Computers in Biology and Medicine 166 (2023) 107555

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

The enlightening role of explainable artificial intelligence in medical & healthcare domains: A systematic literature review☆

Subhan Ali [a], Filza Akhlaq [b], Ali Shariq Imran [a,*], Zenun Kastrati [c], Sher Muhammad Daudpota [b], Muhammad Moosa [a]

[a] Department of Computer Science, Norwegian University of Science & Technology (NTNU), Gjøvik, 2815, Norway
[b] Department of Computer Science, Sukkur IBA University, Sukkur, 65200, Sindh, Pakistan
[c] Department of Informatics, Linnaeus University, Växjö, 351 95, Sweden

ARTICLE INFO

Keywords:
Explainable
Artificial intelligence
Machine learning
Deep learning
Medical
Healthcare

ABSTRACT

In domains such as medical and healthcare, the interpretability and explainability of machine learning and artificial intelligence systems are crucial for building trust in their results. Errors caused by these systems, such as incorrect diagnoses or treatments, can have severe and even life-threatening consequences for patients. To address this issue, Explainable Artificial Intelligence (XAI) has emerged as a popular area of research, focused on understanding the black-box nature of complex and hard-to-interpret machine learning models. While humans can increase the accuracy of these models through technical expertise, understanding how these models actually function during training can be difficult or even impossible. XAI algorithms such as Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) can provide explanations for these models, improving trust in their predictions by providing feature importance and increasing confidence in the systems. Many articles have been published that propose solutions to medical problems by using machine learning models alongside XAI algorithms to provide interpretability and explainability. In our study, we identified 454 articles published from 2018–2022 and analyzed 93 of them to explore the use of these techniques in the medical domain.

1. Introduction

Despite our tendency for having unrealistic short-term expectations for Artificial Intelligence (AI), the future looks promising. Recent advancements in different fields of AI, especially in Machine Learning, are the big reason why AI is gearing to take a central role in our lives. We are just scratching the surface in utilizing deep learning to solve major issues in areas such as e-commerce, the airline industry, warfare, medical diagnoses, and almost all other aspects of human life. AI has made remarkable advancements in the past decade, largely due to unprecedented funding, as well as AI experts' promises to convert narrow AI to artificial general intelligence which can pass the Turing test in every routine task that humans can do seamlessly.

Since the emergence of AI, humans have been fearful that it could take full control and dominate us. This fear is compounded by the fact that it is often difficult to fully understand how AI algorithms operate. The recent revival of neural networks has shown remarkable results, but they function like a black box. A well-trained neural network can mimic human behavior, but the way it updates weights and biases through gradient descent during each iteration is not fully understood, leading to limited control over the algorithm. This is a concerning issue, as we may know what the algorithm is doing, but we cannot explain how it is doing that.

To address the concerns about the opacity of AI algorithms, a new field called Explainable Artificial Intelligence (XAI) has emerged. It encompasses a range of tools and frameworks aimed at helping humans understand and interpret the workings of AI models. The value of XAI in providing insight into the workings of AI algorithms is invaluable across all fields, but it is especially crucial in the medical and healthcare domain where human lives are at stake.

☆ This work was supported in part by the Department of Computer Science (IDI), Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology (NTNU), Gjøvik, Norway; and in part by the Curricula Development and Capacity Building in Applied Computer Science for Pakistani Higher Education Institutions (CONNECT) Project NORPART-2021/10502, funded by DIKU.

* Corresponding author.
E-mail addresses: subhan.ali@ntnu.no (S. Ali), filza.bscsf19@iba-suk.edu.pk (F. Akhlaq), ali.imran@ntnu.no (A.S. Imran), zenun.kastrati@lnu.se (Z. Kastrati), sher@iba-suk.edu.pk (S.M. Daudpota), muhammad.moosa@ntnu.no (M. Moosa).

https://doi.org/10.1016/j.compbiomed.2023.107555
Received 7 May 2023; Received in revised form 13 August 2023; Accepted 28 September 2023
Available online 4 October 2023
0010-4825/© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1.1. Background

AI and XAI have made great advancements in the medical and healthcare domains. Contributions are being made at a fast pace in the XAI field as a whole, as well as in XAI for the medical and healthcare domain. XAI can eliminate the barrier of distrust between clinicians and AI results when it is used in the medical domain.

XAI is a field that provides explanations for the results derived from AI models, or the way the model reached that result or decision. The goal of XAI is to create transparent and trustworthy AI systems that can be integrated into human decision-making in a supplementary way. Although XAI is a relatively new field, it has gained a lot of attention in recent years due to the need for AI and more specifically, transparent AI in various fields.

Healthcare and medicine are broad categories as they include diagnosis, prevention, and treatment of individuals with diseases. There are several domains where it gets tedious for a clinician to manually examine the results, i.e., examinations of X-rays, Magnetic Resonance Imaging (MRI), Computed Tomography (CT) scans, ultrasounds, etc. Diagnosis is not only limited to diagnosing image data but text data as well. For diagnosing mental health problems, there are many studies that have used textual data in order to diagnose depression and other mental health issues [1–4]. Similarly, prevention and treatment also become laborious for healthcare practitioners. Prevention in fact requires an early diagnosis and treatment requires an accurate diagnosis. Both of which can be achieved if trustworthy and transparent AI models are used to help the diagnosis.

The major contributions of this article are as follows:

- Conducted an extensive Systematic Literature Review (SLR) on XAI for medical and healthcare published articles
- Identified widely used models and datasets taxonomy of the domain.
- Reported literature of 93 studies employing rigorous filtering criteria following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework.
- Discussed limitations, advantages, and future research directions.

The rest of the article is structured as follows: after an introduction to XAI in the medical domain, Section 2 discusses recent surveys in the domain, followed by Section 3 to discuss the material and methodology of the manuscript. Section 4 presents the results of the study and finally Section 7 concludes the study.

1.1.1. Uncertainty of CNN models prediction

In the last few years, Convolutional Neural Networks (CNNs) have shown remarkable performance in several medical and healthcare applications, including the classification of COVID-19 X-ray images and the diagnosis of COVID-19. However, it is essential to consider the uncertainty associated with CNN predictions to ensure reliable and trustworthy results. This section explores the role of uncertainty-aware CNN models in the medical and healthcare domain.

One notable study, conducted by Gour et al. gour2022uncertainty, proposes an uncertainty-aware CNN model specifically designed for COVID-19 X-ray image classification. The authors recognized the importance of uncertainty estimation in this critical task and developed a framework that not only focuses on accurate predictions but also quantifies the uncertainty related with each prediction. By incorporating uncertainty into the model, they aimed to provide more reliable and interpretable results for medical professionals.

Another relative research paper by Shamsi et al. ss2021khosravi presents an uncertainty-aware transfer learning-based framework for COVID-19 diagnosis. The authors identified the challenges associated with limited labeled data and leveraged transfer learning techniques to improve the model's performance. Additionally, they included uncertainty estimation in order to provide insights into the reliability of the model's predictions. The framework aimed to assist healthcare professionals by providing not only accurate diagnoses but also information about the uncertainty associated with those diagnoses.

Both studies highlight the significance of considering uncertainty in CNN models for medical and healthcare applications. Uncertainty estimation allows for a more comprehensive understanding of the model's predictions, helping healthcare professionals in making informed decisions. Specifically, one aspect of uncertainty that has gained attention is the uncertainty associated with CNN models' predictions, which can provide valuable insights into the reliability and reliance level of the results.

Uncertainty in CNN models' predictions can be attributed to various factors. One factor is the inherent complexity of medical data, including variations and overlaps in different diseases or conditions. Additionally, limited or imbalanced training data can contribute to uncertainty, as the model may encounter instances that differ significantly from the training distribution. Furthermore, ambiguity or noise in the input data can also introduce uncertainty into the predictions.

To address the uncertainty associated with CNN models' predictions, several approaches have been proposed. These include Bayesian neural networks, Monte Carlo dropout, and ensemble methods. These techniques allow the model to generate multiple predictions or probability distributions, providing a measure of uncertainty along with the final prediction. By considering the uncertainty, medical professionals can make more informed decisions and better understand the limitations of the model.

In conclusion, incorporating uncertainty-aware CNN models in medical and healthcare domains is crucial for reliable and trustworthy predictions. The studies discussed gour2022uncertainty and ss2021 khosravi exemplify the efforts made to quantify uncertainty in the context of COVID-19 X-ray image classification and diagnosis. Uncertainty estimation in CNN models' predictions helps healthcare professionals interpret the results, make informed decisions, and understand the limitations and confidence level associated with the model's outputs. By incorporating uncertainty-aware approaches, the medical and healthcare community can leverage the benefits of CNN models while ensuring the reliability and transparency of their predictions.

2. Related surveys

XAI and Healthcare both are hot topics for research nowadays. Since researchers are actively contributing in both fields, there are numerous surveys that lie under the domain of XAI for medical and healthcare. Out of 11 surveys that we found, four of them belonged to sub-fields of the healthcare domain for example, a survey on epilepsy detection [5], X-ray Image Analysis [6], predictive modeling in healthcare [7] and clinical decision support system [8]. Three of them were discussing benefits and/or application of XAI in medical and healthcare but were not directly linked with the healthcare domain such as [9] has discussed augmentation approaches used in XAI for medical informatics, [10] has investigated interactive visualization which can be beneficial for XAI in different domains such as medical, agriculture, etc, [11] has found the application of XAI in different fields, i.e., Natural Language Processing (NLP), biomedical and malware classification, and lastly a mapping study which found interpretability techniques used in medicine using medicine [12].

To the best of our knowledge, we found three surveys that were related to XAI for the medical or healthcare domain [13–15]. Korica et al. [13] have presented a synthesized taxonomy for categorizing explainability methods and a summary of gaps, challenges, and opportunities for applying XAI in the medical industry through a conducted field survey. Chakrobartty et al. [14] have done a literature survey on the same topic, they have covered 22 studies, and they searched on PubMed only published during the 2008–2020 period. They have tried to find the existing techniques and methods used in the medical domain. The limitation of their work is to use only one database to

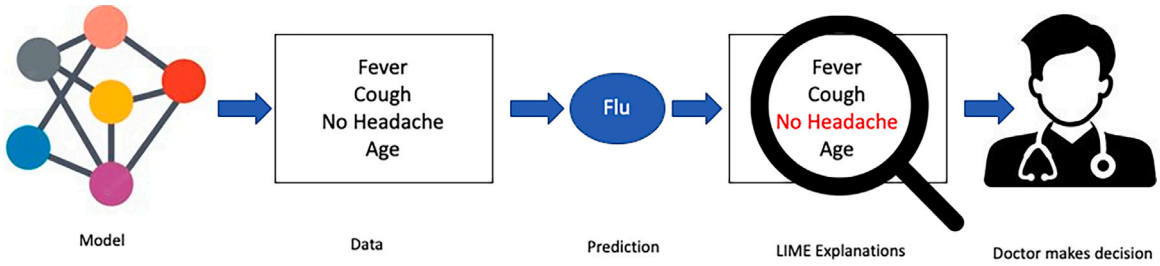

Fig. 1. This is an example of how LIME can help doctors taking the decision. While explaining individual predictions LIME can help identify the important features considered in predictions so doctors can take appropriate decisions.

fetch papers i.e. PubMed. Nazar et al. [15] have done a SLR on the use of AI, XAI, and Human–Computer Interaction (HCI) in the medical domain. They have covered 135 publications, published during the 2016–2021 timeframe and fetched from various data engines. [15] collectively examines the applications and challenges of XAI, AI, and HCI for medical and healthcare.

It is interesting to notice that all the surveys considered for this study, haven been conducted in a specific domain related to Medical and XAI. However, the survey that we are conducting comprises of almost all the medical domains as well as all the XAI algorithms that can possibly correlate with medical and healthcare.

The XAI algorithms used in the publications are explained in detail in the following subsections.

2.1. LIME (local interpretable model-agnostic explanations)

LIME was introduced in the year 2016 by Ribeiro et al. LIME was introduced in the year 2016 by Ribeiro et al. [16]. This innovative methodology serves as a model-agnostic XAI algorithm. The term "model-agnostic" implies that LIME can be applied universally to explain the workings of a wide array of machine learning or deep learning algorithms, regardless of their specific characteristics or complexity.

LIME operates by generating localized explanations centered around individual predictions, utilizing interpretable models. These explanations shed light on the factors contributing to a particular prediction made by a given machine learning or deep learning model. The underlying concept is to estimate the behavior of the complex original model within a smaller, more comprehensible model, known as an interpretable model, specifically tailored to the prediction in question.

This capacity makes LIME highly versatile and adaptable, allowing it to provide explanations not only for prevalent models like deep learning neural networks, random forests, and gradient boosting but also for any other believable machine learning model, due to this property it is referred to as "model-agnostic". For a practical illustration of LIME's utility, consider Fig. 1 which serves as an exemplary scenario. In this context, the primary model has predicted that a patient is afflicted with the flu. However, through the application of XAI techniques, LIME has explained that the presence of symptoms such as sneezing and headache, gleaned from the patient's medical history, has contributed to the model's "flu" diagnosis. This graphic depiction illustrates how a medical professional can make well-informed decisions by leveraging the insights provided by the XAI, particularly in the context of a single prediction.

It is interesting to note that LIME is intentionally engineered to expound upon individual predictions. The key advantage derived from its model-agnostic nature is its capability to seamlessly integrate with an extensive spectrum of models, even when dealing with intricate predictions in high-dimensional feature spaces.

2.2. SHAP (SHapley additive exPlanations)

Lundberg et al. [17] introduced SHAP, a groundbreaking methodology aimed at unifying the realm of model interpretability. The primary objective of SHAP is to provide a comprehensive solution for rendering complex models interpretable, thereby facilitating a broader community of researchers in comprehending the inner workings of machine learning or deep learning models.

In the intricate landscape of XAI, the challenge of selecting the most suitable algorithm for a specific model type proved to be a formidable task. To surmount this hurdle, Lundberg and colleagues devised SHAP, an ingenious framework that bestows importance values upon individual features in the context of a particular prediction [17]. By explaining the importance of each feature, SHAP contributes to the identification of key factors exerting the most substantial influence on a given prediction.

SHAP distinguishes itself as a versatile, all-encompassing XAI algorithm, poised to harmoniously interface with a diverse array of deep learning or tree-based ML algorithms. Notably, its efficacy transcends the boundaries of model intricacies and types, rendering it applicable to a wide spectrum of scenarios.

