Journal Pre-proof

Explainable, trustworthy, and ethical machine learning for healthcare: A survey

Khansa Rasheed, Adnan Qayyum, Mohammed Ghaly, Ala Al-Fuqaha, Adeel Razi, Junaid Qadir

PII: S0010-4825(22)00756-9
DOI: https://doi.org/10.1016/j.compbiomed.2022.106043
Reference: CBM 106043

To appear in: *Computers in Biology and Medicine*

Received date : 22 March 2022
Revised date : 15 August 2022
Accepted date : 20 August 2022

Please cite this article as: K. Rasheed, A. Qayyum, M. Ghaly et al., Explainable, trustworthy, and ethical machine learning for healthcare: A survey, *Computers in Biology and Medicine* (2022), doi: https://doi.org/10.1016/j.compbiomed.2022.106043.

ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

(2022) –29

Journal Logo

Explainable, Trustworthy, and Ethical Machine Learning for Healthcare: A Survey

Khansa Rasheed[1], Adnan Qayyum[1], Mohammed Ghaly[2], Ala Al-Fuqaha[3], Adeel Razi[4,5,6,7], Junaid Qadir[8]

[1] IHSAN Lab, Information Technology University of the Punjab (ITU), Lahore, Pakistan

[2] Research Center for Islamic Legislation and Ethics (CILE), College of Islamic Studies, Hamad Bin Khalifa University (HBKU), Doha, Qatar

[3] Information and Computing Technology Division, College of Science and Engineering, Hamad Bin Khalifa University (HBKU), Doha, Qatar

[4] Turner Institute for Brain and Mental Health, Monash University, Clayton, Australia

[5] Monash Biomedical Imaging, Monash University, Clayton, Australia

[6] Wellcome Centre for Human Neuroimaging, UCL, London, United Kingdom

[7] CIFAR Azrieli Global Scholars program, CIFAR, Toronto, Canada

[8] Department of Computer Science and Engineering, College of Engineering, Qatar University, Doha, Qatar.

Abstract

With the advent of machine learning (ML) and deep learning (DL) empowered applications for critical applications like healthcare, the questions about liability, trust, and interpretability of their outputs are raising. The black-box nature of various DL models is a roadblock to clinical utilization. Therefore, to gain the trust of clinicians and patients, we need to provide explanations about the decisions of models. With the promise of enhancing the trust and transparency of black-box models, researchers are in the phase of maturing the field of eXplainable ML (XML). In this paper, we provided a comprehensive review of explainable and interpretable ML techniques for various healthcare applications. Along with highlighting security, safety, and robustness challenges that hinder the trustworthiness of ML, we also discussed the ethical issues arising because of the use of ML/DL for healthcare. We also describe how explainable and trustworthy ML can resolve all these ethical problems. Finally, we elaborate on the limitations of existing approaches and highlight various open research problems that require further development.

Keywords: Explainable Machine Learning, Interpretable Machine Learning, Trustworthiness, Healthcare.

1. Introduction

In recent years, various machine learning (ML) techniques have been widely applied to different healthcare applications. In particular, deep learning (DL) based methods have provided state-of-the-art performance for various healthcare tasks including medical image reconstruction [1], management of electronic health records [2], cancer segmentation [3], disease prediction [4], clinical imaging [5], image retrieval [6], and computational biology [7]. DL models have a complex architecture that consist of multiple layers of neurons. These neuronal layers are connected through non-linear activation functions. These complex and dense DL models produce more accurate results than conventional ML techniques. However, these models have black-box nature and lack an underlying theoretical foundation behind their decisions [8]. Therefore, despite the significant performance of DL-based healthcare ML systems, building trust of clinicians and patients is quite difficult because entrusting the decisions of black-box systems that are not explainable can be life-threatening [9]. To get the benefit of ML/DL empowered healthcare their decisions should be interpretable and explainable in a human understandable way. Figure 1 illustrates the essential traits of ML models required for clinical imple-

*Corresponding author: Junaid Qadir

Email Addresses: khansa.rasheed@itu.edu.pk (K. Rasheed); adnan.qayyum@itu.edu.pk (A. Qayyum); mghaly@hbku.edu.qa (M. Ghaly); aalfuqaha@hbku.edu.qa (A.Al-Fuqaha); adeel.razi@monash.edu (A. Razi); jqadir@qu.edu.qa (J. Qadir)

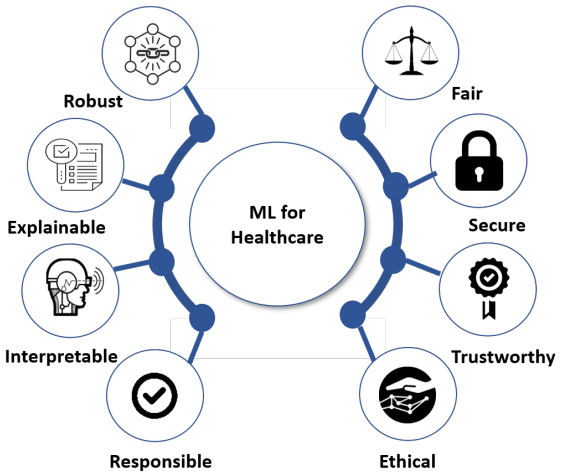

Figure 1: Illustration of essential traits of ML models for clinical implementation.

mentation.

Over the last few years, considerable attention has been devoted to the interpretability, explainability, and trustworthiness of ML/DL models. Among others, two eminent groups of researchers working in this area are: (1) Fairness, Accountability, and Transparency in Machine Learning (FAT-ML) [10] and (2) the Defense Advanced Research Projects Agency (DARPA), explainable AI program [11]. FAT-ML comprises a group of academic researchers with a prime focus on equipping machine algorithms used for social and commercial decision-making with fairness and explainability. This group arranges conferences annually to bring together interested researchers and participants from all over the world. DARPA[1] organized a group of civilians and military researchers in 2017 intending to develop new methodologies for making ML models explainable [11].

Industries with AI/ML products are also contributing to developing XML methods. Microsoft having Azure ML services, H_2O.ai having driverless intelligent products [17], Kyndi serving government, financial, and healthcare sectors with its AI platform are a few of the famous industries working on explainable ML. Fair Isaac Corporation (FICO), a data analytics company, held a challenge in 2018 on explainable ML[2]. The challenge was a collaboration between Google, FICO, and academics of different universities. The challenge aimed to open future directions in the area of explainable algorithms.

We must build safety and trust in ML-based applications by explaining a few questions, i.e., what patterns of features has the ML/DL model learned? Why is the selected model producing better results than other models (for a particular problem at hand)? These explanations are required to convince the clinicians that a particular ML/DL-based algorithm is the best and most powerful tool for disease prediction and diagnosis, which can facilitate their routine practice without causing harm to patients. The explained results will also help patients to understand ML/DL predictions and will help in gaining their trust and satisfaction (being efficiently diagnosed by these algorithms). Thus, for clinical implementation of the ML/DL models, we need transparency, interpretability, and risk understanding.[3]

In addition, mapping of complexly distributed heterogeneous medical data into arbitrary high dimensional space is a major challenge for researchers. With the explainable machine decisions, it would be easier to manage the diverse data for relevant results. Explainable ML (XML) is a solution to these problems for moving towards more transparent ML decisions. Note that the terms explainable and interpretable are sometimes used interchangeably in the literature. However, these two terms are distinct and have domain-specific definitions. Montavon et al. [18] defined interpretation as a mapping of abstract ideas into the human-understandable domain. They discriminate the term interpretation from the explanation by defining the explanation as features of the interpretable domain that contributed to produce the decisions of ML algorithms.

Contributions of this paper: Due to the immense importance of explainable, trustworthy ML decisions, and ethical use of ML for healthcare, multiple surveys cover these topics. Below we outline specific contributions of this paper, which are in contrast to the existing works (a comparison is presented in Table 1).

1. To the best of our knowledge, no existing review or survey provides an in-depth analysis of explainable, trustworthiness, and ethical aspects of using ML/DL models while highlighting their applications and their importance for the medical domain.

2. We propose a pipeline to attain an explainable ML framework for healthcare that involves development, testing, and deployment phases. This pipeline showed the use of different explanation methods to explain and validate data and models.

3. We highlight various security, safety, and robustness challenges associated with the ML/DL that obfuscate their trustworthiness in healthcare applications.

4. We also discuss ethical challenges related to the use of ML/DL in healthcare applications and elaborate upon the use of explainable and trustworthy ML to resolve these ethical problems.

5. Finally, we discuss the limitations of the existing state-of-the-art approaches and highlight various open research problems that require further development.

For instance, Adadi et al. [12] provided a review of explainable artificial intelligence (XAI) techniques and partly described the applications in transportation, healthcare, legal, finance, and military domains. Arrieta et al. [19] have provided

[1] https://www.darpa.mil/attachments/XAIProgramPortfolio.pdf

[2] https://community.fico.com/s/explainable-machine-learning-challenge

[3] https://www.vanderschaar-lab.com/from-black-boxes-to-white-boxes/

Table 1: Comparison of this paper with existing surveys. Legends: √= discussed, ×= not discussed, ≈ = partially discussed, **ML** = explanation of conventional ML methods applied in healthcare, **DL** = explanation of DL methods applied in healthcare

Reference	Year	Scope				Methods			Challenges	Future Directions
		Healthcare	Focused Application(s)	ML	DL	Explainable/ Interpretable	Trustworthy	Ethics		
Holzinger et al. [9]	2017	√	Segmentation of medical images and omic data	×	≈	√	×	×	×	≈
Adadi et al. [12]	2018	≈	Trends of explainable approaches	√	≈	√	×	×	√	≈
Tjoa et al. [13]	2019	√	Categorization of XAI methods and partially discussed application for healthcare.	≈	≈	√	×	×	√	√
Singh et al. [14]	2020	√	Detection and predict--ion of disease using medical imaging.	×	√	√	×	×	×	≈
Char et al. [15]	2020	√	Identification of ethical problems for healthcare application.	√	≈	≈	×	√	√	√
Adadi et al. [16]	2020	√	Partially discussed XML applications for healthcare	×	≈	√	×	×	×	√
This paper	2021	√	All most all healthcare applications	√	√	√	√	√	√	√

Table 2: List of Acronyms

AM	Activation Maximization
CAM	Class Activation Maps
CNN	Convolutional Neural Network
DARPA	Defence Advanced Research Projects Agency
DeconvNet	Deconvolutional Network
DeepLIFT	Deep Learning Important Features
DL	Deep Learning
DNN	Deep Neural Network
DT	Decision Tree
EMANET	Evidence Activation Mapping
FA	Feature Attributes
FAT-ML	Fairness, Accountability, and Transparency in ML
FICO	Fair Isaac Corporation
GAM	General Additive Model
GB	Guided Back Propagation
GWAS	Genome-Wide Association Studies
HSCNN	Deep Hierarchical Semantic Convolutional Neural Network
IG	Integrated Gradient
LIME	Local Interpretable Model-Agnostic Explanations
LRP	Layer-wise Relevance Propagation
ML	Machine Learning
M-LAP	Multi-Layers Average Pooling
P2V	Patient2Vec
PDP	Partial Dependence Plot
PET	Positron Emission Tomography
RF	Random Forest
SA	Sensitivity Analysis
SHAP	Shapley Additive Explanations
XML	Explainable ML

a brief overview of the concept of explainability, future opportunities in the field and the research challenges. Amitojdeep Singh et al. [14] have briefly described the explainable methods for DL and applications of these methods in medical image analysis. This review is unique because it comprehensively provides the application of each interpretable and trustworthy method in support of healthcare applications besides the forenamed contributions. The comparison of this paper with existing surveys is presented in Table 1.

Organization of paper: The organization of this paper is as follows: Section 2 presents challenges encountered in developing clinically effective explainable and trustworthy ML. Section 3 provides a brief background of explainable and interpretable ML with the description of why we need XML models for healthcare, what characteristics healthcare XML models should have, and how to evaluate the quality of explained results. In Section 4, we describe the notion of safe, robust, and trustworthy XML for healthcare along with a comprehensive overview of XML approaches applied in the literature for explaining decisions of healthcare applications for sustaining trust in ML applications. In Section 5, we discussed the requirement of ML ethics for healthcare along with the history of medical ethics, various ethical challenges related to healthcare, and principles of healthcare ethics. Insights and pitfalls are discussed in Section 6 and various future directions are provided in Section 7. Finally, we conclude the paper in Section 8. A list of acronyms used in the paper is available in Table 2.

2. Challenges

For the sake of trustworthy and secure models for clinical settings, researchers are developing the tools and techniques for XML. Despite their efforts, many issues still exist, causing chal-

lenges for effective XML. A few such challenges are described below.

2.1. Lack of Formal Definitions

The explanation of the model structure or decision has no formal definition and is defined according to the problem at hand (as we discussed in Section 3). The same is the case for XML for healthcare applications. There is also the need for defining terms like feature relevance, feature importance, saliency maps, heatmaps, etc., because there is no consistency in the use of these terms.

2.2. Lack of Standardized Representation Methods

All visualization-based explanations produce saliency maps or heatmaps that highlight the areas of images more participating in predictions. However, it is not yet standardized whether the radiologists or neurologists are interested in these explanations or not. It is also not evident how the end-user (i.e., a patient or a clinician) will interpret the explanations. Moreover, it may be difficult for new or untrained clinicians to understand the language of explained results. Also, there is a possibility that the medical experts may be unable to understand the explained risk factors and estimated probabilistic explanations [20]. There must be a platform connecting the medical experts with XML researchers so that they can communicate for the standardized representations of explanations [21]. Another challenge is to quantify how much explanation is required to make the decision understandable to non-technical end-users like patients, which is equally important to gain their trust in these applications.

2.3. Lack of Standardized Requirements for XML

Researchers have developed some initial guidelines about the requirements of an XML model. However, these guidelines are generic. Requirements for explaining the decisions of animal image tagging will be different from medical image tagging. The current field of medical XML lacks requirement guidelines for designing, measuring, and testing explanations. These guidelines are required to build more explicit and systematic ways for generating explanations of how the black-box models predict or detect a particular disease [22].

2.4. What Clinicians Want: Accuracy vs. Explainability

The complex non-linear structure of DL models is one of the reasons causing decisions that are difficult to explain. This challenge is not limited to healthcare XML. However, due to the multi-dimensional nature of medical data, DL algorithms are crucial to avoid for obtaining precise results. It leads to less explained results or algorithm-centric explanations. One possible solution to this problem is to design inherently explainable techniques that can produce accurate results with complex medical data [23]. The other possible solution is considering the preference of the end-user.

2.5. What and Hows of the Explained Results

Feature maps of medical image data produce reconstructed images containing highlighted relevant features for decision-making. However, answers to questions like what to do with these partially reconstructed images, how can we guarantee that the combinations of features highlighted by the XML are robust to perturbations, and how researchers can use the internally highlighted parameters to recover input data that is not yet considered. The reverse image analysis will help analyze complex medical data. This analysis can leverage the clinicians to understand the hidden mechanism of many life-threatening diseases like COVID-19, breast cancer, Zaire Ebola, and human immunodeficiency viruses (HIV).

2.6. Validation of Explanations

The measures to validate the quality of produced explanations are not adequate. In particular, one major problem is the unavailability of a metric for comparing the generated explanations using different methods. For example, to explain the detection of glioma tumors, various XML techniques have been implemented (discussed in Section 4.4), but no one compared which method produced the better explanation of the tumor detection. Similarly, for healthcare applications, clinicians may need different measures to validate the explained results. There does not exist any standard method for measuring the quality of explained healthcare decisions. Also, there is no measure to check which explanations should be preferred from the different explanations produced by the same method [24].

2.7. Lack of Theoretical Understanding

Applied DL for medical applications lacks theoretical fundamentals for working with the randomness of data. Field experts tried to overcome this gap by applying mathematical techniques for dealing with random artifacts and noise in medical data. However, due to the unavailability of sound fundamental laws and models, we can not produce explanations of DL up to the required scale. These issues are also causing challenges for developing self-explained generalized DL for medical applications [8]. In addition, this black-box nature of the DL also poses a major challenge in developing trustworthiness [25].

2.8. Lack of Causality

DL is designed to produce precise results by learning the hidden patterns that generate data. The problem arises due to the use of these techniques for healthcare tasks where decisions should be based on causal links. However, DL is not efficient in inferring causal relations between decisions and data. It leads to the generation of inadequate results, which cause unsatisfactory or incomplete explanations. Moreover, XML should answer the cause-effect scenarios, i.e., the decision of the model will change from A to B if the doctor replaces treatment C with D [26]. These causal links are required for taking fair decisions. Moreover, Castro et al. emphasized the need for a causal relationship between images and their annotations [27].

2.9. Ethical Constraints

To gain the trust of clinicians and patients, explanations of black-box models must ensure the ethical balance between end-users and XML. In particular, an explanation should contain the complete information and not misguide the end-user [28]. XML should explain the reasons for the error in results to increase fairness and reliability. Unfortunately, there are no criteria for assessing the exactitude and comprehensiveness of explanations. Due to the unavailability of these measures, the application of XML in clinical settings may have adverse effects. Moreover, understanding how the explanations impact the dignity and well-being of patients is also an ethical requirement, i.e., data reconstruction from explanations can be used negatively [29].

2.10. Security Challenges

Notwithstanding the state-of-the-art performance of ML and DL-based methods, many recent studies have highlighted the vulnerabilities of these systems towards adversarial ML attacks [30]. Moreover, such attacks have been already realized on ML/DL-based medical systems [31]. Beyond adversarial ML, many security challenges hinder the deployment of ML/DL in actual clinical settings, a detailed overview of these challenges can be found in [32]. These challenges raise many concerns about thy safety of ML/DL empowered systems, therefore, the robustness of ML/DL models is crucial in developing trustworthiness and transparency in ML/DL empowered healthcare applications. The excellent performance of an ML/DL cannot be evidence of its safety, which is simply the determination of how safe is the ML/DL empowered system for humans, i.e., patients. On the other hand, it is equally important that the ML/DL-based techniques should be trusted by both clinicians and patients.

