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[1]

[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

***Abstract*— With the advent of machine learning (ML) applications in daily life, the questions about liability, trust, and interpretability of their outputs are raising, especially for healthcare applications. The black-box nature of ML models is a roadblock for clinical utilization. Therefore, to gain the trust of clinicians and patients, researchers need to provide explanations of how and why the model is making a specific decision. 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 provide a comprehensive review of explainable and interpretable ML techniques implemented for providing the reasons behind their decisions for various healthcare applications. Along with highlighting various security, safety, and robustness challenges that hinder the trustworthiness of ML we also discussed the ethical issues of healthcare ML and describe how explainable and trustworthy ML can resolve these ethical problems. Finally, we elaborate on the limitations of existing approaches and highlight various open research problems that require further development.**

***Index Terms*—Explainable Machine Learning, Interpretable Machine Learning, Trustworthiness, Healthcare.**

I. 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 consists 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 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 real benefit of ML/DL empowered predictive or diagnostic healthcare, it is highly desirable that ML/DL decisions should be interpretable and explainable in human understandable way. Figure 1 is the illustration of essential traits of ML models required for clinical implementation.

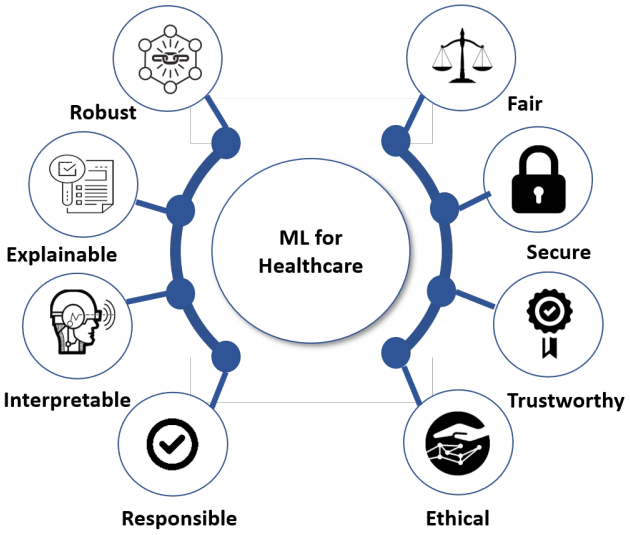

Fig. 1: Illustration of essential traits of ML models for clinical implementation.

Over the last few years, considerable research attention has been devoted to 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

Corresponding Author; Email: junaid.qadir@itu.edu.pk

(DARPA), explainable AI program [11]. FAT-ML comprises of 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 is holding conferences annually for bringing 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 taking an interest in developing XML methods. Microsoft having Azure ML services, H_2O.ai having driverless intelligent products [17], Kyndi serving to 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.

ML/DL techniques are transforming the healthcare research, however, we must build safety and trust in ML-based applications by explaining a few important questions, i.e., what patterns of features the ML/DL model has learned? Why the selected model is 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, there are multiple surveys that cover these topic. However, our review is unique from previous reviews in the following aspects:

1) To the best of our knowledge, there is no review that covers these topics in-depth while highlighting the link and applications of explainable and trustworthy methods for the medical domain.
2) Our review also proposes a pipeline from development to deployment to attain an explainable ML framework for healthcare.
3) We highlight various security, safety, and robustness challenges that hinder the trustworthiness of ML.
4) We also focus on the the ethical issues of healthcare ML and describe how explainable and trustworthy ML can resolve these ethical problems.
5) Finally, we elaborate on the limitations of existing 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 a brief overview of the concept of explainability and the available future opportunities in the field, along with 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 aforementioned contributions. The comparison of this paper with existing surveys is presented in Table I.

Organization of paper: The organization of this paper is as follows: Section II, presents challenges encountered in developing clinically effective explainable and trustworthy ML. Section III 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 IV, 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 V, 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 VI and various future directions are provided in Section VII. Finally, we conclude the paper in Section VIII. List of acronyms used in the paper is provided in Table II.

II. Challenges

For the sake of trustworthy and secure models for clinical settings, researchers are developing the tools and techniques for XML. Despite their efforts, there exist many issues that are causing challenges for effective XML. A few such challenges are described below.

[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 I: Comparison of this paper with existing surveys. Legends: $\surd$= discussed, $\times$= not discussed, $\approx$ = 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	$\surd$	Segmentation of medical images and omic data	$\times$	$\approx$	$\surd$	$\times$	$\times$	$\times$	$\approx$
Adadi et al. [12]	2018	$\approx$	Trends of explainable approaches	$\surd$	$\approx$	$\surd$	$\times$	$\times$	$\surd$	$\approx$
Tjoa et al. [13]	2019	$\surd$	Categorization of XAI methods and partially discussed application for healthcare.	$\approx$	$\approx$	$\surd$	$\times$	$\times$	$\surd$	$\surd$
Singh et al. [14]	2020	$\surd$	Detection and predict--ion of disease using medical imaging.	$\times$	$\surd$	$\surd$	$\times$	$\times$	$\times$	$\approx$
Char et al. [15]	2020	$\surd$	Identification of ethical problems for healthcare application.	$\surd$	$\approx$	$\approx$	$\times$	$\surd$	$\surd$	$\surd$
Adadi et al. [16]	2020	$\surd$	Partially discussed XML applications for healthcare	$\times$	$\approx$	$\surd$	$\times$	$\times$	$\times$	$\surd$
This paper	2021	$\surd$	All most all healthcare applications	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$

TABLE II: 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. 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 III). This is also the case for XML for healthcare applications. There is also the need for defining terms like feature relevance, feature importance, saliency maps, heatmaps, etc. As there is no consistency in the use of these terms.

B. 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 standardized yet 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 making the decision understandable to the non-technical end-users like patients, which also equally important to gain their trust in these applications.

C. Lack of Standardized Requirements for XML

Researchers have developed some initial guidelines about the requirements of a good XML model, however, these guidelines are generic. Requirements for explaining the decisions of animal image tagging will be different from the medical image tagging. The current field of medical XML lacks requirement guidelines about 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].

D. What Clinicians Want: Accuracy vs. Explainability

It is a well-known problem of XML that the simple ML is easy to explain with less accurate results and complex DL produce more accurate results with fewer explanations due to their complex non-linear structures. This challenge is not

limited to the healthcare XML. However, due to the multi-dimensional nature of medical data, DL algorithms are crucial to avoid for precise results which 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 taking the preference of the end-user into consideration.

E. What and Hows of the Explained Results

Feature maps of medical image data produce reconstructed images, which consist of highlighted areas that represent human-understandable features. However, the answers of 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, 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).

F. 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 comparison of generated explanations using different methods. For examples, to explain the detection of glioma tumors, various XML techniques have been implemented (discussed in Section IV-D) but no one compared which method produced the better explanation of the tumor detection. Similarly, for each healthcare application, clinicians may need different measures for the validation of 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].

G. Lack of Theoretical Understanding

Applied DL for medical applications lack 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 present in medical data. However, due to the unavailability of sound fundamental laws and models, the explanation of DL cannot be produced 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 is yet a major challenge in developing trustworthiness [25].

H. Lack of Causality

DL is designed to produce precise results by learning the hidden patterns that generate data. The problem arises with the application of these techniques for healthcare tasks where decisions should be based on the causal links. However, DL is not efficient to infer causal relations between decisions and data. This 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 of causal relationship between images and their annotations [27].

I. Ethical Constraints

For gaining 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 exactitude and comprehensiveness of explanations. Due to the unavailability of these measures, the application of XML in clinical settings may have adverse affects. 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].

J. Security Challenges

Notwithstanding the state of the art performance of ML, particularly, 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, there are many security challenges that hinder the practical 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. As the excellent performance of a 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 the clinicians and patients.

III. Explainable & Interpretable ML

XML is a research field first explored by Van et al. in 2004 [33]. They described that their system has the ability to explain the behavior of AI algorithms in simulation games application. Although 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 [34]. With increasing employment of AI/ML methods in industry, medicine, education, and defense systems, the explanation

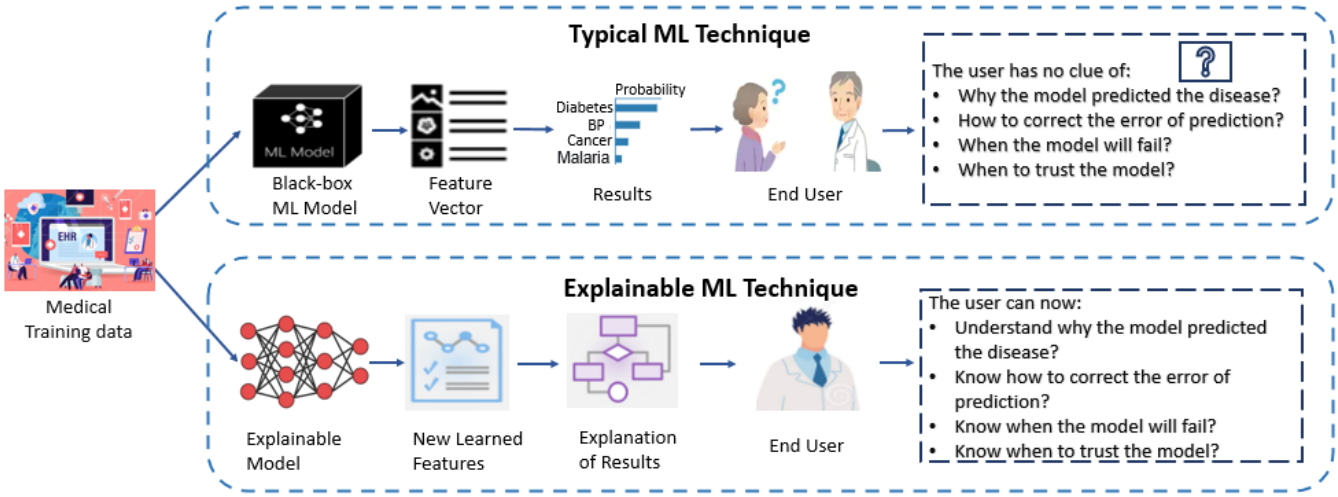

Fig. 2: Depiction of how the explainable ML technique is different from the typical ML technique.

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 puts social and ethical obligation for explainable decisions.

A. Explainable Vs Interpretable ML

In the research area of ML and AI, explainability and interpretability are often used synonymously. These terms are very closely intertwined it is worth noting the differences. Interpretability is about the understanding of the causality of learned features and the effects of causes within a system. One can also understand it as the extent to which humans can predict how the algorithm or model behaves with a slight change in input or algorithmic parameters. On the other hand, explainability is to explain the internal mechanics of a ML system in human-understandable terms. It is facile to overlook the subtle difference between these two terms. However, one can remember it like this: interpretability is being able to understand the mechanics of an algorithm without necessarily knowing why. Explainability is being able to explain what is transpiring.