Moreover, when confronted with multifaceted scenarios wherein a multitude of features coalesce, SHAP's efficacy shines through. It has been demonstrated that SHAP can yield superior results compared to alternative methodologies in such scenarios. This capability highlights SHAP's prowess in disentangling intricate relationships and facilitating a more nuanced comprehension of the factors driving predictions.

In summation, Lundberg and his collaborators' introduction of SHAP has addressed a critical need within the XAI landscape. By providing a unified approach to model interpretability, SHAP empowers researchers to unravel the enigmatic inner workings of complex machine learning or deep learning models, transcending the limitations of conventional XAI methodologies.

2.3. CAM (class activation mapping)

CAM emerges as a specialized tool tailored to fulfill the eager appetite for interpretability within the realm of deep learning-based computer vision models. Designed with a specific focus on the complex complexities of computer vision, CAM serves as an illuminating XAI technique, offering insights into the enigmatic decision-making processes of neural networks, particularly the formidable Convolutional Neural Networks (CNNs) [18]. At its core, CAM coordinates the generation of class activation maps through the integration of global average pooling. This computational maneuver unveils the dominant regions within an image that have precipitated a prediction made by a neural network, particularly the potent Convolutional Neural Networks. The resultant class activation maps demystify the convolutional neural network's inner workings, shedding light on the salient features and distinct regions that the network has honed in on during its decision-making process.

Notably, CAM's efficacy extends beyond its foundational role in the computer vision domain. It carves out a versatile domain in the expansive field of medical imaging, where deep learning-based models hold immense promise. By harnessing CAM's elucidative prowess, predictions spawned by deep learning models utilized in the intricate realm of medical imaging can be unraveled and contextualized. This critical

application of CAM within medical imaging fortifies the understanding of predictive outcomes, fostering a symbiotic relationship between cutting-edge technology and informed medical decision-making.

In a nutshell, CAM stands as a testament to the resourcefulness of XAI, catering specifically to the demanding terrain of computer vision models. Its utilization as a potent instrument for deciphering deep learning-based models' predictions, both in the visually intricate realm of computer vision and the vital domain of medical imaging, underscores its pivotal role in unraveling the latent complexities of modern neural networks.

2.4. Grad-CAM (gradient-weighted class activation mapping)

Grad-CAM stands as a direct descendant of the CAM methodology, extending and amplifying the prowess of interpretability for intricate predictions churned out by deep learning-based models. The ingenious concept underlying CAM serves as the foundational bedrock upon which Grad-CAM is built, albeit with a novel twist that leverages the potent tool of gradients [19]. At its essence, Grad-CAM propels the realm of XAI into a new era by harnessing the raw power of gradients. This technique ushers in a new era of interpretability for complex predictions emerging from deep learning models, casting a radiant light on the convoluted decision-making processes of these models.

In practice, Grad-CAM undertakes the task of crafting coarse localization maps that delineate the critical regions within an image. These maps serve as visual waypoints, directing the observer's attention to the pivotal areas of the image that have contributed substantively to the model's prediction regarding a specific concept. This marks a significant leap forward compared to the predecessor CAM, as Grad-CAM refines the art of visualization, offering a more nuanced and accurate representation of the regions that wield the most profound influence on the prediction.

One of Grad-CAM's remarkable attributes is its aptitude for dealing with images boasting high resolutions. The technique's proficiency shines when faced with these intricate, information-rich images, making it a potent tool for scenarios demanding a granular understanding of complex predictions.

Worthy of note is the common thread linking CAM and Grad-CAM—their shared reliance on gradients to unearth the crux of interpretability. Both methodologies employ gradients as guiding beacons, illuminating the path toward the most crucial regions within a given input image that have propelled the model's prediction. This shared reliance underscores the pivotal role of gradients in unraveling the enigmatic decision-making processes of deep learning-based models.

In summary, Grad-CAM represents a exquisite evolution of the CAM lineage, empowered by the impressive tool of gradients. By adeptly amalgamating visualization and gradient analysis, Grad-CAM paves the way for a more profound and accurate comprehension of complex deep learning-based model predictions. Its invaluable utility in navigating high-resolution images elevates it to a critical position within the ever-expanding toolkit of eXplainable Artificial Intelligence, furthering the frontiers of model interpretability.

2.5. Counterfactual explanations

Counterfactual explanation, also known as "what-if" explanations are a type of explanation which are used to understand the result/prediction or outcome of an AI or ML algorithm. Counterfactuals as the name suggest, they analyze alternative scenarios referred to as counterfactuals, which are hypothetical situations where some input variables/features are changed while keeping the rest of the model or system unchanged. By systematically altering the input features and observing the resulting changes in the output, counterfactual explanations help to identify the key features or factors that influenced the prediction. The paper titled "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR" by Wachter, Mittelstadt, and Russell was published in 2017 [20]. This paper discusses the importance of counterfactual explanations for automated decision-making systems and their relevance in the context of the European Union's General Data Protection Regulation (GDPR).

2.6. Anchors

Anchors [21] are rule-based explanations that are mostly used for highlighting the most important features/predictions done by an ML algorithm. The main idea behind anchors is to find a minimal set of conditions that, when satisfied, are likely to guarantee a specific prediction. These conditions are expressed as simple and intuitive rules, making them accessible to non-experts. Anchors focus on providing explanations for individual predictions rather than explaining the overall behavior of a model. Anchors work by generating concise and interpretable "if-then" rules that explain why a machine learning model made a specific prediction for a given instance. The process involves iteratively perturbing features and observing the model's response to identify the minimal set of conditions that are both necessary and sufficient for the prediction. Anchors provide local explanations by focusing on individual predictions, offering a transparent and understandable way to understand the model's decision-making process.

2.7. Influence functions

Influence Functions [22] are statistical tools used for finding the influence of the individual training sample or parameters on the outcome of a machine learning model. Influence functions measure how sensitive a model's predictions or parameters are to changes in the training data. They enable the identification of influential training examples that have a significant impact on the model's behavior. By quantifying the influence of each example, influence functions provide insights into which training instances contribute the most to the model's decision-making process. The calculation of influence functions involves computing the derivatives of the model's predictions or parameters with respect to changes in individual training examples. These derivatives capture how small perturbations in the training data affect the model's output. By analyzing these derivatives, one can determine the influence of each example on the model's behavior.

3. Materials and methodology

In this study, we are aiming to investigate the state-of-the-art on XAI for medical and healthcare domains. We are performing a systematic review using mapping study or scoping study technique [23]. In this study, we have done a comprehensive review of the literature in the research domain and have identified the techniques, datasets, performance metrics, and algorithms used in the literature. This study follows the proposed guidelines by Kitchenham et al. [24] and it includes the following phases:

1. Specifying research questions
2. Search strategy
3. Identification of primary studies
4. Data extraction
5. Threat to the validity

3.1. Research questions

The key research question for this study was to find state-of-the-art technologies, algorithms, and datasets evaluation metrics for XAI in the medical and healthcare domain. To do an in-depth analysis for this systematic mapping review, the key question is further divided into four research questions which are mentioned in Table 1. These research questions will clearly show the direction and road map for this study and will help the readers to understand the structure of this work.

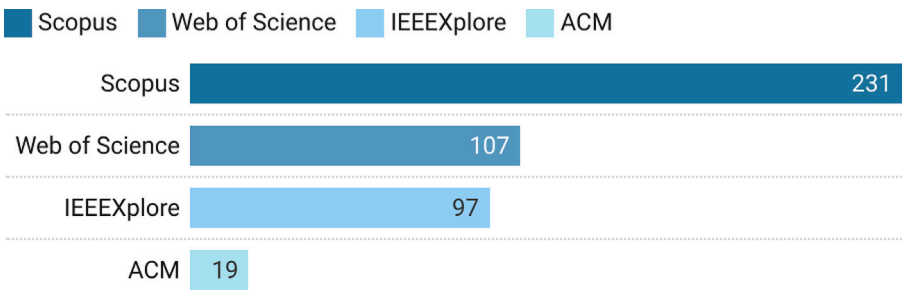

Fig. 2. Database vs No. of publications.

Table 1
Research questions.

RQ#	Research question
RQ1	What are the most common XAI algorithms/methods/tools used by researchers for the medical and healthcare domains?
RQ2	What challenges and limitations have been faced or undertaken by researchers in these domains?
RQ3	Which datasets are being prominently and mostly used in research on explainable artificial intelligence for medical and healthcare?
RQ4	Which performance metrics are commonly considered in XAI research in the medical and healthcare domains?

3.2. Search strategy

For searching literature on XAI for medical and healthcare domains we searched the literature in four well-known online databases, i.e., Web of Science, Scopus, IEEE Xplore, and ACM Digital Library. For forming our search string we used three groups of keywords as mentioned in Table 2. In group A we were also considering "explainability" and "interpretable" keywords as well but these keywords did not return good quality works and therefore we dropped it from the query. To be inclusive and structured we used the same keywords on all four databases and searched the literature by searching those keywords in the title, abstract, and author keywords. By hitting search, we fetched 454 results, the distribution of which can be seen in Fig. 2.

We got 231 papers from Scopus which is the highest number, 107 from Web of Science, 97 from IEEE Xplore, and 19 papers from ACM Digital Library.

3.3. Identification of primary studies

The search string was applied to different digital databases to fetch the relevant results. The data is extracted by applying the search string on the title, keywords, and abstract along with applying year, article type (only conference and article papers), and language filters, as a result, we got 454 articles. After removing duplicates we were left with 239 articles. Then we applied the exclusion criteria mentioned in Table 3, we applied exclusion criteria while reading the abstract and title of the literature, after removing 87 papers including 11 survey papers, 152 papers were selected for full-text screening. While full-text screening we also applied inclusion criteria mentioned in Table 3, we included papers having scientific rigor, credibility, and relevance. There was doubt about including 17 papers so we used an anonymous majority voting mechanism and three authors participated in that, we removed two papers based on majority voting. After majority voting and full-text screening, 93 studies were selected for systematic review using PRISMA protocol. Fig. 3 is the PRISMA diagram which shows the process from fetching the whole data to filtering the relevant data.

3.3.1. Identification

We retrieved 454 studies in total when we searched four well-known databases (Scopus, Web of Science, IEEE Xplore, and ACM Digital Library). Fig. 4 shows the number of studies found initially and studies that were selected from each database. Scopus was the main contributor so we selected 66 publications out of 231. Similarly, from IEEE Xplore 19 studies were considered out of 96, and from WoS only six studies were considered as most of the studies were removed because of duplication with Scopus. Lastly, we selected two studies from ACM which were relevant to our topic.

3.3.2. Screening

In this step, we filtered out studies according to our inclusion and exclusion criteria, as shown in Table 3. We first discarded the duplicate papers, then the papers which were not in the English language, and the papers which were out of our time span (2018–2022). Further, we discarded the papers which were not related to the medical domain and papers that were not related to XAI, Machine Learning (ML), Deep Learning (DL), and AI. We also excluded the survey papers which were a total of 11 in number, as well as the papers that did not fall under the quality assessment criteria that were carefully set. The assessment criteria are comprised of three factors that are defined in the following subsection.

Scientific rigor. If an appropriate research methodology has been applied in the paper, it is considered to be scientifically rigorous.

Credibility. If the research is believable and the findings are accurate and well presented, that paper is considered credible.

Relevance. If the findings of a paper are relevant to the academic community and actors in the medical/healthcare domain, it is considered to be relevant.

3.3.3. Eligibility

Exclusion criteria implementation reduced the 87 papers. Then we included the studies based on scientific rigor, credibility, and relevance. Out of 454 fetched articles, 93 studies passed all the phases hence we included them in this study.

3.3.4. Included studies

Finally, 93 studies were selected throughout the study and review. We had 152 studies for full-text screening, during full-text screening, we were extracting data and also checking if those studies meet the quality assessment criteria defined in 3.3.2. We removed those studies that either failed any of the three criteria (i.e., scientific rigor, credibility, and relevance). Also, we found 11 survey papers that were later excluded from this work. There were three papers removed due to having paid access. Moreover, there was one paper that was abstract, so we also removed that. There were many studies that were related to XAI, but we excluded them because they did not belong to the medical domain. Some were in the medical domain but they were not really related to XAI, they had just used the keyword XAI in the author's

Table 2
Query formation.

Group A: Explainable Artificial Intelligence-related keywords	XAI
Group B: Machine Learning related keywords	Machine learning OR artificial intelligence OR deep learning
Group C: Medical related keywords	Medical OR healthcare
Query	(Group A) AND (Group B) AND (Group C)

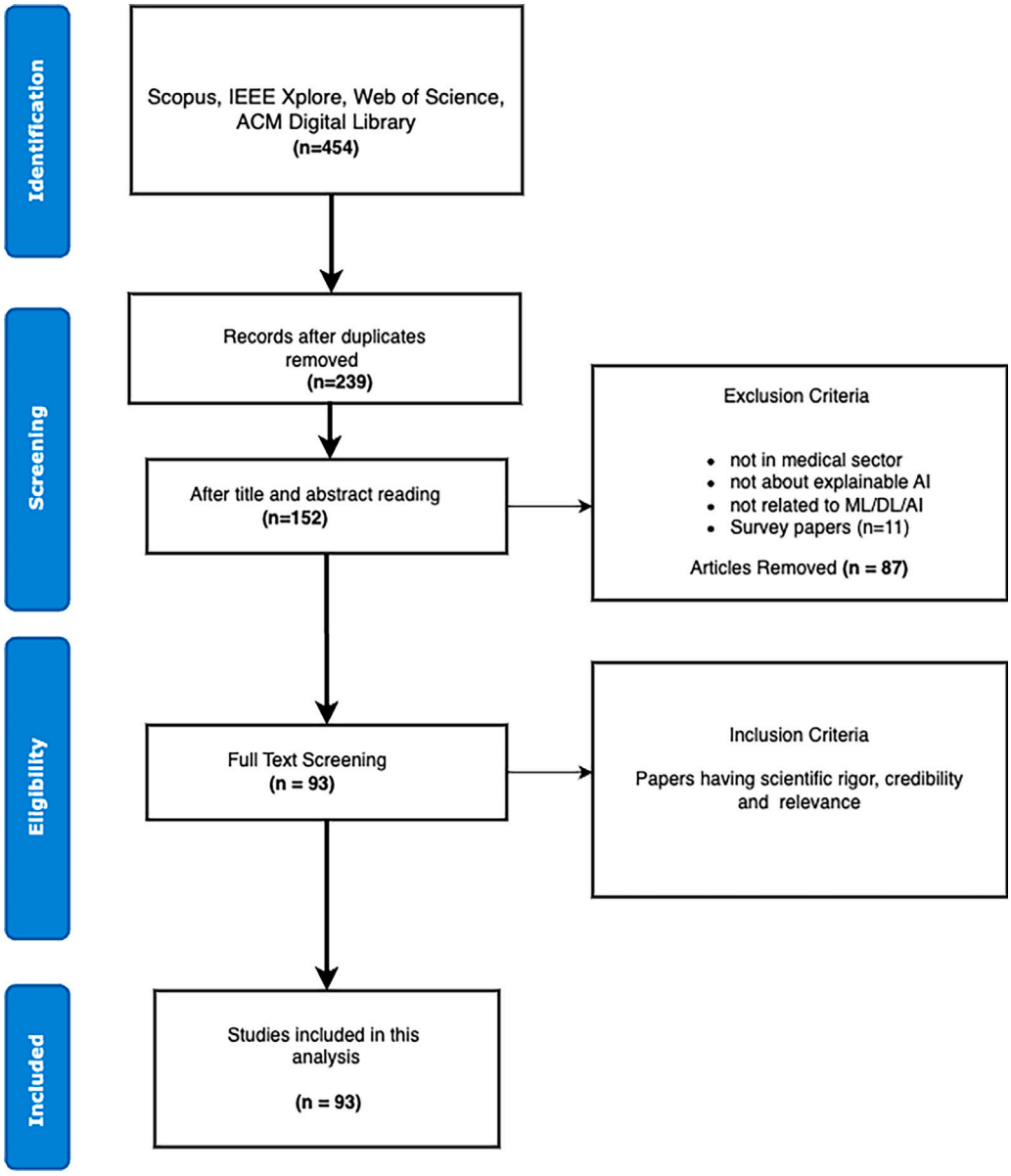

Fig. 3. PRISMA diagram.