3. Explainable ML

The problem of explaining intelligent algorithms to humans is known since the 1970s, however, the work in this research area slowed down due to advances in ML techniques [33]. XML is a research field first explored by Van et al. in 2004 [34]. They described, that their developed system can explain the behavior of the algorithm, Full spectrum command, which is used by the U.S. Army. Their XAI system allows the user to click on any AI-controlled soldier in the playback window and access a pop-up menu of questions that can be asked of that soldier. However, their developed XAI system can not provide detailed and deep explanations like why each task should be approached in a specified way.

With the increasing employment of AI/ML methods in industry, medicine, education, and defense systems, the explanation of the machine decisions is crucial to avoid unwanted circumstances, specifically, for healthcare applications. For example, applications like medicine suggestion, disease prognosis or prediction, and mortality prediction demand explainable decisions for ethical reasons and for making these applications socially acceptable.

3.0.1. Definitions of Explainability in Literature

Explainable and interpretable ML has no formal and generally applicable definition. Some of the definitions introduced and used by researchers are the following:

- *Explainability:* DARPA defines explainability as producing explainable models while maintaining high prediction results that help users to understand and trust the decisions of artificial systems [11]. FAT-ML cleared its goals by stating XML as a procedure to ensure that machine decisions and the data driving those decisions should be explainable to humans in non-technical terms [10]. FICO said that XML is a shift toward converting the black box of ML to a white box. The organization defined XML as a challenge to develop techniques that provide a trustworthy explanation with high accuracy to meet the needs of end-users. Leilani et al. [24] stated the term as a science of perceiving what a model did or might have done.

Predictions in the medical field should not be based on blind faith since the consequences can be tragic. By explanation of prediction, we mean providing textual or visual features that provide a contextual interpretation of the correlation between the components of the instance and the prediction results of the model. The idea of XML is illustrated in Figure 2. It is clear that if understandable explanations are given, a doctor is far better prepared to make a decision using these explainable models. In this example, a small list of conditions with corresponding weights is an explanation for taking the decision. Humans typically have foreknowledge of the problem domain, which they will use to believe or deny a prediction if they understand the explanation of results by the algorithm.

3.1. Taxonomy of XML

To explain the decisions and behavior of ML, different explaining models should be developed and implemented. Here we describe the categorization of XML approaches based on their complexity, scope, and employment.

- *Intrinsic Model:* This method is used to design explainable models by reducing the complexity of the ML and adopting simple architectures that are inherently explainable.

- *Post-hoc Models:* It is a technique to analyze complex high-performance black-box ML after the training process. To derive the explanation of these models, reverse problem techniques are usually applied.

- *Model-specific Explanation:* Techniques for model-specific explanations are restricted to specific types of models. For example, the explanation of learned weights of regression or linear model is limited to the specific model. Moreover, the explanation of intrinsic models is model-specific by definition.

- *Model-agnostic Explanation:* These can be usually applied to any ML model after the training. Agnostic models cannot access the internal architecture and weights of the

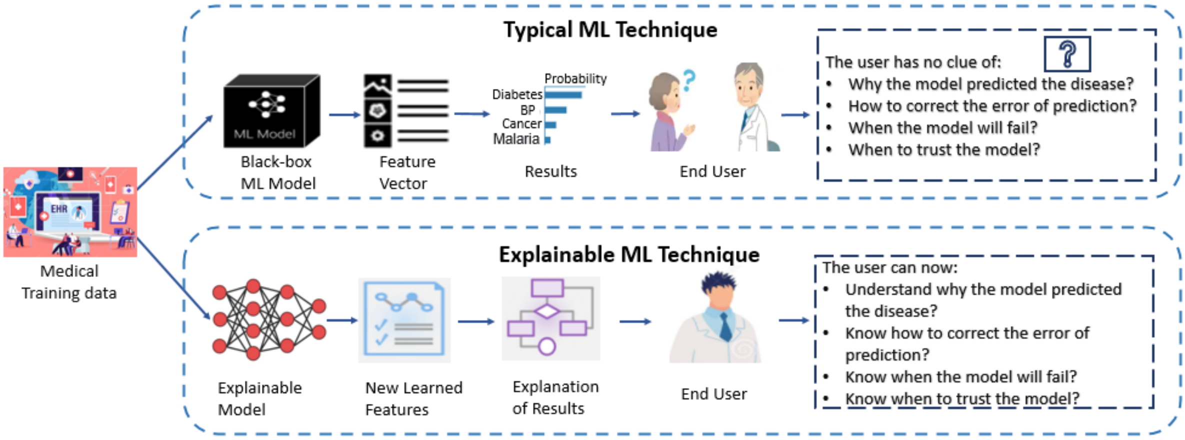

Figure 2: Depiction of how the explainable ML technique is different from the typical ML technique.

ML technique. For the post-hoc models, agnostic explanations are sometimes drawn by using simplification techniques to reduce the complexity.

- *Surrogate Methods:* In this method, different explainable models are designed to analyze the ML black-box. The explanation of black-box models is produced by comparing the decisions of surrogate models and the decision of the black-box model.
- *Visualization Methods:* These explanation methods use visual graphics like activation maps or heatmaps to explain some parameters of architecture of the black-box model.

Based on the mechanism of the explanation model, explainable methods have two broad categories white-box explanation and black-box explanation. The white-box learning model produces explanations for individual output. In this technique, the model identifies the portion of features that are significant for the prediction [35]. Another approach used for white-box explanation is the gradient computation of the prediction with respect to individual input samples to find out the prediction relevant features [36]. White-box explanation mostly provides the model-specific explanation. The black-box methodology provides local explanations of a model for a prediction [37]. However, this mechanism lacks the ability to describe all representations learned by the model.

3.2. Local vs Global Explainability

Explanation of the model will be global if the explanation provides details about which feature is contributing the most to all predictions of the model. It is an average across all the predictions. Global explainability provides information about 'what extent each feature contributes to how the model makes its predictions over all of the data.' On the other hand, local explainability helps in answering the question, "for this particular example, why did the model make this particular decision?"[4]

3.3. Need of XML for Healthcare

Explanation of results is not only necessary for financial gains and ethical challenges but is also desirable for clinical practice if end users (patients or doctors) want to learn, understand, and efficiently manage ML algorithms. Based on the literature reviewed, the following factors are the reason for the necessity of XML models in the research area of healthcare.

3.3.1. To explicate data

Contamination of clinical data and its complex and multivariate nature can lead to bias in the data that the model can learn. Learning biased information in the medical domain can pose life-threatening risks. Explanations derived from the XML allow the visualization of the relation of features affecting the outcome. Thus, the explanation provides a fair analysis of model architecture and learned parameters [38].

3.3.2. To pick the best model

Many design choices, not just the selection of the classification or prediction algorithm but innumerable variations in each stage of pre-processing of medical data during model development, alter the model. There can be countless algorithms with high predictive results. It can be a case that the model with higher performance and accuracy is the worst one, which limits the understanding of the end user in the real-time clinical practice as per the so-called 'Rashomon effect' [39]. The explanation of each algorithm reveals entirely different aspects of the disease learned by the model. These explanations of the results can help researchers and developers to pick between high-performance models.

3.3.3. To enhance clinical use of ML

With the availability of an enormous amount of medical data and advanced ML techniques, research and publications on healthcare are also growing. However, the employment of these algorithms for clinical practice or the use of patients is still distant. The primary reason for this gap is the unexplained results

[4] https://towardsdatascience.com/a-look-into-global-cohort-and-local-model-explainability-973bd449969f

of algorithms and sometimes the poor performance of the algorithm. The explainable techniques allow the researchers or end-users to get involved in improving the performance of the algorithm and to trust the prediction results [40].

3.3.4. To facilitate end-users

ML and XML algorithms are designed to aid the medical staff, not for replacing the medical experts [41]. Medical-related decisions and their explanations have a direct influence on the results of treatment and the survival of patients. So, these intelligent systems still require human supervision to avoid any adverse effects. There can be cases where ML can guide healthcare staff to improve or correct their decisions about treatments. This human-machine combination is a powerful tool to facilitate the patients and develop high-quality treatments [21]. Explanations of these systems are required to gain insights into ML decisions. These insights can help improve the prescribed medicines, facilities provided to patients in hospitals [42], and health monitoring systems [43].

3.4. Enhancing the Clinical Practice of ML: A Framework of Effective XML for Healthcare

It is now evident that the explainability of black-box models is required to attain fair and trustworthy healthcare decisions. Researchers have started developing techniques to build explainable models. However, the field of XML for healthcare has many directions to improve. In this section, we formulate the pipeline for the explainability of data-driven healthcare applications. We discuss the need for explainability at each stage from development to clinical deployment of algorithms.

3.4.1. Unfolding the hidden aspects of data

ML techniques learn patterns of data to make decisions. Any bias in the data, subjectivity, redundancy, or problem in data representation causes misleading results. To produce trustworthy and fair results, we should start with the data explanation. Consider the work of Caruana et al. [44] who built classifiers to classify pneumonia patients as high or low risk for in-hospital death. Their best model gave the results that a patient with asthma has a low risk of in-hospital mortality when admitted for pneumonia. However, the opposite is true. On further investigating the counterintuitive results, the authors realized that asthmatic patients admitted for pneumonia were provided more timely treatment compared to non-asthmatic patients, which led to increased survival success. Thus it was the timely treatment and not the fact that the patient had asthma that reduced the risk of in-hospital mortality. Another example of data bias could arise from patients being denied access to medical care due to not having health insurance. If ML learns from such biased data, it will generate biased results. Similar is the case for data leakage, which can mislead the model learning and testing [45]. To avoid these problems, researchers need to develop a data explanation method that interrogates all dependencies of the target on acquired data.

3.4.2. Explaining the structure of black-box

The problem of explaining black-box models can be further divided into two subproblems. The first subproblem (*model-based explanations*) is to explain the logic of the black-box model in an interpretable human-understandable way, while the second subproblem (result explanations) is to explain the input-output relevance used by the model to make decisions [46]. The model-based explanation methods are well developed and implemented for healthcare applications (further discussed in Sec 4.4). These models very well mimic the behavior of black-box models in terms of logic learning and provide global interpretability. Some ML techniques are inherently explainable due to their simple structure, like decision trees and random forests. However, many black-box models require other models that mimic their work for the explanation.

3.4.3. Explaining the results

Explaining the structure and logic of a model can be complicated for some non-technical medical end-users. In this case, only the explanation of why the model is making this decision can be helpful. This explanation usually consists of the feature relevance for output. Contrary to the local explanation for a single patient, a global explanation is required for generalization purposes.

3.4.4. Measuring the effectiveness of explanations

The quality of explanation depends on the training and validation of the model. In general, the training and validation data sets are divided into 80:20. As a result, 20% of the data is reserved for validation. The ratio varies according to the size of the data. In the scenario when the data size is significantly huge, we can also use a 90:10 data split ratio in which the validation data set comprises 10% of the data. Furthermore, when we divided the data set into three parts, namely the training data set, the validation data set, and the test data set. We train the model with the training data set, evaluate its performance with the validation data set, and improve its performance with the training and validation data sets. Finally, the test set is used to evaluate the model generalization performance. It is important to note that the test set is kept secret during the model training and model assessment process stages, i.e., it is not visible to the model. In such cases, we can divide the data in a 70:20:10 ratio. Where 10% of the data set can be reserved as test data to evaluate the performance of the model.

Due to the non-monolithic and subjective nature of explainability, the evaluation of explanations is a complicated task. There are no sound traces of the best measurement for evaluating the XML, nor we could say anything about how much the model is explainable. Despite the increasing research on the said topic, few researchers focused on the problem of evaluating XML. Some of the main approaches adopted by the healthcare researchers for the evaluation are described next. Figure 3 is the depiction of these steps required for explaining the black-box models. We note that these approaches are not limited to the evaluation of healthcare XML.

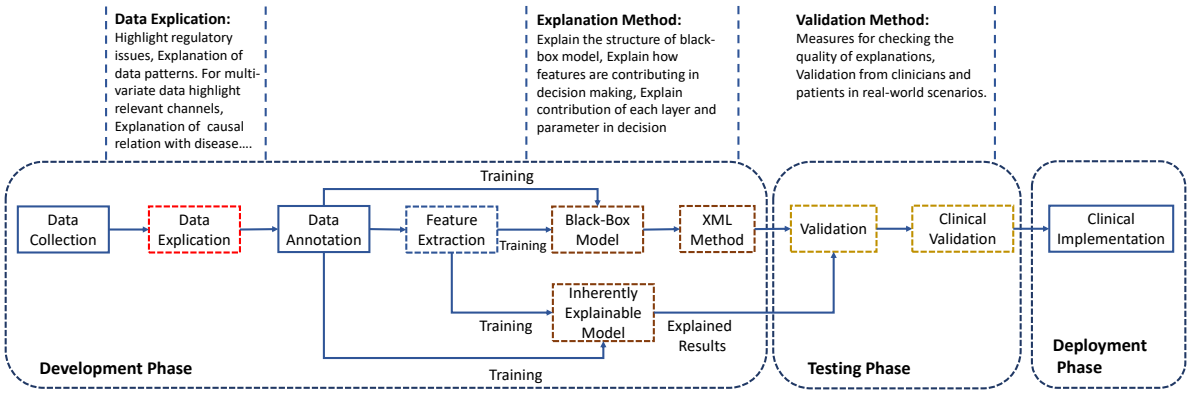

Figure 3: The pipeline for explaining the black-box models.

Application-based Evaluation: Place the explanation into the product or application and get it tested by the end customer, which is usually a domain expert. This technique helps in evaluating the explanation in real-time practical scenarios. For example, consider the ML-based medical data annotation software that places markers on the diseased regions of data. In the clinical application, the clinician would test the annotation software to evaluate the model. The clinician can explain the same decision and can evaluate the explanation and performance of data annotating software [47].

Human-based Evaluation: This technique is similar to application-based evaluation. However, the difference is that it does not require a costly experimental environment and domain expert for testing. One can test the explanations with laypersons, and it helps to generalize the findings as the larger number of testers (laypersons) are easily available. This evaluation approach was applied by Mohseni et al. for evaluating the explanations on image and text data [48]. The authors distinguish between two types of human-subject involvement for evaluating the explanations, i.e., *feedback setting* and *feed-forward setting*. Participants submit feedback on actual explanations in a feedback context, and experimenters use this input to measure the quality of the explanations. In the feed-forward setting, however, no explanations are available. Instead, people generate examples of reasonable explanations that serve as a benchmark for explanations generated by algorithms.

Function-based Evaluation: This approach does not require humans in the loop (layperson or domain expert). It works appropriately when human-based or application-based evaluations have already been performed.

3.5. Characteristics of XML for Healthcare

The goal is to explain the decisions of the ML methods applied for the detection and prediction of diseases, and to achieve this goal research community relies on the explanation method. An XML technique usually explains in a human-understandable way how the feature of data relates to prediction results, i.e., what features of X-ray images a model learns to detect the fractures. Robnik et al. [49] listed some properties of good quality XML method. These properties are mostly required for explanations of any black-box model, however, we are describing these in terms of the healthcare domain. The only limitation is that there is no definite method to calculate these properties. Figure 4 depicts these properties.

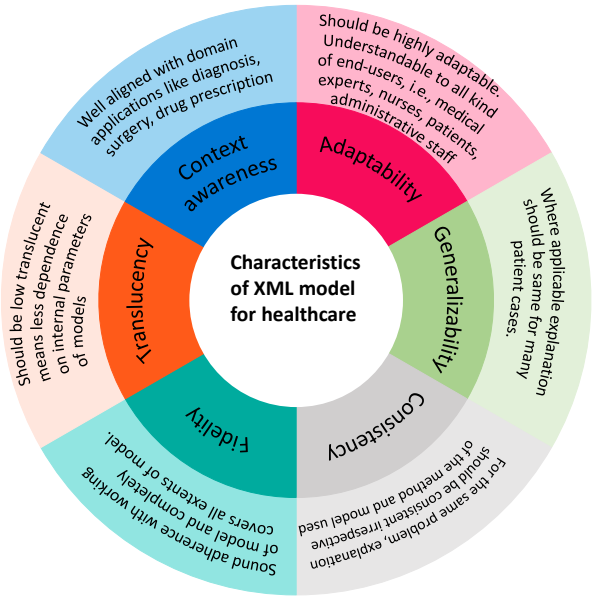

Figure 4: Illustration of characteristics of XML model for healthcare applications.

- *Domain adaptable outputs:* It is how an explainable model represents its explanation according to the application domain and end-users. The explanation could be an if-then

scenario, decision trees, or formulated mathematically or in a natural text language. For medical end-users (i.e., clinicians, radiologists, pathologists, neurologists, and patients), it is more likely that they do not have enough knowledge of understanding complex mathematical explanations. So, for them, rule-based, textual, or visualization-based explanations are required.

- *Translucency of XML:* It represents how much the explaining method depends on the internal architecture of the ML model, i.e., learnable parameters. The more the dependency is, the more translucent the explanation will be. High translucency allows the explanation method to gather information from more internal parameters of a model. Lower translucency has the advantage of more compact explanation results. To get a generalized model for clinical use low translucency is desirable. On the other hand, for a patient-specific application, the explanation should be more detailed and accurate which requires high translucency.

- *Adoptability:* It illustrates the variety of ML methods for which an explanation technique can be applied. The techniques with low translucency are applicable to a wide range of ML methods. Explanation methods of complex deep neural networks (DNN) have high translucency and thus can not be applicable to other models.

The above-mentioned characteristics could be used to select, design, and compare the architectures of XML methods for healthcare applications. A good explanation should be accurate, especially in the case of disease prediction. A low value of accuracy is acceptable if the performance of the ML model is also low. Fidelity is a property of explanation that shows the precision of approximating the decision of the ML method.