1) Definitions 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 their 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 towards 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.
- *Interpretability*: Miller et al. [35] defined interpretability as the extent to which humans can understand the source of the resolution while interpretability in ML is defined by Kim et al. [36] as the extent to which humans can continuously predict results from the model. Thus in ML, a system will be more interpretable if the person can easily predict and resolve the model behavior. Similarly, a model is considered more interpretable than another if it is easy to predict the outcome of the prior than the latter. However, recently Doshi-Velez et al. [37] explain interpretability as a capacity to disclose or to introduce in justifiable terms to a human. Molnar et al. [38] defines interpretable ML as the strategies and the models used to explain the prediction of the outcome in a humanly comprehensible way. So interpretability is all about how a person understands and infers from the model just by looking at it.

From now on in the paper, we will use only the explainability word because in literature interpretable and explainable methods are described as a single concept. 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 the 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 of 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.

B. 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 explaining method is used to design the explainable models by reducing the complexity of the ML. Another aspect is, inherently or intrinsically XML are the ML methods that are explainable because of their simplistic architectures.
- *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 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 [39]. 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 [40]. White-box explanation mostly provides the model-specific explanation. The black-box methodology provides local explanations of a model for a prediction [41]. However, this mechanism lacks in describing all representations learned by the model.

C. Need of XML for Healthcare

Explanation of results are not only necessary for financial gains and ethical challenges but are 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.

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 of biased information leads to the life-risking results in the medical domain. 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 [42].

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, will alter the model slightly. 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 and can limit the understanding of end user in the real-time clinical practice. It is called the Rashomon effect [43]. 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) To enhance clinical use of ML: With the availability of an enormous amount of medical data and advanced ML techniques, research and publication 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 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 [44].

4) To facilitate end-users: ML and XML algorithms are designed to aid the medical staff, not for replacing the medical experts [45]. Medical-related decisions and their explanations have a direct influence on the results of treatment and 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 [46], and health monitoring systems [47].

D. 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.

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 explanation of data. We can take the work of Caruana et al. [48] as an example. They built classifiers for pneumonia patients to classify them as

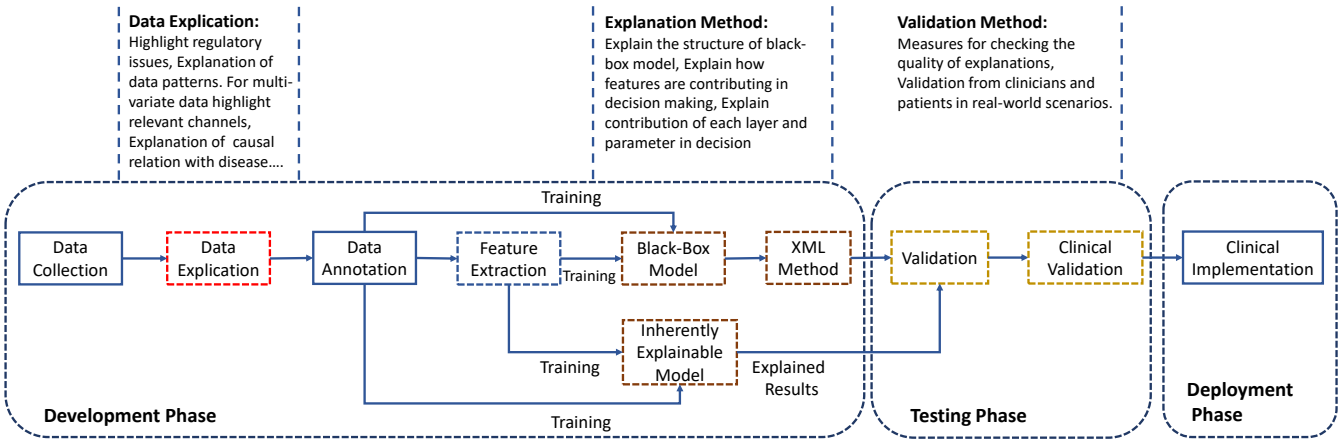

Fig. 3: The pipeline for explaining the black-box models.

high or low risk of 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. When they further investigated, they came to know that the asthma patients, when admitted due to pneumonia, were provided more timely treatment than the patients without asthma, which is why incurring a survival success.

Another example can be the data of patients who were denied to take medical care because of no health insurance. If ML learns from that data, it will generate biased results. Similar is the case for data leakage, which can mislead the model learning and testing [49]. To avoid these problems, researchers need to develop a data explanation method that interrogates all dependencies of the target on acquired data.

2) Explaining the structure of black-box: The problem of explaining the black-box can be further divided into two categories. First category is explaining the logic of the black-box in a human-understandable way (model-based explanations), and the second category is explaining the input-output relevance used by the model to make decisions (explanation of results) [50]. The model-based explanation methods are well developed and implemented for healthcare applications (further discussed in Sec IV-D). 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 forest. However, many black-box models require other models that mimic their work for the explanation.

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. In contrary to the local explanation, for a single patient, a global explanation is required for generalization purposes.

4) Measuring the effectiveness of explanations: Due to the non-monolithic concept and subjective nature of explainability, 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 approaches opted by the healthcare researchers for the evaluation are the following. These approaches are not limited to the evaluation of healthcare XML. Figure 3 is the depiction of these steps required for explaining the black-box models.

a) 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.

b) 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 more number of testers (laypersons) are easily available. This evaluation approach was applied by Mohseni et al. for evaluating the explanations using image and text data.

c) 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.

E. Characteristics of XML for Healthcare

The goal is to explain the decisions of the ML methods applied for detection and prediction of diseases, and to achieve this goal research community relies on 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. [51] 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 healthcare domain. The only limitation is that there is no definite method to calculate these properties. Figure 4 depicts these properties.

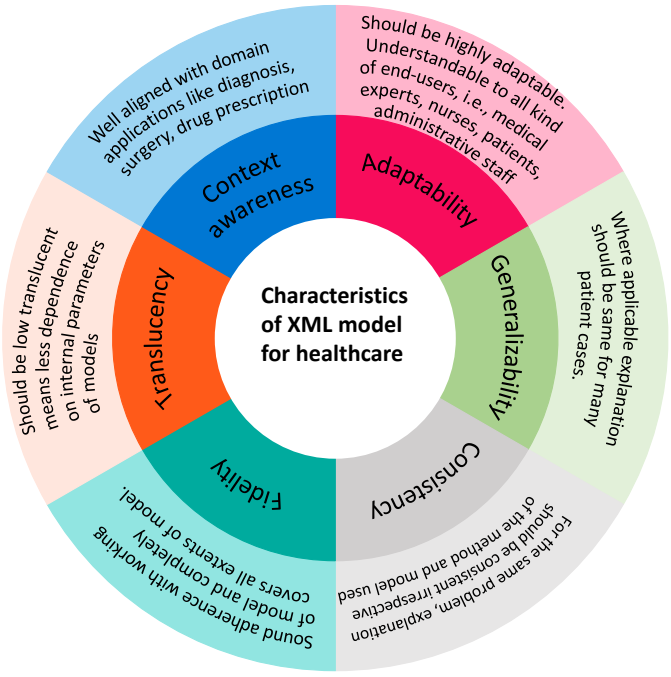

Fig. 4: Illustration of characteristics of XML model for healthcare applications.

- *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, in the from of mathematical formulation, or a natural text language. For end-users of the medical field, 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. Low translucency has the advantage of more compact explanation results. To get a generalized model for clinical use low translucency is desirable and for a patient-specific application, the explanation should be more detailed and accurate that can lead towards 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 thus can not be applicable for other models.

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

F. Explaining DL Techniques

With advanced DL techniques, deep neural networks (DNNs), convolutional neural networks (CNNs), and recurrent neural networks are widely employed for healthcare applications, i.e., epilepsy seizure prediction [52], segmentation of brain tumor [53], Alzheimer detection [54], genomics [55], and medical prescriptions [56]. These techniques are precise in terms of performance but their decisions are difficult to explain because of their complex model architecture. Saliency methods and feature attribution (FA) are the two broad categories of methods applied for explanation of DL 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 the 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 the model and visualize these gradient as heatmaps. Several techniques have been proposed to improve the visualization quality of these heatmaps, i.e., SmoothGrad [57], class activation maps (CAM) [58], and gradient weighted class activation mapping (GradCAM) [59]. The signal method is used to highlight the patterns of data that activate the neurons of higher layers. This can be done by back-propagating a signal from the last layer of the network to the input layer. DeConvNet [39], Guided BackProp [60], and PatternNet [61] 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 [62] and integrated gradients [63] are famous attribution methods.

A local interpretable model-agnostic explanation (LIME) technique was proposed by Marco et al. [64] to address the issue of explaining 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 more noisy compared to the heatmaps generated using LRP [65]. Samek et al. also provided a brief survey on the post hoc methods for explaining the DL models with theoretical background of each methods [66].

IV. Safe, Robust, and Trustworthy ML for Healthcare

The lack of transparency of ML techniques, particularly DL is yet another key challenge that hinders the practical deployment of these methods into critical applications like healthcare. It is crucial for a typical ML/DL empowered healthcare system to be fully trusted by both the clinicians and patients for getting the real impact of such systems. Moreover, unlike other domains, healthcare has unique challenges that involve 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 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.

A. Principles of Trustworthy AI for Healthcare

The literature refers to two sets of popular principles that have been outlined by the Organisation for Economic Co-operation and Development (OECD) [67] and European Commission's AI High-Level Expert Group (HLEG) [68] that can be used for sustaining trust and trustworthiness in AI. 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 for building sustainable trustworthy AI systems for healthcare and respect for 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 majority of 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 that respect ethical values along with focusing on the legal and regulatory considerations.

B. 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 kinds of 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 we refer the interested readers to our recent detailed work on the security and robustness 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 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, and privacy breaching attacks, etc.