Table 3
Inclusion and exclusion criteria.

Inclusion criteria	• Papers having scientific rigor, credibility, and relevance
Exclusion criteria	• Not in the medical sector • Not about explainable AI • Not related to ML/DL/AI • Excluded editorial materials, book chapters and reviews, letters, and retracted publications • Survey Papers (n = 11)

keywords or abstract. Table 4, 5–7 provide a detailed summary of all studies included.

Among the selected 93 publications, 59 are journal articles while 34 are conference or proceeding papers. This information is visualized in Fig. 5.

Fig. 6 shows the papers published with respect to years in journals and conferences, respectively. We can see that 35 journal articles and 12 conferences were published in 2022, 18 journal papers and 10 conference papers in 2021, and 6 articles and nine conference papers in 2020, while in 2019 only 3 conference papers were published but no journal articles. This figure also shows an increasing trend in terms of the number of publications hence we deduced that XAI is taking popularity with time.

Fig. 7 shows the proportion of papers published by different publishers. This information is evident in IEEE being the leading publisher

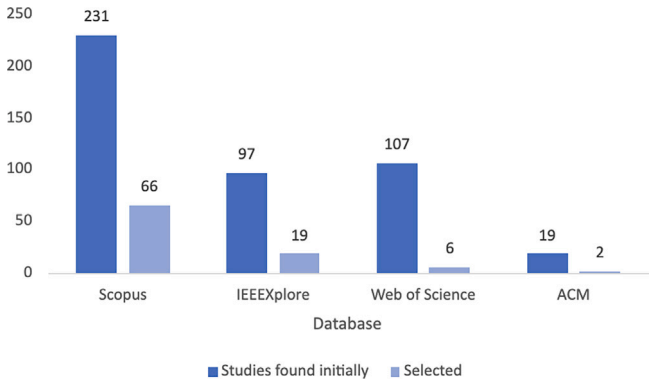

Fig. 4. Studies found initially and selected.

Table 4
Summary of studies considered 1/4.

Study	Years	Type	Publisher	Application Area	Citations
[25]	2021	Journal Article	IEEE	Neonatal Intensive Care Unit (NICU)	112
[26]	2021	Conference Paper	IEEE	Lesion Localization	25
[27]	2022	Journal Article	MDPI	Advanced Stage Epithelial Ovarian Cancer	17
[28]	2020	Conference Paper	IEEE	Pathological voices discrimination	14
[29]	2022	Journal Article	MDPI	Prevention Musculoskeletal Symptoms	14
[30]	2022	Journal Article	MDPI	Breast tomosynthesis examination	14
[31]	2022	Journal Article	MDPI	"Clinical COVID-19 Diagnosis with Routine Blood Tests"	11
[32]	2022	Journal Article	MDPI	Medical XAI applications in diagnosis and surgery	10
[33]	2021	Journal Article	IEEE	Parkinson's Disease	9
[34]	2021	Journal Article	Korean Physical Society	Hemorrhage prediction	9
[35]	2021	Journal Article	MDPI	Breast Cancer	8
[36]	2022	Journal Article	ELSEVIER	ECG data of cardiac disorders	6
[37]	2021	Journal Article	IEEE	Stress Disorder	6
[38]	2021	Conference Paper	IEEE	Lung Cancer and COVID-19 prediction	6
[39]	2021	Conference Paper	IEEE	Lung cancer prediction	6
[40]	2019	Conference Paper	IEEE	N/A	6
[41]	2020	Journal Article	IEEE	COVID-19	5
[42]	2022	Journal Article	Nature Research	"Auto-labeling of chest X-ray images based on quantitative similarity"	5
[43]	2019	Conference Paper	Springer	N/A	5
[44]	2022	Journal Article	ELSEVIER	UX effectiveness of the Local Interpretable Model-Agnostic Explanations	4
[45]	2022	Journal Article	ELSEVIER	Early detection of Acute Kidney Injury	4
[46]	2022	Conference Paper	IEEE	Pediatric pulmonary health evaluation	4

Fig. 5. Conference papers vs Journal articles.

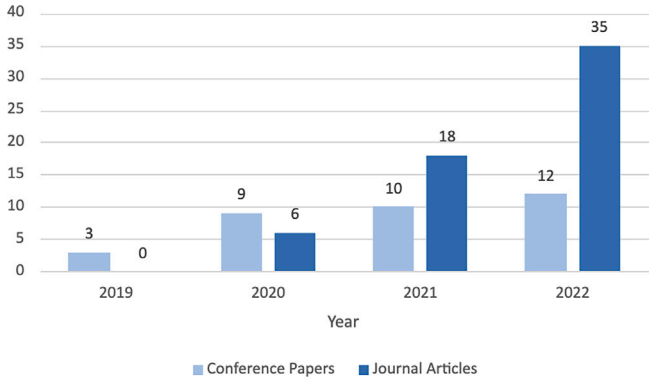

Fig. 6. Journal and Conference papers published with respect to years.

Table 5
Summary of studies considered 2/4.

[47]	2022	Conference Paper	IEEE	Polyp classification	4
[48]	2021	Conference Paper	IEEE	Breast Cancer Prediction	4
[49]	2021	Journal Article	IEEE	Feature extraction	4
[50]	2020	Journal Article	IEEE	Cancer computer-aided diagnosis system	4
[51]	2021	Conference Paper	Springer	Autism Detection	4
[52]	2022	Conference Paper	IEEE	Explainable Image Retrieval	3
[53]	2021	Journal Article	IEEE	Classification of healthy and unhealthy neonates	3
[54]	2020	Conference Paper	IEEE	Heart Failure Prediction	3
[55]	2022	Journal Article	ACM	Alzheimer's Disease	2
[56]	2022	Journal Article	ELSEVIER	ECG monitoring healthcare system	2
[57]	2022	Journal Article	IEEE	Breast Cancer	2
[58]	2021	Journal Article	IEEE	Mental Health	2
[59]	2022	Conference Paper	SPIE	Determining severity grade of Diabetic Retinopathy	2
[60]	2020	Journal Article	BioMed Central Ltd	Classification	1
[61]	2022	Journal Article	ComSIS Consortium	Medical Image Segmentation	1
[62]	2021	Journal Article	IEEE	COVID-19	1
[63]	2022	Journal Article	MDPI	Brain Tumor Detection	1
[64]	2022	Journal Article	ACM	Clinical Gait Analysis	0
[65]	2021	Journal Article	ACM	Brain Tumor Detection	0
[66]	2020	Journal Article	ACM	N/A	0
[67]	2020	Conference Paper	CEUR-WS	Feature importance	0
[68]	2022	Journal Article	ELSEVIER	N/A	0
[69]	2022	Journal Article	ELSEVIER	Threat detection in Internet of Medical Things networks	0
[70]	2021	Journal Article	ELSEVIER	Pneumonia classification	0
[71]	2021	Journal Article	ELSEVIER	DR severity classification	0
[72]	2021	Journal Article	ELSEVIER	Multi-modal causability	0
[73]	2021	Journal Article	ELSEVIER	Classification of ACS patients	0
[74]	2022	Journal Article	Frontiers Media S.A.	Deep Learning in Neuroimaging	0
[75]	2022	Journal Article	Hindawi Limited	Leukemia Diagnosis	0

Table 6
Summary of studies considered 3/4.

[76]	2022	Journal Article	IEEE	COVID-19	0
[77]	2022	Journal Article	IEEE	COVID-19	0
[78]	2022	Journal Article	IEEE	Alzheimer's Disease	0
[79]	2022	Journal Article	IEEE	N/A	0
[80]	2022	Journal Article	IEEE	Diabetes	0
[81]	2022	Journal Article	IEEE	Alzheimer's Disease	0
[82]	2022	Journal Article	IEEE	Arrhythmia	0
[83]	2022	Journal Article	IEEE	Hepatic Steatosis	0
[84]	2022	Conference Paper	IEEE	Real-time detection of COVID-19 using CXR	0
[85]	2022	Conference Paper	IEEE	Medical Image Captioning	0
[86]	2022	Journal Article	IEEE	Explainability of glaucoma predictions	0
[87]	2022	Journal Article	IEEE	Diagnosis of Paratuberculosis in Histopathological Images	0
[88]	2021	Journal Article	IEEE	Alzheimer's Disease and Mild Cognitive Impairment Diagnosis	0
[89]	2021	Journal Article	IEEE	Alzheimer's Patient	0
[90]	2021	Conference Paper	IEEE	Readmission Prediction	0
[91]	2021	Conference Paper	IEEE	N/A	0
[92]	2021	Conference Paper	IEEE	"heart disease classification"	0
[93]	2021	Conference Paper	IEEE	Pneumonia and COVID-19 classification	0
[94]	2020	Journal Article	IEEE	Neuroscience	0
[95]	2019	Conference Paper	IEEE	Stroke prediction	0
[96]	2022	Journal Article	MDPI	Diabetic Retinopathy	0
[97]	2021	Journal Article	MDPI	Chronic Wound Classification	0
[98]	2020	Conference Paper	SPIE	Retinal OCT image classification	0
[99]	2020	Conference Paper	SPIE	Breast Cancer	0

for research in this domain. 40 papers were published in IEEE journals while Springer, Elsevier, and MDPI remained second, third, and fourth choice respectively for the researchers. While Fig. 8 shows the proportion of papers published in journals and conferences with respect to publishers. From Fig. 8 we can see that 23 of 40 papers published in IEEE were journal articles and 17 papers were conference papers. The information about other publishers is also available in this figure to help future researchers in the selection of a venue for their publication. It can also be deduced from this figure that researchers who wish to publish their articles at Conferences can keep IEEE, Springer, and International Society for Optics and Photonics (SPIE) conferences as their first choice as the majority of conference papers are published in these venues.

3.4. Data extraction

In this section, we explain the data extraction process. The data was extracted from the 93 papers, for which a tabulated Microsoft Excel spreadsheet was used to log the data. A unique identifier was assigned to each article that was made up of the initial letter of a source followed by a serial number. We extracted fields such as "Datasets used" in the study, dataset link, and source code link were also logged if they were mentioned in the paper. In addition, "Algorithms used" in the study, "Impact factor" of the publishing source from the year of publishing, "Evaluation metrics" in case any experiments were performed and assessed, followed by 'Application Area' that defines the medical domain in which the work was applicable, "Limitations" of the study proposed,

Table 7
Summary of studies considered 4/4.

[100]	2020	Conference Paper	SPIE	N/A	0
[101]	2022	Journal Article	Springer	brain tumors	0
[102]	2022	Conference Paper	Springer	Detection of thoracolumbar fractures using X-rays	0
[103]	2022	Conference Paper	Springer	Detecting medical instruments in endoscopy images	0
[104]	2022	Conference Paper	Springer	Allowing interactive correction of both decisions and explanations by the experts	0
[105]	2022	Conference Paper	Springer	Evaluation of LIME vs SHAP on the Melanoma dataset	0
[106]	2022	Conference Paper	Springer	Breast Cancer Detection	0
[107]	2022	Conference Paper	Springer	Prediction of heart diseases	0
[108]	2021	Journal Article	Springer	N/A	0
[109]	2021	Conference Paper	Springer	retinal diseases classification	0
[110]	2020	Conference Paper	Springer	Medical	0
[111]	2020	Conference Paper	Springer	detect malaria in cell images	0
[98]	2020	Journal Article	Springer	Retinal OCT image classification	0
[112]	2020	Conference Paper	Springer	measuring the accuracy of image classification for breast cancer screening.	0
[113]	2022	Journal Article	Tech Science Press	Classification of ASD and Non-ASD patients	0
[114]	2022	Journal Article	Tech Science Press	classification of GI tract diseases	0
[115]	2022	Journal Article	Turkiye Klinikleri	Gastrointestinal disease classification	0
[116]	2022	Journal Article	Springer	Predicting depressive symptoms	0

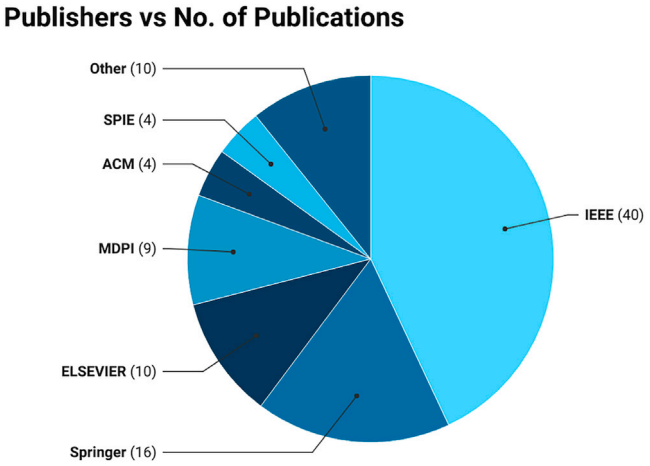

Fig. 7. Publishers vs No. of publications.