3.6. Explaining Techniques

With advancement of techniques, ML models such as DNNs, convolutional neural networks (CNNs), and recurrent neural networks (RNNs) are widely employed for healthcare applications such as epilepsy seizure prediction [50], segmentation of brain tumor [51], Alzheimer detection [52], genomics [53], and medical prescriptions [54]. These techniques are precise in terms of performance, but their decisions are difficult to explain due to multiple reasons, including the complexity of the model architecture. Based on the level of transparency, these models are divided into three following categories:

3.6.1. White-Box Models

These models have a clear inner logic, workings process, and programming steps. These models are inherently interpretable with a high level of transparency. Decision trees (DTs) are the most common example of White-Box models, while other examples are linear regression models, Bayesian Networks, and Fuzzy Cognitive Maps. In this paper, these models are discussed under the heading of intrinsic explanation methods in the section 4.4.1.

3.6.2. Black-Box Models

Black-Box models are typically more accurate than White-Box models. However, their inner working is not easily interpretable. Deep or shallow neural networks are the most common examples of Black-Box models. Saliency maps and feature attribution methods are commonly used for explaining the decisions of Black-Box models.

Saliency methods produce the explanation by presenting important feature maps of each data sample. Gradient-based saliency methods reveal how the output of the model changes with a small change in input. These methods are computationally efficient because of a single pass of input (forward and backward) through the network. The simplest way is to take the gradient of the input sample with respect to the output of model and visualize these gradients as heatmaps. Several techniques have been proposed to improve the visualization quality of these heatmaps, i.e., SmoothGrad [55], class activation maps (CAM) [56], and gradient weighted class activation mapping (GradCAM) [57]. The signal method is used to highlight the patterns of data that activate the neurons of higher layers. It can be done by back-propagating a signal from the last layer of the network to the input layer. DeConvNet [35], Guided BackProp [58], and PatternNet [59] are commonly applied signal-based saliency techniques. The feature attribution method decomposes each value produced from each neuron of the output layer according to the contributions made by the individual dimensions of an input sample.

Deep Taylor decomposition [60] and integrated gradients (IG) [61] are two other popular attribution methods. Deep Taylor decomposition method estimates the importance of single pixels in image classification tasks using the heatmaps. This method does not require hyper-parameters tuning and can be applied directly to existing neural networks without retraining. This method can help explain the decisions of algorithms in which the availability of enough amount of medical data is not possible due to surgical limitations. However, we can implement this method only on the image data. The IG method is applicable for images and text data. It requires no modification to the original network and just needs a few calls to the standard gradient operator.

A local interpretable model-agnostic explanation (LIME) technique was proposed by Marco et al. [62] to address the issue of explaining the results of black-box models. LIME produces an explanation list that shows the contribution of individual features to the prediction. This local explanation allows the end-user to determine which feature is important for the precise prediction and how the perturbation in feature affects the prediction results. Samek et al. compared the explanation quality of two methods: sensitivity analysis (SA) and layer-wise relevance propagation (LRP). These methods generate values for each feature of the input sample according to the contribution of features in predicting the output. They showed that the heatmaps produced by SA are much noisier compared to the heatmaps generated using LRP [63]. Samek et al. provide a survey with theoretical background for the post hoc methods for explaining DL models [64]. The authors also provided insights on implementation best practices, i.e., how we can include ex-

planation methods into the standard ML models.

3.6.3. Grey-Box Model

The Grey-Box model is a hybrid of the Black-Box and White-Box models, which aims to simultaneously provide accuracy and interpretability [65]. The idea behind the development of a Grey-Box model is to provide an ensemble of Black-Box and White-Box models for acquiring the benefits of both and building a more efficient global composite model. In general, we can consider any ensemble of ML models containing both Black and White-Box models, like neural networks and linear regression, as a Grey-Box. We don't particularly focus on the Grey-Box model as our paper's focus is on the interpretability of Black-Box models along with some discussion on the interpretability of White-Box models.

4. Safe, Robust, and Trustworthy ML for Healthcare

The lack of transparency of ML techniques, particularly DL, is yet another challenge that hinders the practical deployment of these methods in critical applications like healthcare. It is crucial for a typical ML/DL empowered healthcare system to be fully trusted by the clinicians and patients to realize the full potential of such systems. Moreover, unlike other domains, healthcare has unique challenges, e.g., legal, regulatory, and ethical challenges that need to be considered while integrating ML/DL based algorithms into actual clinical settings while ensuring that the deployed systems are safe, robust, and free from algorithmic bias.

We have discussed such challenges in detail in an earlier section, in this section, we will describe the notion of safe, robust, and trustworthy ML for healthcare.

4.1. Principles of Trustworthy AI for Healthcare

To sustain the trustworthiness in AI literature refers to two sets of popular principles that have been outlined by the Organisation for Economic Co-operation and Development (OECD) [66] and European Commission's AI High-Level Expert Group (HLEG) [67]. The OECD defines the following five complementary principles for implementing trustworthy AI.

1. Inclusive growth, sustainable development, and well-being
2. Human-centred values and fairness
3. Transparency and explainability
4. Robustness, security and safety
5. Accountability

These principles in the OECD framework argue for a human-centered approach to building trustworthy AI systems for healthcare that respect human dignity, values, autonomy, fairness, and explainability. On a similar note, the AI HLEG defines the following guidelines to develop trustworthy AI systems.

1. Human agency and oversight
2. Technical robustness and safety
3. Privacy and data governance
4. Transparency
5. Diversity, non-discrimination, and fairness
6. Environmental and societal well-being
7. Accountability

It is worth noting that the guidelines in both aforementioned frameworks, mainly focus on the AI aspects of robustness, safety, security, explainability, and fairness and are therefore the key requirements for building trustworthy AI systems. Moreover, these principles are human-focused and value-based, which respect ethical values along with focusing on the legal and regulatory considerations.

4.2. Secure, Safe, and Robust ML for Healthcare

The literature suggests that ML systems are not safe, secure, and robust. Such vulnerabilities can be exploited by adversaries for misleading the AI-empowered system to get desired outcomes. In the literature, different attacks have been proposed ranging from privacy attacks to targeted adversarial attacks. In this section, we will focus on the implications of security and robustness issues while building trustworthy AI systems and refer the interested readers to recent detailed work on the security and robustness of ML/DL models for healthcare applications [32]. An abstraction of safe, robust, and trustworthy ML outlining challenges like privacy and adversarial attacks in ML/DL pipeline for healthcare applications is shown in Figure 5. From the figure, it is evident that the whole ML pipeline suffers from different vulnerabilities that can be exploited by malicious actors to get the intended outcomes. For example, an adversary can realize an adversarial attack on the underlying Ml/DL classifier to increase the misclassification error (untargated attack) and can influence the classifier to classify a specific input (containing adversarial noise) to an intended class label (targeted attack). In addition, an adversary can extract the privacy-related information from the deployed model by exploiting the query-response pair (such an attack is known as a model extraction attack). Moreover, from Figure 5, we can see that ML/DL models lack explainability, interpretability, and robustness and are not privacy-aware. Therefore, trustworthy ML can only be possible by addressing challenges related to privacy, fairness, explainability, security, and robustness.

The safe and robust ML is a broad term and we define the robustness of the ML/DL models along three dimensions, i.e., robustness to security threats, robustness to distribution shifts, and data imperfections. We further note that security threats can be of many kinds, e.g., evasion attacks, adversarial attacks, privacy breaching attacks, etc.

4.2.1. Robustness to Security Attacks

Adversarially Robust ML: In recent years, adversarial ML attacks have been shown to be a real threat to the clinical deployment of ML/DL models. For instance, Mirsky et al. [68]

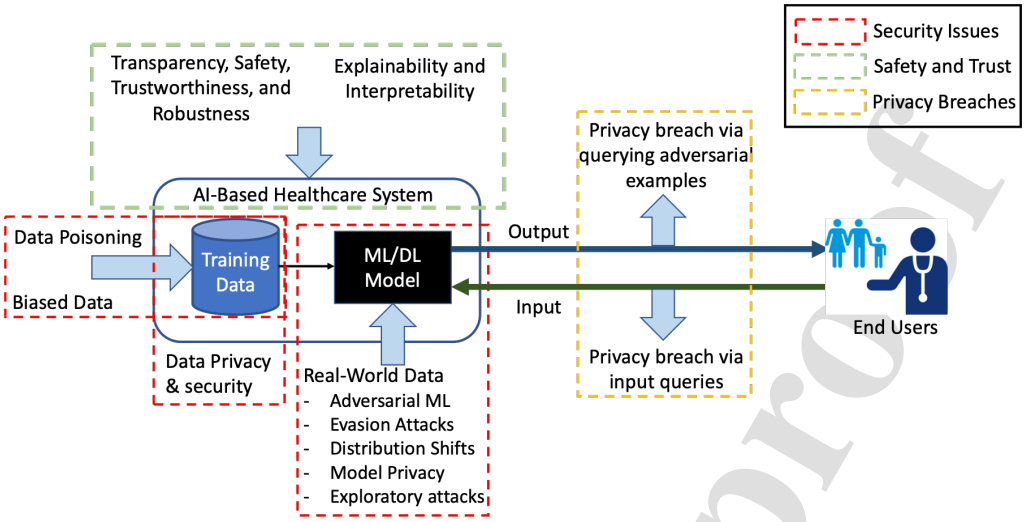

Figure 5: An abstraction of safe, robust, and trustworthy ML for healthcare applications.

demonstrated the real implications of adversarial ML by realizing an adversarial attack in an active hospital network. Specifically, they showed that CT scans generated by a DL-based image reconstruction model can be manipulated to add or remove medical evidence (e.g., removing lung cancer from CT scans of patients having it and injecting lung cancer into normal CT scans). Furthermore, they also showed that three expert radiologists were susceptible to their attack, thus highlighting the threat of adversarial ML in an actual clinical environment. Similarly, in the literature, the threat of adversarial ML has been successfully highlighted for different medical applications. In [31], the authors demonstrated the success of white-box (utilizing full knowledge) and black-box (no knowledge) adversarial ML attacks on three medical image classification tasks, i.e., diabetic retinopathy classification, skin cancer detection, and phenomena detection in chest X-ray scans. Similarly, Paschali et al. [69] evaluated the robustness of three different state-of-the-art models each for medical image classification and segmentation for the tasks of skin cancer detection and semantic segmentation of brain MRI. Han et al. [70] demonstrated the threat of adversarial ML attacks for ECG classification and highlighted that 1000 different adversarial examples can be created from the original ECG signal.

In [71], the authors argued that adversarial attacks in medical images are due to the noise inherent in the technology of their formation, e.g., CT and MRI scanners. Therefore, robustness to adversarial attacks can be a road map toward developing safe and trustworthy ML-based healthcare applications. Adversarial robustness can be defined as the survivability of ML-based systems against adversarial attacks. In this line, three types of adversarial defense methods have been proposed in the literature, i.e., modifying data, modifying model, and adding an auxiliary model. Taxonomy and detailed description of such methods can be found in [72].

Privacy Preserving ML: Preserving the privacy of patients is one of the key challenges in data-driven healthcare and is a matter of high concern in building trust in AI-based systems. Privacy preservation indicates that the ML model should not reveal any confidential information about the data owners (i.e., from whom data has been generated and collected) either during training or at inference time (c.f. Figure 5). On the other side, the users (i.e., patients and clinicians) expect that the AI system is safe and respects their privacy. Privacy attacks on data integrity can be of two types: learning about confidential information, and malicious data use [32]. Similarly, privacy information can be unveiled by querying the deployed ML model (i.e., at inference time). Therefore, the development of appropriate defense strategies to withstand privacy attacks is crucial to ensure safe and trustworthy ML in healthcare applications.

Different techniques can be use for preserving privacy—e.g., using cryptographic approaches (like homomorphic encryption [73], and multi-party computation [74]); differential privacy [75]; and federated learning [76]. In addition, hybrid approaches can also be developed, for instance, the use of differential privacy in federated learning settings is proposed in [77]. Federated learning works by training local models at each client's side and then sharing the learned parameters with the server. Then the server performs global averaging (after receiving parameter updates from each client) and shares the updated parameters with all participating clients in the network (this process is repeated until desired criteria are achieved). In this way, the privacy of the data is preserved as it is not shared with either the server or any client. Differential privacy ensures privacy by introducing random noise in the training data, thus making it hard to infer privacy-related information from the trained models. In homomorphic encryption, the data is first encrypted using some cryptographic technique and then all computations required for model training are performed on the encrypted data [78]. Multi-party computation enables the pri-

vacy of the data by allowing multiple entities to send secret inputs that are then used for performing all computations.

4.2.2. Robustness to Distributional Shifts and Data Imperfections

Data distribution shifts (which refers to the divergence of training and testing data) are yet another major challenge that hinders the practical deployment of ML/DL models in realistic clinical settings [79]. As it is highly expected that the distribution of real-world data encountered by the deployed model is different from the one it was trained which is usually trained in controlled settings with rather good data to achieve better performance. This issue results in the reduced performance of the developed ML system in an actual clinical environment and on the other hand, it also fails to gain the trust of end-users (i.e., clinicians and patients), due to increased false positives and false negatives rate that can lead to life-threatening consequences as well. In addition, the real-world data contains imperfections (e.g., missing observations or variables) and is imbalanced (uneven distribution of samples across different classes). These data imperfections will eventually result in biased training of the ML models and will increase the false positives and negatives. Therefore, to build the trust of end-users in ML-based systems, the development of generalized approaches that can mitigate these issues is required. As the life-critical nature of healthcare applications demands that the developed ML systems should be safe and robust and should remain safe and robust over time. Moreover, the literature suggests that the difference in data distributions can be leveraged to craft adversarial examples [80]. Also, it has been shown that adversarial robustness is closely related to robustness to certain kinds of distributional shifts. Therefore, the literature recommends that future adversarial defenses should consider evaluating the robustness of their methods to distributional shifts as well [81].

4.3. Trade-off between Accuracy, Explainability, and Robustness

One has to pay a cost for developing explainable, robust, trustworthy, and accurate ML/DL models, as shown in Figure 6. In [82], an analysis of the trade-off between the accuracy and adversarial robustness of 18 well-known ImageNet classifiers with different metrics is presented. The authors noted that a clear trade-off existed between accuracy and robustness. Similar observations were noted in [83], where the authors quantified this trade-off and argued that adversarial robustness is incompatible with standard accuracy. Tsipras et al. demonstrated that there exists provably a trade-off between adversarial robustness and the accuracy of the model even in a concrete simplistic setting [83]. The authors argue that this behavior is a reflection of the robust models that tend to learn fundamentally different feature representations than the standard (non-robust) models. Also, these differences may also provide unexpected benefits, e.g., learned representations from adversarially robust classifiers seem to be more aligned with human perception and data characteristics. As adversarial perturbations used for robust training of models contains such properties that are expected to be consistent with human perception [83].

Robust models pay the cost of accuracy and can be more explainable and interpretable as compared to complex models having high accuracy but low explainability. In practice, the higher the accuracy of the predictive model, the less explainable/interpretable it becomes. This highlights that solely getting high accuracy from an ML/DL model may get us in real trouble. A few studies have focused on addressing this trade-off [84], however, such methods are not generalizable and applicable to all domains, in particular, task-specific ML/DL applications.

4.4. Applications of XML Models for Trustworthy Healthcare

The Translucency, credibility, and explainability of ML models are requirements for the clinical application of these models. Transparency of decisions of these models can help clinicians trust and rely on ML/DL prediction algorithms. Moreover, interpretable and explainable AI models are required for answering questions about accountability and transparency of their decisions and outcomes. These questions are particularly important for domains like healthcare, where failing to provide accountable and transparent AI predictions will limit the potential transnational impact of AI and at the same time, will reduce the trust of end-users in AI-based medical interventions. The European General Data Protection Regulation (GDPR) emphasizes that explainability and accountability are necessary for the application of ML/DL models in any domain, especially, in the medical domain [9]. The explainability of black-box models can assure reliable and ethical use in the medical field. Transparency of ML models can help to eradicate myths by explaining what features a model learned for making certain predictions and can help in building the trust of end-users [14]. Relevant literature argues that explainable ML can be a potential step towards trustworthy ML by building trust of clinicians in AI-based systems [85]. We describe next the applications of XML models in the medical domain:

4.4.1. Intrinsic XML Models

As noted previously, intrinsic explainable models are those models that are understandable and interpretable due to their simple architecture. The following are applications of inherently XML models and Table 3 provides a summary of different model-intrinsic explanation methods for healthcare applications.