1) Robustness to Security Attacks:

a) 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, an adversarial attack for manipulating CT scans in an active hospital network is presented in [69]. Similarly, the threat of adversarial ML is highlighted for different medical applications, e.g., medical image classification [31], medical image segmentation [70], and as well as using time series medical signals [71]. In [72], the authors argued that adversarial attacks in medical images are due to the noise inherent in the technology of their formation. Robustness to adversarial attacks can be a road map towards developing safe and trustworthy ML-based healthcare applications. The adversarial robustness can be defined as the survivability of ML-based systems to 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 auxiliary model, a taxonomy of such methods can be found in [73].

b) Privacy Preserving ML: Preserving the privacy of the 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 refers to 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 at training and at inference time. 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, i.e., learning about the confidential information and malicious use of data [32]. Similarly, privacy information can also be unveiled by querying the deployed ML model. Therefore, the development of appropriate defense strategies to withstand privacy attacks is crucial to ensure safe and trustworthy ML in healthcare applications. In this regard, different techniques can be use for

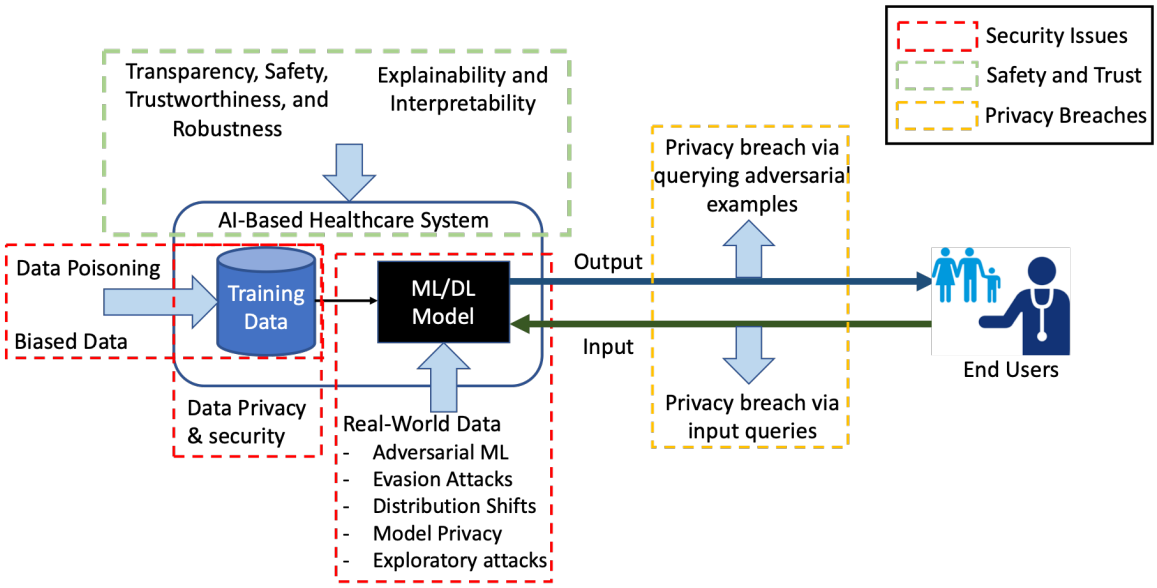

Fig. 5: An abstraction of safe, robust, and trustworthy ML for healthcare applications.

preserving privacy, e.g., using cryptographic approaches (like homomorphic encryption [74], and multi-party computation [75], etc.), differential privacy [76], and federated learning [77]. In addition, hybrid approaches can also be developed, for instance, the use of differential privacy in federated learning settings is proposed in [78].

2) Robustness to Distributional Shifts and Data Imperfections: Data distribution shifts (which refers to the divergence of training and testing data) is 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. 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. In addition, the real-world data contains imperfections and is imbalanced, e.g., for example, incomplete data due to missing observations or variables and uneven distribution of samples across different classes, respectively. These data imperfections will eventually result in biased training and will increase the false positives and negatives. Therefore, to build trust 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. The difference in data distributions can be leveraged to craft adversarial examples [80]. Moreover, adversarial robustness is closely related to robustness to certain kinds of distributional shifts. In this context, the literature recommends that future adversarial defenses should consider evaluating the robustness of their methods to distributional shifts [81].

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

It is worth noting that one has to pay a cost for developing explainable, robust, trustworthy, and accurate ML/DL models, as shown in Figure 6. Robust models pay the cost of accuracy and can be more explainable and interpretable as compared to the complex models having high accuracy with low explainability. Therefore, the higher the accuracy of the predictive model, the less explainable/interpretable it becomes. Moreover, it has been provably demonstrated that there exists a trade-off between adversarial robustness and the accuracy of the model even in a concrete simplistic setting [82]. In addition, the trade-off analysis between the accuracy and adversarial robustness of eighteen well known ImageNet classifiers with different metrics is presented in [83]. Furthermore, the authors suggested that there exists a clear trade-off between accuracy and robustness. 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.

D. Applications of XML Models for Trustworthy Healthcare

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 to trust and rely on ML/DL prediction algorithms. Moreover, the 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 impact of AI. According to new directions of the European general data protection regulation (GDPR), explainability and accountability are necessary for the application

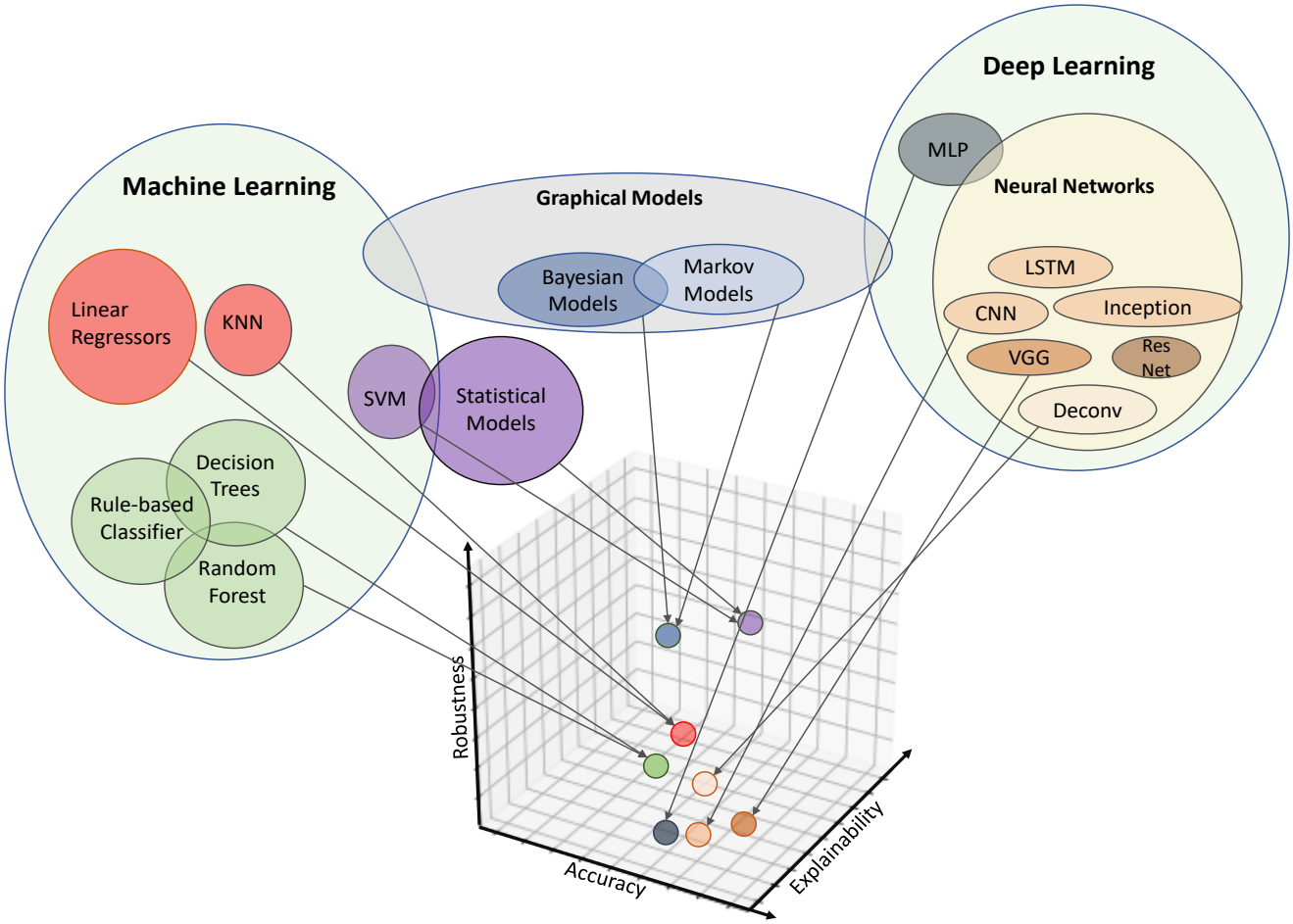

Fig. 6: Illustration of the trade-offs between accuracy, explainability, and robustness of intelligent models.

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 the predictions and can help in building the trust of end-users [14]. Explainable ML can be a potential step towards trustworthy ML by building trust of clinicians in AI system [85]. Below we described the application of XML models in the medical domain:

1) Intrinsic XML Models: As we described earlier that intrinsic models are explainable due to their simple architecture and are understandable by themselves. The following are applications of inherently XML models and Table III is the summary of model-intrinsic explanation methods for healthcare applications.

a) 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 DTs are not feasible for producing human-understandable explanations. Dlaeen et al. [87] implemented DTs to predict Alzheimer's disease of 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 nodes. However, they have not evaluated the quality of the explanation.

A DT shows considerable change in output with a small perturbation in the input, which considerably affects the performance of these models. By averaging large numbers of DTs for constructing ensemble DTs is a solution for improving the performance. Gibbons et al. [86] used a hybrid approach for incorporating both the benefit of individual DT and the efficiency of ensemble DTs. They worked on the development of a computerized adaptive diagnostic (CAD) system for the diagnosis of major depressive disorder (MDD) using random forest and DTs. They worked with the data of 656 patients and achieved a sensitivity of 95% and specificity of 87%. Suresh et al. [88] proposed the use of radial basis function (RBF) network and DTs for the detection of lesions in mammograms. DT algorithm learns the suitable attributes of data in a top-down search manner. They selected the best attributes by constructing and evaluating the different structures of DTs. They also compared their algorithm with k-nearest neighbors (K-NN), support vector machine (SVM), and naive Bayes classifier and concluded that DTs outperformed all these classifiers. Generalization of algorithms is required for clinical employment. However, DTs have a secondary generalization

TABLE III: 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. [46]		Question classification in health care	Coarse and fine-grained classes from cloud questionnaire
RF	2017	Wang et al. [90]	An ensemble of large numbers of DTs, used mainly for regression or classification	Epilepsy detection	EEG signals
	2019	Byeon et al. [91]		Alzheimer's patients depression detection	Social demographic factors, Health status, Behaviors, Living style, Economic activity
	2019	Kaur et al. [47]		Healthcare monitoring system	Breast cancer, Diabetes, Heart disease, Spect-heart, Thyroid, Surgery, Dermatology, Liver disorder
	2020	Simsekler et al. [92]		Evaluation of patient safety culture	Continuous and Categorical variables for patient safety
	2020	Iwendi et al. [93]		Covid death and recovery rate detection	Categorical variables in dataset such as fatigue ,fever ,cough.
GAM	2015	Caruana et al. [94]	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. [95]		Effect of age and diagnosis -specific cohort of HIV patients on psychosocial activities and behavioral activities	Various attributes related to psychosocial and behavioural outcomes
	2020	Dastoorpoor et al. [96]		Effect of air pollution on pregnancy	Various air pollutants data
	2020	Yang et al. [97]		Study air pollution effect on TB cases	Pulmonary TB and air pollutants data

property that is why these algorithms are not appropriate for healthcare applications.

b) 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 textual format. The other differences are the order of rules (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] implemented the association rule technique using 23 attributes of cardiovascular data for the detection of heart diseases. They generated the rules which map the attributes to classes to identify features that are provoking the disease. They used confidence, lift, and support as parameters for the generation of rules. For the validation of generated rules, they implemented the accuracy metric.