Table 8
Elements of the study.

Elements	Description
Study ID	Source of paper and serial number
Impact factor	Impact factor of the journal if published in one.
Objectives	Objectives of the conducted study or experiment
Algorithms used	Which AI or XAI algorithms were used?
Tools used	Any additional tools that were used after XAI methods.
Application area	In which medical domain does the study intend to be applicable
Evaluation metrics	Which evaluation metrics were used if any experiments were conducted
Limitations	Any limitations of the study

and "Tools used" represents the additional tools to XAI models if any were used. Table 8 provides a description of each element.

3.5. Threats to validity

Search String: The query that was originally made, included the words, "XAI", "interpretability", "explainability", and "medical and healthcare". As a result, we got around 1600 articles which were too many, and most of which were not relevant to the topic under our study. Therefore, we omitted the words "explainability" and "interpretability" from our search query, and went ahead with the remaining words. This omitting might have caused us to miss some valuable articles related to our domain of study.

Selection of databases: The databases from which we selected articles for our study were Web of Science, Scopus, ACM Digital Library, and IEEE for the sake of credibility and quality. Since our domain is medical and healthcare, it is possible that Pubmed (a famous site for medical research), that was missed in this study, might have some esteemed research.

Language Barrier: The papers were also filtered out on the basis of language and only papers written in the English language were selected for this study. We might have lost some valuable research due to the language barrier as well.

Time frame of the studies selected: We have only considered the studies from the past five years, there may have been some studies before that time that could be of a beneficial contribution to our domain of study.

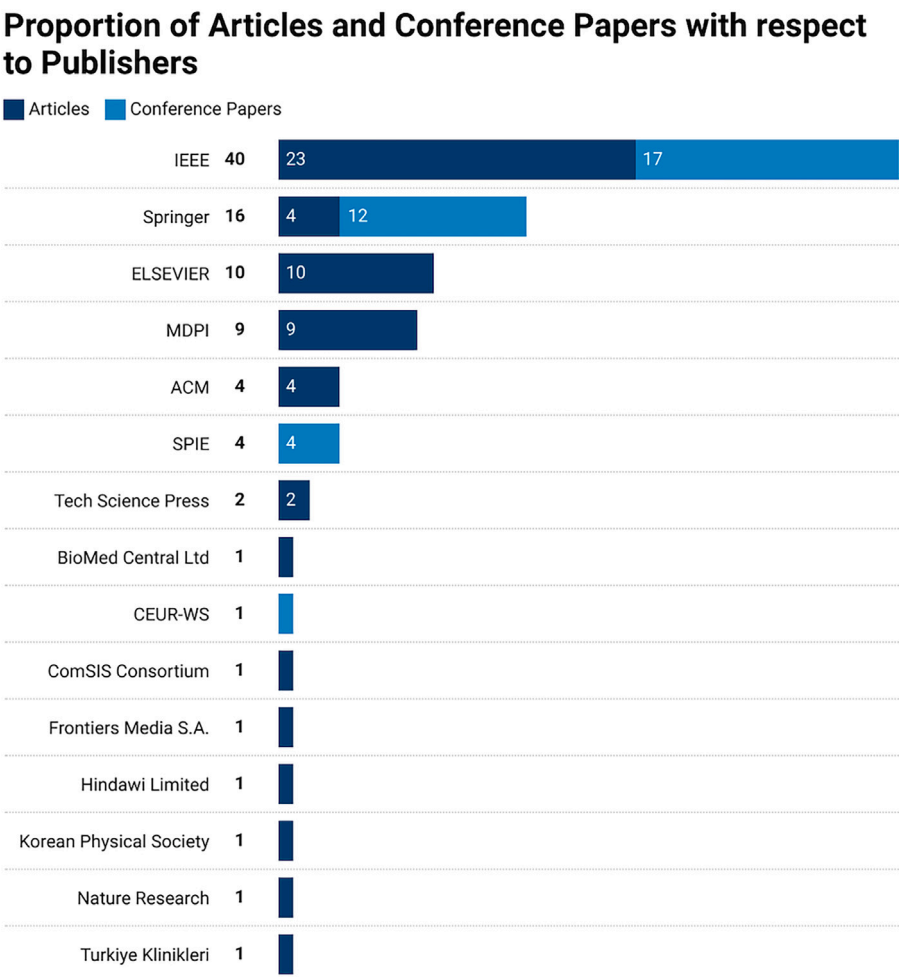

Fig. 8. Articles vs Conference papers across publishers.

4. Results and discussions

In this section, we discuss our answers for the RQs presented in Table 1.

4.1. RQ1. What are the most common XAI algorithms/methods/tools used by researchers for the medical and healthcare domains?

There are multiple algorithms found in the systematic literature review of XAI in the medical and healthcare domain. We have distributed these algorithms in two parts. One is related to algorithms of XAI and the other is for ML algorithms.

4.1.1. XAI algorithms

LIME and SHAP are the two top most used algorithms in XAI [117]. Although both methods LIME and SHAP can come up with similar results, we have seen this in most of the papers (i.e., 13) which are using both methods. The purpose of using LIME and SHAP together is to validate explanations.

SHAP and Grad-CAM are the second most used methods in these papers. Since both SHAP and LIME are used for getting an explanation of predictions made by models. These are the models which use images, tabular or textual data [118]. There were nine studies that used just LIME method, and ten that used SHAP and Grad-CAM respectively. It is also observed that LIME is a bit flexible to implement and it also generates noisy dataset. So, it means that LIME only needs one observation to be calculated [16]. On the other hand, SHAP needs to be more structured. SHAP needs multiple observations and an entire sample to get its result calculated [17]. The precise summary of the XAI algorithms used can be seen in Table 9.

4.1.2. Machine learning & deep learning algorithms

Table 10 provides information about machine learning and deep learning-based models used with XAI algorithms for research in medical and healthcare. From this table, we can infer that deep learning algorithms like CNN, Deep Neural Network (DNN), VGG-16, and ResNet are widely used for building models, while XAI is used in junction with deep learning models for better interpretation or explainability. This information can be helpful for future researchers to decide which ML/DL models provide better performance while combining XAI algorithms with them for applications in the medical domain.

Along with the XAI algorithms explained above, we have found a portion of researchers that have used some additional tools in order to take assistance in representing the outcomes of XAI algorithms' results as well as to better explain their research. Since the domain here is of medical and healthcare, most of the researchers were made to detect medical abnormalities. As many detections are done through X-rays or some other sort of visual representation of the problematic area of the human body. Although not all of the studies have used or at least mentioned the additional tools, but those which have, most of them have used heat maps, attention maps, or activation maps in order to visually highlight the areas that are important for the abnormality that is under study.

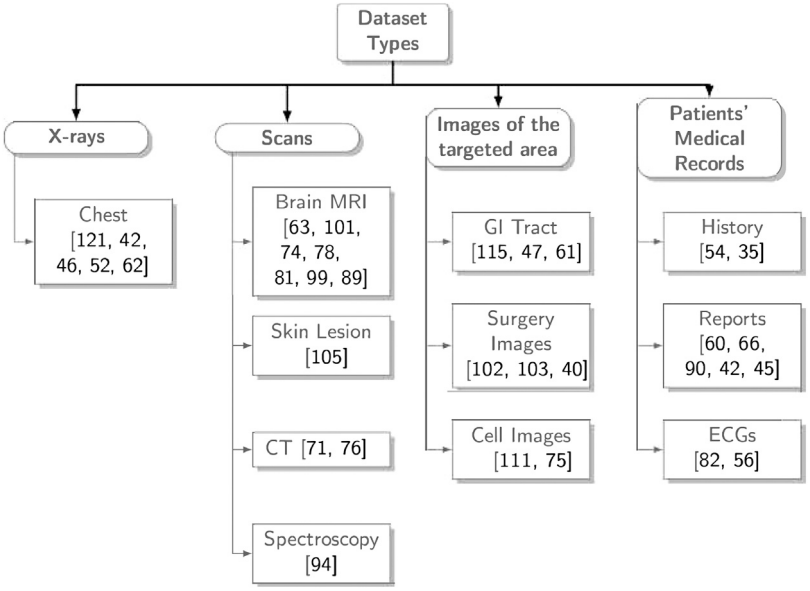

Fig. 9. Taxonomy for the types of datasets mostly used. Numbers in brackets represent the articles where those datasets are used.

Table 9
XAI algorithms used by publications.

Algorithms	Publications	Percentage
LIME	[25,32,38,39,50,55,57,58,61,66,73,75,76,79–82, 85,86,89,97,102,103,106,111,113,114,116]	30.1%
SHAP	[25,28,32,33,35,38,39,41,47,50,55,57,58,66,73, 76,82,85,91,98,100,102,106,108,113,119]	27.95%
Grad-CAM	[26,33,34,47,59,82,87,93,109,114]	0.11%
CAM	[53,65,76,102]	0.04%

4.2. RQ2. What challenges and limitations have been faced or undertaken by researchers in these domains?

In most of the studies, they did not mention the challenges and limitations. There were only 19 studies that came up with some challenges and limitations. The most common challenge we observed was about the data. Some studies mentioned that data was not enough to improve the performance of the model. In [30] they came up with the challenge of large-scale dataset as they were unable to train and test their model with large-scale dataset acquired from different cohorts. In [56] they faced the challenge of reduced accuracy of aggregated model. Because for collection and analysis of data they used multiple devices, this resulted in an increase in chances of data poisoning attacks. For one study, language was a big challenge as they implemented the data set of only Norwegian text. But, in some studies evaluation was considered as the biggest challenge faced. They lacked the experts relating to XAI and medical fields.

4.3. RQ3. Which datasets are being prominently and mostly used in research on explainable artificial intelligence for medical and healthcare?

In this literature review, the datasets utilized in various studies play a pivotal role in driving advancements in AI research for medical applications. Among the extensive list of datasets presented in Tables 11–13, three datasets stand out as the most commonly used. To gain insights into the types of these datasets, we can refer to Fig. 9, which provides a taxonomy of dataset types used in the reviewed studies. Let us delve deeper into the dimensionality and quality of these datasets to understand the significance and potential challenges when applied to medical AI research. Understanding the dimensionality and quality of these datasets is essential for assessing the potential benefits and challenges associated with explainability in medical AI research.

4.3.1. MIMIC dataset

The MIMIC dataset stands out as a widely utilized and versatile dataset in the medical domain, known for its comprehensive electronic health records (EHRs) of intensive care unit patients. To achieve explainability in AI models using MIMIC, the dimensionality of the data needs to be considered carefully. Analyzing the multitude of patient attributes present in each record will provide insights into the complexity of medical data being handled by XAI techniques. Ensuring the quality of MIMIC is crucial for building reliable and transparent models. With EHRs, data accuracy and completeness are critical factors that can impact the explainability of AI algorithms. Furthermore, addressing potential biases present in the EHRs, such as demographic variations or medical conditions prevalence, becomes essential to develop fair and interpretable models in the medical XAI context.

4.3.2. Chest X-ray datasets

The chest X-ray datasets, such as CheXpert, are extensively used in medical XAI research for diagnosing various pulmonary anomalies. Understanding the dimensionality of the radiographic features and the richness of expert annotations in these datasets is essential for designing explainable AI models for chest X-ray analysis. In medical diagnosis, explainability is crucial to gaining trust and acceptance from medical professionals. Ensuring the quality of the chest X-ray datasets, including accurate annotations and addressing potential confounding factors, will pave the way for interpretable AI models that can provide meaningful insights to radiologists and aid in early disease detection and diagnosis.

4.3.3. Kvasir dataset

The Kvasir dataset is prominently used in medical XAI research for analyzing gastrointestinal endoscopy images. Understanding the dimensionality and diversity of endoscopic findings present in the Kvasir dataset is vital for building explainable AI models for gastroenterological applications. Interpretable AI in gastroenterology can assist clinicians in making informed decisions and improving patient care. Ensuring the quality of the Kvasir dataset, including reliable annotations and addressing potential challenges in endoscopic imaging, will enable the development of transparent and trustworthy AI systems for diagnosing gastrointestinal conditions.

Table 10
ML/DL algorithms used in publications.

Algorithms	Publications	Percentage
CNN	[26,32,36,46,47,49,53,56,61,62,64,65,70,71,74,77,81,86–89,98,99,101–104,109,111,113,114]	33.33%
DNN	[27,63,93,108]	0.04%
KNN	[28,31,32,41,89,107]	0.06%
XGBoost	[27,28,35,38,39,41,44,54,58,73,79,80,89,106]	0.15%
SVM	[25,28,38,51,63,64,74,80,83,88,106,114]	0.13%
VGG-16	[26,30,32,53,62,75,78,81,87,93,97,105]	0.13%
ResNet	[26,30,34,36,40,75,81,85,87,102,104,109,112]	0.14%
Decision Tree	[28,44,54,67,79,80,91,106,107]	0.1%
Random Forest	[25,28,31,38,44,54,58,80,91,92,95,106,107]	0.13%
Ada Boost	[28,54,58,60]	0.04%
LRP	[46,58,78,98,100,119]	0.06%
DeepLIFt	[33,98,100,119]	0.04%
GBP	[98,100,119]	0.03%
MLP	[31,64,111]	0.03%
ReLU	[33,77,82,101]	0.04%
ANN	[58,88]	0.02%
Logistic Regression	[44,106,107]	0.03%
ANFIS, COBA, CUBA	[86]	0.01%
Naive Bayes	[106]	0.01%
SeNet	[36]	0.01%
CovNet	[76]	0.01%
U-Net	[76,96]	0.02%
Dense-Sharp, APN, NSAM	[43]	0.01%
TabNet, DFS, Bayesian Network	[48]	0.01%
ChexNet	[70]	0.01%
CIU	[113]	0.01%

Table 11
Dataset Summary 1/3.

Study	Dataset	Link	Modality
[104]	Medical MNIST	Link	Image
[104]	Fashion MNIST	Link	Image
[85]	ImageCLEFmedical 2021	Link	Image
[105]	Skin Lesion Images for Melanoma Classification	Link	Image
[59]	Fine-Grained Annotated Diabetic Retinopathy (FGADR) dataset	Link	Image
[106]	INbreast	Link	Image and Text
[45]	MIMIC-IV database	Link	Text
[46]	COVID-19 chest X-ray (CXR) dataset collected from Northern Italy	Link	Image
[46]	Pediatric Pneumonia dataset of Labeled Optical Coherence Tomography (OCT) and Chest XRay Images from with directories CNV, DME, DRUSEN, and NORMAL	Link	Image
[46]	Chest X-ray (Pneumonia, COVID-19, Tuberculosis) Dataset	Link	Image
[75]	Leukemia Classification Dataset	Link	Image
[86]	Fundus Images of Glaucoma Patients	Link	Image
[115]	Kvasir-v2	Link	Image
[47]	Kvasir-SEG	Link	Image
[52]	COVID-19 chest X-ray dataset	Link	Image
[52]	ISIC 2017 skin lesion dataset	Link	Image
[61]	WCE video dataset (collected from gastroenterology department of PSRI)	Closed	Video
[113]	ASD Screening Dataset	Link	Text
[114]	Kvasir	Link	Image
[116]	A large dataset of text was obtained from a public Norwegian information website: ung.no.	N/A	Text
[73]	Datasets from Western Australia Department of Health were used.	Closed	Not specified
[65]	RadioPaedia database	Link	Image
[53]	Thermograms were obtained from Selcuk University's NICU (Turkey) for this study	Closed	Image
[90]	not mentioned	Closed	N/A
[71]	OPHDIAT Image-level evaluation dataset	Closed	Image
[34]	Felipe Kitamura's CT dataset	Link	Image
[97]	Chronic wound data repository was collected from eKare Inc. data repository	Closed	Image
[26]	ProstateX	Link	Image
[92]	Heart disease database from UCI	Link	Medical records
[108]	Collected from patients with SARS-CoV-2 infection from UK Health Research Authority	Closed	Not specified
[48]	Shanghai Ruijin Hospital's mammography data set	Closed	Image
[93]	Images containing Pneumonia lesions	Closed	Image

4.4. RQ4. Which performance metrics are commonly considered in XAI research in the medical and healthcare domains?

The performance of machine learning-based algorithms is important because it tells the accuracy of predictions done by the algorithm and it helps build trust in the model. The performance of AI models used for medical diagnoses and healthcare devices is critical because the failure of AI models used in the medical domain can be life-threatening [123]. In Table 14 we have presented performance metrics used in the publications we are considering for this review. Most of the papers have used evaluation metrics used for AI and ML algorithms because XAI is applied to explain the AI/ML models used in these research publications.

From data presented in Table 14 we can say that accuracy, precision, recall, and F1-Score are the performance metrics that are mostly

Table 12
Dataset Summary 2/3.

[38]	Artificial Simulacrumhealth dataset	Link	Medical records
[38]	UK Covid-19 patients data	Closed	Not specified
[39]	Data from the Simulacrum dataset from NCRAS, England was used	Link	Medical Records
[51]	Toddler Dataset	Link	Medical records
[109]	Retinal Fundus Image Quality Assessment (RFIQA) dataset	Link	Image
[109]	EyePacs dataset	Link	Image
[28]	Pathological voice samples of people with vocal cord polyp and paralysis were obtained from an unknown source	Closed	Audio
[54]	Dataset of Heart Failure Patients from UCI	Link	Medical records
[67]	Cervical Cancer Risk Factors	Link	Medical records
[112]	Two in-house mammogram datasets: "Data A" and "Data B". Data A was gathered from four medical centers and Data B was acquired from a separate single medical center.	Closed	Image
[110]	UCI's Cleveland heart disease database and the Framingham Heart Study Repository	Link	Medical records
[111]	Cell Images for detecting Malaria	Link	Image
[99]	T1-weighted DCE-MRI scans from six institutions were collected and used	[120]	Image
[40]	Cholec80	Link	Video
[43]	LIDC-IDRI dataset	Link	Image
[62]	Chest Xray Dataset collected by[121]	Link	Image
[76]	CT volume data from four hospitals in China (Private), CC-CCII	Open	Image
[78]	MRI Alzheimer brain image dataset	Open	Image
[33]	SPECT image dataset	Open	Image
[88]	EEG dementia diagnosis dataset by[122]	Open	Medical records
[79]	sEMG Dataset	Closed	Image
[57]	Breast Cancer Dataset by the University of California	Link	Medical records
[81]	T1 weighted MRI Dataset	Link	
[37]	An automated regular expression based searching was used to find potential veterans with PTSD from twitter	Not specified	Text

Table 13
Dataset Summary 3/3.

[58]	A web-based survey conducted from July 13 to July 17, 2020 was used to collect data that is available on Kaggle	Link	Image
[82]	MIT-BIH Arrhythmia database	Open	Medical records
[94]	Infants' Functional near-infrared spectroscopies (fNIRS) were used	Link	Image
[83]	National Health and Nutrition Exam Survey (NHANES) III	Link	Text
[89]	MRI images of AD and microarray gene expression were used	Link	Image
[25]	Alzheimers Dataset	Link	Image
[32]	Breast Cancer Wisconsin (Diagnostic) Dataset	Link	Medical records
[35]	Patient clinical information for TCGA breast-invasive carcinoma cohort (BRCA) from two projects on the cbioPortal were used.	Link	Medical records
[35]	Clinical information for 1101 patients from Firehouse Legacy	Link	Medical records
[96]	APTOS 2019 Blindness Detection Dataset	Link	Medical records
[55]	Dementia Prediction w/ Tree-based Models Dataset	Link	Medical records

used for AI/ML algorithms used in combination with XAI for research in the medical/healthcare domain. This information can help researchers interested in XAI in the future to decide on evaluation metrics for use in their research for benchmarking. These evaluation metrics can be used to check the performance of AI models accompanied by XAI algorithms. Furthermore, researchers can also refer to related studies to check how these performance metrics were used to evaluate the experiments and how experiments performed in this direction can be improved with the help of these performance metrics.

Fig. 10 shows the frequency of performance metrics utilized to evaluate three Explainable Artificial Intelligence (XAI) techniques—LIME, SHAP, and Grad-CAM—in the context of medical and healthcare applications. The numbers represent the count of papers that have employed each specific metric for evaluation. Accuracy was the most commonly used metric, with 15 papers assessing it for LIME, 10 for SHAP, and 3 for Grad-CAM. Precision and Recall were also frequently used, with 5 papers using Precision for LIME, 3 for SHAP, and 1 for Grad-CAM, and 5 papers employing Recall for LIME, 3 for SHAP, and 2 for Grad-CAM. F1-Score was used in 4 papers for LIME, 3 for SHAP, and 2 for Grad-CAM. Area Under Curve (AUC), a metric suitable for binary classification tasks, was assessed in 1 paper for LIME, 3 for SHAP, and 3 for Grad-CAM. Sensitivity and Specificity were both used in 2 papers for LIME, 3 for SHAP, and 3 for Grad-CAM. These numbers demonstrate that various performance metrics have been widely applied to comprehensively evaluate the XAI techniques, providing a comprehensive analysis of their effectiveness in medical and healthcare scenarios.

4.4.1. *Metrics discussion:*

In this section, we provide a comprehensive discussion on the performance metrics utilized to evaluate the Explainable Artificial Intelligence (XAI) techniques in the context of medical and healthcare applications. The selection of appropriate performance metrics is crucial in assessing the effectiveness and interpretability of AI models, which play a vital role in critical domains like healthcare.

Firstly, the metrics of Accuracy, Precision, Recall, and F1-Score have been widely used in evaluating the XAI techniques' performance. Accuracy measures the overall correctness of predictions, while Precision quantifies the ratio of true positive predictions to total positive predictions. Recall, also known as Sensitivity, represents the ability to correctly identify positive instances. The F1-Score is the harmonic mean of Precision and Recall, providing a balanced performance measure. These metrics are essential in evaluating the AI models' interpretability and their capability to identify relevant features in medical data, thereby aiding in informed decision-making by healthcare professionals.

Secondly, the Area Under the Curve (AUC) has been employed as a performance metric, particularly in binary classification problems. AUC represents the ability of AI models to distinguish between positive and negative instances. In the medical context, where identifying critical conditions accurately is crucial, AUC serves as a valuable metric to gauge the model's effectiveness in making accurate predictions and achieving high interpretability.

Thirdly, the discussion extends to Sensitivity and Specificity, which are vital in evaluating the XAI techniques' ability to correctly identify

Table 14
Performance Metrics and their usage.

Performance metrics	Publications	Count
Accuracy	[28,32,34,39,41,43–46,48,51,53–56,61–63,65,67,69,75–81,84,86–88,91,92,95,96,102,104,106,107,111,113–115]	44
Precision	[27,28,36,38,39,44,46,54,56,58,60–62,67,69,86,87,91,93,97,113–116]	24
Recall	[27,28,36,38,39,44,46,54–56,58,61,67,69,86,91,93,97,111,113,114,116]	22
F1-Score	[25,27,28,36,38,44,46,54,56,58,61,69,75,79,86,87,91–93,97,102,111,113,114,116]	25
Area Under Curve (AUC)	[27,28,30,36,41,45,54,55,70,91,97,102,109,114]	14
Sensitivity	[33,34,41,51,53,54,63,87,98,102,115]	11
Specificity	[33,34,41,51,53,54,59,63,87,115]	10
Receiver Operating Characteristic (ROC)	[31,54,70,71,90,97,115]	7
Computational Runtime	[65,98,100,114]	4
PR Curves	[61,71,115]	3
Similarity	[42,82,98]	3
Matthews correlation coefficient	[28,87,91]	3
Dice coefficient index	[59,63]	2
Average	[26,31]	2
PSNR	[49,65]	2
Root Mean Squared Error	[100,103]	2
Confidence	[42]	1
Training Time	[87]	1
AUPRC	[36]	1
Standard Deviation	[64]	1
Zero Rule Baseline	[64]	1
BLEU Score	[85]	1
Jacard Index Metrics	[59]	1
Mean Average Precision	[59]	1
Mean IOU	[47]	1
Dice Loss	[47]	1
Loss	[75]	1

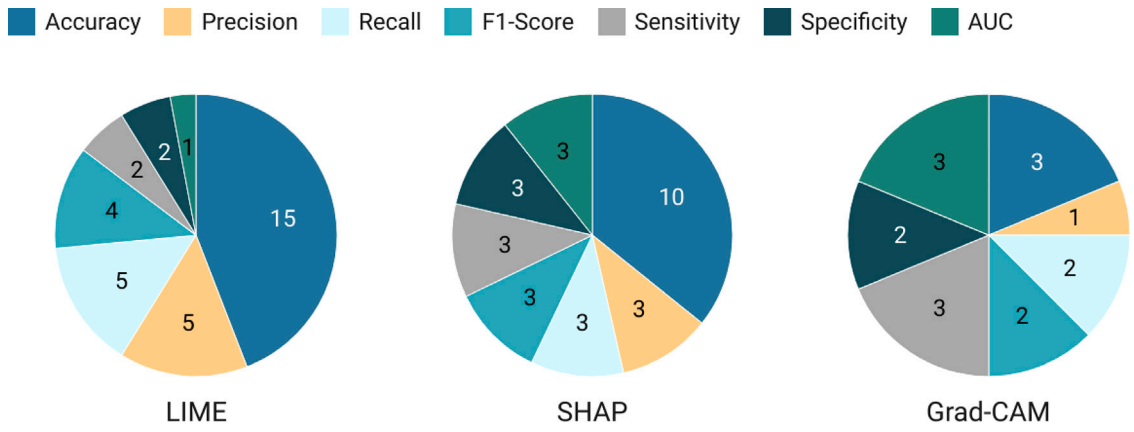

Fig. 10. Performance used with XAI algorithms.

positive and negative instances, respectively. In medical and healthcare applications, where false negatives or false positives can have severe consequences, these metrics play a pivotal role in assessing the models' reliability and interpretability.

Finally, we acknowledge potential issues and limitations associated with the use of certain metrics. For example, while Accuracy is a widely used metric, it may not be the most appropriate choice in cases of class imbalance, where a skewed distribution of data affects its interpretability. It is essential to consider such limitations when interpreting the results and selecting the most suitable metrics for a given medical and healthcare application.

In summary, the selection of performance metrics is critical in evaluating the effectiveness and interpretability of XAI techniques in medical and healthcare applications. The metrics of Accuracy, Precision, Recall, F1-Score, AUC, Sensitivity, and Specificity provide valuable insights into the models' reliability and their ability to make accurate predictions while being interpretable. Acknowledging potential issues and limitations associated with certain metrics ensures a comprehensive evaluation, enabling the effective deployment of XAI in critical healthcare scenarios.

5. Open challenges and future directions

From the literature, we have found that most papers have used performance metrics that are commonly used for AI models for model evaluation, i.e., accuracy, recall, precision, and F1-Score but there are no performance metrics specific to evaluate the results of XAI algorithms. Researchers have used AI performance metrics for the evaluation of XAI results, hence suggesting specific evaluation metrics for XAI can be interesting to work upon for future researchers. Moreover, an XAI performance evaluation method could also be proposed that is dedicated to the medical and healthcare field. It should be achieved in collaboration with medical experts.

Another promising future direction is to assess the individual contribution of each XAI technique when more than one technique is being exploited. It is important to highlight the fact that when multiple explanation techniques are being used in conjunction then it is not necessary that each interpretation technique should have an equal contribution to the final result. There should be an assignment of weight to each XAI technique in the explanation, the weighted combination will be helpful to differentiate the contribution of each XAI technique, and the amount of contribution in the final explanation will determine the weight of the explanation technique being used. Investigating individual contributions of each XAI technique when used together thus can be an interesting topic to work on.

Explaining the features and their importance through XAI could also help in understanding the models used in the medical and healthcare domains, but looking for the process behind the decision-making of models used in medical can also be a potential future direction. Moreover, due to XAI explanations, it is easy to determine the region of interest for different medical imaging problems. Medical experts can use such explanations with their field knowledge and propose alternative diagnosis methods for different diseases.

Most of the pre-trained models such as Inception, VGG16, etc, restrict the image to be of a certain size in order to use them for classification. Whereas, medical images, such as X-rays, scans, and MRIs, lose their quality and meaning when resized. A pre-trained model could be proposed that does not require resizing images in order to perform classification on them.