Support Vector Machine (SVM): The SVM algorithm finds the best hyperplane that can split data points into different classes. For the classification of data, a feature map is used to transform the inputs into a higher dimensional feature space. The trade-off between a smooth decision boundary and correct classification of the training data is made through the strictly positive regularization constant. Belle et al. investigated the explainability of SVM with linear, polynomial, and RBF kernels [99]. They showed that the explainability of an SVM depends on the values of chosen parameters i.e., the degree of the polynomial kernel, the width of the RBF kernel, and the regularization constant. According to their findings, when several combinations of parameter values yield the same cross-validation performance, combinations with a lower polynomial degree or

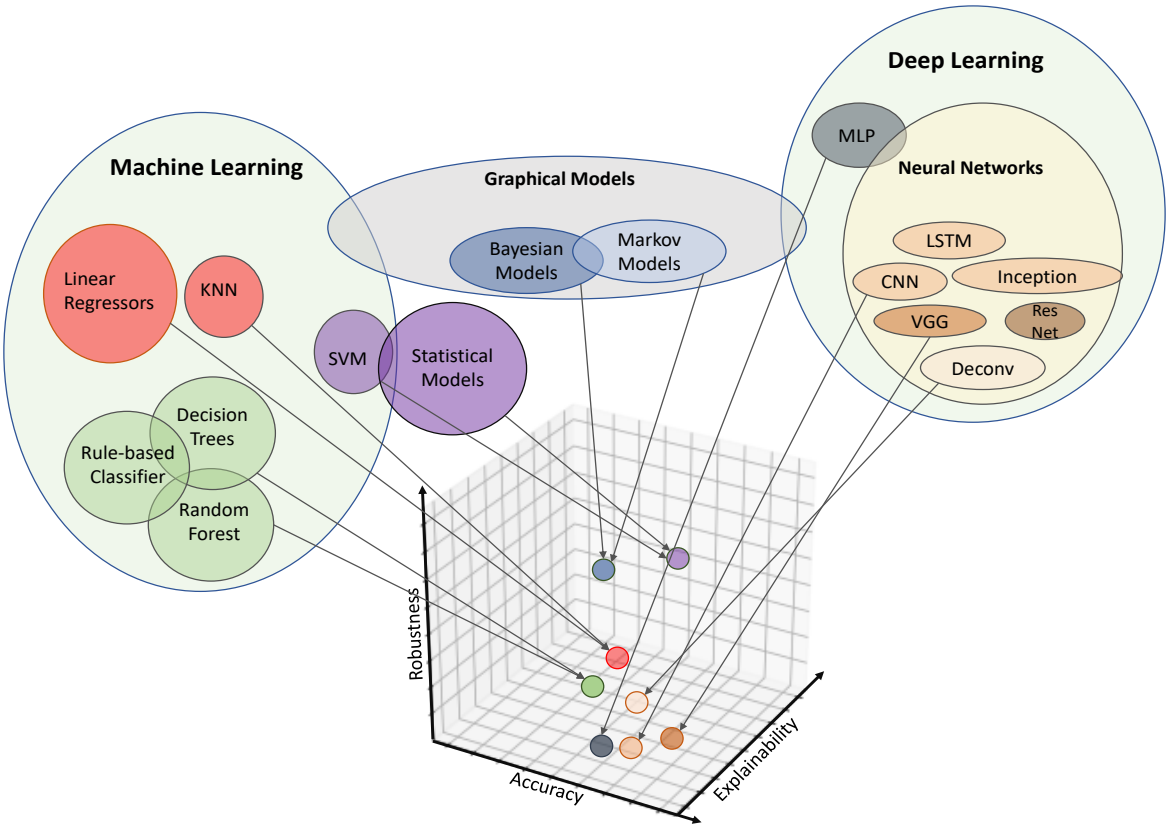

Figure 6: Illustration of the typical trade-offs between accuracy, explainability, and robustness of intelligent models. *These trends are indicative and can vary for different tasks and data.*

a larger kernel width have a higher chance of being explainable [99]. Eslami et al. conducted a study for the detection of autism spectrum disorder (ASD) from functional magnetic resonance imaging (fMRI) and used a hybrid of DL and SVM to perform explainable classification. The SVM was used as a classifier on the features of a DL model and the visualization of the decision boundary explained the model [100]. However, for large and complex medical data sets SVM algorithm is not suitable because SVM under-performs where the number of features for each data point exceeds the number of training data samples. Usually, the medical data has a low signal-to-noise ratio and SVM does not perform very well when the data set has more noise [101].

Decision Tree (DT): These are the self-explanatory surrogate models that use the if-then rule for the explanation of decisions. However, for complex high-dimensional medical data in which decision variables are non-linearly related to each other DTs are not feasible for producing human-understandable explanations. For instance, Dlaeen et al. [87] implemented DTs to predict Alzheimer's disease in seventeen patients. They used gender, age, genetic causes, brain injury, and vascular disease as data attributes and measured information gain of attributes for the selection of DT nodes. However, the quality of the explanation provided by their proposed DT-based approach has not been evaluated.

Note that DTs are highly vulnerable to minute changes in input, which considerably affects the performance of these models, e.g., they show considerable change in output with a small perturbation in the input. To overcome this performance drop, ensemble DTs are constructed by averaging large numbers of DTs. In this regard, Gibbons et al. [86] used a hybrid approach for leveraging both the benefit of individual DT and the efficiency of ensemble DTs. They proposed a computerized adaptive diagnostic system for the diagnosis of major depressive disorder using random forest and DTs. For performance evaluation, they used data collected from 656 patients and achieved a sensitivity of 95% and specificity of 87%. Similarly, Suresh et al. [88] proposed the use of a radial basis function (RBF) network and DTs for the detection of lesions in mammograms. In their proposed approach, the DT model was used to learn suitable attributes of data in a top-down search manner, specifically, they selected the best attributes by constructing and evaluating different structures of DTs. They also compared their algo-

Table 3: Summary of model-intrinsic explanation methods for healthcare applications.

Explaining Method	Year	Reference	Description	Application	Modality
DT	2013	Gibbons et al. [86]	The self-explanatory surrogate the model uses if-then logic for decision	MDD detection	Psychiatric and non psychiatric attributes
	2014	Dlaeen et al. [87]		Alzheimer's disease detection	Gender, Age, Genetic causes, Brain injury, Vascular disease
	2020	Suresh et al. [88]		Breast cancer detection	Mammographic images
Rule-Lists	2016	Khare et al. [89]	Textual format explanation using if-then logic for decision making	Cardiovascular disease detection	Various attributes of patients
	2019	Agrawal et al. [90]		Question classification in health care	Coarse and fine-grained classes from cloud questionnaire
RF	2017	Wang et al. [91]	An ensemble of large numbers of DTs, used mainly for regression or classification	Epilepsy detection	EEG signals
	2019	Byeon et al. [92]		Alzheimer's patients depression detection	Social demographic factors, Health status, Behaviors, Living style, Economic activity
	2019	Kaur et al. [43]		Healthcare monitoring system	Breast cancer, Diabetes, Heart disease, Spect-heart, Thyroid, Surgery, Dermatology, Liver disorder
	2020	Simsekler et al. [93]		Evaluation of patient safety culture	Continuous and Categorical variables for patient safety
	2020	Iwendi et al. [94]		Covid death and recovery rate detection	Categorical variables in dataset such as fatigue ,fever ,cough.
GAM	2015	Caruana et al. [95]	The output is modeled as the weighted sum of random nonlinear functions of data features	Pneumonia risk prediction	Various attributes of patients
	2019	Sagaon et al. [96]		Effect of age and diagnosis -specific cohort of HIV patients on psychosocial activities and behavioral activities	Various attributes related to psychosocial and behavioral outcomes
	2020	Dastoorpoor et al. [97]		Effect of air pollution on pregnancy	Various air pollutants data
	2020	Yang et al. [98]		Study air pollution effect on TB cases	Pulmonary TB and air pollutants data

rithm with k-nearest neighbors (K-NN), support vector machine (SVM), and naive Bayes classifier and concluded that DTs outperformed all these classifiers. As discussed above, generalization of algorithms is required for clinical deployment to avoid the phenomenon of distribution shifts. However, the literature shows that DTs cannot generalize to variations not seen in the training set [102]. Therefore, DTs-based systems are not appropriate for healthcare applications.

Rule-Lists: Rule-based XML models produce explanations using if-then rules or other complex rules. These are different from the DTs as they generate the explanations in the textual format. Other differences between these models as compared to DTs include rules ordering (rules are ordered according to their properties) and the generation of mutually exclusive rules (different rules that are generated by the same attributes). Khare et al. [89] proposed an association rule technique using 23 attributes of cardiovascular data for detecting heart diseases. Their method works by generating such rules that map the attributes to classes for the identification most appropriate features influencing the model prediction (i.e., provoking a specific disease). They used confidence, lift, and support as parameters for generating rules. Accuracy was used for the validation of generated rules.

With the emerging use of natural language processing (NLP) and ML, automatic answering to healthcare-related questions is a conspicuous technique. Classification of questions is required for the generation of answers. Agrawal et al. [90] implemented a rule-based algorithm for the question classification system (QCS). They extracted rules after the preprocessing of 427 health-based questions to classify them into 9 question types (i.e., classes). The extracted rules were validated in terms of accuracy and their proposed rule-based method achieved an accuracy of 80.7% by correctly classifying 345 questions.

Random Forest (RF): RF is an ensemble of large numbers of DTs, that are used for regression or classification problems. Each DT in RF performs the classification and the final prediction of RF is measured based on the most occurring class. Wang et al. [91] evaluated RF, C4.5 algorithm of DT, SVM-based RF, and SVM-based DT algorithms for the detection of epileptic seizures using the Bonn university dataset. To detect the seizures, they classified EEG signals into different groups. They concluded that the RF algorithm outperformed all other classifiers with an accuracy of 98.6% for two-class, 96% for three-class, and 82.6% for five-class classification experiments.

In the literature, RF has been also used beyond general classification and regression. For instance, Simsekler et al. [93]

implemented an RF algorithm to estimate the association between the safety culture dimensions and grades of patient safety by using the HSOPSC dataset from 677 U.S. hospitals. As the safety of a patient is necessary to ensure the quality of medical facilities provided by a healthcare unit, which is a patient safety culture of a hospital or clinic. The authors considered 12 variables of safety culture and identified the importance of each variable for the safety of patients using an RF algorithm. The quality of explanations of safety variables was measured in terms of mean absolute percentage error (MAPE), mean absolute error (MAE), and mean square error (MSE). In a recent study, Iwendi et al. [94] used an RF model with the AdaBoost algorithm to analyze the severity of COVID-19, death, or recovery rate for a patient. RF has been used for the prediction of depression in Alzheimer's patients [92], healthcare monitoring systems [43], and prediction of medical expenditures [103].

General Additive Model (GAM): ML regression models produce predictions by adding weighted features. GAM works by modeling the output as the weighted sum of random nonlinear functions of data features and to approximate these non-linear functions, a combination of spline functions is used. Chang et al. [104] performed a comparative analysis of different GAM algorithms quantitatively and qualitatively. They concluded that GAMs that only use a few variables to make predictions can miss patterns in the data and can be unfair to rarely occurring data samples. Therefore, GAMs cannot be used for such medical applications that have high feature sparsity. In the literature, GAM is extensively used in different health-related applications such as environmental research [97], pneumonia risk prediction [95], research on the distribution of species [105], and the effect of age and a diagnosis-specific cohort of HIV patients on psychosocial and behavioral activities [96]. Yang et al. [98] studied how tuberculosis (TB) cases changed with air pollution in the Wulumuqi. They obtained the air quality and TB patients data of slightly more than two years duration. They founded that $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, and O_3 were the dominant pollutants in the air data. They used a GAM model to study the relation between the aforementioned pollutants and the number of TB cases. However, their analysis was based on an assumption that the number of patients followed the Poisson distribution. Moreover, to encounter the linear and non-linear features of data, they used the natural cubic spline. With statistical validation of results, they concluded that with the 1 mg/m^3 increase in $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, and O_3 particles number of TB patients increased by 0.09%, 0.08%, 0.58%, 0.42%, 6.9%, and 0.57%, respectively.

4.4.2. Model-Agnostic Explainability

As the name specifies, model-agnostic explanations are flexible in terms of applications of models and representation. Table 4 summarizes the model-agnostic explanation methods for healthcare applications, and below we provide a brief description of these methods when applied to explaining healthcare decisions.

Partial Dependence Plot (PDP): These plots provide visual explanations by showing the partial effects the input features have on the prediction of a black-box model. These plots also help in visualizing the type of relation (linear or non-linear) between the label and data features. Yang et al. [106] predicted the mortality of COVID-19 patients using age, time to the hospital, gender, and any chronic disease as attributes. They plotted partial dependencies to check the effect of each attribute on the prediction of mortality. They showed that the age of a patient is the most important factor and the second important factor is how much time the patient has spent in the hospital.

Class Activation Maps (CAM): These models are used to explain the decisions of CNNs by highlighting the class-relevant areas of images. However, CAMs are only applicable for specific CNN architecture, i.e., CNN must have a dense and global averaging pooling layer after the last convolutional layer. Vikash Gupta et al. [107] detected acute proximal femoral fractures in elderly people using radiographic data. They detected the fractures using VGG16 and used CAM to localize the fractures. Kumar et al. [108] proposed a novel model named mosquito-net for the classification of malaria cells and explained the decisions of the model using CAM and GradCAM (a variant of CAM). Similarly, Irvin et al. [109] used the GradCAM to provide the visual explanation of active pleural effusion areas of chest radiograph which were indicated by the CNN model. Sebastian et al. [110] used CAM to evaluate the errors of their CNN model proposed for the multiclass labeling of ECG signals. Pereira et al. [111] explained the brain tumor grading decision of their proposed CNN classifier using CAM. Izadyyazdanabadi et al. [112] integrated multiscale activation maps (MLCAM) with the CNN model to locate the attributes of glioma tumors. In the literature, it has been shown that CAM suffers from the gradient saturation problem due to which it fails to localize relevant regions to overcome this issue, gradient-free CAM has been proposed [113]. Also, CAM is noisy and can induce a loss of spatial information in making explanations of model predictions.

Layer-wise Relevance Propagation (LRP): This technique works by backpropagating the output decision to the input layer to estimate the relevance of each attribute. Yang et al. [114] proposed the use of LRP to select the features with high relevance to model-making predictions regarding the decision of therapy of patients. They also evaluated the quality of explanation from the expert clinicians and found that the features, which are highlighted by the LRP have relevance and largely agree with clinical knowledge and guidelines. Chlebus et al. [115] implemented an LRP algorithm for explaining the decisions of semantic networks used for segmenting liver tumors. They highlighted MRI segments that were most relevant for the classification of tumors. Böhle et al. [116] implemented LRP and guided backpropagation (GB) for explaining the decisions of CNN that they used for the classification of Alzheimer's disease (AD). They evaluated the quality of generated explanations of both techniques by measuring Atlas-based evaluation metrics. To measure the quality, they analyzed the heatmaps and the underlying CNN model. Specifically, they measured the importance of different brain areas by calculating the sum of AD importance per area, size-normalized AD importance

metric, and gain—ratio of values with respect to the average healthy controls. They concluded that LRP generates more relevant explanations by describing why any individual patient has the disease. Jo et al. [117] used LRP to highlight the areas of positron emission tomography (PET) that highly contribute to the classification of Alzheimer's disease using 3D-CNN. Note that heatmap-based explanation methods require a ground truth for the validation and in cases where the ground truth is not available, explanations are qualitatively evaluated using visual assessment, which itself is subjective. Also,the heatmap-based explanations are generally algorithm-dependent.

Local Interpretable Model-Agnostic Explanations (LIME): This technique generates explanations by apportioning an image data sample into superpixels (groups of pixels having similar features) that provided contextual details about the local part of an image. Samples of perturbated images are then generated by tweaking the values of randomly selected superpixels. The algorithm provides information about how perturbation in features affects the prediction. The significance of every superpixel for the prediction is measured as weighted values, i.e., positive values show a high impact on a correct prediction, and negative values show less or no impact on prediction. Sousa et al. [118] used LIME to explain the CNN and VGG16 models that were trained for the task of metastases detection from the histology whole slide images (WSI) patches. They evaluated the explanations by cross-checking the highlighted areas with medical annotations on the same images and generally found both in agreement.

Zafar et al. [123] pointed out the problem of instability of generated explanations due to the addition of perturbation and random feature selection in the medical computer-aided diagnosis (CAD) systems. Furthermore, they proposed the use of hierarchical clustering (HC) and KNN to group the data and for the selection of relevant feature clusters. The proposed algorithm was named deterministic LIME (DLIME) and was used to explain the decisions of three medical image classification models, i.e., breast cancer, liver disease, and hepatitis detection. The performance evaluation showed that the proposed DLIME performs better than the standard LIME. Kitamura et al. [119] proposed to use LIME for the explanation of CNN-based diabetic nephropathy (DN) detection model that was trained using immunofluorescent images. LIME successfully highlighted learned patterns (by CNN model) of peripheral lesion of DN glomeruli for DN detection.

Deep Learning Important FeaTures (DeepLIFT): DeepLIFT provides the explanations of black box models by identification of the saliency of input data. The algorithm measures the saliency according to how sensitive the prediction of the algorithm is to input features in comparison to their reference value. Reference value problem specific which is selected based on the problem at hand. Sharma et al. [124] proposed DeepLIFT for Genome-Wide Association Studies (GWAS), which focused on studying genetic variants caused by common diseases. They proposed the use of DeepLIFT to explain interactions that a normal GWAS would not identify and showed that diabetes genetic risk factors are identifiable using DL techniques.

SHapley Additive exPlanations (SHAP): SHAP explains the prediction of a data sample by calculating the contribution of each feature to the prediction of the algorithm. The SHAP uses coalitional game theory to calculate Shapley values. Shapley values show the distribution of prediction among features. Tseng et al. [120] studied the effect of intraoperative variables on the cardiac surgery-associated acute kidney injury. They used various ML algorithms logistic regression (LR), SVM, RF, extreme gradient boosting (XGboost), and RF + XGboost to solve the problem. Using SHAP values, they concluded that intraoperative urine output, IV fluid infusion, blood product transfusion, and dynamic changes of hemodynamic features are significant causes of injury. They also stated that these factors were not revealed using traditional techniques. Daping Yu et al. [125] detected lung cancer from the copy number variation (CNV) derived cell-free DNA (cfDNA) using an extreme gradient boosting (XGBoost) algorithm. They showed the contribution of each plasma feature using SHAP. They concluded that a high concentration of cfDNA in plasma and CNV in chromosomes affected the pathogenesis of cancer cases.

Sensitivity Analysis (SA): SA is an effective and powerful algorithm to understand the stability of black box models by examining the effect of perturbations in input on the prediction of the model. If the model outcome has changed notably with perturbations, it shows us that the feature has a high contribution to the prediction. Couteaux et al. [126] proposed an explanation method based on the DeepDreams concept for explaining the classification of tumors using data of liver computed tomography (CT). Their proposed method used the SA of each feature by maximizing the neuron activation using gradient ascent. They showed that the network is sensitive to intensity and sphericity in coherence with domain information.

Guided Back Propagation (GBP): GBP is also known as guided saliency. GBP uses the concept of both vanilla backpropagation and DeconvNets to explain the decisions of DL models. The only difference is that the positive error signals are backpropagated and negative gradients are set to zero. Moreover, like vanilla backpropagation algorithm, it limits itself to positive inputs. Pianpanit et al. [121] proposed a 3D-CNN architecture for Parkinson's disease (PD), and to explain the detection they implemented and compare six different explainable methods, i.e., saliency map, GBP, Grad-CAM, Guided Grad-CAM, DeepLIFT, and SHAP, and showed that GBP among all the methods produced the best explanations. DeepLIFT and SHAP produced the second best explanations by distinguishing between features of healthy and PD patients. These three methods performed better in PD diagnosis by correctly analyzing the absorption of ^{123}I-Ioflupane in the dopamine depletion region of single-photon emission computed tomography (SPECT) of PD patients. They evaluated the quality of produced explanations using the Dice coefficient measure.

Integrated Gradient (IG): IG is a DL technique that uses the input feature significance to visualize the model prediction. IG works by calculating the gradient of model output with its input attributes. IG does not require any changes to the primordial

Table 4: Summary of model-agnostic explanation methods for healthcare applications. **Legends: N/M** = Not mentioned.

Explaining Method	Year	Reference	Description	Black Box Model	Application	Modality
PDP	2020	Yang et al. [106]	Highlight the partial effects the input features have on the prediction of a black-box model	XG-Boost	Mortality rate in COVID-19	Age, Gender, Time to hospital
CAM	2020	Vikash Gupta et al. [107]	Highlight the class relevant areas of input data.	VGG16	Fracture detection	X-Rays
	2020	Sebastian et al. [110]		CNN	ECG classification	ECG signals
	2018	Pereira et al. [111]		CNN	Grading of brain tumor	MRI
GradCAM	2019	Irvin et al. [109]	Generates weighted gradient CAM by computing gradients of output as it goes towards last layer.	CNN	Detection of different diseases	Chest X-Rays
	2020	Aayush Kumar et al. [108]		Mosquito-net	Malaria detection	Blood samples
MLCAM	2018	Izadyyazdanabadi et al.[112]	Generates the maps of discriminating features of data.	CNN	Brain tumor detection	MRI
LRP	2018	Yang et al. [114]	Back-propagates the output decision to the input layer to estimate the relevance of each attribute.	LSTM	Cancer therapy decision prediction	N/M
	2019	Chlebus et al. [115]		Semantic segmentation network	Liver tumor classification	MRI
	2019	Böhle et al. [116]		CNN	Alzheimer's disease classification	MRI
	2020	Taeho Jo et al. [117]		3D-CNN	Alzheimer's disease classification	PET
LIME	2019	Sousa et al. [118]	Decompose the data based on similar features and tweak randomly selected features to measure output dependence.	CNN VGG16	Detection of metastases	WSI patches
	2020	Kitamura et al. [119]		CNN	detection of diabetic nephropathy	immunofluorescent images
DeepLIFT	2020	Yang et al. [106]	Uses a reference value and measures the reference values of all neurons using a forward and backward pass	–	Genetic variants caused by diseases	Single-Nucleotide Polymorphisms
SHAP	2020	Tseng et al. [120]	Uses coalitional game theory to calculate Shapley values that show the distribution of prediction among features	LR, SVM, RF, XGboost, RF + XGboost	Detection of cardiac surgery-associated acute kidney injury.	Various disease related features
GBP	2019	Theerasarn et al. [121]	Backpropagates the positive error signals by setting negative gradients to zero and limits itself to positive inputs.	3D-CNN	Detection of Parkinson's disease.	SPECT
AM	2019	Borjali et.al [122]	Generate explanations by maximize the activation of neurons tweaking the input.	CNN	Detection of hip implant misplacement.	X-rays

deep neural network. IG can be utilized for any kind of model and data type, i.e., image. This algorithm works on two axioms sensitivity and implementation variance. Simple drug development classification of toxic and non-toxic drugs is not enough and to solve the problem of toxic drugs, a chemist needs the structural element which is causing the problem. Preuer et al. [127] demonstrated that IG can identify these elements from the classified drug using CNN.

Activation Maximization (AM): AM aims to maximize the activation of neurons. In the AM model weights and output remain the same while by changing the input we maximize the activation of the neuron. Borjali et al. [122] trained the CNN model for orthopedic application in observing hip implant misplacement using the X-rays dataset. The explainability of this CNN model at a lower level is done using AM, which was used to visualize the outputs of the model.

Deep Hierarchical Semantic Convolutional Neural Network (HSCNN): Shen et al. [128] proposed an interpretable DL model named hierarchical semantic convolutional neural network (HSCNN) to detect the malignant pulmonary nodule that appeared on computed tomography (CT) scan. Their proposed model provided two types of outputs, one was low-level semantic features which were used by radiologists, and also explained how the model detected the malignant nodules. The second level of output was the malignancy prediction score. They also compared the performance of their proposed model with CNN and concluded that the HSCNN outperformed the CNN with a high prediction score and explainable results.

Patient2Vec (P2V): The extensive use of electronic health records (EHR) in the clinical system provides a large amount of data for healthcare. Jinghe et al. [129] presented P2V to explain the unexplored EHR dataset for predicting disease correlation, health outcome, and health history of new patients. P2V is a recurrent convolutional neural network used to explain the longitudinal EHR dataset customized for each patient. The implementation of P2V improves the predictive model working efficiency and also increases the explainability of these models.

The proposed model was used to explain the importance of each diagnostic product, medication, and treatment procedure.

Evidence Activation Mapping (EMANet): Lia et al. [130] proposed a CNN model for glaucoma diagnosis. The proposed CNN architecture is not only able to detect the diseases but also shows transparency by highlighting the affected area detected by the system. The system consists of CNN as the backbone for feature extraction and uses multilayers average pooling (M-LAP) to overcome the gap problem between the information interpretability and localization while evidence activation mapping is used for the verification.

5. Ethical ML for Healthcare

The integration of AI/ML into healthcare practice and clinical applications promises to provide substantial improvements to the healthcare sector. To name a few, it can improve care quality, cut the overall costs, reduce or even eliminate diagnostic errors and improve the process of predicting disease. In response, private companies are incorporating ML-based technologies into healthcare decision-making, creating tools that assist clinicians and developing algorithms designed to perform independently of them. Clinicians and researchers are prophesying that knowledge of ML for analyzing heterogeneous medical data will be a primary requirement for future physicians and that ML models might compete or even replace clinicians in fields that involve analysis of images, such as radiology and anatomical pathology [131]. However, incorporating the ML techniques into the healthcare system also raises serious ethical challenges and complex questions that need to be seriously considered to make a robust and well-balanced assessment of possible benefits and expected harms [132].

To set the scene for those who are not specialists in bioethics, this section will start by (a) providing a concise overview of bioethics as a scholarly discipline and its methodological approaches, with a focus on the so-called "principlism" and the widely known four principles, namely beneficence, non-maleficence, autonomy, and justice [133]. It is noteworthy that explicability is a newly proposed principle within the particular AI context, which has the same meaning outlined above in this paper [134]. In the remaining part of this section, we will (b) review the key works that examined the interplay of AI/ML and bioethics and (c) analyze the main bioethical issues and challenges posed by the implementation of AI/ML applications in the healthcare sector.

5.1. Historical Overview

That practicing medicine or providing healthcare should be tied to, and governed by, certain sets of moral principles and values is one of the widely agreed-upon facts throughout human history. The Hippocratic oath is one of the earliest and widely known codes of ethics for medical professionals. The oath established various principles of medical ethics and its purport continues to be the subject of modern studies, which examine its possible relevance to modern bioethical discussions [135]. World religions like Judaism, Christianity, and Islam also brought their insights to ethicize the physician's work. A good representative example here is the work of the 9^{th}-century physician Ishaq b. Ali al-Ruhawi, who lived in the golden age of the Islamic civilization and wrote one of the most popular works on medical ethics, entitled *Adab al-Tabib* (Ethics of the physician) [136], [137]. In 1803, the physician, Thomas Percival, published a report on the necessities and expectations of medical staff to assure ethical medical practice [138]. This code of medical ethics was adapted for the first time in 1847 [139] and is now broadly accepted and practiced throughout the world as an ethical code for the medical domain.

Owing to a wide range of diverse factors, not only related to the breathtaking biomedical advancements but to various intellectual and sociopolitical changes, the twentieth century, especially from the second half onwards, witnessed the history-making shift from the pre-modern "medical ethics" to the modern "biomedical ethics" or simply "bioethics". The American oncologist Van Rensselaer Potter (1911-2001) was the first to use the term "bioethics" in the title of his book *Bioethics: Bridge to the Future*, published in 1971. Potter proposed introducing a new discipline, which he named Bioethics, to address the basic problems of human flourishing by creating an interdisciplinary discourse between the two cultures of humanities and sciences [140].

One of the important milestones in modern bioethics is the so-called "Belmont Report[5]; produced in 1979 by the US National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The report charted the basic ethical principles and guidelines that should govern the conduct of biomedical and behavioral research involving human subjects. The report identified three main bioethical principles, namely respect for persons, beneficence, and justice. Parallel to these developments, the two renowned American bioethicists, Tom Beauchamp and James Childress, published the first edition of their seminal work *Principles of Biomedical Ethics*. The authors introduced four principles, namely autonomy, beneficence, nonmaleficence, and justice [133]. Their principle-based theory, which later came to be known as principlism, proved to be one of the most seminal contributions to the modern field of bioethics, as demonstrated by the number of subsequent editions and printings of their book, the eighth edition was published in 2019, and by the global discussions around this theory [141]. Besides the famous principlist approach to bioethics, there are other important approaches in modern bioethics, including virtue ethics, casuistry, narrative ethics, feminist approach, and care ethics. Each of these approaches has its own proponents and opponents who debate on the added value of each approach and its possible drawbacks [142]. Figure 7 illustrates the history of the development of bioethics over time.

Besides these foundational publications for modern bioethics, the atrocities of the two world wars and the associated ethical violations in conducting medical research

[5]The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research `https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html`

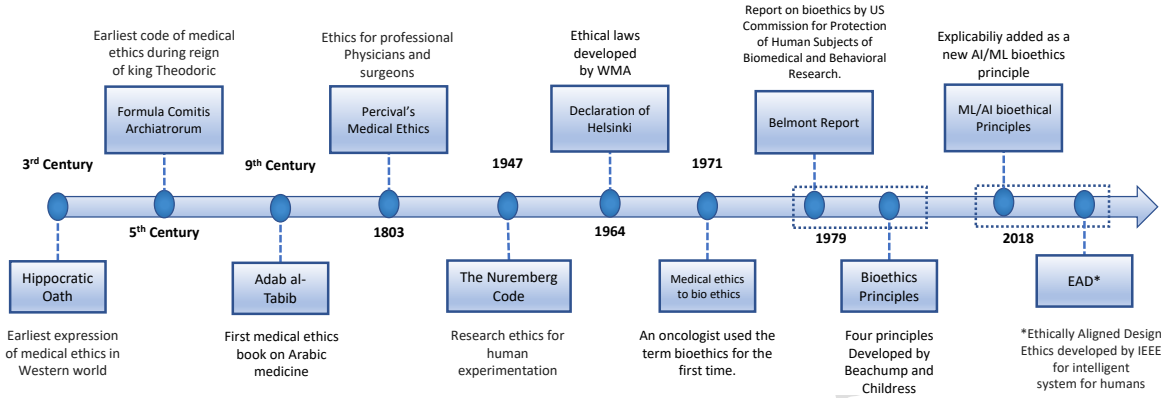

Figure 7: History of the development of bioethics over the time.

on human subjects also resulted in issuing of several codes and documents to regulate research experiments and trials on humans. The Nuremberg code, drafted in 1947, is one of the main examples in this regard. It consisted of ten points under the title of "Permissible Medical Experiments", including consent of patients, patient's right to end the experiment at any stage, high expertise of researcher, and avoiding unnecessary mental and physical suffering [143]. In 1964, the declaration of Helsinki was developed by the World Medical Association (WMA). This declaration consisted of ethical principles and regulations for the physicians. Respect for each patient, right to self-determination, a thorough evaluation of possible risks and benefits, and beneficence of society and mankind are a few of the principles stated in this declaration [144]. As they were produced at earlier dates, none of the aforementioned foundational works, codes or documents paid special attention to the ethical challenges triggered by the implantation of AI/ML technologies into healthcare. However, these works and the bioethical approaches they introduced and theorized remain essential for developing a robust analysis of related challenges and questions. Additionally, some of the recently published bioethical works examined a number of ethical questions, which are specific to the interplay of AI/ML and bioethics. These key works will be reviewed below.

5.2. Key Works

The field of healthcare is increasingly representing one of the main applied areas of AI/ML technologies. This fact is reflected in the growing number of publications in this research area. Due to space availability, we will not be able to provide a comprehensive review of all the relevant publications. Instead, we will focus on a number of the key works in this emerging field, especially those published as book-length studies or thematic issues in reputable journals. Individual journal articles or book chapters will be referred to only when they relate to the examined books and/or thematic issues.

Some of the relatively early works in this area focused more on issues related to the conventional computerization and digitalization of healthcare. However, they occasionally touched upon bioethical issues within the particular context of AI and ML. In *Ethics, computing, and medicine: Informatics and the transformation of health care*, published in 2007 [145], a group of interdisciplinary authors examined the ethical issues related to health informatics. A distinct chapter was dedicated to "Ethical and Legal Issues in Decision Support" [146]. *The Digital Doctor*, a New York Times science bestseller and published in 2015, by Robert Wachter (University of California San Francisco), also serves as a good example in this regard [147]. Similar issues were also examined in the edited volume *Smart Health: Open Problems and Future Challenges*, published in 2015 [148].

One of the main contributors to the discourse on AI-driven healthcare is the American cardiologist and professor of genomics, Eric Topol, a well-known high-tech enthusiast. Between 2012 and 2019, Topol wrote what can be called a trio on the revolutionization of medicine by using available digital, smart and AI-based technologies. In his *The Creative Destruction of Medicine: How the Digital Revolution Will Create Better Health Care*, published in 2012 [149], and *The Patient Will See You Now: The Future of Medicine Is in Your Hands*, published in 2015 [150], the focus was more on the benefits of using available digital technologies, especially those offered by smartphones. In 2019, Topol crowned this trio by publishing *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*, where AI technologies were introduced as the main driver of the promised revolutionization of medicine [151]. He also outlined his ideas in this area in an article published in 2019 in *Nature Medicine* [152]. Besides simplifying the scientific and technical information that would otherwise be unintelligible to the non-specialist reader, Topol touched upon, and sometimes seriously examined, some of the ethical questions and challenges triggered by the promised revolutionization of medicine, including those related to the privacy of people, confidentiality of information and security of data. Topol, a paid adviser to AI health companies, is also sometimes criticized for adopting a market-driven discourse that is similar to the one propagated by tech-giants like Google and Facebook [153].

In 2020, *The American Journal of Bioethics* published a thematic issue entitled "Planning for the known unknown: Machine learning for human healthcare systems" [154]. The contributions to this thematic issue, made by a number of interdisciplinary experts, provided useful frameworks that can help future researchers critically examine the ethical concerns of the AI Health Care Applications (HCA). Important ethical questions related to the concepts of explainability, auditability, and accountability were also addressed in this issue. The edited volume *Artificial Intelligence in Healthcare*, published in 2020, provided an extensive overview of the current state of the art in this field and outlined what is achievable in near future. Besides discrete references to ethics throughout the work, the last chapter was dedicated to "Ethical and legal challenges of artificial intelligence-driven healthcare" [155]. The important reference work, *The Oxford Handbook of Ethics of AI*, published in 2020 as well, included a distinct chapter on " The ethics of AI in biomedical research, patient care and public health" [156].

One of the latest relevant publications in this area is *Machine Learning and AI for Healthcare: Big Data for Improved Health Outcomes*, whose second edition was published in 2021. Besides introducing the basic terminology, concepts, and applications of AI technologies in healthcare, the book also discussed various ethical issues and a distinct chapter was dedicated to "Machine Learning and AI Ethics" [157].

5.3. Main Ethical Questions and Challenges

The integration of AI/ML technologies into healthcare promises to bring great benefits including efficiency and access to all stakeholders such as patients, physicians, and healthcare service providers [154]. But at the same time, these applications raise various ethical challenges and complex questions that need to be seriously examined. Below, we give an analytical and systematic overview of these issues.

5.3.1. Data related ethical concerns

As outlined above, the main thrust of AI/ML applications in healthcare is to maximize the benefits (principle of beneficence) and minimize the harms (principle of non-maleficence) for as many stakeholders as possible, especially the patients. To achieve this noble aim, AI technologies are highly dependent on vast amounts of data from which these technologies will "learn" how to make predictions and decisions. The "automation" of these AI-based tools for algorithmic decision-making provides no guarantee that we will have more ethically-committed outcomes. This is because the input of big data is actually a record of human actions, which are not free from biases and injustices. Thus, the behavior of machine learning systems is simply mirroring and echoing human behavior, including its moral failures even if we claim that we do not do them intentionally [154].

Against this background, the quality of training data has a high impact on the performance of ML algorithms. ML models learn the latent variable of data to deduce the predictions. So it is required to consider the problems with the data first while developing efficient models. Here we discuss the ethical problems related to the medical datasets:

Imbalanced Datasets Imbalanced class data is a common data-related problem that occurs in the supervised training of ML/DL models. This problem arises due to the non-uniform distribution of samples among classes. Training the model on such imbalanced data results in outcomes that are biased to certain categories. Biases in outcomes of models used for healthcare services may have profound consequences. One of the famous examples in this regard is the Google Health study, published in Nature, which argued that an AI system can outperform radiologists at predicting cancer. The lack of adequately described methodologies and computer code behind the study, weakens its scientific worth. The work was then accused of breaking transparency and reproducibility criteria. [158], [159]. Haibe-Kains et al. identified the challenges to making this work transparent and reproducible and provided solutions with implications for the broader field [160].

Data Bias Other than the biased outcomes due to class imbalance, the biases in data also lead to biased outcomes. In order to realize the impactful significance of ML/DL methods, it is highly required that the ML/DL models should produce fair outcomes that are bias-free. Here we will discuss the various facts and circumstances that are affecting fair healthcare data collection and causing data bias. For example, researchers have shown that the model predicts that black people have strong immunity and are healthier as compared to equally sick white people [161]. The algorithm is biased because it utilizes health costs as a measure of medical needs. Because less money is spent on Black patients with the same level of need, the algorithm incorrectly predicts that Black patients are healthier than equally sick White patients.

The dependence of models' learning on skin tone, face structure, or nationality is problematic for healthcare applications. Another problem is that ML-based healthcare products are manufactured by Western companies and these products are developed and tested on Caucasian data. This problem can be resolved by ensuring diversity in the collection of data around the globe. The healthcare datasets are mostly biased towards males because clinical trials held for collecting the data have large data samples of male patients. This bias causes ML models to show more precision for males in contrast to females. It is important to take into account that the healthcare datasets must represent both genders equally [162]. It is a common practice that more healthcare facilities are available for wealthy people which makes it less likely that low-income people are able to access advanced technological treatments. This bias in the availability of facilities is also reflected in the data and can cause the biased decision of ML algorithms [163]. Similar is the case for geographical biases, where fewer healthcare facilities are provided in rural areas and under-developing countries [164]. Explanation of the data, as we proposed in the pipeline of explainable ML presented in Figure 3, in the first place is required to check for these biases.

5.3.2. Privacy

Protecting the privacy of patients and the confidentiality of their data is one of the fundamentals of ethical healthcare. This

principle is also translated into legal codification. For example, the health insurance and portability and accountability (HIPAA) act assures the privacy of the medical data of patients. HIPAA's policy standards are designed to improve the healthcare systems and mandate it for all healthcare organizations to protect medical information [165].

In the healthcare context, privacy is defined as keeping the information of patients protected from unauthorized access. However, ML algorithms require access to as much data as possible to improve the precision and accuracy of the outcome. The amount and type of the needed data are increasing over time to the extent of seriously blurring the boundaries between what is "medical", which should be shared with one's physician, and what is "personal" and thus one has the right to keep it private. How AI-based healthcare or the so-called "deep medicine" would deal with a disease like depression is an apt example in this regard. To achieve the potential of "deep medicine", the scope of the to-be collected data should be wide enough to include speech, the intonation of voice, reaction times from keyboard use, GPS data, social media usage, distinctive facial attributes in one's selfies, etc. [151], [153]. To make the situation more complex, conducting a proper analysis of all these data would necessitate giving access not only to one's physician but to many other experts in various areas. Against this backdrop, special attention should be given to the privacy requirements, e.g., determining which data is needed, for what purpose, and with access to whom. Various factors can put people's privacy at risk, and we highlight here two of them:

Unprotected Data Sharing: With the advanced technologies, the records, and reports of patients are converted into electronic health records (EHR). These records are available online via the cloud servers. Techniques based on Internet-of-Things (IoT) are widely used in healthcare systems for real-time monitoring of critical patients. However, this ability leads to data breaching through tracking and monitoring of patients' routines which dishonors the patients' privacy. An unprotected data sharing technique may lead to breaching healthcare data and hackers can access confidential information like email accounts, messages, and reports of patients. A systematic review focused on the ethical issues related to the use of IoT is presented in [166].

Misuse of Medical Data: Online prognosis and diagnosis systems are trending these days. Many websites provide cloud-hosted ML/DL-based healthcare facilities that allow users to get the recommendation through an online healthcare system based on their EHRs. These websites also provide free data storing facilities and are not always concerned about the privacy of the users' data. Consequently, they might unethically trade the record or data of patients to other companies. Considering the sensitive nature of medical data and the requirement for protecting the privacy of patients, there is a need to design a system that protects against such data breaches. It must be considered while developing a system that patient data cannot be inferred by examining the outputs of the ML/DL model [167]. Thus, it is crucial to manage and protect the personal information of the patients. Concerned medical staff and researchers should be aware of risks linked with the breach of patient data and their legal responsibilities in processing the data. Because of the particular significance of addressing the data-related concerns, different countries have developed policies and laws [168].

5.3.3. Informed Consent

As outlined above, respect for persons and autonomy are among the widely agreed upon principles in modern bioethics. Obtaining informed consent from the patient before exposing him/her to any medical intervention is one of the practical applications of these principles [142]. As it is clear from its very term, the consent of the patient should be "informed" in nature. In other words, the patient's consent should be premised on sufficient information about the medical procedure, especially efficacy, safety, possible benefits, and expected harms.

The black-box nature of ML models, as outlined in this paper, is a serious obstacle to getting the necessary informed consent from the patients. Due to this black-box nature, neither the patient nor even the clinician will be able to understand the rationale behind the conclusions or recommendations made by the AI technologies. To address this concern, the European GDPR has introduced rules for the decisions and methods based on data-driven approaches to provide an ethical framework [169]. According to the GDPR rules, it is the right of an individual to understand why the model is taking a specific decision and the underlying mechanism of decisions concerning the individual. This step limits the implementation of ML models for clinical applications because of the use of patient data. That is why improving the explainability and interpretability of the black-box models represents an ethical requirement in order to facilitate proper informed consent.

Until this ideal situation is in place, where both accuracy and explainability of ML-based healthcare systems can be achieved, a number of ethical considerations should be in order. At the minimum level, the patient should be properly informed about the black-box nature of the ML applications and all related pros and cons of these applications should be made clear. Additionally, ML-based medical interventions cannot be judged indiscriminately; without considering the morally significant differences and nuances. For instance, consenting to the use of ML-based interventions as the only available tool to treat an incurable and life-threatening disease will not be the same as consenting to an intervention meant for enhancing specific physical traits rather than treating a serious health condition.

Other concerns related to the doctrine of informed consent have to do with the surveillance of public health, which also raises ethical issues like invasion of privacy, data protection, autonomy, freedom, equity, and liability [168]. To avoid these ethical issues, it is necessary to do preventive ethical assessments of developing AI technology for medical use. Lack of ethical regulations, as well as inadequate or no training for such surveillance operations, poses ethical challenges [170]. Due to the availability of implantable devices, it is now possible to monitor patients without their consent. Despite all these ethical issues, Lee et al. [171] provided ethical justification about the surveillance of public health without any explicit consent is ethically justifiable if principles of contemporary clinical and public health ethics are taken into the account. However, it is

also not guaranteed that the data collected for a specific objective will always be used for the same purpose. As it has been shown, the data can be used for any other purpose by doing slight changes. Additionally, merging datasets of two different experiments can be used for the modeling of a third type of experiment [172]. Therefore, the explicit and targeted consent of patients is required for the data collection through IoT and for ML/DL empowered personalized medical systems [173].

5.3.4. Care Ethics

As mentioned above, modern bioethics has other approaches besides principlism. Some of these non-principlist approaches can provide fresh insights into some of the ethical questions triggered by AI/ML-based healthcare systems. The care ethics approach, which focuses on the domain of intimate human relationships rather than the abstract application of rules [142], serves as a good example in this regard. The points discussed below are meant to just give representative examples of how the care ethics approach can be of benefit and relevance to the ethical discussions on AI/ML-driven healthcare.

The issues that can be discussed within this approach go beyond the question of solely measuring the efficacy and safety of certain applications or calculating their possible health-related benefits and harms. For instance, there is a concern about how these developments would negatively affect the job security of the medical staff, who may be replaced by AI devices that can relentlessly work and possibly more efficiently than humans and without complaints. In response, different voices stress that the AI tools are meant to support, facilitate, and enhance the human work of healthcare providers but not to replace them. On the other hand, some optimist voices argue that integrating AI systems into healthcare will make the healthcare profession more humane, by improving the physician-patient relationship [151], [155].

Additionally, some researchers expressed specific concerns about the negative impact of certain applications on the desired intimate inter-human relations, especially in the healthcare sector. One of the famous examples is the so-called "carebots"; employed to offload caregiving to a machine. Even if this automation of caregiving will not result in causing medical harm to the patient or job cuts in the healthcare staff, replacing human care will still have social costs, e.g., exchanging feelings and emotions among humans will cease to be part of caregiving [174]. It is to be noted that this concern was a point of heated discussions among early pioneers in the ethics of computer science. For instance, the computer scientist, Joseph Weizenbaum, wrote in the 1970s that it is immoral to use computer systems for substituting a human function, which involves interpersonal respect, understanding, and love, even if they proved to be technically successful [175]. Figure 8 illustrates the overview of the explainable, trustworthy, and ethical ML methods used for healthcare in literature.

6. Potential Pitfalls

The recent advancements in technology have made it possible to acquire, save, and share high-resolution medical images. Such data is being massively generated by many healthcare facilities on a daily basis, which has a significant potential to enable data-driven healthcare. In this regard, researchers are developing learning-based methods using such large-scale datasets, particularly, DL-based methods have provided state-of-the-art performance in many medical image analysis tasks [176]. However, despite their significant performance, these models are black-box and lack theoretical understanding behind their decisions. Their black-box nature makes them susceptible to many vulnerabilities such as adversarial attacks, biased decisions, and not being able to generalize out of distribution samples, etc. Thus raising concerns about the robustness and trustworthiness of ML methods is crucial, because of their practice in life-critical applications like healthcare. To circumvent this issue, the explainability of black-box models and considering ethical constraints, is proposed in the literature. However, the developed explanation methods have unique challenges and limitations associated with them, which are described below.

6.1. Vulnerability to Input Changes

In clinical settings, it is highly desirable that the explanation of a particular method should be similar for the same disease across different patients, which are geographically dispersed and have unique characteristics (i.e., generalized explanations for a particular type of disease for different patients). However, it has been shown in the literature that the explanation methods are vulnerable to input changes. For instance, Ghorbani et al. [177] demonstrated that a minor change in the input sample caused significant fluctuations in the output representations generated by XML. In addition, the inherent bias in the input (medical) data (e.g., class imbalance) can be reflected in the outputs of the models, i.e., the model might prefer a specific class as compared to other classes, and this bias might influence the explanations of the model [178].

6.2. Sub-optimal Explanations

In the literature, visualization-based methods are widely applied to explain the decisions of ML/DL methods. However, it is not evident that these explanations are the optimal requirement of medical experts. Weerts et al. [179] examined how the explanations produced from SHAP influence human performance for alert processing tasks. They conducted a human-based study to evaluate whether decision-making tasks can be improved by presenting explanations. They showed that SHAP explanations of class probability did not improve the decision-making. Similarly, Mohseni et al. [48] conducted a human-grounded study and evaluated the performance of the LIME algorithm by comparing the explanation produced by LIME with the weighted explanations generated by ten human experts. Their results showed that LIME highlights some attributions which were irrelevant to the explanations produced by humans. Therefore, without using the sound quality measuring technique, the use of these explanation methods should be avoided for making healthcare decisions.

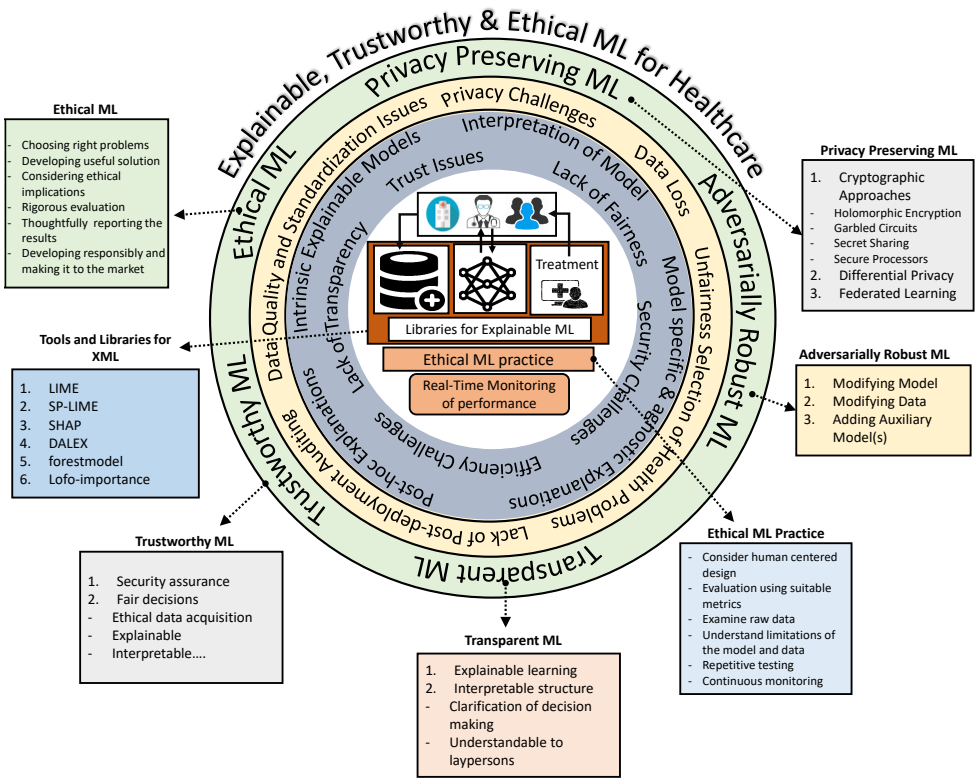

Figure 8: Overview of the explainable, trustworthy, and ethical ML models for healthcare.

6.3. Dependence on Data and Model

The literature suggests that the explanations generated by some gradient-based methods are dependent on the model architecture and data generation procedure [180]. As these explanations depend on the choice of reference point, a slight change in the reference point of the gradient will significantly change the explanation, thus causing confusion that will eventually lead to misleading results or interpretation.

6.4. Accountability Attribution

It is without a doubt that the deployment of ML for clinical practice will aid the clinicians. However, it is not clear yet who will be responsible in case an algorithm shows wrong outputs. Whether the clinicians will be responsible (because they are the ones making final decisions) or will the institutes force clinicians to rely on the decisions of ML? Researchers developing the algorithms can also be responsible for bad decisions [181]. This situation becomes even more complex when we consider all stakeholders in the loop. This blame game will eventually foster "epistemic vices" such as "dogmatism" or "gullibility" [182].

6.5. Rigorously Evaluating the Method

It has been emphasized in the literature that rigorous evaluation of the ML method should be performed to ensure that no unintended label leakage can occur between the datasets used in the model training [183]. Label leakage can possibly arise in subtle ways, e.g., an algorithm may learn the inherent noise instead of learning the diagnostic parameters. Another important aspect is to identify and validate the scope of model performance in both cases, i.e., where it succeeds in accurately diagnosing and where it fails. Moreover, it has been argued in the literature that traditional statistical performance metrics like the area under the curve may not be sufficient for evaluating the models making clinical decisions [183]. Therefore, clinically relevant metrics should be developed to evaluate such models. In addition to using quantitative metrics, qualitative measures can be used to identify whether the model is reliable and relevant for the intended task. Randomized controlled validation should be performed to evaluate the model efficacy in a real-time environment. The silent mode testing can be effective for identifying the errors in the real-time settings [184].

7. Future Research Directions

The motivation and the need for explainable, trustworthy, secure, and robust ML/DL methods applied in healthcare is clear. In this section, we discuss some future research opportunities in this field.

7.1. Explaining Medical Data

ML techniques build their decisions on the latent variables which are learned from the data. Medical data is one of the most difficult data to handle due to its complex, multi-variate, and sometimes non-stationary and scarce nature. The dependence of latent variables on each other can cause misleading patterns and due to this issue, the ML-based decision-making will be misleading. The literature suggests that the data should be thoroughly scrutinized before the model development to ensure that it is appropriate for the problem being modeled [183]. Moreover, it is imperative to understand how and for what purpose this medical data was collected. In addition, bias in the data is also a major challenge to handle and that can eventually lead to algorithmic bias [185]. These biases are hard to undo and their elimination has unintended consequences on the results [186]. The presence of these subtle biases in medical data decreases model reliability, especially when they are not corrected during model development [187, 188]. Therefore, to develop explainable, reliable, robust, and trustworthy algorithms, it is highly required to explain the dependence and relevance of data variables and patterns first (before feeding the data to ML algorithms).

7.2. Representation Techniques for Explanation

It is well established that the explanation of the ML/DL techniques is required to gain the trust of clinicians in ML/DL-empowered healthcare solutions. However, it is necessary to understand how these explanations are presented to them, i.e., explanations should be comprehensible to clinicians. The representation of the explanations needs the adoption of knowledge from other fields. For example, human-computer interaction (HCI) is a well-developed technique to empower users. XML researchers should incorporate the knowledge and techniques from the HCI to better represent the explanations. Therefore, developing efficient representation techniques for explanations of ML/DL methods remains an open research problem.

7.3. Generalized Explanations

As discussed in Section 6, the explanations produced by the data-dependent explanation models are vulnerable to the change in inputs and may vary from patient to patient and even for the same patient for the same disease. This issue should be resolved by developing robust, efficient, and generalized explainable models. As we discussed in Section 3, the explanations of the DL models are model-specific in nature, therefore, it is also required to develop inherently explainable and generalized explainable methods for the DL algorithms in the future.

7.4. Adversarially Robust ML

To attain explainable, trustworthy, safe, and robust ML/DL methods, it is very important to address the challenges like adversarial ML attacks. Over the past few years, it has been shown that ML/DL methods can be easily fooled and desired outcomes can be obtained [30, 31]. The critical nature of healthcare applications provides significant motivation for the malicious actors to defame the ML/DL-based system and to get the desired outcomes. In the literature, a wide variety of adversarial ML attacks have been already proposed and the research on developing respective defense methods is very limited [72]. This highlights that there is an utmost need for developing adversarially robust ML/DL techniques. Moreover, the clinical impact of ML/DL advancements is only completely possible by overcoming challenges like adversarial ML attacks.

7.5. Interdisciplinary Development Workforce

The advancements in ML/DL techniques have a great potential to revolutionize healthcare. However, to get the real benefit of these advancements, challenges like ethical issues are needed to be effectively addressed. In this regard, a few studies suggested involving all types of stakeholders in the ML/DL method development process that may include clinicians, policymakers, data scientists, ML researchers, and hospital staff, to name a few [189, 183]. Such an interdisciplinary development workforce will enable collaboration between the knowledge experts (i.e., clinicians and ML researchers) and healthcare service providers which will eventually improve productivity and outcomes.

8. Conclusions

In this paper, we have built upon existing literature on the explainable, trustworthy, and ethical machine learning (ML) for healthcare and have provided a comprehensive review of these emerging topics. In addition, we have highlighted the interconnection among them along with their relevance and applicability for healthcare applications. We highlighted various challenges that are hindering the successful deployment of ML and deep learning (DL) techniques in healthcare applications and formulated the pipeline for the development of clinically implementable and explainable ML methods. We also elaborated upon different security, safety, robustness, and ethical challenges which are the key barrier to the development of trustworthy ML/DL-based healthcare applications. Furthermore, we have discussed in detail, how explainable ML can be used to address such challenges. Finally, we have identified the limitations of existing methods and highlighted various open research issues that require further development.

9. Acknowledgments

This publication was made possible by NPRP grant #[13S-0206-200273] from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

[1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, Medical image analysis 42 (2017) 60–88.

[2] C. Xiao, E. Choi, J. Sun, Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review, Journal of the American Medical Informatics Association 25 (10) (2018) 1419–1428.

[3] S. Trebeschi, J. J. van Griethuysen, D. M. Lambregts, M. J. Lahaye, C. Parmar, F. C. Bakers, N. H. Peters, R. G. Beets-Tan, H. J. Aerts, Deep learning for fully-automated localization and segmentation of rectal cancer on multiparametric MR, Scientific reports 7 (1) (2017) 1–9.

[4] J. Betancur, F. Commandeur, M. Motlagh, T. Sharir, A. J. Einstein, S. Bokhari, M. B. Fish, T. D. Ruddy, P. Kaufmann, A. J. Sinusas, et al., Deep learning for prediction of obstructive disease from fast myocardial perfusion SPECT: a multicenter study, JACC: Cardiovascular Imaging 11 (11) (2018) 1654–1663.

[5] J.-G. Lee, S. Jun, Y.-W. Cho, H. Lee, G. B. Kim, J. B. Seo, N. Kim, Deep learning in medical imaging: general overview, Korean journal of radiology 18 (4) (2017) 570–584.

[6] A. Qayyum, S. M. Anwar, M. Awais, M. Majid, Medical image retrieval using deep convolutional neural network, Neurocomputing 266 (2017) 8–20.

[7] C. Angermueller, T. Pärnamaa, L. Parts, O. Stegle, Deep learning for computational biology, Molecular systems biology 12 (7) (2016) 878.

[8] E. Begoli, T. Bhattacharya, D. Kusnezov, The need for uncertainty quantification in machine-assisted medical decision making, Nature Machine Intelligence 1 (1) (2019) 20–23.

[9] A. Holzinger, C. Biemann, C. S. Pattichis, D. B. Kell, What do we need to build explainable AI systems for the medical domain?, arXiv preprint arXiv:1712.09923.

[10] M. FAT, Fairness, accountability, and transparency in machine learning, Retrieved December 24 (2018) 2018.

[11] D. Gunning, Explainable artificial intelligence (XAI), Defense Advanced Research Projects Agency (DARPA), nd Web 2 (2).

[12] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160.

[13] E. Tjoa, C. Guan, A survey on explainable artificial intelligence (XAI): towards medical XAI, arXiv preprint arXiv:1907.07374.

[14] A. Singh, S. Sengupta, V. Lakshminarayanan, Explainable deep learning models in medical image analysis, arXiv preprint arXiv:2005.13799.

[15] D. S. Char, M. D. Abràmoff, C. Feudtner, Identifying ethical considerations for machine learning healthcare applications, The American Journal of Bioethics 20 (11) (2020) 7–17.

[16] A. Adadi, M. Berrada, Explainable AI for healthcare: From black box to interpretable models, in: Embedded Systems and Artificial Intelligence, Springer, 2020, pp. 327–337.

[17] P. Hall, M. Kurka, A. Bartz, Using H2O driverless AI.

[18] G. Montavon, W. Samek, K.-R. Müller, Methods for interpreting and understanding deep neural networks, Digital Signal Processing 73 (2018) 1–15.

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

[20] A. D. Jeffery, L. L. Novak, B. Kennedy, M. S. Dietrich, L. C. Mion, Participatory design of probability-based decision support tools for in-hospital nurses, Journal of the American Medical Informatics Association 24 (6) (2017) 1102–1110.

[21] M. A. Ahmad, C. Eckert, A. Teredesai, Interpretable machine learning in healthcare, in: Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics, 2018, pp. 559–560.

[22] C. Wierzynski, The challenges and opportunities of explainable AI, Intel. com 12.

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

[24] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: An approach to evaluating interpretability of machine learning, arXiv preprint arXiv:1806.00069 (2018) 118.

[25] F. Gille, A. Jobin, M. Ienca, What we talk about when we talk about trust: Theory of trust for AI in healthcare, Intelligence-Based Medicine 1 (2020) 100001.

[26] M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, R. Ranganath, Opportunities in machine learning for healthcare, arXiv preprint arXiv:1806.00388.

[27] D. C. Castro, I. Walker, B. Glocker, Causality matters in medical imaging, Nature Communications 11 (1) (2020) 1–10.

[28] L. Floridi, Establishing the rules for building trustworthy AI, Nature Machine Intelligence 1 (6) (2019) 261–262.

[29] S. R. Meikle, J. C. Matthews, V. J. Cunningham, D. L. Bailey, L. Livieratos, T. Jones, P. Price, Parametric image reconstruction using spectral analysis of pet projection data, Physics in Medicine & Biology 43 (3) (1998) 651.

[30] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, arXiv preprint arXiv:1312.6199.

[31] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, I. S. Kohane, Adversarial attacks on medical machine learning, Science 363 (6433) (2019) 1287–1289.

[32] A. Qayyum, J. Qadir, M. Bilal, A. Al-Fuqaha, Secure and robust machine learning for healthcare: A survey, IEEE Reviews in Biomedical Engineering 14 (2021) 156–180. doi:10.1109/RBME.2020.3013489.

[33] J. D. Moore, W. R. Swartout, Explanation in expert systems: A survey, Tech. rep., University of Southern California (USC) Information Sciences Institute (1988).

[34] M. Van Lent, W. Fisher, M. Mancuso, An explainable artificial intelligence system for small-unit tactical behavior, in: Proceedings of the national conference on artificial intelligence, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004, pp. 900–907.

[35] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European conference on computer vision, Springer, 2014, pp. 818–833.

[36] L. M. Zintgraf, T. S. Cohen, T. Adel, M. Welling, Visualizing deep neural network decisions: Prediction difference analysis, arXiv preprint arXiv:1702.04595.

[37] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in neural information processing systems, 2017, pp. 4765–4774.

[38] E. Ratti, M. Graves, Explainable machine learning practices: opening another black box for reliable medical ai, AI and Ethics (2022) 1–14.

[39] K. G. Heider, The Rashomon effect: When ethnographers disagree, American Anthropologist 90 (1) (1988) 73–81.

[40] J. Petch, S. Di, W. Nelson, Opening the black box: the promise and limitations of explainable machine learning in cardiology, Canadian Journal of Cardiology.

[41] I. Y. Chen, E. Pierson, S. Rose, S. Joshi, K. Ferryman, M. Ghassemi, Ethical machine learning in healthcare, Annual Review of Biomedical Data Science 4 (2021) 123–144.

[42] S. Levin, S. Barnes, M. Toerper, A. Debraine, A. DeAngelo, E. Hamrock, J. Hinson, E. Hoyer, T. Dungarani, E. Howell, Machine-learning-based hospital discharge predictions can support multidisciplinary rounds and decrease hospital length-of-stay, BMJ Innovations 7 (2).

[43] P. Kaur, R. Kumar, M. Kumar, A healthcare monitoring system using random forest and internet of things (iot), Multimedia Tools and Applications 78 (14) (2019) 19905–19916.

[44] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad, Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '15, Association for Computing Machinery, New York, NY, USA, 2015, p. 1721–1730. doi:10.1145/2783258.2788613.
URL https://doi.org/10.1145/2783258.2788613

[45] S. Kaufman, S. Rosset, C. Perlich, O. Stitelman, Leakage in data mining: Formulation, detection, and avoidance, ACM Trans. Knowl. Discov. Data 6 (4). doi:10.1145/2382577.2382579.
URL https://doi.org/10.1145/2382577.2382579

[46] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (5). doi:10.1145/3236009.
URL https://doi.org/10.1145/3236009

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

[48] S. Mohseni, J. E. Block, E. D. Ragan, A human-grounded evaluation benchmark for local explanations of machine learning, arXiv preprint arXiv:1801.05075.

[49] M. Robnik-Šikonja, M. Bohanec, Perturbation-based explanations of prediction models, in: Human and machine learning, Springer, 2018,

pp. 159–175.
[50] K. Rasheed, A. Qayyum, J. Qadir, S. Sivathamboo, P. Kwan, L. Kuhlmann, T. O'Brien, A. Razi, Machine learning for predicting epileptic seizures using eeg signals: A review, IEEE Reviews in Biomedical Engineering (2020) 1–1doi:10.1109/RBME.2020.3008792.
[51] A. Işın, C. Direkoğlu, M. Şah, Review of MRI-based brain tumor image segmentation using deep learning methods, Procedia Computer Science 102 (2016) 317–324.
[52] J. Islam, Y. Zhang, A novel deep learning based multi-class classification method for alzheimer's disease detection using brain MRI data, in: International Conference on Brain Informatics, Springer, 2017, pp. 213–222.
[53] J. Zou, M. Huss, A. Abid, P. Mohammadi, A. Torkamani, A. Telenti, A primer on deep learning in genomics, Nature genetics 51 (1) (2019) 12–18.
[54] C. Rong, X. Li, X. Sun, H. Sun, Chinese medicine prescription recommendation using generative adversarial network, IEEE Access 10 (2022) 12219–12228.
[55] E. Galinkin, Robustness and usefulness in ai explanation methods, arXiv preprint arXiv:2203.03729.
[56] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2921–2929.
[57] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.
[58] M. Suzuki, K. Masuda, H. Asakuma, K. Takeshita, K. Baba, Y. Kubo, K. Ushijima, S. Uchida, T. Akagi, Deep learning predicts rapid oversoftening and shelf life in persimmon fruits, The Horticulture Journal (2022) UTD–323.
[59] P.-J. Kindermans, K. T. Schütt, M. Alber, K.-R. Müller, D. Erhan, B. Kim, S. Dähne, Learning how to explain neural networks: Patternnet and patternattribution, arXiv preprint arXiv:1705.05598.
[60] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, K.-R. Müller, Explaining nonlinear classification decisions with deep taylor decomposition, Pattern Recognition 65 (2017) 211–222.
[61] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, arXiv preprint arXiv:1703.01365.
[62] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.
[63] W. Samek, T. Wiegand, K.-R. Müller, Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models, arXiv preprint arXiv:1708.08296.
[64] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, K.-R. Müller, Explaining deep neural networks and beyond: A review of methods and applications, Proceedings of the IEEE 109 (3) (2021) 247–278.
[65] T. P. Bohlin, Practical grey-box process identification: theory and applications, Springer Science & Business Media, 2006.
[66] K. Yeung, Recommendation of the council on artificial intelligence (oecd), International Legal Materials 59 (1) (2020) 27–34.
[67] A. Holzinger, M. Dehmer, F. Emmert-Streib, R. Cucchiara, I. Augenstein, J. Del Ser, W. Samek, I. Jurisica, N. Díaz-Rodríguez, Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence, Information Fusion 79 (2022) 263–278.
[68] Y. Mirsky, T. Mahler, I. Shelef, Y. Elovici, Ct-gan: Malicious tampering of 3d medical imagery using deep learning, in: 28th {USENIX} Security Symposium ({USENIX} Security 19), 2019, pp. 461–478.
[69] M. Paschali, S. Conjeti, F. Navarro, N. Navab, Generalizability vs. robustness: adversarial examples for medical imaging, arXiv preprint arXiv:1804.00504.
[70] X. Han, Y. Hu, L. Foschini, L. Chinitz, L. Jankelson, R. Ranganath, Deep learning models for electrocardiograms are susceptible to adversarial attack, Nature medicine 26 (3) (2020) 360–363.
[71] A. Vatian, N. Gusarova, N. Dobrenko, S. Dudorov, N. Nigmatullin, A. Shalyto, A. Lobantsev, Impact of adversarial examples on the efficiency of interpretation and use of information from high-tech medical images, in: 2019 24th Conference of Open Innovations Association (FRUCT), IEEE, 2019, pp. 472–478.
[72] A. Qayyum, I. Aneeqa, M. Usama, W. Iqbal, J. Qadir, Y. Elkhatib, A. Al-Fuqaha, Securing machine learning (ML) in the cloud: A systematic review of cloud ML security, Frontiers in Big Data.
[73] H. Takabi, E. Hesamifard, M. Ghasemi, Privacy preserving multi-party machine learning with homomorphic encryption, in: 29th Annual Conference on Neural Information Processing Systems (NIPS), 2016.
[74] D. Bogdanov, L. Kamm, S. Laur, V. Sokk, Implementation and evaluation of an algorithm for cryptographically private principal component analysis on genomic data, IEEE/ACM transactions on computational biology and bioinformatics 15 (5) (2018) 1427–1432.
[75] N. Phan, X. Wu, H. Hu, D. Dou, Adaptive Laplace mechanism: Differential privacy preservation in deep learning, in: 2017 IEEE International Conference on Data Mining (ICDM), IEEE, 2017, pp. 385–394.
[76] A. Qayyum, K. Ahmad, M. A. Ahsan, A. Al-Fuqaha, J. Qadir, Collaborative federated learning for healthcare: Multi-modal covid-19 diagnosis at the edge, arXiv preprint arXiv:2101.07511.
[77] O. Choudhury, A. Gkoulalas-Divanis, T. Salonidis, I. Sylla, Y. Park, G. Hsu, A. Das, Differential privacy-enabled federated learning for sensitive health data, arXiv preprint arXiv:1910.02578.
[78] H. Ali, R. T. Javed, A. Qayyum, A. AlGhadhban, M. Alazmi, A. Alzamil, K. AlUtaibi, J. Qadir, Spam-das: Secure and privacy-aware misinformation detection as a service.
[79] C. S. Perone, P. Ballester, R. C. Barros, J. Cohen-Adad, Unsupervised domain adaptation for medical imaging segmentation with self-ensembling, NeuroImage 194 (2019) 1–11.
[80] G. S. Kuntla, X. Tian, Z. Li, Security and privacy in machine learning: A survey., Issues in Information Systems 22 (3).
[81] N. Ford, J. Gilmer, N. Carlini, D. Cubuk, Adversarial examples are a natural consequence of test error in noise, arXiv preprint arXiv:1901.10513.
[82] D. Su, H. Zhang, H. Chen, J. Yi, P.-Y. Chen, Y. Gao, Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 631–648.
[83] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, A. Madry, Robustness may be at odds with accuracy, stat 1050 (2018) 11.
[84] J. Gao, X. Wang, Y. Wang, X. Xie, Explainable recommendation through attentive multi-view learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 3622–3629.
[85] A. F. Markus, J. A. Kors, P. R. Rijnbeek, 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 (2021) 103655.
[86] R. D. Gibbons, G. Hooker, M. D. Finkelman, D. J. Weiss, P. A. Pilkonis, E. Frank, T. Moore, D. J. Kupfer, The cad-mdd: A computerized adaptive diagnostic screening tool for depression, The Journal of clinical psychiatry 74 (7) (2013) 669.
[87] A.-D. Dana, A. Alashqur, Using decision tree classification to assist in the prediction of alzheimer's disease, in: 2014 6th International Conference on Computer Science and Information Technology (CSIT), IEEE, 2014, pp. 122–126.
[88] A. Suresh, R. Udendhran, M. Balamurgan, Hybridized neural network and decision tree based classifier for prognostic decision making in breast cancers, Soft Computing 24 (11) (2020) 7947–7953.
[89] S. Khare, D. Gupta, Association rule analysis in cardiovascular disease, in: 2016 Second International Conference on Cognitive Computing and Information Processing (CCIP), IEEE, 2016, pp. 1–6.
[90] S. Agrawal, N. Mishra, Question classification for health care domain using rule based approach, in: International Conference on Innovative Data Communication Technologies and Application, Springer, 2019, pp. 410–419.
[91] G. Wang, Z. Deng, K.-S. Choi, Detection of epilepsy with electroencephalogram using rule-based classifiers, Neurocomputing 228 (2017) 283–290.
[92] H. Byeon, Developing a random forest classifier for predicting the depression and managing the health of caregivers supporting patients with alzheimer's disease, Technology and Health Care 27 (5) (2019) 531–544.
[93] M. C. E. Simsekler, A. Qazi, M. A. Alalami, S. Ellahham, A. Ozonoff, Evaluation of patient safety culture using a random forest algorithm, Reliability Engineering & System Safety 204 (2020) 107186.
[94] C. Iwendi, A. K. Bashir, A. Peshkar, R. Sujatha, J. M. Chatterjee, S. Pa-

supuleti, R. Mishra, S. Pillai, O. Jo, Covid-19 patient health prediction using boosted random forest algorithm, Frontiers in public health 8 (2020) 357.

[95] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad, Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission, in: Proceedings of the 21st ACM SIGKDD international conference on knowledge discovery and data mining, 2015, pp. 1721–1730.

[96] L. Sagaon-Teyssier, A. Vilotitch, M. Mora, G. Maradan, V. Guagliardo, M. Suzan-Monti, R. Dray-Spira, B. Spire, A generalized additive model to disentangle age and diagnosis-specific cohort effects in psychological and behavioral outcomes in people living with hiv: the french cross-sectional anrs-vespa2 survey, BMC public health 19 (1) (2019) 1–10.

[97] M. Dastoorpoor, N. Khanjani, A. Moradgholi, R. Sarizadeh, M. Cheraghi, F. Estebsari, Prenatal exposure to ambient air pollution and adverse pregnancy outcomes in ahvaz, iran: a generalized additive model, International Archives of Occupational and Environmental Health (2020) 1–16.

[98] Y. Jiandong, Z. Mengxi, C. Yanggui, M. Li, Y. Rayibai, L. Yaoqin, L. Pengwei, P. Yujiao, X. Ran, R. Baolin, A study on the relationship between air pollution and pulmonary tuberculosis based on the general additive model in wulumuqi, china, International Journal of Infectious Diseases.

[99] V. Van Belle, B. Van Calster, S. Van Huffel, J. A. Suykens, P. Lisboa, Explaining support vector machines: a color based nomogram, PloS one 11 (10) (2016) e0164568.

[100] T. Eslami, J. S. Raiker, F. Saeed, Explainable and scalable machine learning algorithms for detection of autism spectrum disorder using fmri data, in: Neural Engineering Techniques for Autism Spectrum Disorder, Elsevier, 2021, pp. 39–54.

[101] D. Anguita, A. Ghio, N. Greco, L. Oneto, S. Ridella, Model selection for support vector machines: Advantages and disadvantages of the machine learning theory, in: The 2010 international joint conference on neural networks (IJCNN), IEEE, 2010, pp. 1–8.

[102] Y. Bengio, O. Delalleau, C. Simard, Decision trees do not generalize to new variations, Computational Intelligence 26 (4) (2010) 449–467.

[103] S. M. Mohnen, A. H. Rotteveel, G. Doornbos, J. J. Polder, Healthcare expenditure prediction with neighbourhood variables–a random forest model, Statistics, Politics and Policy 11 (2) (2020) 111–138.

[104] C.-H. Chang, S. Tan, B. Lengerich, A. Goldenberg, R. Caruana, How interpretable and trustworthy are gams?, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 95–105.

[105] A. Guisan, T. C. Edwards Jr, T. Hastie, Generalized linear and generalized additive models in studies of species distributions: setting the scene, Ecological modelling 157 (2-3) (2002) 89–100.

[106] R. Yang, Who dies from covid-19? post-hoc explanations of mortality prediction models using coalitional game theory, surrogate trees, and partial dependence plots, medRxiv.

[107] V. Gupta, M. Demirer, M. Bigelow, M. Y. Sarah, S. Y. Joseph, L. M. Prevedello, R. D. White, B. S. Erdal, Using transfer learning and class activation maps supporting detection and localization of femoral fractures on anteroposterior radiographs, in: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), IEEE, 2020, pp. 1526–1529.

[108] A. Kumar, S. B. Singh, S. C. Satapathy, M. Rout, Mosquito-net: A deep learning based cadx system for malaria diagnosis along with model interpretation using gradcam and class activation maps, Expert Systems (2021) e12695.

[109] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al., Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 590–597.

[110] S. D. Goodfellow, D. Shubin, R. W. Greer, S. Nagaraj, C. McLean, W. Dixon, A. J. Goodwin, A. Assadi, A. Jegatheeswaran, P. C. Laussen, et al., Rhythm classification of 12-lead ECGs using deep neural network and class-activation maps for improved explainability.

[111] S. Pereira, R. Meier, V. Alves, M. Reyes, C. A. Silva, Automatic brain tumor grading from MRI data using convolutional neural networks and quality assessment, in: Understanding and interpreting machine learning in medical image computing applications, Springer, 2018, pp. 106–114.

[112] M. Izadyyazdanabadi, E. Belykh, C. Cavallo, X. Zhao, S. Gandhi, L. B. Moreira, J. Eschbacher, P. Nakaji, M. C. Preul, Y. Yang, Weakly-supervised learning-based feature localization for confocal laser endomicroscopy glioma images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 300–308.

[113] H. Jung, Y. Oh, Towards better explanations of class activation mapping, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1336–1344.

[114] Y. Yang, V. Tresp, M. Wunderle, P. A. Fasching, Explaining therapy predictions with layer-wise relevance propagation in neural networks, in: 2018 IEEE International Conference on Healthcare Informatics (ICHI), IEEE, 2018, pp. 152–162.

[115] G. Chlebus, N. Abolmaali, A. Schenk, H. Meine, Relevance analysis of MRI sequences for automatic liver tumor segmentation, arXiv preprint arXiv:1907.11773.

[116] M. Böhle, F. Eitel, M. Weygandt, K. Ritter, Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification, Frontiers in aging neuroscience 11 (2019) 194.

[117] T. Jo, K. Nho, S. L. Risacher, A. J. Saykin, Deep learning detection of informative features in tau pet for alzheimer's disease classification, BMC bioinformatics 21 (21) (2020) 1–13.

[118] I. Palatnik de Sousa, M. Maria Bernardes Rebuzzi Vellasco, E. Costa da Silva, Local interpretable model-agnostic explanations for classification of lymph node metastases, Sensors 19 (13) (2019) 2969.

[119] S. Kitamura, K. Takahashi, Y. Sang, K. Fukushima, K. Tsuji, J. Wada, Deep learning could diagnose diabetic nephropathy with renal pathological immunofluorescent images, Diagnostics 10 (7) (2020) 466.

[120] P.-Y. Tseng, Y.-T. Chen, C.-H. Wang, K.-M. Chiu, Y.-S. Peng, S.-P. Hsu, K.-L. Chen, C.-Y. Yang, O. K.-S. Lee, Prediction of the development of acute kidney injury following cardiac surgery by machine learning, Critical Care 24 (1) (2020) 1–13.

[121] T. Pianpanit, S. Lolak, P. Sawangjai, A. Ditthapron, P. Leelaarporn, S. Marukatat, E. Chuangsuwanich, T. Wilaiprasitporn, Neural network interpretation of the parkinson's disease diagnosis from spect imaging, arXiv preprint arXiv:1908.11199.

[122] A. Borjali, A. F. Chen, O. K. Muratoglu, M. A. Morid, K. M. Varadarajan, Deep learning in orthopedics: How do we build trust in the machine?, Healthcare Transformation.

[123] M. R. Zafar, N. Khan, Deterministic local interpretable model-agnostic explanations for stable explainability, Machine Learning and Knowledge Extraction 3 (3) (2021) 525–541.

[124] D. Sharma, A. Durand, M.-A. Legault, L.-P. L. Perreault, A. Lemaçon, M.-P. Dubé, J. Pineau, Deep interpretability for gwas, arXiv preprint arXiv:2007.01516.

[125] D. Yu, Z. Liu, C. Su, Y. Han, X. Duan, R. Zhang, X. Liu, Y. Yang, S. Xu, Copy number variation in plasma as a tool for lung cancer prediction using extreme gradient boosting (xgboost) classifier, Thoracic Cancer 11 (1) (2020) 95–102.

[126] V. Couteaux, O. Nempont, G. Pizaine, I. Bloch, Towards interpretability of segmentation networks by analyzing deepdreams, in: Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support, Springer, 2019, pp. 56–63.

[127] K. Preuer, G. Klambauer, F. Rippmann, S. Hochreiter, T. Unterthiner, Interpretable deep learning in drug discovery, in: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer, 2019, pp. 331–345.

[128] S. Shen, S. X. Han, D. R. Aberle, A. A. Bui, W. Hsu, An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification, Expert systems with applications 128 (2019) 84–95.

[129] J. Zhang, K. Kowsari, J. H. Harrison, J. M. Lobo, L. E. Barnes, Patient2vec: A personalized interpretable deep representation of the longitudinal electronic health record, IEEE Access 6 (2018) 65333–65346.

[130] W. Liao, B. Zou, R. Zhao, Y. Chen, Z. He, M. Zhou, Clinical interpretable deep learning model for glaucoma diagnosis, IEEE journal of biomedical and health informatics 24 (5) (2019) 1405–1412.

[131] Z. Obermeyer, E. J. Emanuel, Predicting the future—big data, machine learning, and clinical medicine, The New England journal of medicine 375 (13) (2016) 1216.

[132] D. S. Char, N. H. Shah, D. Magnus, Implementing machine learning in

health care—addressing ethical challenges, The New England journal of medicine 378 (11) (2018) 981.
[133] T. L. Beauchamp, J. F. Childress, et al., Principles of biomedical ethics, Oxford University Press, USA, 2001.
[134] J. Cowls, L. Floridi, Prolegomena to a white paper on an ethical framework for a good AI society, Available at SSRN 3198732.
[135] S. H. Miles, The Hippocratic Oath and the ethics of medicine, Oxford University Press, 2005.
[136] P. Prioreschi, A History of Medicine. Byzantine and Islamic Medicine (2001).
[137] M. Levey, Medical ethics of medieval Islam with special reference to al-ruhāwī's" practical ethics of the physician", Transactions of the American Philosophical society (1967) 1–100.
[138] I. Waddington, The development of medical ethics—a sociological analysis, Medical History 19 (1) (1975) 36–51.
[139] F. A. Riddick, The code of medical ethics of the american medical association (2003).
[140] V. R. Potter, Bioethics: bridge to the future.
[141] M. Ghaly, Islamic perspectives on the principles of biomedical ethics, Vol. 1, World Scientific, 2016.
[142] R. M. Veatch, L. K. Guidry-Grimes, The basics of bioethics, Routledge, 2019.
[143] M. R. Marrus, The nuremberg doctors' trial in historical context, Bulletin of the History of Medicine 73 (1) (1999) 106–123.
[144] W. M. Association, et al., World medical association declaration of helsinki. ethical principles for medical research involving human subjects., Bulletin of the World Health Organization 79 (4) (2001) 373.
[145] K. W. Goodman, Ethics, computing, and medicine: informatics and the transformation of health care, Cambridge University Press, 1998.
[146] E. S. Berner, Clinical decision support systems, Vol. 233, Springer, 2007.
[147] R. M. Wachter, The digital doctor: hope, hype, and harm at the dawn of medicine's computer age, McGraw-Hill Education New York, 2015.
[148] A. Holzinger, C. Röcker, M. Ziefle, Smart health: open problems and future challenges, Vol. 8700, Springer, 2015.
[149] E. J. Topol, D. Hill, The creative destruction of medicine: How the digital revolution will create better health care, Basic Books New York, 2012.
[150] E.-Y. Kim, Patient will see you now: The future of medicine is in your hands, Healthcare Informatics Research 21 (4) (2015) 321–323.
[151] E. Topol, Deep medicine: how artificial intelligence can make healthcare human again, Hachette UK, 2019.
[152] E. J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nature medicine 25 (1) (2019) 44–56.
[153] L. Engelmann, Into the deep-AI and total pathology: Deep medicine: How artificial intelligence can make healthcare human again, by eric topol, new york: Basic books, 2010, 400 pp., $17.99 (paperback), isbn 9781541644649. (2020).
[154] J. H. Chen, A. Verghese, Planning for the known unknown: Machine learning for human healthcare systems, The American Journal of Bioethics 20 (11) (2020) 1–3.
[155] A. Bohr, K. Memarzadeh, Artificial intelligence in healthcare, Elsevier Science & Technology, 2020.
[156] A. Blasimme, E. Vayena, The ethics of AI in biomedical research, patient care and public health, Patient Care and Public Health (April 9, 2019). Oxford Handbook of Ethics of Artificial Intelligence, Forthcoming.
[157] A. Panesar, Machine learning and AI for healthcare, Springer, 2019.
[158] S. M. McKinney, A. Karthikesalingam, D. Tse, C. J. Kelly, Y. Liu, G. S. Corrado, S. Shetty, Reply to: Transparency and reproducibility in artificial intelligence, Nature 586 (7829) (2020) E17–E18.
[159] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. S. Corrado, A. Darzi, et al., International evaluation of an AI system for breast cancer screening, Nature 577 (7788) (2020) 89–94.
[160] B. Haibe-Kains, G. A. Adam, A. Hosny, F. Khodakarami, L. Waldron, B. Wang, C. McIntosh, A. Goldenberg, A. Kundaje, C. S. Greene, et al., Transparency and reproducibility in artificial intelligence, Nature 586 (7829) (2020) E14–E16.
[161] Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations, Science 366 (6464) (2019) 447–453.
[162] D. M. West, J. R. Allen, Turning Point: Policymaking in the Era of Artificial Intelligence, Brookings Institution Press, 2020.
[163] M. Anderson, A. Perrin, Barriers to adaption and attitudes towards technology, Tech adoption climbs among older adults.
[164] N. C. Arpey, A. H. Gaglioti, M. E. Rosenbaum, How socioeconomic status affects patient perceptions of health care: a qualitative study, Journal of primary care & community health 8 (3) (2017) 169–175.
[165] D. Box, D. Pottas, Improving information security behaviour in the healthcare context, Procedia Technology 9 (2013) 1093–1103, cENTERIS 2013 - Conference on ENTERprise Information Systems / ProjMAN 2013 - International Conference on Project MANagement/ HCIST 2013 - International Conference on Health and Social Care Information Systems and Technologies. doi:https://doi.org/10.1016/j.protcy.2013.12.122.
URL https://www.sciencedirect.com/science/article/pii/S2212017313002764
[166] H. Atlam, G. Wills, IoT Security, Privacy, Safety and Ethics, 2019, pp. 1–27. doi:10.1007/978-3-030-18732-3_8.
[167] M. Meingast, T. Roosta, S. Sastry, Security and privacy issues with health care information technology, in: 2006 International Conference of the IEEE Engineering in Medicine and Biology Society, 2006, pp. 5453–5458. doi:10.1109/IEMBS.2006.260060.
[168] B. Mittelstadt, B. Fairweather, M. Shaw, N. Mcbride, The ethical implications of personal health monitoring, International Journal of Technoethics 5 (2014) 37–60. doi:10.4018/ijt.2014070104.
[169] P. Voigt, A. Von dem Bussche, The eu general data protection regulation (gdpr), A Practical Guide, 1st Ed., Cham: Springer International Publishing 10 (2017) 3152676.
[170] C. Klingler, D. S. Silva, C. Schuermann, A. A. Reis, A. Saxena, D. Strech, Ethical issues in public health surveillance: a systematic qualitative review, BMC public health 17 (1) (2017) 1–13.
[171] L. M. Lee, C. M. Heilig, A. White, Ethical justification for conducting public health surveillance without patient consent, American journal of public health 102 (1) (2012) 38–44.
[172] T. Wu, J. Chung, J. Yamata, J. Richman, The ethics (or not) of massive government surveillance, The Ethics (or Not) of Massive Government Surveillance.
[173] B. Mittelstadt, Ethics of the health-related internet of things: a narrative review, Ethics and Information Technology 19 (3) (2017) 157–175.
[174] J. Donath, Ethical issues in our relationship with artificial entities, The Oxford Handbook of Ethics of AI (2020) 53.
[175] J. Weizenbaum, Computer power and human reason: From judgment to calculation.
[176] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, M. K. Khan, Medical image analysis using convolutional neural networks: a review, Journal of medical systems 42 (11) (2018) 226.
[177] A. Ghorbani, A. Abid, J. Zou, Interpretation of neural networks is fragile, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 3681–3688.
[178] S. Wang, T. Zhou, J. Bilmes, Bias also matters: Bias attribution for deep neural network explanation, in: International Conference on Machine Learning, PMLR, 2019, pp. 6659–6667.
[179] H. J. Weerts, W. van Ipenburg, M. Pechenizkiy, A human-grounded evaluation of shap for alert processing, arXiv preprint arXiv:1907.03324.
[180] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity checks for saliency maps, arXiv preprint arXiv:1810.03292.
[181] T. Grote, P. Berens, On the ethics of algorithmic decision-making in healthcare, Journal of medical ethics 46 (3) (2020) 205–211.
[182] Q. Cassam, Vices of the Mind: From the Intellectual to the Political, Oxford University Press, 2018.
[183] J. Wiens, S. Saria, M. Sendak, M. Ghassemi, V. X. Liu, F. Doshi-Velez, K. Jung, K. Heller, D. Kale, M. Saeed, et al., Do no harm: a roadmap for responsible machine learning for health care, Nature medicine 25 (9) (2019) 1337–1340.
[184] B. Nestor, M. McDermott, G. Chauhan, T. Naumann, M. C. Hughes, A. Goldenberg, M. Ghassemi, Rethinking clinical prediction: why machine learning must consider year of care and feature aggregation, arXiv preprint arXiv:1811.12583.
[185] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, D. King, Key challenges for delivering clinical impact with artificial intelligence, BMC medicine 17 (1) (2019) 195.
[186] S. Latif, A. Qayyum, M. Usama, J. Qadir, A. Zwitter, M. Shahzad,

Caveat emptor: the risks of using big data for human development, IEEE Technology and Society Magazine 38 (3) (2019) 82–90.
[187] S. Saria, A. Subbaswamy, Tutorial: safe and reliable machine learning, arXiv preprint arXiv:1904.07204.
[188] I. Y. Chen, P. Szolovits, M. Ghassemi, Can AI help reduce disparities in general medical and mental health care?, AMA journal of ethics 21 (2) (2019) 167–179.
[189] J. He, S. L. Baxter, J. Xu, J. Xu, X. Zhou, K. Zhang, The practical implementation of artificial intelligence technologies in medicine, Nature medicine 25 (1) (2019) 30–36.

Explainable Machine Learning, Interpretable Machine Learning, Trustworthiness, Healthcare.

Manuscript ID: CIBM-D-22-01543

Revised Highlights of the paper

- Provides in-depth review of explainable, trustworthy, and ethical ML for healthcare
- Presents a pipeline to explain and validate data and ML models for healthcare
- Various ML-related security, safety, robustness, & ethical challenges are presented
- Use of explainable & trustworthy ML is presented to resolve above mentioned issues
- Limitations of existing methods and various open research issues are highlighted

Conflict of Interest Statement

Editor-in-Chief

Elsevier Computers in Biology and Medicine

Subject: Declaration of Conflict of Interests

Dear Editor,

I am very pleased to submit our paper entitled “Explainable, Trustworthy, and Ethical Machine Learning for Healthcare: A Survey” authored by Khansa Rasheed, Adnan Qayyum, Mohammed Ghaly, Ala Al-Fuqaha, Adeel Razi, and Junaid Qadir for consideration for publication in Elsevier Computers in Biology and Medicine.

We have no conflict of interest to disclose.

Regards,

Junaid Qadir

Professor of Computer Engineering, Qatar University

IEEE and ACM Senior Member

ACM Distinguished Speaker

Corresponding author, on behalf of all authors.