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. [46] implemented a rule-based algorithm for the question classification system (QCS). They extracted reules after the preprocessing of 427 health-based questions to classify these into 9 question types. For the validation of extracted rules, they measured the accuracy of the algorithm. Their rule-based algorithm classified 345 questions correctly and achieved 80.7% accuracy.

c) Random Forest (RF): This algorithm is an ensemble of large numbers of DTs, broadly used for regression or classification. This algorithm generates several DTs, each DT in RF perform the classification. The final output of the algorithm is measured based on the most occurring class. Wang et al. [90]. implemented an 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 the accuracy of 98.6% for two-class, 96% for three-class, and 82.6% for five-class classification experiments.

Saftey of a patient is necessary for ensuring the quality of medical facilities provided by a healthcare unit. The way to provide safety to the patients is a patient safety culture of a hospital or clinic. Simsekler et al. [92] 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. They studied 12 variables of safety culture and using an RF algorithm they checked the importance of each variable for the safety of patients. They measured mean absolute percentage error (MAPE), mean absolute error (MAE), and mean square error (MSE) for checking the quality of explanation of safety variables.

In recent work, Iwendi et al. [93] used an RF model with the AdaBoost algorithm to the severity of Covid-19, death, or recovery rate for a patient. Due to their simple and self-explanatory structure, these models are also implemented for predicting the depression of Alzheimer's patients [91], healthcare monitoring systems [47], prediction of medical

expenditures [98].

d) General Additive Model (GAM): ML regression models produce predictions by adding weighted features. GAM modeled the output as the weighted sum of random nonlinear functions of data features. A combination of spline functions is used to approximate these non-linear functions. GAM is extensively used in health-related environmental research [96], pneumonia risk prediction [94], research on the distribution of species [99], and the effect of age and a diagnosis-specific cohort of HIV patients on psychosocial and behavioral activities [95]. Yang et al. [97] studied how tuberculosis (TB) cases changed with air pollution in the Wulumuqi. They obtained the air quality and TB patients data of slightly larger than two years duration. They found that $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, and O_3 were the dominant pollutants in the air data. They implemented GAM to study the relation between these pollutants and the number of TB cases. They assumed that the number of patients followed the Poisson distribution. 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.

2) Model-Agnostic Explainability: These explanations are flexible in terms of applications of models and representation. Table IV Summaries the model-agnostic explanation methods for healthcare applications and following are the details of these methods for explaining healthcare decisions:

a) 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. [100] 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.

b) 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. [101] detected the acute proximal femoral fractures in elderly people using radiographic data. They detected the fractures using VGG16 and used CAM to localize the fractures. Aayush Kumar et al. [102] classified the malaria cells by proposing a mosquito-net and explained the decision using the GradCAM (a variant of CAM) and CAM. Sebastian et al. [103] used CAM to evaluate the errors of their CNN model proposed for the multiclass labeling of ECG signals. Pereira et al. [104] explained the brain tumor grading decision of a proposed CNN classifier using CAM. Irvin et al. [105] used the GradCAM to provide the visual explanation of active pleural effusion areas of chest radiograph which were indicated by the CNN model. Izadyyazdanabadi et al. [106] integrated multiscale activation maps (MLCAM) with the CNN model to locate the attributes of glioma tumors.

c) Layer-wise Relevance Propagation (LRP): This technique backpropagates the output decision to the input layer to estimate the relevance of each attribute. Yang et al. [107] proposed the use of LRP to select the features with high relevance for predicting the decision of therapy of patients. They also evaluated the quality of explanation from the expert clinicians. Chlebus et al. [108] implemented an LRP algorithm for explaining the decisions of semantic networks used for segmenting the liver tumors. They highlighted MRI segments that were most relevant for the classification of tumors. Böhle et al. [109] implemented LRP and guided backpropagation (GB) for explaining the decisions of CNN that they used for the classification of Alzheimer's disease. They evaluated the quality of generated explanations of both techniques by measuring Atlas-based evaluation metrics. They concluded that LRP generates more relevant explanations by describing why any individual patient has the disease. Taeho Jo et al. [110] implemented LRP to highlight the areas of tau positron emission tomography (PET) that highly contribute to the classification of Alzheimer's disease using 3D-CNN.

d) 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 the 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 in a correct prediction and negative values show less or no impact in prediction. Sousa et al. [111] implemented LIME for generating the explanation of how CNN and VGG16 detect the metastases from the histology WSI patches. They evaluated the explanations by cross-checking the highlighted areas with expert pathologists and showed that the algorithm used the human-like approach for explanation generation, at least for this application.

Zafar et al. [116] pointed out the problem of instability of generated explanations due to perturbation addition and random feature selection in the medical computer-aided diagnosis (CAD) systems. They proposed the use of hierarchical clustering (HC) and KNN to group the data and for the selection of relevant feature clusters. They named the proposed algorithm deterministic LIME (DLIME). They implemented the DLIME algorithm for explaining the decisions of three medical domains, i.e., breast cancer, liver disease, hepatitis detection. They concluded that the DLIME shows better results than LIME. Kitamura et al. [112] detected diabetic nephropathy (DN) using a CNN network from the immunofluorescent images. They analyzed the decision of CNN using LIME and described that CNN learned the patterns of peripheral lesion of DN glomeruli for DN detection.

e) Deep Learning Important FeaTures (DeepLIFT): DeepLIFT provides the explanations of black box models by

TABLE IV: 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. [100]	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. [101]	Highlight the class relevant areas of input data.	VGG16	Fracture detection	X-Rays
	2020	Sebastian et al. [103]		CNN	ECG classification	ECG signals
	2018	Pereira et al. [104]		CNN	Grading of brain tumor	MRI
GradCAM	2019	Irvin et al. [105]	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. [102]		Mosquito-net	Malaria detection	Blood samples
MLCAM	2018	Izadyyazdanabadi et al.[106]	Generates the maps of discriminating features of data.	CNN	Brain tumor detection	MRI
LRP	2018	Yang et al. [107]	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. [108]		Semantic segmentation network	Liver tumor classification	MRI
	2019	Böhle et al. [109]		CNN	Alzheimer's disease classification	MRI
	2020	Taeho Jo et al. [110]		3D-CNN	Alzheimer's disease classification	PET
LIME	2019	Sousa et al. [111]	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. [112]		CNN	detection of diabetic nephropathy	immunofluorescent images
DeepLIFT	2020	Yang et al. [100]	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. [113]	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. [114]	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 [115]	Generate explanations by maximize the activation of neurons tweaking the input.	CNN	Detection of hip implant misplacement.	X-rays

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. The selection of reference value is based on the problem at hand. Sharma et al. [117] implemented 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 that revealed diabetes genetic risk factors are identifiable using DL techniques.

f) 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. [113] 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 described that the 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. [118] 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.

g) 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. [119] implemented 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 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.

h) 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. While like vanilla backpropagation the algorithm limits itself to positive inputs. Theerasarn et al. [114] proposed 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. The GBP produced best explanations among all other methods. DeepLIFT and SHAP produced 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 dice coefficient measure.

i) 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 deep neural network. IG can be implemented for any kind of model, i.e., image. This algorithm works on two axioms sensitivity and implementation variance. In the drug development classification of toxic and non-toxic drugs is not enough. To resolve the problem of toxic drugs a chemist need the structural element which is causing the problem. Preuer et al. [120] demonstrated that IG can identify these elements from the classified drug using CNN.

j) 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 [115] trained the CNN model for orthopedic application in observing hip implant misplacement using X-rays dataset. The explainability of this CNN model at a lower level is done using AM. AM is used to visualize the classification of the model.

k) Deep Hierarchical Semantic Convolutional Neural Network (HSCNN): Shen et al. [121] proposed an interpretable deep network 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 provide the explanation of 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 out-performed the CNN with a high prediction score and explainable results.

l) Patient2Vec (P2V): The extensive use of electronic health record (EHR) in the clinical system provides a large amount of data for healthcare. Jinghe et al. [122] 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 increase the explainability of these models. They used the proposed model to explain the importance of each diagnostic product, medication, and treatment procedure.

m) Evidence Activation Mapping (EMANet): Lia et al. [123] proposed a CNN based model for glaucoma diagnosis named as EAMNet. The proposed architecture not only able to detect the diseases but also show transparency by highlighting the affected area detected by the system. The system consists of CNN as the backbone for feature extraction and uses multi-layers average pooling (M-LAP) to overcome the gap problem between the information interpretability and localization while evidence activation mapping is used for the verification.

V. 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 [124]. However, incorporating the ML techniques into the healthcare system also raises serious ethical challenges and complex questions that need to be seriously considered in order to make a robust and well-balanced assessment of possible benefits and expected harms [125].

For the purpose of setting 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 focus on the so-called "principlism" and the widely known four principles, namely beneficence, non-maleficence, autonomy, and justice [126]. It is to be noted here that explicability is a newly proposed principle, within the particular AI context, which has the same meaning outlined above in this paper [127]. 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.