Also, when the dataset is small, there is a chance of the model overfitting. XAI explanations can help understand the underlying, important features of a particular prediction. XAI can help identify features that are not so important and those that are creating noise and problems for training data so those problematic features can be avoided in order to avoid model over-fitting. Potential researchers can study this area in depth to make accurate conclusions about the contributions of XAI for avoiding model over-fitting problems.

Labeling accuracy and efficiency are directly related to the quality of the initial training set. Some studies faced limitations in the feature selection method, which was found to be slightly more time-consuming, impacting the overall efficiency of the methodology. Additionally, the lack of prospective quality-of-life data in certain studies posed a limitation in understanding the long-term impacts of AI applications. One common limitation across several studies was the reliance on specific types of data, such as 2-D MRI images or solely using GRF signals for classification. This limited data variety may restrict the models' generalizability and applicability to a broader range of medical scenarios. However, some studies attempted to mitigate this limitation by training and testing their models on large-scale datasets acquired from different cohorts.

The use of federated learning raised concerns about data integrity and authentication due to the distributed nature of data collection and analysis across multiple devices. Furthermore, the assumption of homogeneous data and devices in some proposed frameworks may not hold true in real-world scenarios, potentially reducing the accuracy of aggregated models. For applications like telesurgical operations, researchers faced challenges in minimizing errors during virtual surgery control, ensuring real-time operation with minimal latency, and maintaining privacy and security during remote surgical procedures.

In certain studies, models encountered difficulties in generating captions for some words due to data preprocessing and the presence of unknown words, leading to limitations in caption generation capabilities. In the context of acute kidney injury (AKI) prediction, limitations arose from missing laboratory parameters, the absence of recorded etiology in the databases used, and the exclusive use of a single-centered dataset without external validation. An external multicenter validation is suggested to enhance the reliability of the models. While some studies claimed the adaptability of their models to other languages using translated features, it remains essential to consider potential limitations and differences in language-specific data.

Restrictions imposed by pre-trained models, such as fixed input sizes, required resizing of input data and may have impacted model performance in certain applications. Proposed solutions for foreground/background separation were observed to be limited to binary or multilabel classification, presenting challenges in directly applying them to multiclass classification scenarios. To improve model performance, researchers suggested utilizing larger training datasets and fine-tuning hyperparameters. However, the interpretability of models could be limited by equivalent weights assigned to explanations in certain combination frameworks. Scalability issues were observed with certain techniques, such as LORE, which hindered their application in larger-scale experiments. The data-hungriness of some medical AI models presented a significant challenge, making it difficult to transfer learned models from one task to another.

Lastly, limitations in literature coverage and the lack of proper evaluation of explainability in many medical XAI applications were noted, potentially hindering the adoption and understanding of these models by medical experts. The field of XAI currently lacks benchmark datasets. In other AI fields, such as classification, clustering, and segmentation, benchmark datasets are readily available, facilitating progress. However, in XAI, there is currently no benchmark dataset that researchers can use to establish a common platform and agreed-upon performance metrics. This absence of a benchmark dataset impedes the progress of the field. Therefore, the development of a proper benchmark dataset could prove to be a watershed moment in the field of XAI.

6. Practical implications of XAI in healthcare

Explainable Artificial Intelligence (XAI) has gained increasing attention in the healthcare domain due to its potential to improve patient outcomes and enhance medical decision-making. In this section, we discuss the practical implications of employing XAI techniques in healthcare settings, highlighting how its implementation has positively impacted medical practices.

6.1. Improved patient outcomes and informed decision-making:

One of the key practical implications of XAI in healthcare is its ability to improve patient outcomes by providing transparent insights into the decision process of AI models. XAI techniques, such as LIME, SHAP, and Grad-CAM, have enabled healthcare professionals to gain a deeper understanding of the features contributing to model predictions. This interpretability empowers clinicians to make more informed and confident decisions, leading to accurate diagnoses, optimized treatment plans, and better patient care.

6.2. Bias detection and mitigation:

Another practical implication of XAI in healthcare is its role in identifying potential biases in AI models. By revealing biases in the data and model outputs, XAI allows healthcare systems to address issues related to fairness and equity. This ensures that AI-powered healthcare interventions are more inclusive and provide equitable treatment for all patient groups.

Personalized Medicine and Tailored Treatment Plans: XAI has facilitated the adoption of personalized medicine in healthcare. Through the transparent and interpretable nature of AI models, XAI enables healthcare professionals to tailor treatment plans to individual patients based on their unique medical history and characteristics. This personalized approach has led to improved treatment efficacy and patient satisfaction.

6.3. Reduced medical errors and early disease detection:

Documented cases and empirical evidence from various healthcare institutions indicate that the implementation of XAI has contributed to reducing medical errors and early detection of diseases. The interpretability of AI models allows healthcare providers to catch potential errors and identify anomalies in medical data, leading to timely interventions and better patient outcomes.

6.4. Case studies and empirical evidence:

A wealth of case studies and empirical evidence exists, demonstrating the practical impact of XAI in healthcare. These documented cases showcase instances where the implementation of XAI has significantly improved medical decision-making, diagnostic accuracy, and patient care.

The practical implications of XAI in healthcare are far-reaching and hold immense potential for transforming medical practices. The transparency and interpretability of AI models provided by XAI techniques have improved patient outcomes, facilitated personalized medicine, and reduced medical errors. This section highlights the real-world benefits of incorporating XAI in healthcare settings and underscores its significance in revolutionizing medical practices.

7. Conclusion

AI and ML in particular have been advancing quite rapidly in the last decade, however relying completely on the results of AI-based algorithms in sensitive fields is difficult, especially in the medical field. Thus, the field of XAI was introduced, which aims to explain the results derived by the AI or ML models. It explains how the model has reached a certain conclusion. This increases the credibility of the models to be used by medical practitioners to aid in their manual practices. In this SLR, we have targeted the articles from the last five years that have discussed or used XAI for the said domain. We ended up with a total of 93 studies after a thorough selection process.

We carried out information such as the most common algorithms being used in the domain of XAI for medical and healthcare, which included both ML and XAI algorithms. LIME was the most talked about and used in most of the studies. We have also discussed LIME and how it works since it was being prominently used. After LIME, came the SHAP, CAM, and GradCAM which we have also discussed in the Related Surveys Section 2. In addition to that, we observed the limitations and challenges of the proposed study. Moreover, we proposed to find out the datasets that are being used most commonly for these studies, and we discovered that not much can be said as there was a lot of variation found. However, COVID-19 X-rays were used more commonly, from different regions of the world.

As there are only a few hospitals or clinics that make their data available for research, and even if they do make it available, it is only shared privately with the researchers. Researchers in the medical domain experience a very common issue which is the lack of medical image data. They have to use various techniques to combat that, such as using pre-trained models and doing synthetic image generation.

CRediT authorship contribution statement

Subhan Ali: Conceptualization, Methodology, Data collection, Original draft preparation, Writing – review & editing. **Filza Akhlaq:** Data collection, Original draft preparation, Writing – review & editing. **Ali Shariq Imran:** Conceptualization, Methodology, Review & editing, Supervision. **Zenun Kastrati:** Review & editing, Supervision. **Sher Muhammad Daudpota:** Review & editing, Supervision. **Muhammad Moosa:** Data collection, Methodology, Writing.

Declaration of competing interest

The authors declare that there are no known competing interests that could affect this work.

References

[1] D. William, D. Suhartono, Text-based depression detection on social media posts: A systematic literature review, Procedia Comput. Sci. 179 (2021) 582–589.

[2] J.M. Havigerová, J. Haviger, D. Kučera, P. Hoffmannová, Text-based detection of the risk of depression, Front. Psychol. 10 (2019) 513.

[3] T. Al Hanai, M.M. Ghassemi, J.R. Glass, Detecting depression with audio/text sequence modeling of interviews, in: Interspeech, 2018, pp. 1716–1720.

[4] J. Ye, Y. Yu, Q. Wang, W. Li, H. Liang, Y. Zheng, G. Fu, Multi-modal depression detection based on emotional audio and evaluation text, J. Affect. Disord. 295 (2021) 904–913.

[5] P. Rathod, S. Naik, Review on epilepsy detection with explainable artificial intelligence, in: 2022 10th International Conference on Emerging Trends in Engineering and Technology-Signal and Information Processing, ICETET-SIP-22, IEEE, 2022, pp. 1–6.

[6] M. Miró-Nicolau, G. Moyà-Alcover, A. Jaume-i Capó, Evaluating explainable artificial intelligence for X-ray image analysis, Appl. Sci. 12 (2022) 4459.

[7] C.C. Yang, Explainable artificial intelligence for predictive modeling in healthcare, J. Healthc. Inform. Res. 6 (2022) 228–239.

[8] A.M. Antoniadi, Y. Du, Y. Guendouz, L. Wei, C. Mazo, B.A. Becker, C. Mooney, Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: A systematic review, Appl. Sci. 11 (2021) 5088.

[9] L. Caroprese, E. Vocaturo, E. Zumpano, Argumentation approaches for explanaible AI in medical informatics, Intell. Syst. Appl. 16 (2022) 200109.

[10] J. Ooge, K. Verbert, Explaining artificial intelligence with tailored interactive visualisations, in: 27th International Conference on Intelligent User Interfaces, 2022, pp. 120–123.

[11] S.M. Mathews, Explainable artificial intelligence applications in NLP, biomedical, and malware classification: A literature review, in: Intelligent Computing-Proceedings of the Computing Conference, Springer, 2019, pp. 1269–1292.

[12] H. Hakkoum, I. Abnane, A. Idri, A systematic map of interpretability in medicine, in: Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies - HEALTHINF, INSTICC, SciTePress, 2022, pp. 719–726, http://dx.doi.org/10.5220/0010968700003123.

[13] P. Korica, N.E. Gayar, W. Pang, Explainable artificial intelligence in healthcare: Opportunities, gaps and challenges and a novel way to look at the problem space, in: International Conference on Intelligent Data Engineering and Automated Learning, Springer, 2021, pp. 333–342.

[14] S. Chakrobartty, O. El-Gayar, Explainable artificial intelligence in the medical domain: A systematic review, 2021.

[15] M. Nazar, M.M. Alam, E. Yafi, M. Mazliham, A systematic review of human-computer interaction and explainable artificial intelligence in healthcare with artificial intelligence techniques, IEEE Access (2021).

[16] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[17] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017).

[18] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.

[19] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.

[20] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, Harv. JL Tech. 31 (2017) 841.

[21] M.T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, 2018.

[22] P.W. Koh, P. Liang, Understanding black-box predictions via influence functions, in: International Conference on Machine Learning, PMLR, 2017, pp. 1885–1894.

[23] M. Saleemi, M. Anjum, M. Rehman, eServices classification, trends, and analysis: A systematic mapping study, IEEE Access 5 (2017) 26104–26123.

[24] B. Kitchenham, O.P. Brereton, D. Budgen, M. Turner, J. Bailey, S. Linkman, Systematic literature reviews in software engineering—A systematic literature review, Inf. Softw. Technol. 51 (2009) 7–15.

[25] G. Marvin, M.G.R. Alam, Explainable feature learning for predicting neonatal intensive care unit (NICU) admissions, in: 2021 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health, BECITHCON, IEEE, 2021, pp. 69–74.

[26] M.A. Gulum, C.M. Trombley, M. Kantardzic, Improved deep learning explanations for prostate lesion classification through grad-CAM and saliency map fusion, in: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems, CBMS, IEEE, 2021, http://dx.doi.org/10.1109/cbms52027.2021.00099.

[27] A. Laios, E. Kalampokis, R. Johnson, S. Munot, A. Thangavelu, R. Hutson, T. Broadhead, G. Theophilou, C. Leach, D. Nugent, D.D. Jong, Factors predicting surgical effort using explainable artificial intelligence in advanced stage epithelial ovarian cancer, Cancers 14 (2022) 3447, http://dx.doi.org/10.3390/cancers14143447.

[28] N. Seedat, V. Aharonson, Y. Hamzany, Automated and interpretable m-health discrimination of vocal cord pathology enabled by machine learning, in: 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering, CSDE, IEEE, 2020, http://dx.doi.org/10.1109/csde50874.2020.9411529.

[29] N. Mollaei, C. Fujao, L. Silva, J. Rodrigues, C. Cepeda, H. Gamboa, Human-centered explainable artificial intelligence: Automotive occupational health protection profiles in prevention musculoskeletal symptoms, Int. J. Environ. Res. Public Health 19 (2022) 9552, http://dx.doi.org/10.3390/ijerph19159552.

[30] S.M. Hussain, D. Buongiorno, N. Altini, F. Berloco, B. Prencipe, M. Moschetta, V. Bevilacqua, A. Brunetti, Shape-based breast lesion classification using digital tomosynthesis images: The role of explainable artificial intelligence, Appl. Sci. 12 (2022) 6230, http://dx.doi.org/10.3390/app12126230.

[31] V. Sargiani, A.A.D. Souza, D.C.D. Almeida, T.S. Barcelos, R. Munoz, L.A.D. Silva, Supporting clinical COVID-19 diagnosis with routine blood tests using tree-based entropy structured self-organizing maps, Appl. Sci. 12 (2022) 5137, http://dx.doi.org/10.3390/app12105137.

[32] Y. Zhang, Y. Weng, J. Lund, Applications of explainable artificial intelligence in diagnosis and surgery, Diagnostics 12 (2022) 237.

[33] T. Pianpanit, S. Lolak, P. Sawangjai, T. Sudhawiyangkul, T. Wilaiprasitporn, Parkinson's disease recognition using SPECT image and interpretable AI: A tutorial, IEEE Sens. J. 21 (2021) 22304–22316.

[34] K.H. Kim, H.-W. Koo, B.-J. Lee, S.-W. Yoon, M.-J. Sohn, Cerebral hemorrhage detection and localization with medical imaging for cerebrovascular disease diagnosis and treatment using explainable deep learning, J. Korean Phys. Soc. 79 (2021) 321–327, http://dx.doi.org/10.1007/s40042-021-00202-2.

[35] D. Chakraborty, C. Ivan, P. Amero, M. Khan, C. Rodriguez-Aguayo, H. Başağaoğlu, G. Lopez-Berestein, Explainable artificial intelligence reveals novel insight into tumor microenvironment conditions linked with better prognosis in patients with breast cancer, Cancers 13 (2021) 3450.

[36] A. Anand, T. Kadian, M.K. Shetty, A. Gupta, Explainable AI decision model for ECG data of cardiac disorders, Biomed. Signal Process. Control 75 (2022) 103584, http://dx.doi.org/10.1016/j.bspc.2022.103584.

[37] M.A.U. Alam, D. Kapadia, Laxary: A trustworthy explainable twitter analysis model for post-traumatic stress disorder assessment, in: 2020 IEEE International Conference on Smart Computing, SMARTCOMP, IEEE, 2020, pp. 308–313.

[38] M. Kapcia, H. Eshkiki, J. Duell, X. Fan, S. Zhou, B. Mora, ExMed: An AI tool for experimenting explainable AI techniques on medical data analytics, in: 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence, ICTAI, IEEE, 2021, http://dx.doi.org/10.1109/ictai52525.2021.00134.

[39] J. Duell, X. Fan, B. Burnett, G. Aarts, S.-M. Zhou, A comparison of explanations given by explainable artificial intelligence methods on analysing electronic health records, in: 2021 IEEE EMBS International Conference on Biomedical and Health Informatics, BHI, IEEE, 2021, http://dx.doi.org/10.1109/bhi50953.2021.9508618.

[40] D.R. Chittajallu, B. Dong, P. Tunison, R. Collins, K. Wells, J. Fleshman, G. Sankaranarayanan, S. Schwaitzberg, L. Cavuoto, A. Enquobahrie, XAI-CBIR: Explainable AI system for content based retrieval of video frames from minimally invasive surgery videos, in: 2019 IEEE 16th International Symposium on Biomedical Imaging, ISBI 2019, IEEE, 2019.

[41] V.C. Pezoulas, A. Liontos, E. Mylona, C. Papaloukas, O. Milionis, D. Biros, C. Kyriakopoulos, K. Kostikas, H. Milionis, D.I. Fotiadis, Predicting the need for mechanical ventilation and mortality in hospitalized COVID-19 patients who received heparin, in: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC, IEEE, 2022, pp. 1020–1023.

[42] D. Kim, J. Chung, J. Choi, M.D. Succi, J. Conklin, M.G.F. Longo, J.B. Ackman, B.P. Little, M. Petranovic, M.K. Kalra, M.H. Lev, S. Do, Accurate auto-labeling of chest X-ray images based on quantitative similarity to an explainable AI model, Nature Commun. 13 (2022) http://dx.doi.org/10.1038/s41467-022-29437-8.

[43] J. Yang, R. Fang, B. Ni, Y. Li, Y. Xu, L. Li, Probabilistic radiomics: Ambiguous diagnosis with controllable shape analysis, in: Lecture Notes in Computer Science, in: Lecture notes in computer science, Springer International Publishing, Cham, 2019, pp. 658–666.

[44] J. Dieber, S. Kirrane, A novel model usability evaluation framework (MUsE) for explainable artificial intelligence, Inf. Fusion 81 (2022) 143–153, http://dx.doi.org/10.1016/j.inffus.2021.11.017.

[45] C. Hu, Q. Tan, Q. Zhang, Y. Li, F. Wang, X. Zou, Z. Peng, Application of interpretable machine learning for early prediction of prognosis in acute kidney injury, Comput. Struct. Biotechnol. J. 20 (2022) 2861–2870, http://dx.doi.org/10.1016/j.csbj.2022.06.003.

[46] G. Marvin, M.R. Alam, Explainable augmented intelligence and deep transfer learning for pediatric pulmonary health evaluation, in: 2022 International Conference on Innovations in Science, Engineering and Technology, ICISET, IEEE, 2022, http://dx.doi.org/10.1109/iciset54810.2022.9775845.

[47] S.M. Javali, R.S. Upadhyayula, T. De, Comparative study of xAI layer-wise algorithms with a robust recommendation framework of inductive clustering for polyp segmentation and classification, in: 2021 International Seminar on Machine Learning, Optimization, and Data Science, ISMODE, IEEE, 2022, http://dx.doi.org/10.1109/ismode53584.2022.9743003.

[48] D. Chen, H. Zhao, J. He, Q. Pan, W. Zhao, An causal XAI diagnostic model for breast cancer based on mammography reports, in: 2021 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2021, http://dx.doi.org/10.1109/bibm52615.2021.9669648.

[49] G. Dong, Y. Ma, A. Basu, Feature-guided CNN for denoising images from portable ultrasound devices, IEEE Access 9 (2021) 28272–28281, http://dx.doi.org/10.1109/access.2021.3059003.

[50] L.V. Utkin, A.A. Meldo, M.S. Kovalev, E.M. Kasimov, A simple general algorithm for the diagnosis explanation of computer-aided diagnosis systems in terms of natural language primitives, in: 2020 XXIII International Conference on Soft Computing and Measurements, SCM, IEEE, 2020, pp. 202–205.

[51] M. Biswas, M.S. Kaiser, M. Mahmud, S.A. Mamun, M.S. Hossain, M.A. Rahman, An XAI based autism detection: The context behind the detection, in: Brain Informatics, Springer International Publishing, 2021, pp. 448–459, http://dx.doi.org/10.1007/978-3-030-86993-9_40.

[52] B. Hu, B. Vasu, A. Hoogs, X-MIR: Explainable medical image retrieval, in: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, IEEE, 2022, http://dx.doi.org/10.1109/wacv51458.2022.00161.

[53] A.H. Ornek, M. Ceylan, Explainable artificial intelligence (XAI): Classification of medical thermal images of neonates using class activation maps, Trait. Signal 38 (2021) 1271–1279, http://dx.doi.org/10.18280/ts.380502.

[54] P.A. Moreno-Sanchez, Development of an explainable prediction model of heart failure survival by using ensemble trees, in: 2020 IEEE International Conference on Big Data, Big Data, IEEE, 2020, http://dx.doi.org/10.1109/bigdata50022.2020.9378460.

[55] G. Loveleen, B. Mohan, B.S. Shikhar, J. Nz, M. Shorfuzzaman, M. Masud, Explanation-driven HCI model to examine the mini-mental state for alzheimer's disease, ACM Trans. Multimed. Comput. Commun. Appl. (2022).

[56] A. Raza, K.P. Tran, L. Koehl, S. Li, Designing ECG monitoring healthcare system with federated transfer learning and explainable AI, Knowl.-Based Syst. 236 (2022) 107763, http://dx.doi.org/10.1016/j.knosys.2021.107763.

[57] R. Larasati, Explainable AI for breast cancer diagnosis: Application and user's understandability perception, in: 2022 International Conference on Electrical, Computer and Energy Technologies, ICECET, IEEE, 2022, pp. 1–6.

[58] A.U. Hussna, I.I. Trisha, I.J. Ritun, M.G.R. Alam, COVID-19 impact on students' mental health: Explainable AI and classifiers, in: 2021 International Conference on Decision Aid Sciences and Application, DASA, IEEE, 2021, pp. 847–851.

[59] H. Kheradfallah, J.J. Balaji, V. Jayakumar, M.A. Rasheed, V. Lakshminarayanan, Annotation and segmentation of diabetic retinopathy lesions: An explainable AI application, in: K.M. Iftekharuddin, K. Drukker, M.A. Mazurowski, H. Lu, C. Muramatsu, R.K. Samala (Eds.), Medical Imaging 2022: Computer-Aided Diagnosis, SPIE, 2022, http://dx.doi.org/10.1117/12.2612576.

[60] J. Hatwell, M.M. Gaber, R.M.A. Azad, Ada-WHIPS: Explaining AdaBoost classification with applications in the health sciences, BMC Med. Inform. Decis. Mak. 20 (2020) http://dx.doi.org/10.1186/s12911-020-01201-2.

[61] A. Singh, H. Pannu, A. Malhi, Explainable information retrieval using deep learning for medical images, Comput. Sci. Inf. Syst. 19 (2022) 277–307, http://dx.doi.org/10.2298/csis201030049s.

[62] P. Saxena, S.K. Singh, G. Tiwary, Y. Mittal, I. Jain, An artificial intelligence technique for COVID-19 detection with explainability using lungs X-Ray images, in: 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics, ICDCECE, IEEE, 2022, pp. 1–6.

[63] S. Maqsood, R. Damaševičius, R. Maskeliūnas, Multi-modal brain tumor detection using deep neural network and multiclass SVM, Medicina 58 (2022) 1090, http://dx.doi.org/10.3390/medicina58081090.

[64] D. Slijepcevic, F. Horst, S. Lapuschkin, B. Horsak, A.-M. Raberger, A. Kranzl, W. Samek, C. Breiteneder, W.I. Schöllhorn, M. Zeppelzauer, Explaining machine learning models for clinical gait analysis, ACM Trans. Comput. Healthc. 3 (2021) 1–27, http://dx.doi.org/10.1145/3474121.

[65] A. Kumar, R. Manikandan, U. Kose, D. Gupta, S.C. Satapathy, Doctor's dilemma: Evaluating an explainable subtractive spatial lightweight convolutional neural network for brain tumor diagnosis, ACM Trans. Multimed. Comput. Commun. Appl. 17 (2021) 1–26, http://dx.doi.org/10.1145/3457187.

[66] C. Panigutti, A. Perotti, D. Pedreschi, Doctor XAI, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, ACM, New York, NY, USA, 2020.

[67] U. Pawar, D. O'Shea, S. Rea, R. O'Reilly, Incorporating explainable artificial intelligence (XAI) to aid the understanding of machine learning in the healthcare domain, in: AICS, 2020, pp. 169–180.

[68] B.H.M. van der Velden, H.J. Kuijf, K.G.A. Gilhuijs, M.A. Viergever, Explainable artificial intelligence (XAI) in deep learning-based medical image analysis, Med. Image Anal. 79 (2022) 102470.

[69] I.A. Khan, N. Moustafa, I. Razzak, M. Tanveer, D. Pi, Y. Pan, B.S. Ali, XSRU-IoMT: Explainable simple recurrent units for threat detection in internet of medical things networks, Future Gener. Comput. Syst. 127 (2022) 181–193, http://dx.doi.org/10.1016/j.future.2021.09.010.

[70] H. Liz, M. Sánchez-Montañés, A. Tagarro, S. Domínguez-Rodríguez, R. Dagan, D. Camacho, Ensembles of convolutional neural network models for pediatric pneumonia diagnosis, Future Gener. Comput. Syst. 122 (2021) 220–233, http://dx.doi.org/10.1016/j.future.2021.04.007.

[71] G. Quellec, H.A. Hajj, M. Lamard, P.-H. Conze, P. Massin, B. Cochener, ExplAIn: Explanatory artificial intelligence for diabetic retinopathy diagnosis, Med. Image Anal. 72 (2021) 102118, http://dx.doi.org/10.1016/j.media.2021.102118.

[72] A. Holzinger, B. Malle, A. Saranti, B. Pfeifer, Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI, Inf. Fusion 71 (2021) 28–37, http://dx.doi.org/10.1016/j.inffus.2021.01.008.

[73] I.R. Ward, L. Wang, J. Lu, M. Bennamoun, G. Dwivedi, F.M. Sanfilippo, Explainable artificial intelligence for pharmacovigilance: What features are important when predicting adverse outcomes? Comput. Methods Programs Biomed. 212 (2021) 106415, http://dx.doi.org/10.1016/j.cmpb.2021.106415.

[74] J. Smucny, G. Shi, I. Davidson, Deep learning in neuroimaging: Overcoming challenges with emerging approaches, Front. Psychiatry 13 (2022) http://dx.doi.org/10.3389/fpsyt.2022.912600.

[75] W.H. Abir, M.F. Uddin, F.R. Khanam, T. Tazin, M.M. Khan, M. Masud, S. Aljahdali, Explainable AI in diagnosing and anticipating leukemia using transfer learning method, in: M.Z. Asghar (Ed.), Comput. Intell. Neurosci. 2022 (2022) 1–14, http://dx.doi.org/10.1155/2022/5140148.

[76] Q. Ye, J. Xia, G. Yang, Explainable AI for COVID-19 CT classifiers: An initial comparison study, in: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems, CBMS, IEEE, 2021, pp. 521–526.

[77] C.K. Leung, E.W.R. Madill, J. Souza, C.Y. Zhang, Towards trustworthy artificial intelligence in healthcare, in: 2022 IEEE 10th International Conference on Healthcare Informatics, ICHI, IEEE, 2022, pp. 626–632.

[78] K.M. Sudar, P. Nagaraj, S. Nithisaa, R. Aishwarya, M. Aakash, S.I. Lakshmi, Alzheimer's disease analysis using explainable artificial intelligence (XAI), in: 2022 International Conference on Sustainable Computing and Data Communication Systems, ICSCDS, IEEE, 2022, pp. 419–423.

[79] A. Vijayvargiya, P. Singh, R. Kumar, N. Dey, Hardware implementation for lower limb surface EMG measurement and analysis using explainable AI for activity recognition, IEEE Trans. Instrum. Meas. 71 (2022) 1–9.

[80] P. Nagaraj, V. Muneeswaran, A. Dharanidharan, K. Balananthanan, M. Arunkumar, C. Rajkumar, A prediction and recommendation system for diabetes mellitus using XAI-based lime explainer, in: 2022 International Conference on Sustainable Computing and Data Communication Systems, ICSCDS, IEEE, 2022, pp. 1472–1478.

[81] H.A. Shad, Q.A. Rahman, N.B. Asad, A.Z. Bakshi, S.M.F. Mursalin, M.T. Reza, M.Z. Parvez, Exploring Alzheimer's disease prediction with XAI in various neural network models, in: TENCON 2021-2021 IEEE Region 10 Conference, TENCON, IEEE, 2021, pp. 720–725.

[82] P. Singh, A. Sharma, Interpretation and classification of arrhythmia using deep convolutional network, IEEE Trans. Instrum. Meas. 71 (2022) 1–12.

[83] R. Deo, S. Panigrahi, Explainability analysis of black box SVM models for hepatic steatosis screening, in: 2022 IEEE Healthcare Innovations and Point of Care Technologies, HI-POCT, IEEE, 2022, pp. 22–25.

[84] S.H.P. Abeyagunasekera, Y. Perera, K. Chamara, U. Kaushalya, P. Sumathipala, O. Senaweera, LISA : Enhance the explainability of medical images unifying current XAI techniques, in: 2022 IEEE 7th International Conference for Convergence in Technology, I2CT, IEEE, 2022, http://dx.doi.org/10.1109/i2ct54291.2022.9824840.

[85] D. Beddiar, M. Oussalah, S. Tapio, Explainability for medical image captioning, in: 2022 Eleventh International Conference on Image Processing Theory, Tools and Applications, IPTA, IEEE, 2022, http://dx.doi.org/10.1109/ipta54936.2022.9784146.

[86] M.S. Kamal, N. Dey, L. Chowdhury, S.I. Hasan, K.C. Santosh, Explainable AI for glaucoma prediction analysis to understand risk factors in treatment planning, IEEE Trans. Instrum. Meas. 71 (2022) 1–9, http://dx.doi.org/10.1109/tim.2022.3171613.

[87] T. Yiğit, N. Şengöz, Ö. Özmen, J. Hemanth, A.H. Işık, Diagnosis of paratuberculosis in histopathological images based on explainable artificial intelligence and deep learning, Trait. Signal 39 (2022) 863–869, http://dx.doi.org/10.18280/ts.390311.

[88] M. Sidulova, N. Nehme, C.H. Park, Towards explainable image analysis for Alzheimer's disease and mild cognitive impairment diagnosis, in: 2021 IEEE Applied Imagery Pattern Recognition Workshop, AIPR, IEEE, 2021, pp. 1–6.

[89] M.S. Kamal, A. Northcote, L. Chowdhury, N. Dey, R.G. Crespo, E. Herrera-Viedma, Alzheimer's patient analysis using image and gene expression data and explainable-AI to present associated genes, IEEE Trans. Instrum. Meas. 70 (2021) 1–7.

[90] J. Jiang, S. Hewner, V. Chandola, Explainable deep learning for readmission prediction with tree-GloVe embedding, in: 2021 IEEE 9th International Conference on Healthcare Informatics, ICHI, IEEE, 2021, http://dx.doi.org/10.1109/ichi52183.2021.00031.

[91] P. Dwivedi, A.A. Khan, S. Mugde, G. Sharma, Diagnosing the major contributing factors in the classification of the fetal health status using cardiotocography measurements: An AutoML and XAI approach, in: International Conference on Electronics, Computers and Artificial Intelligence, IEEE, 2021.

[92] U. Pawar, C.T. Culbert, R. O'Reilly, Evaluating hierarchical medical workflows using feature importance, in: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems, CBMS, IEEE, 2021, http://dx.doi.org/10.1109/cbms52027.2021.00075.

[93] R. Corizzo, Y. Dauphin, C. Bellinger, E. Zdravevski, N. Japkowicz, Explainable image analysis for decision support in medical healthcare, in: 2021 IEEE International Conference on Big Data, Big Data, IEEE, 2021, http://dx.doi.org/10.1109/bigdata52589.2021.9671335.

[94] M. Kiani, J. Andreu-Perez, H. Hagras, M.L. Filippetti, S. Rigato, A type-2 fuzzy logic based explainable artificial intelligence system for developmental neuroscience, in: 2020 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE, IEEE, 2020, pp. 1–8.

[95] N. Prentzas, A. Nicolaides, E. Kyriacou, A. Kakas, C. Pattichis, Integrating machine learning with symbolic reasoning to build an explainable AI model for stroke prediction, in: 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering, BIBE, IEEE, 2019.

[96] M. Obayya, N. Nemri, M.K. Nour, M. Al Duhayyim, H. Mohsen, M. Rizwanullah, A. Sarwar Zamani, A. Motwakel, Explainable artificial intelligence enabled TeleOphthalmology for diabetic retinopathy grading and classification, Appl. Sci. 12 (2022) 8749.

[97] S. Sarp, M. Kuzlu, E. Wilson, U. Cali, O. Guler, The enlightening role of explainable artificial intelligence in chronic wound classification, Electronics 10 (2021) 1406, http://dx.doi.org/10.3390/electronics10121406.

[98] A. Singh, S. Sengupta, J.B. J., A.R. Mohammed, I. Faruq, V. Jayakumar, J. Zelek, V. Lakshminarayanan, What is the optimal attribution method for explainable ophthalmic disease classification? in: Ophthalmic Medical Image Analysis, Springer International Publishing, 2020, pp. 21–31.

[99] Z. Papanastasopoulos, R.K. Samala, H.-P. Chan, L. Hadjiiski, C. Paramagul, M.A. Helvie, C.H. Neal, Explainable AI for medical imaging: Deep-learning CNN ensemble for classification of estrogen receptor status from breast MRI, in: H.K. Hahn, M.A. Mazurowski (Eds.), Medical Imaging 2020: Computer-Aided Diagnosis, SPIE, 2020.

[100] A. Singh, A.R. Mohammed, J. Zelek, V. Lakshminarayanan, Interpretation of deep learning using attributions: Application to ophthalmic diagnosis, in: M.E. Zelinski, T.M. Taha, J. Howe, A.A. Awwal, K.M. Iftekharuddin (Eds.), Applications of Machine Learning 2020, SPIE, 2020.

[101] R.A. Zeineldin, M.E. Karar, Z. Elshaer, J. Coburger, C.R. Wirtz, O. Burgert, F. Mathis-Ullrich, Explainability of deep neural networks for MRI analysis of brain tumors, Int. J. Comput. Assist. Radiol. Surg. 17 (2022) 1673–1683.

[102] F. Cabitza, A. Campagner, L. Famiglini, E. Gallazzi, G.A.L. Maida, Color shadows (Part I): Exploratory usability evaluation of activation maps in radiological machine learning, in: Lecture Notes in Computer Science, Springer International Publishing, 2022, pp. 31–50, http://dx.doi.org/10.1007/978-3-031-14463-9_3.

[103] J. Stodt, C. Reich, N. Clarke, Explainable AI with domain adapted FastCAM for endoscopy images, in: Computational Science – ICCS 2022, Springer International Publishing, 2022, pp. 57–64, http://dx.doi.org/10.1007/978-3-031-08757-8_6.

[104] E. Slany, Y. Ott, S. Scheele, J. Paulus, U. Schmid, CAIPI in practice: Towards explainable interactive medical image classification, in: IFIP Advances in Information and Communication Technology, Springer International Publishing, 2022, pp. 389–400, http://dx.doi.org/10.1007/978-3-031-08341-9_31.

[105] S. Hurtado, H. Nematzadeh, J. García-Nieto, M.-Á. Berciano-Guerrero, I. Navas-Delgado, On the use of explainable artificial intelligence for the differential diagnosis of pigmented skin lesions, in: Bioinformatics and Biomedical Engineering, Springer International Publishing, 2022, pp. 319–329, http://dx.doi.org/10.1007/978-3-031-07704-3_26.

[106] M. Rodriguez-Sampaio, M. Rincón, S. Valladares-Rodriguez, M. Bachiller-Mayoral, Explainable artificial intelligence to detect breast cancer: A qualitative case-based visual interpretability approach, in: Artificial Intelligence in Neuroscience: Affective Analysis and Health Applications, Springer International Publishing, 2022, pp. 557–566, http://dx.doi.org/10.1007/978-3-031-06242-1_55.

[107] G. Lokesh, T.K. Tejaswy, Y.S. Meghana, M.K. Rao, Medical report analysis using explainable AI, in: Lecture Notes in Electrical Engineering, Springer Nature Singapore, 2022, pp. 1083–1090, http://dx.doi.org/10.1007/978-981-16-7985-8_113.

[108] B.V. Patel, S. Haar, R. Handslip, C. Auepanwiriyakul, T.M.-L. Lee, S. Patel, J.A. Harston, F. Hosking-Jervis, D. Kelly, B. Sanderson, B. Borgatta, K. Tatham, I. Welters, L. Camporota, A.C. Gordon, M. Komorowski, D. Antcliffe, J.R. Prowle, Z. Puthucheary, A.A. Faisal, Natural history, trajectory, and management of mechanically ventilated COVID-19 patients in the United Kingdom, Intensive Care Med. 47 (2021) 549–565, http://dx.doi.org/10.1007/s00134-021-06389-z.

[109] S.M. Muddamsetty, M.N.S. Jahromi, T.B. Moeslund, Expert level evaluations for explainable AI (XAI) methods in the medical domain, in: Pattern Recognition. ICPR International Workshops and Challenges, Springer International Publishing, 2021, pp. 35–46, http://dx.doi.org/10.1007/978-3-030-68796-0_3.

[110] S.S. Samuel, N.N.B. Abdullah, A. Raj, Interpretation of SVM using data mining technique to extract syllogistic rules, in: Lecture Notes in Computer Science, Springer International Publishing, 2020, pp. 249–266, http://dx.doi.org/10.1007/978-3-030-57321-8_14.

[111] C. Meske, E. Bunde, Transparency and trust in human-AI-interaction: The role of model-agnostic explanations in computer vision-based decision support, in: Artificial Intelligence in HCI, Springer International Publishing, 2020, pp. 54–69, http://dx.doi.org/10.1007/978-3-030-50334-5_4.

[112] L. Ness, E. Barkan, M. Ozery-Flato, Improving the performance and explainability of mammogram classifiers with local annotations, in: Interpretable and Annotation-Efficient Learning for Medical Image Computing, Springer International Publishing, 2020, pp. 33–42, http://dx.doi.org/10.1007/978-3-030-61166-8_4.

[113] J.P. S, R. Saranya, V. Indragandhi, R.R. Singh, V. Subramaniyaswamy, Y. Teekaraman, S. Urooj, N. Alwadai, Autism spectrum disorder prediction by an explainable deep learning approach, Comput. Mater. Contin. 71 (2022) 1459–1471, http://dx.doi.org/10.32604/cmc.2022.022170.

[114] T. Saeed, C.K. Loo, M.S.S. Kassim, Ensembles of deep learning framework for stomach abnormalities classification, Comput. Mater. Contin. 70 (2022) 4357–4372, http://dx.doi.org/10.32604/cmc.2022.019076.

[115] M.A. Ayidzoe, Y.U. Yongbin, P.K. Mensah, J. Cai, F.U. Bawah, Visual interpretability of capsule network for medical image analysis, Turk. J. Electr. Eng. Comput. Sci. 30 (2021) 978–995.

[116] M.Z. Uddin, K.K. Dysthe, A. Følstad, P.B. Brandtzaeg, Deep learning for prediction of depressive symptoms in a large textual dataset, Neural Comput. Appl. 34 (2021) 721–744, http://dx.doi.org/10.1007/s00521-021-06426-4.

[117] A. Das, P. Rad, Opportunities and challenges in explainable artificial intelligence (xai): A survey, 2020, arXiv preprint arXiv:2006.11371.

[118] C. Panati, S. Wagner, S. Brüggenwirth, Feature relevance evaluation using grad-CAM, LIME and SHAP for deep learning SAR data classification, in: 2022 23rd International Radar Symposium, IRS, IEEE, 2022, pp. 457–462.

[119] A. Singh, A.R. Mohammed, J. Zelek, V. Lakshminarayanan, Interpretation of deep learning using attributions: Application to ophthalmic diagnosis, in: M.E. Zelinski, T.M. Taha, J. Howe, A.A. Awwal, K.M. Iftekharuddin (Eds.), Applications of Machine Learning 2020, SPIE, 2020, http://dx.doi.org/10.1117/12.2568631.

[120] W. Lingle, B.J. Erickson, M.L. Zuley, R. Jarosz, E. Bonaccio, J. Filippini, N. Gruszauskas, Radiology data from the cancer genome atlas breast invasive carcinoma [tcga-brca] collection, Cancer Imaging Arch. 10 (2016) K9.

[121] J.P. Cohen, P. Morrison, L. Dao, K. Roth, T.Q. Duong, M. Ghassemi, COVID-19 image data collection: Prospective predictions are the future, 2020, arXiv:2006.11988, URL: https://github.com/ieee8023/covid-chestxray-dataset.

[122] M. Cejnek, I. Bukovsky, O. Vysata, Adaptive classification of EEG for dementia diagnosis, in: 2015 International Workshop on Computational Intelligence for Multimedia Understanding, IWCIM, 2015, pp. 1–5, http://dx.doi.org/10.1109/IWCIM.2015.7347075.

[123] T. Wang, X. Li, Y. Li, X. Hu, The importance of performance evaluation in medical AI, 2020, arXiv preprint arXiv:2012.06368.

Ali Shariq Imran received a master's degree in software engineering and computing from the National University of Science and Technology (NUST), Pakistan, in 2008 and a Ph.D. in computer science from the University of Oslo (UiO), Norway, in 2013. He is associated as an Associate Professor with the Department of Computer Science (IDI), Norwegian University of Science and Technology (NTNU), Norway. With over 15 years of teaching and research experience, he devised innovative ways to design effective multimedia learning objects and integrate the teaching–research nexus frameworks at the graduate level. He served as a commission member of the Ministry of Education of Macedonia in setting up Mother Theresa University in Skjope. He leads a capacity-building project called CONNECT (https://norpart-connect.com) funded by the Higher Education Commission of Norway, DIKU, under the NORPART scheme as a coordinator and three Erasmus+ KA2 projects (PhDICTKES (https://phdictkes.eu), RAPID, and TKAEDiT) as a project manager at NTNU, along with an Excited mini-project funded by NTNU. Dr. Ali is also leading a research group on Deep NLP (http://deep-nlp.net) and specializes in applied deep learning research to address various multi-modality media analysis application areas for audio-visual and text processing. He has co-authored over 100 peer-reviewed journals and conference publications and has served as an editor and reviewer for many reputed journals. He is a member of the Intelligent Systems and Analytics (ISA) academic discipline at NTNU and an IEEE/ACM Member.

Zenun Kastrati received a master's degree in computer science through the EU TEMPUS Programme developed and implemented jointly by the University of Pristina, Kosovo, the Universite de La Rochelle, France, and the Institute of Technology Carlow, Ireland, and the Ph.D. degree in computer science from the Norwegian University of Science and Technology (NTNU), Norway, in 2018. He is currently associated with the Department of Informatics at Linnaeus University, Sweden. His research interests lie in the field of artificial intelligence with a special focus on NLP, machine/deep learning, and sentiment analysis. He is the author of more than 60 peer-reviewed journals and conferences and has served as a reviewer for many reputed journals over the years.

Sher Muhammad Daudpota received his Masters and Ph.D. degrees from Asian Institute of Technology, Thailand in the year 2008 and 2012, respectively. His research areas include deep learning, natural language processing, video and signal processing. He is author of more than 35 peer reviewed journal and conference publications. Presently he is serving as a Professor of Computer Science at Sukkur IBA University, Pakistan. