Stakeholders in explainable AI,” Sep. 2018. [Online]. Available: https://arxiv.org/abs/1810.00184

[11] A. B. Arrieta, N. Diaz-Rondrigez, J. D. Ser, A. Bennetod, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, “Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” Dec. 2019. [Online]. Available: https://doi.org/10.1016/j.inffus.2019.12.012

[12] M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kastner, E. Schmidt, A. Sesing,¨ and K. Baum, "What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research," *Artificial Intelligence*, vol. 296, p. 103473, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370221000242

[13] E. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," *IEEE Transactions on Neural Networks and Learning Systems*, vol. PP, Sep. 2020. [Online]. Available: http://dx.doi.org/10.1109/TNNLS.2020.3027314

[14] M. Robnik-Sikonja and M. Bohanec, *Perturbation-Based Explanations of Prediction Models*, Jun. 2018, pp. 159–175.

[15] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining Explanations: An Overview of Interpretability of Machine Learning," in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 2018, pp. 80–89.

[16] N. Lavrac, E. Keravnou, and B. Zupan, "Intelligent Data Analysis In Medicine," *Encyclopedia of Library and Information Science: Volume 68-(Supplement 31)*, p. 209, Nov. 2000.

[17] S. L. Salzberg, "C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993," *Machine Learning*, vol. 16, no. 3, p. 235–240, Sep. 1994. [Online]. Available: http://dx.doi.org/10.1007/BF00993309

[18] I. Bichindaritz and S. Montani, "Guest Editorial: Advances in Case-Based Reasoning in the Health Sciences," *Artificial Intelligence in Medicine*, vol. 51, no. 2, p. 75–79, Feb. 2011. [Online]. Available: https://doi.org/10.1016/j.artmed.2011.01.001

[19] E. Shortliffe, "Computer-based medical consultations: MYCIN," *Artificial Intelligence - AI*, vol. 388, Oct. 1976.

[20] R. Davis, "Interactive transfer of expertise: Acquisition of new inference rules," *Artificial Intelligence*, vol. 12, no. 2, pp. 121–157, 1979. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0004370279900146

[21] W.J. Clancey, "GUIDON," Nov. 1983. [Online]. Available: https://apps.dtic.mil/sti/tr/pdf/ADA139999.pdf

[22] W. J. Clancey and R. Letsinger, "NEOMYCIN: Reconfiguring a Rule-Based Expert System for Application to Teaching," in *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, ser. IJCAI'81. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1981, p. 829–836.

[23] W. J. Clancey, "From GUIDON to NEOMYCIN and HERACLES in twenty short lessons," *AI Magazine*, vol. 7, no. 3, pp. 40–40, 1986.

[24] J. Schoenborn, R. Weber, W. Aha, J. Cassens, and K.-D. Althoff, "Explainable Case-Based Reasoning: A Survey," Feb. 2021.

[25] J.-B. Lamy, B. Sekar, G. Guezennec, J. Bouaud, and B. Seroussi, "Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach," *Artificial Intelligence in Medicine*, vol. 94, pp. 42–53, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365718304846

[26] J. Crabbe and M. van der Schaar, "Explaining Time Series Predictions with Dynamic Masks," 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2106.05303

[27] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[28] M. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, J. DeNero, M. Finlayson, and S. Reddy, Eds. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 97–101. [Online]. Available: https://aclanthology.org/N16-3020

[29] ——, "Anchors: High-Precision Model-Agnostic Explanations," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 04 2018. [Online]. Available: https://doi.org/10.1609/aaai.v32i1.11491

[30] G. Montavon, W. Samek, and K.-R. Muller, "Methods for interpreting and understanding¨ deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–15, 2018. [Online]. Available: https://doi.org/10.1016/j.dsp.2017.10.011

[31] G. Ras, N. Xie, M. van Gerven, and D. Doran, "Explainable Deep Learning: A Field Guide for the Uninitiated," *Journal of Artificial Intelligence Research*, vol. 73, pp. 329–397, 01 2022. [Online]. Available: https://doi.org/10.1613/jair.1.13200

[32] Z. Papanastasopoulos, R. K. Samala, H.-P. Chan, L. Hadjiiski, C. Paramagul, M. A. Helvie, and C. H. Neal, "Explainable AI for medical imaging: deep-learning CNN ensemble for classification of estrogen receptor status from breast MRI," in *Medical Imaging 2020: Computer-Aided Diagnosis*, H. K. Hahn and M. A. Mazurowski, Eds., vol. 11314, International Society for Optics and Photonics. SPIE, 2020, p. 113140Z. [Online]. Available: https://doi.org/10.1117/12.2549298

[33] A. Mahendran and A. Vedaldi, "Visualizing Deep Convolutional Neural Networks Using Natural Pre-images," *International Journal of Computer Vision*, vol. 120, no. 3, p. 233–255, May 2016. [Online]. Available: http://dx.doi.org/10.1007/s11263-016-0911-8

[34] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2014. [Online]. Available: https://doi.org/10.48550/arXiv.1312.6034

[35] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K. Muller, *Layer-Wise Relevance Propagation: An Overview*, ser. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer Verlag, 2019, pp. 193–209, publisher Copyright: © Springer Nature Switzerland AG 2019. [Online]. Available: https://doi.org/10.1007/978-3-030-28954-6 10

[36] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626. [Online]. Available: https://doi.org/10.48550/arXiv.1610.02391

[37] S. Lundberg, G. Erion, H. Chen, A. DeGrave, J. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, “From Local Explanations to Global Understanding with Explainable AI for Trees,” *Nature Machine Intelligence*, vol. 2, 01 2020.

[38] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, “Not Just a Black Box: Learning Important Features Through Propagating Activation Differences,” 05 2016. [Online]. Available: https://doi.org/10.48550/arXiv.1605.01713

[39] E. Strumbelj and I. Kononenko, “Explaining prediction models and individual predictions with feature contributions,” *Knowledge and Information Systems*, vol. 41, pp. 647–665, 2014. [Online]. Available: https://doi.org/10.1007/s10115-013-0679-x

[40] P. Chen, W. Dong, J. Wang, X. Lu, U. Kaymak, and Z. Huang, “Interpretable clinical prediction via attention-based neural network,” *BMC Medical Informatics and Decision Making*, vol. 20, 07 2020. [Online]. Available: https://www.researchgate.net/publication/342806167_Interpretable_clinical_prediction_via_attention-based_neural_network

[41] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres, “Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV),” 2018. [Online]. Available: https://doi.org/10.48550/arXiv.1711.11279

[42] K. Framling, “Explaining results of neural networks by contextual importance and utility,” in *Proceedings of the AISB*, vol. 96, 1996. [Online]. Available: https://www.researchgate.net/profile/KaryFraemling/publication/228897070_Explaining_results_of_neural_networks_by_contextual_importance_and_utility/links/00b7d51cc05fe6f010000000/Explaining-results-of-neuralnetworks-by-contextual-importance-and-utility.pdf

[43] S. Knapic, A. Malhi, R. Saluja, and K. Framling, “Explainable Artificial Intelligence for Human Decision Support System in the Medical Domain,” *Machine Learning and Knowledge Extraction*, vol. 3, no. 3, pp. 740–770, 2021. [Online]. Available: https://www.mdpi.com/2504-4990/3/3/37

[44] N. Prentzas, C. Pattichis, and A. Kakas, “Integrating Machine Learning with Symbolic Reasoning to Build an Explainable AI Model for Stroke Prediction,” Nov. 2019.

[45] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[46] H. Deng, “Interpreting tree ensembles with inTrees,” Aug. 2014. [Online]. Available: https://doi.org/10.48550/arXiv.1408.5456

[47] A. C. Kakas, P. Moraitis, and N. I. Spanoudakis, “Gorgias: Applying argumentation,” *Argument & Computation*, vol. 10, no. 1, pp. 55–81, 2019.

[48] H. Ren, A. B. Wong, W. Lian, W. Cheng, Y. Zhang, J. He, Q. Liu, J. Yang, C. J. Zhang, K. Wu, and H. Zhang, "Interpretable Pneumonia Detection by Combining Deep Learning and Explainable Models With Multisource Data," *IEEE Access*, vol. 9, pp. 95872–95883, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9458276

[49] A. Malhi, S. Knapic, and K. Framling, "Explainable agents for less bias in human-agent decision making," in *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, D. Calvaresi, A. Najjar, M. Winikoff, and K. Framling, Eds. Cham: Springer International Publishing, 2020, pp. 129–146.

[50] T. M. Mitchell, *Machine Learning*. McGraw-Hill, 2021.

[51] F. Chollet and J. J. Allaire, Deep Learning with R, 1st ed. USA: Manning Publications Co., 2018.

[52] S. Bharati, M. R. H. Mondal, and P. Podder "A Review on Explainable Artificial Intelligence for Healthcare: Why, How, and When?," 2023. [Online]. Available: https://arxiv.org/ftp/arxiv/papers/2304/2304.04780.pdf