A. 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 most 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 the modern bioethical discussions [128]. World religions like Judaism, Christianity, and Islam also brought their own 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

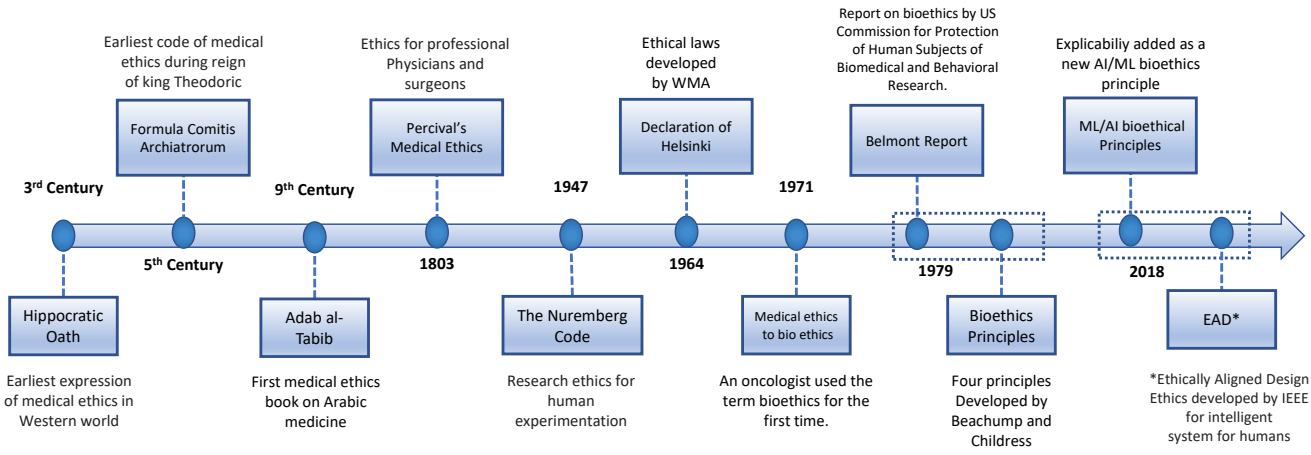

Fig. 7: History of the development of bioethics over the time.

one of the most popular works on medical ethics, entitled *Adab al-Tabib* (Ethics of the physician) [129],[130]. Back in 1803, the physician, Thomas Percival, published a report on the necessities and expectations of medical staff to assure ethical medical practice [131]. This code of medical ethics was adapted for the first time in 1847 [132], 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 the idea of introducing a new discipline, which he named Bioethics, to address the basic problems of human flourishing by creating interdisciplinary discourse between the two cultures of humanities and sciences [133].

One of the important milestones in modern bioethics is the so-called "Belmont Report[4]; 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. Exactly 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 [126]. 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 [134]. Besides the famous principlist approach to bioethics, there are other important approaches in modern bioethics, including virtue ethics, casuistry and 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 [135]. Figure 7 illustrates the history of 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 on human subjects also resulted in issuing a number of 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 [136]. 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 [137]. Because of 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 the ethical questions, which are specific to the interplay of AI/ML and bioethics. These key works will be reviewed below.

[4]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

B. 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 were more focused 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 [138], 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" [139]. *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 [140]. Similar issues were also examined in the edited volume *Smart Health: Open Problems and Future Challenges*, published in 2015 [141].

One of the main contributors to the discourse on AI-driven healthcare is the American cardiologist and professor of genomics, Eric Topol, who is a well-known high-tech enthusiast. Between 2012 and 2019, Topol wrote what can be called a trio on the revolutionization of medicine by making use of 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 [142], and *The Patient Will See You Now: The Future of Medicine Is in Your Hands*, published in 2015 [143], 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 drive of the promised revolutionization of medicine [144]. He also outlined his ideas in this area in an article published in 2019 in *Nature Medicine* [145]. 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 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 [146].

In 2020, *The American Journal of Bioethics* published a thematic issue entitled "Planning for the known unknown: Machine learning for human healthcare systems" [147]. The contributions to this thematic issue, made by a number of interdisciplinary experts, provided useful frameworks that are meant to help future researchers to 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 art in this field and outlined what is achievable in the 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" [148]. 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" [149].

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" [150].

C. Main Ethical Questions and Challenges

The possible integration of AI/ML technologies into healthcare is characterized by the promise of potentials that will be of great benefits to many involved stakeholders, including patients, and physicians. Thus, the great potential is optimizing the overall quality, efficiency and access in healthcare system [147]. On the other hand, these applications concurrently raise various ethical challenges and complex questions that need to be seriously examined. Below, we give an analytical and systematic overview of these issues.

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, the AI technologies are highly dependent on huge 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 [147].

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:

a) Imbalance Datasets: Imbalance class data is one of the common data-related problems that occur 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 study was later accused of violating the principles of transparency and reproducibility [151], [152].

b) 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 that cause the data bias. For example, researchers have shown that the model predicts that black people have a strong immunity and are healthier as compared to equally sick white people because of the fact that less money is spent on the healthcare facilities of black people in comparison to white people [153]. Dependence of models' learning on the 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 [154]. 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 [155]. Similar is the case for geographical biases, where fewer healthcare facilities are provided in rural areas and under-developing countries [156]. 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.

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 [157].

In the healthcare context, privacy is defined as keeping the information of patients protected from unauthorized access. However, ML algorithms requires 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 by 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. [144], [146]. To make the situation more complex, conducting proper analysis for 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 who should have access to it. Various factors can put people's privacy at risk, and we highlight here two of them:

a) 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 patient's routines which dishonors the patients' privacy. An un-protected 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 [158].

b) 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 facility 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 [159]. 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 about processing the data. Because of the particular significance of addressing the data-related concerns, different countries have developed policies and laws [160].

3) Informed Consent: As outlined above, the respect for persons and autonomy are among the widely-agreed upon principles in modern bioethics. Obtaining an informed consent from the patient before exposing him/her to any medical intervention is one of the practical applications of these principles [135]. 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 get 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 general data protection regulation (GDPR) has introduced rules for the decisions and methods based on data-driven approaches to provide an ethical framework [161]. 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, as discussed in detail in Section III, represent an ethical requirement in order to facilitate obtaining a 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 for using ML-based interventions as the only available tool to treat an incurable and life-threatening disease will not be the same as consenting for 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 [160]. Lack of ethical guidelines and fewer or no proper training of such surveillance programs raises ethical concerns [162]. Due to the availability of implantable devices, it is now possible to monitor the patients without their consent. Despite all these ethical issues, Lee et al. [163] 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 [164]. Therefore, the explicit and targeted consent of patients is required for the data collection through IoT and for ML/DL empowered personalized medical systems [165].

4) Care Ethics: As mentioned above, modern bioethics have other approaches besides principlism. Some of these non-principlist approaches can provide fresh insights for 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 [135], 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 the AI systems into healthcare will make the healthcare profession more humane, by improving the physician-patient relationship [144], [148].

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 [166]. 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 [167]. Figure 8 illustrates the overview of the explainable, trustworthy, and ethical ML methods used for healthcare in literature.

VI. Potential Pitfalls

The recent advancements in the 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 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 the state-of-the-art performance in many medical image analysis tasks [168]. 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 being not able to generalize out of distribution samples, etc. Thus raising concerns about robustness and trustworthiness of M methods is of high importance, 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.

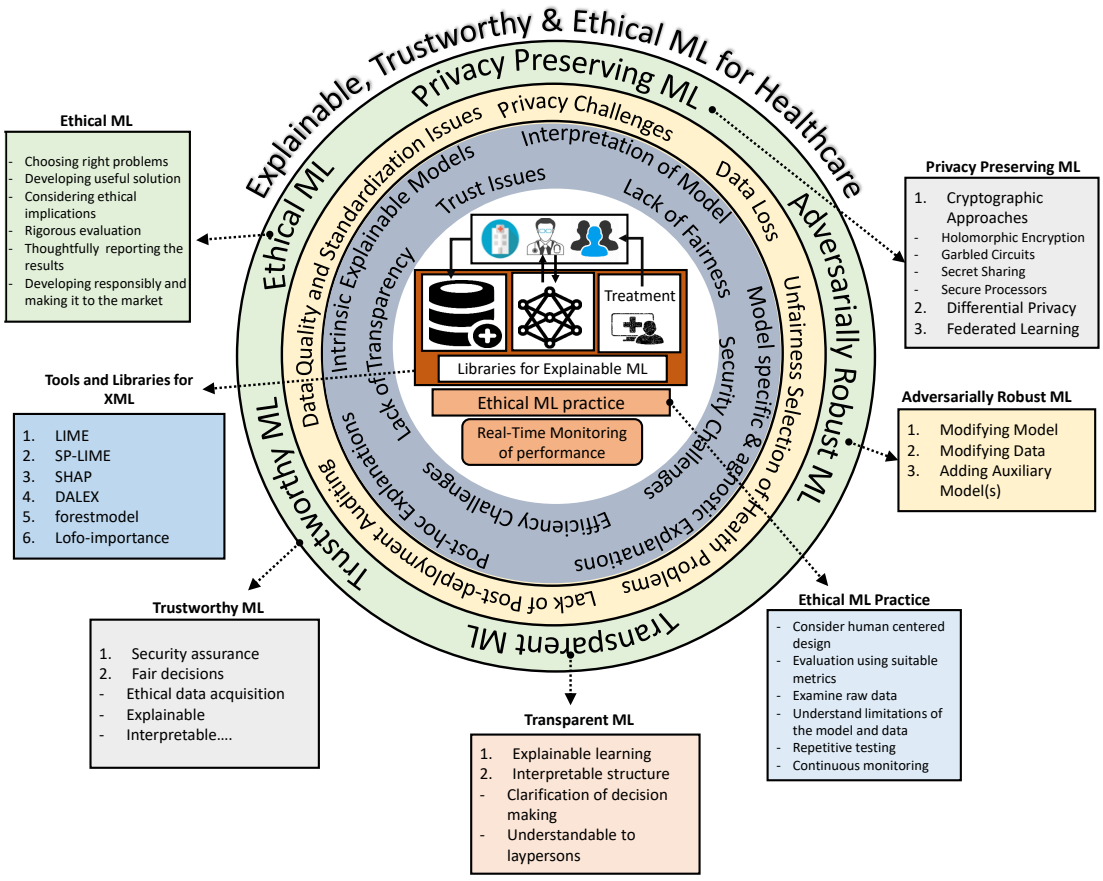

Fig. 8: Overview of the explainable, trustworthy, and ethical ML models for healthcare.

A. 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 disease type 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. [169] demonstrated that a small change in the input sample caused large 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 model's outputs i.e., model might prefer a specific class as compared to other classes and this bias might influence the explanations of the model [170].

B. 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 the medical experts. Weerts et al. [171] 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 to class probability did not improve the decision-making. Similarly, Mohseni et al. [172] 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 10 humans 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.

C. Dependence on Data and Model

The literature suggests that explanations generated by some gradient-based methods are dependent on the model architecture and data generation procedure [173]. As these explanations are dependent on the choice of reference point and a slight change in the reference point of gradient will significantly change the explanation thus causing confusion that will eventually lead to misleading results or interpretation.

D. Accountability Attribution

There is no doubt that the deployment of ML for clinical practice will aid the clinicians. However, it is not clear yet that who will be responsible in case an algorithm shows

wrong outputs? Whether the clinicians will be responsible because they are the ones for making final decisions or institutes forcing clinicians to rely on the decisions of ML? Researchers developing the algorithms can also be responsible for bad decisions [174]. 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" [175].

E. 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 [176]. 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 to accurately diagnose 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 [176]. Therefore, clinically relevant metrics should be developed for evaluating 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 [177].

VII. 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.

A. 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 [176]. Moreover, it is very important 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 [178]. These biases are hard to undo and their elimination have unintended consequences on the results [179]. The presence of these subtle biases in medical data decreases model reliability, especially when they are not corrected during model development [180], [181]. 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).

B. 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 very important to understand that how these explanations are presented to them, i.e., explanations should be completely understandable to clinicians. The representation of the explanations needs 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.

C. Generalized Explanations

As discussed in Section VI, 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 same disease. This issue should be resolved by developing robust, efficient, and generalize explainable models. As we discussed in Section III, that the explanations of the DL models are model-specific in nature, therefore, it is also required to develop inherently explainable and generalize explainable methods for the DL algorithms in future.

D. Adversarially Robust ML

To attain the 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 [73]. 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.

E. 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 need to be effectively addressed. In this regard, a few studies suggested involving all types of stakeholders into the ML/DL method development process that may include clinicians, policymakers, data scientists, ML researchers, and hospital

staff, to name a few [182], [176]. Such an interdisciplinary development workforce will enable collaboration between the knowledge experts (i.e., clinicians and ML researchers) and healthcare service providers that will eventually improve productivity and outcomes.

VIII. 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 deployable and explainable ML methods. We also elaborate upon different security, safety, robustness, and ethical challenges which are the key barrier towards 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 developments.

IX. 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, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[2] C. Xiao, E. Choi, and 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*, vol. 25, no. 10, pp. 1419–1428, 2018.

[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, and H. J. Aerts, "Deep learning for fully-automated localization and segmentation of rectal cancer on multiparametric MR," *Scientific reports*, vol. 7, no. 1, pp. 1–9, 2017.

[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*, vol. 11, no. 11, pp. 1654–1663, 2018.

[5] J.-G. Lee, S. Jun, Y.-W. Cho, H. Lee, G. B. Kim, J. B. Seo, and N. Kim, "Deep learning in medical imaging: general overview," *Korean journal of radiology*, vol. 18, no. 4, pp. 570–584, 2017.

[6] A. Qayyum, S. M. Anwar, M. Awais, and M. Majid, "Medical image retrieval using deep convolutional neural network," *Neurocomputing*, vol. 266, pp. 8–20, 2017.

[7] C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle, "Deep learning for computational biology," *Molecular systems biology*, vol. 12, no. 7, p. 878, 2016.

[8] E. Begoli, T. Bhattacharya, and D. Kusnezov, "The need for uncertainty quantification in machine-assisted medical decision making," *Nature Machine Intelligence*, vol. 1, no. 1, pp. 20–23, 2019.

[9] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable AI systems for the medical domain?" *arXiv preprint arXiv:1712.09923*, 2017.

[10] M. FAT, "Fairness, accountability, and transparency in machine learning," *Retrieved December*, vol. 24, p. 2018, 2018.

[11] D. Gunning, "Explainable artificial intelligence (XAI)," *Defense Advanced Research Projects Agency (DARPA), nd Web*, vol. 2, no. 2, 2017.

[12] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[13] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): towards medical XAI," *arXiv preprint arXiv:1907.07374*, 2019.

[14] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *arXiv preprint arXiv:2005.13799*, 2020.

[15] D. S. Char, M. D. Abràmoff, and C. Feudtner, "Identifying ethical considerations for machine learning healthcare applications," *The American Journal of Bioethics*, vol. 20, no. 11, pp. 7–17, 2020.

[16] A. Adadi and 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, and A. Bartz, "Using H2O driverless AI," 2017.

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

[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*, vol. 58, pp. 82–115, 2020.

[20] A. D. Jeffery, L. L. Novak, B. Kennedy, M. S. Dietrich, and L. C. Mion, "Participatory design of probability-based decision support tools for in-hospital nurses," *Journal of the American Medical Informatics Association*, vol. 24, no. 6, pp. 1102–1110, 2017.

[21] M. A. Ahmad, C. Eckert, and 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*, vol. 12, 2018.

[23] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

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

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

[26] M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, and R. Ranganath, "Opportunities in machine learning for healthcare," *arXiv preprint arXiv:1806.00388*, 2018.

[27] D. C. Castro, I. Walker, and B. Glocker, "Causality matters in medical imaging," *Nature Communications*, vol. 11, no. 1, pp. 1–10, 2020.

[28] L. Floridi, "Establishing the rules for building trustworthy AI," *Nature Machine Intelligence*, vol. 1, no. 6, pp. 261–262, 2019.

[29] S. R. Meikle, J. C. Matthews, V. J. Cunningham, D. L. Bailey, L. Livieratos, T. Jones, and P. Price, "Parametric image reconstruction using spectral analysis of pet projection data," *Physics in Medicine & Biology*, vol. 43, no. 3, p. 651, 1998.

[30] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[31] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287–1289, 2019.

[32] A. Qayyum, J. Qadir, M. Bilal, and A. Al-Fuqaha, "Secure and robust machine learning for healthcare: A survey," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 156–180, 2021.

[33] M. Van Lent, W. Fisher, and 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.

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

[35] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[36] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," *Advances in neural information processing systems*, vol. 29, pp. 2280–2288, 2016.

[37] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[38] C. Molnar, "Interpretable machine learning," 2019, https://christophm.github.io/interpretable-ml-book/.

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

[40] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," *arXiv preprint arXiv:1702.04595*, 2017.

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

[42] A. Bennetot, J.-L. Laurent, R. Chatila, and N. Díaz-Rodríguez, "Towards explainable neural-symbolic visual reasoning," in *NeSy Workshop IJCAI*, 2019.

[43] K. G. Heider, "The Rashomon effect: When ethnographers disagree," *American Anthropologist*, vol. 90, no. 1, pp. 73–81, 1988.

[44] A. Chander, R. Srinivasan, S. Chelian, J. Wang, and K. Uchino, "Working with beliefs: AI transparency in the enterprise." in *IUI Workshops*, 2018.

[45] P. Gupta, "Machine learning: The future of healthcare," *Harvard Science Review*, 2017.

[46] S. Agrawal and 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.

[47] P. Kaur, R. Kumar, and M. Kumar, "A healthcare monitoring system using random forest and internet of things (iot)," *Multimedia Tools and Applications*, vol. 78, no. 14, pp. 19 905–19 916, 2019.

[48] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and 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*, ser. KDD '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 1721–1730. [Online]. Available: https://doi.org/10.1145/2783258.2788613

[49] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," *ACM Trans. Knowl. Discov. Data*, vol. 6, no. 4, Dec. 2012. [Online]. Available: https://doi.org/10.1145/2382577.2382579

[50] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, Aug. 2018. [Online]. Available: https://doi.org/10.1145/3236009

[51] M. Robnik-Šikonja and M. Bohanec, "Perturbation-based explanations of prediction models," in *Human and machine learning*. Springer, 2018, pp. 159–175.

[52] K. Rasheed, A. Qayyum, J. Qadir, S. Sivathamboo, P. Kwan, L. Kuhlmann, T. O'Brien, and A. Razi, "Machine learning for predicting epileptic seizures using eeg signals: A review," *IEEE Reviews in Biomedical Engineering*, pp. 1–1, 2020.

[53] A. Işın, C. Direkoğlu, and M. Şah, "Review of MRI-based brain tumor image segmentation using deep learning methods," *Procedia Computer Science*, vol. 102, pp. 317–324, 2016.

[54] J. Islam and 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.

[55] J. Zou, M. Huss, A. Abid, P. Mohammadi, A. Torkamani, and A. Telenti, "A primer on deep learning in genomics," *Nature genetics*, vol. 51, no. 1, pp. 12–18, 2019.

[56] Z. Liang, G. Zhang, J. X. Huang, and Q. V. Hu, "Deep learning for healthcare decision making with emrs," in *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2014, pp. 556–559.

[57] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[58] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[59] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[60] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[61] P.-J. Kindermans, K. T. Schütt, M. Alber, K.-R. Müller, D. Erhan, B. Kim, and S. Dähne, "Learning how to explain neural networks: Patternnet and patternattribution," *arXiv preprint arXiv:1705.05598*, 2017.

[62] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep taylor decomposition," *Pattern Recognition*, vol. 65, pp. 211–222, 2017.

[63] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," *arXiv preprint arXiv:1703.01365*, 2017.

[64] M. T. Ribeiro, S. Singh, and 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.

[65] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv preprint arXiv:1708.08296*, 2017.

[66] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller, "Explaining deep neural networks and beyond: A review of methods and applications," *Proceedings of the IEEE*, vol. 109, no. 3, pp. 247–278, 2021.

[67] K. Yeung, "Recommendation of the council on artificial intelligence (oecd)," *International Legal Materials*, vol. 59, no. 1, pp. 27–34, 2020.

[68] R. Hamon, H. Junklewitz, and I. Sanchez, "Robustness and explainability of artificial intelligence," *Publications Office of the European Union*, 2020.

[69] Y. Mirsky, T. Mahler, I. Shelef, and 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.

[70] M. Paschali, S. Conjeti, F. Navarro, and N. Navab, "Generalizability vs. robustness: adversarial examples for medical imaging," *arXiv preprint arXiv:1804.00504*, 2018.

[71] X. Han, Y. Hu, L. Foschini, L. Chinitz, L. Jankelson, and R. Ranganath, "Deep learning models for electrocardiograms are susceptible to adversarial attack," *Nature medicine*, vol. 26, no. 3, pp. 360–363, 2020.

[72] A. Vatian, N. Gusarova, N. Dobrenko, S. Dudorov, N. Nigmatullin, A. Shalyto, and 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.

[73] A. Qayyum, I. Aneeqa, M. Usama, W. Iqbal, J. Qadir, Y. Elkhatib, and A. Al-Fuqaha, "Securing machine learning (ML) in the cloud: A systematic review of cloud ML security," *Frontiers in Big Data*, 2020.

[74] H. Takabi, E. Hesamifard, and M. Ghasemi, "Privacy preserving multi-party machine learning with homomorphic encryption," in *29th Annual Conference on Neural Information Processing Systems (NIPS)*, 2016.

[75] D. Bogdanov, L. Kamm, S. Laur, and 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*, vol. 15, no. 5, pp. 1427–1432, 2018.

[76] N. Phan, X. Wu, H. Hu, and 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.

[77] A. Qayyum, K. Ahmad, M. A. Ahsan, A. Al-Fuqaha, and J. Qadir, "Collaborative federated learning for healthcare: Multi-modal covid-19 diagnosis at the edge," *arXiv preprint arXiv:2101.07511*, 2021.

[78] O. Choudhury, A. Gkoulalas-Divanis, T. Salonidis, I. Sylla, Y. Park, G. Hsu, and A. Das, "Differential privacy-enabled federated learning for sensitive health data," *arXiv preprint arXiv:1910.02578*, 2019.

[79] C. S. Perone, P. Ballester, R. C. Barros, and J. Cohen-Adad, "Unsupervised domain adaptation for medical imaging segmentation with self-ensembling," *NeuroImage*, vol. 194, pp. 1–11, 2019.

[80] N. Papernot, P. McDaniel, A. Sinha, and M. Wellman, "Towards the science of security and privacy in machine learning," *arXiv preprint arXiv:1611.03814*, 2016.

[81] N. Ford, J. Gilmer, N. Carlini, and D. Cubuk, "Adversarial examples are a natural consequence of test error in noise," *arXiv preprint arXiv:1901.10513*, 2019.

[82] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," *stat*, vol. 1050, p. 11, 2018.

[83] D. Su, H. Zhang, H. Chen, J. Yi, P.-Y. Chen, and 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.

[84] J. Gao, X. Wang, Y. Wang, and X. Xie, "Explainable recommendation through attentive multi-view learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 3622–3629.

[85] A. F. Markus, J. A. Kors, and 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," *arXiv preprint arXiv:2007.15911*, 2020.

[86] R. D. Gibbons, G. Hooker, M. D. Finkelman, D. J. Weiss, P. A. Pilkonis, E. Frank, T. Moore, and D. J. Kupfer, "The cad-mdd: A computerized adaptive diagnostic screening tool for depression," *The Journal of clinical psychiatry*, vol. 74, no. 7, p. 669, 2013.

[87] A.-D. Dana and 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, and M. Balamurgan, "Hybridized neural network and decision tree based classifier for prognostic decision making in breast cancers," *Soft Computing*, vol. 24, no. 11, pp. 7947–7953, 2020.

[89] S. Khare and 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] G. Wang, Z. Deng, and K.-S. Choi, "Detection of epilepsy with electroencephalogram using rule-based classifiers," *Neurocomputing*, vol. 228, pp. 283–290, 2017.

[91] 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*, vol. 27, no. 5, pp. 531–544, 2019.

[92] M. C. E. Simsekler, A. Qazi, M. A. Alalami, S. Ellahham, and A. Ozonoff, "Evaluation of patient safety culture using a random forest algorithm," *Reliability Engineering & System Safety*, vol. 204, p. 107186, 2020.

[93] C. Iwendi, A. K. Bashir, A. Peshkar, R. Sujatha, J. M. Chatterjee, S. Pasupuleti, R. Mishra, S. Pillai, and O. Jo, "Covid-19 patient health prediction using boosted random forest algorithm," *Frontiers in public health*, vol. 8, p. 357, 2020.

[94] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and 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*, 2015, pp. 1721–1730.

[95] L. Sagaon-Teyssier, A. Vilotitch, M. Mora, G. Maradan, V. Guagliardo, M. Suzan-Monti, R. Dray-Spira, and 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*, vol. 19, no. 1, pp. 1–10, 2019.

[96] M. Dastoorpoor, N. Khanjani, A. Moradgholi, R. Sarizadeh, M. Cheraghi, and 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*, pp. 1–16, 2020.

[97] Y. Jiandong, Z. Mengxi, C. Yanggui, M. Li, Y. Rayibai, L. Yaoqin, L. Pengwei, P. Yujiao, X. Ran, and 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*, 2020.

[98] S. M. Mohnen, A. H. Rotteveel, G. Doornbos, and J. J. Polder, "Healthcare expenditure prediction with neighbourhood variables–a random forest model," *Statistics, Politics and Policy*, vol. 11, no. 2, pp. 111–138, 2020.

[99] A. Guisan, T. C. Edwards Jr, and T. Hastie, "Generalized linear and generalized additive models in studies of species distributions: setting the scene," *Ecological modelling*, vol. 157, no. 2-3, pp. 89–100, 2002.

[100] 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*, 2020.

[101] V. Gupta, M. Demirer, M. Bigelow, M. Y. Sarah, S. Y. Joseph, L. M. Prevedello, R. D. White, and 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.

[102] A. Kumar, S. B. Singh, S. C. Satapathy, and M. Rout, "MOSQUITO-NET: A deep learning based CADx system for malaria diagnosis along with model interpretation using GradCam and class activation maps," *arXiv preprint arXiv:2006.10547*, 2020.

[103] 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."

[104] S. Pereira, R. Meier, V. Alves, M. Reyes, and 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.

[105] 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.

[106] M. Izadyyazdanabadi, E. Belykh, C. Cavallo, X. Zhao, S. Gandhi, L. B. Moreira, J. Eschbacher, P. Nakaji, M. C. Preul, and 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.

[107] Y. Yang, V. Tresp, M. Wunderle, and 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.

[108] G. Chlebus, N. Abolmaali, A. Schenk, and H. Meine, "Relevance analysis of MRI sequences for automatic liver tumor segmentation," *arXiv preprint arXiv:1907.11773*, 2019.

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

[110] T. Jo, K. Nho, S. L. Risacher, and A. J. Saykin, "Deep learning detection of informative features in tau pet for alzheimer's disease classification," *bioRxiv*, 2020.

[111] I. Palatnik de Sousa, M. Maria Bernardes Rebuzzi Vellasco, and E. Costa da Silva, "Local interpretable model-agnostic explanations for classification of lymph node metastases," *Sensors*, vol. 19, no. 13, p. 2969, 2019.

[112] S. Kitamura, K. Takahashi, Y. Sang, K. Fukushima, K. Tsuji, and J. Wada, "Deep learning could diagnose diabetic nephropathy with renal pathological immunofluorescent images," *Diagnostics*, vol. 10, no. 7, p. 466, 2020.

[113] P.-Y. Tseng, Y.-T. Chen, C.-H. Wang, K.-M. Chiu, Y.-S. Peng, S.-P. Hsu, K.-L. Chen, C.-Y. Yang, and O. K.-S. Lee, "Prediction of the development of acute kidney injury following cardiac surgery by machine learning," *Critical Care*, vol. 24, no. 1, pp. 1–13, 2020.

[114] T. Pianpanit, S. Lolak, P. Sawangjai, A. Ditthapron, P. Leelaarporn, S. Marukatat, E. Chuangsuwanich, and T. Wilaiprasitporn, "Neural network interpretation of the parkinson's disease diagnosis from spect imaging," *arXiv preprint arXiv:1908.11199*, 2019.

[115] A. Borjali, A. F. Chen, O. K. Muratoglu, M. A. Morid, and K. M. Varadarajan, "Deep learning in orthopedics: How do we build trust in the machine?" *Healthcare Transformation*, 2020.

[116] M. R. Zafar and N. M. Khan, "DLIME: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems," *arXiv preprint arXiv:1906.10263*, 2019.

[117] D. Sharma, A. Durand, M.-A. Legault, L.-P. L. Perreault, A. Lemaçon, M.-P. Dubé, and J. Pineau, "Deep interpretability for GWAS," *arXiv preprint arXiv:2007.01516*, 2020.

[118] D. Yu, Z. Liu, C. Su, Y. Han, X. Duan, R. Zhang, X. Liu, Y. Yang, and S. Xu, "Copy number variation in plasma as a tool for lung cancer prediction using extreme gradient boosting (xgboost) classifier," *Thoracic Cancer*, vol. 11, no. 1, pp. 95–102, 2020.

[119] V. Couteaux, O. Nempont, G. Pizaine, and 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.
[120] K. Preuer, G. Klambauer, F. Rippmann, S. Hochreiter, and T. Unterthiner, "Interpretable deep learning in drug discovery," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 2019, pp. 331–345.
[121] S. Shen, S. X. Han, D. R. Aberle, A. A. Bui, and W. Hsu, "An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification," *Expert systems with applications*, vol. 128, pp. 84–95, 2019.
[122] J. Zhang, K. Kowsari, J. H. Harrison, J. M. Lobo, and L. E. Barnes, "Patient2vec: A personalized interpretable deep representation of the longitudinal electronic health record," *IEEE Access*, vol. 6, pp. 65 333–65 346, 2018.
[123] W. Liao, B. Zou, R. Zhao, Y. Chen, Z. He, and M. Zhou, "Clinical interpretable deep learning model for glaucoma diagnosis," *IEEE journal of biomedical and health informatics*, vol. 24, no. 5, pp. 1405–1412, 2019.
[124] Z. Obermeyer and E. J. Emanuel, "Predicting the future—big data, machine learning, and clinical medicine," *The New England journal of medicine*, vol. 375, no. 13, p. 1216, 2016.
[125] D. S. Char, N. H. Shah, and D. Magnus, "Implementing machine learning in health care—addressing ethical challenges," *The New England journal of medicine*, vol. 378, no. 11, p. 981, 2018.
[126] T. L. Beauchamp, J. F. Childress *et al.*, *Principles of biomedical ethics*. Oxford University Press, USA, 2001.
[127] J. Cowls and L. Floridi, "Prolegomena to a white paper on an ethical framework for a good AI society," *Available at SSRN 3198732*, 2018.
[128] S. H. Miles, *The Hippocratic Oath and the ethics of medicine*. Oxford University Press, 2005.
[129] P. Prioreschi, "A History of Medicine. Byzantine and Islamic Medicine," 2001.
[130] 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*, pp. 1–100, 1967.
[131] I. Waddington, "The development of medical ethics—a sociological analysis," *Medical History*, vol. 19, no. 1, pp. 36–51, 1975.
[132] F. A. Riddick, "The code of medical ethics of the american medical association," 2003.
[133] V. R. Potter, "Bioethics: bridge to the future," 1971.
[134] M. Ghaly, *Islamic perspectives on the principles of biomedical ethics*. World Scientific, 2016, vol. 1.
[135] R. M. Veatch and L. K. Guidry-Grimes, *The basics of bioethics*. Routledge, 2019.
[136] M. R. Marrus, "The nuremberg doctors' trial in historical context," *Bulletin of the History of Medicine*, vol. 73, no. 1, pp. 106–123, 1999.
[137] 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*, vol. 79, no. 4, p. 373, 2001.
[138] K. W. Goodman, *Ethics, computing, and medicine: informatics and the transformation of health care*. Cambridge University Press, 1998.
[139] E. S. Berner, *Clinical decision support systems*. Springer, 2007, vol. 233.
[140] R. M. Wachter, *The digital doctor: hope, hype, and harm at the dawn of medicine's computer age*. McGraw-Hill Education New York, 2015.
[141] A. Holzinger, C. Röcker, and M. Ziefle, *Smart health: open problems and future challenges*. Springer, 2015, vol. 8700.
[142] E. J. Topol and D. Hill, *The creative destruction of medicine: How the digital revolution will create better health care*. Basic Books New York, 2012.
[143] E. J. Topol, "The patient will see you now: the future of medicine is in your hands," 2015.
[144] E. Topol, *Deep medicine: how artificial intelligence can make healthcare human again*. Hachette UK, 2019.
[145] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature medicine*, vol. 25, no. 1, pp. 44–56, 2019.
[146] 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.
[147] J. H. Chen and A. Verghese, "Planning for the known unknown: Machine learning for human healthcare systems," *The American Journal of Bioethics*, vol. 20, no. 11, pp. 1–3, 2020.
[148] A. Bohr and K. Memarzadeh, *Artificial intelligence in healthcare*. Elsevier Science & Technology, 2020.
[149] A. Blasimme and 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*, 2019.
[150] A. Panesar, *Machine learning and AI for healthcare*. Springer, 2019.
[151] S. M. McKinney, A. Karthikesalingam, D. Tse, C. J. Kelly, Y. Liu, G. S. Corrado, and S. Shetty, "Reply to: Transparency and reproducibility in artificial intelligence," *Nature*, vol. 586, no. 7829, pp. E17–E18, 2020.
[152] 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*, vol. 577, no. 7788, pp. 89–94, 2020.
[153] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.
[154] D. M. West and J. R. Allen, *Turning Point: Policymaking in the Era of Artificial Intelligence*. Brookings Institution Press, 2020.
[155] M. Anderson and A. Perrin, "Barriers to adaption and attitudes towards technology," *Tech adoption climbs among older adults*, 2017.
[156] N. C. Arpey, A. H. Gaglioti, and M. E. Rosenbaum, "How socioeconomic status affects patient perceptions of health care: a qualitative study," *Journal of primary care & community health*, vol. 8, no. 3, pp. 169–175, 2017.
[157] D. Box and D. Pottas, "Improving information security behaviour in the healthcare context," *Procedia Technology*, vol. 9, pp. 1093–1103, 2013, 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. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2212017313002764
[158] H. Atlam and G. Wills, *IoT Security, Privacy, Safety and Ethics*, 03 2019, pp. 1–27.
[159] M. Meingast, T. Roosta, and 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.
[160] B. Mittelstadt, B. Fairweather, M. Shaw, and N. Mcbride, "The ethical implications of personal health monitoring," *International Journal of Technoethics*, vol. 5, pp. 37–60, 07 2014.
[161] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, vol. 10, p. 3152676, 2017.
[162] C. Klingler, D. S. Silva, C. Schuermann, A. A. Reis, A. Saxena, and D. Strech, "Ethical issues in public health surveillance: a systematic qualitative review," *BMC public health*, vol. 17, no. 1, pp. 1–13, 2017.
[163] L. M. Lee, C. M. Heilig, and A. White, "Ethical justification for conducting public health surveillance without patient consent," *American journal of public health*, vol. 102, no. 1, pp. 38–44, 2012.
[164] T. Wu, J. Chung, J. Yamata, and J. Richman, "The ethics (or not) of massive government surveillance," *The Ethics (or Not) of Massive Government Surveillance*, 2015.
[165] B. Mittelstadt, "Ethics of the health-related internet of things: a narrative review," *Ethics and Information Technology*, vol. 19, no. 3, pp. 157–175, 2017.
[166] J. Donath, "Ethical issues in our relationship with artificial entities," *The Oxford Handbook of Ethics of AI*, p. 53, 2020.
[167] J. Weizenbaum, "Computer power and human reason: From judgment to calculation." 1976.
[168] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, "Medical image analysis using convolutional neural networks: a review," *Journal of medical systems*, vol. 42, no. 11, p. 226, 2018.
[169] A. Ghorbani, A. Abid, and J. Zou, "Interpretation of neural networks is fragile," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 3681–3688.
[170] S. Wang, T. Zhou, and J. Bilmes, "Bias also matters: Bias attribution for deep neural network explanation," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6659–6667.
[171] H. J. Weerts, W. van Ipenburg, and M. Pechenizkiy, "A human-grounded evaluation of shap for alert processing," *arXiv preprint arXiv:1907.03324*, 2019.
[172] S. Mohseni, J. E. Block, and E. D. Ragan, "A human-grounded evaluation benchmark for local explanations of machine learning," *arXiv preprint arXiv:1801.05075*, 2018.
[173] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," *arXiv preprint arXiv:1810.03292*, 2018.

[174] T. Grote and P. Berens, "On the ethics of algorithmic decision-making in healthcare," *Journal of medical ethics*, vol. 46, no. 3, pp. 205–211, 2020.
[175] Q. Cassam, *Vices of the Mind: From the Intellectual to the Political*. Oxford University Press, 2018.
[176] 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*, vol. 25, no. 9, pp. 1337–1340, 2019.
[177] B. Nestor, M. McDermott, G. Chauhan, T. Naumann, M. C. Hughes, A. Goldenberg, and M. Ghassemi, "Rethinking clinical prediction: why machine learning must consider year of care and feature aggregation," *arXiv preprint arXiv:1811.12583*, 2018.
[178] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC medicine*, vol. 17, no. 1, p. 195, 2019.
[179] S. Latif, A. Qayyum, M. Usama, J. Qadir, A. Zwitter, and M. Shahzad, "Caveat emptor: the risks of using big data for human development," *IEEE Technology and Society Magazine*, vol. 38, no. 3, pp. 82–90, 2019.
[180] S. Saria and A. Subbaswamy, "Tutorial: safe and reliable machine learning," *arXiv preprint arXiv:1904.07204*, 2019.
[181] I. Y. Chen, P. Szolovits, and M. Ghassemi, "Can AI help reduce disparities in general medical and mental health care?" *AMA journal of ethics*, vol. 21, no. 2, pp. 167–179, 2019.
[182] J. He, S. L. Baxter, J. Xu, J. Xu, X. Zhou, and K. Zhang, "The practical implementation of artificial intelligence technologies in medicine," *Nature medicine*, vol. 25, no. 1, pp. 30–36, 2019.

.

Khansa Rasheed is a research assistant at the IHSAN Lab at the Information Technology University of Punjab (ITU), Lahore, Pakistan. She completed her Masters in Electrical Engineering from ITU in 2020. She completed her undergraduate degree from University of Engineering and Technology, Lahore, Pakistan. She is interested in applying machine learning (ML) for healthcare applications and in ensuring that these applications are secure, trustworthy, human beneficial, and ethical.

Adnan Qayyum is pursuing Ph.D. in computer science with Information Technology University, Lahore, Pakistan. He received the bachelor's degree in electrical (computer) engineering from the COMSATS Institute of Information Technology, Wah, Pakistan, in 2014, and the M.S. degree in computer engineering (signal and image processing) from the University of Engineering and Technology, Taxila, Pakistan, in 2016. His research interests include inverse medical imaging problems, healthcare, and secure and robust ML.

Mohammed Ghaly is Professor of Islam and Biomedical Ethics at the Research Center for Islamic Legislation & Ethics (CILE) at the College of Islamic Studies (CIS) at Hamad bin Khalifa University (HBKU) in Doha, Qatar. He has a Bachelor of Arts degree in Islamic Studies from Al-Azhar University (Egypt) and Master of Arts and PhD degrees in the same specialization from Leiden University, the Netherlands. Dr. Ghaly's main specialization is in the intersection of Islamic ethics and biomedical sciences. He is presently the editor-in-chief of the *Journal of Islamic Ethics* (published by Brill). Ghaly has lectured on Islamic (bio)ethics at many universities worldwide including Imperial College London, Oxford University, University of Oslo, University of Chicago, and Georgetown University. Besides his book, *Islam and Disability: Perspectives in Theology and Jurisprudence* (Routledge, 2010), and two edited volumes, *Islamic Perspectives on the Principles of Biomedical Ethics* (Imperial College & World Scientific, 2016) and *Islamic Ethics and the Genome Question* (Brill, 2019), Ghaly is the single author of more than 30 peer-reviewed publications. He serves on the editorial board of a number of academic journals.

Ala Al-Fuqaha [S'00-M'04-SM'09] received Ph.D. degree in Computer Engineering and Networking from the University of Missouri-Kansas City, Kansas City, MO, USA. He is currently a professor at the Information and Computing Technology division, college of Science and Engineering, Hamad Bin Khalifa University (HBKU). His research interests include the use of machine learning in general and deep learning in particular in support of the data-driven and self-driven management of large-scale deployments of IoT and smart city infrastructure and services, Wireless Vehicular Networks (VANETs), cooperation and spectrum access etiquette in cognitive radio networks, and management and planning of software defined networks (SDN). He is a senior member of the IEEE and an ABET Program Evaluator (PEV). He serves on editorial boards of multiple journals including IEEE Communications Letter and IEEE Network Magazine. He also served as chair, co-chair, and technical program committee member of multiple international conferences including IEEE VTC, IEEE Globecom, IEEE ICC, and IWCMC.

Adeel Razi is an Associate Professor at the Turner Institute for Brain and Mental Health, Monash University, Australia, where he is the Director of Computational Neuroscience Laboratory and the Deputy Lead of the Brain Mapping and Modelling Research Program. He is currently Australian Research Council DECRA Fellow (2017-2020) and has also been awarded NHMRC Investigator (Emerging Leader) Fellowship (2021-2025). He is an Honorary Senior Research Fellow at the Wellcome Centre for Human Neuroimaging of University College London, where he also worked from 2012 to 2018. His research is cross-disciplinary – combining engineering, physics, and machine-learning approaches – motivated by questions grounded in neuroscience leading towards the understanding of how the brain works.

Junaid Qadir [SM'14] is a Professor at the Information Technology University (ITU) of Punjab in Lahore, Pakistan, where he directs the IHSAN Research Lab. His primary research interests are in the areas of computer systems and networking, applied machine learning, using ICT for development (ICT4D); and engineering education. He has published more than 150 peer-reviewed articles at various high-quality research venues including publications at top international research journals including IEEE Communication Magazine, IEEE Journal on Selected Areas in Communication (JSAC), IEEE Communications Surveys and Tutorials (CST), and IEEE Transactions on Mobile Computing (TMC). He was awarded the highest national teaching award in Pakistan—the higher education commission's (HEC) best university teacher award—for the year 2012-2013. He has been appointed as ACM Distinguished Speaker for a three-year term starting from 2020. He is a senior member of IEEE and ACM.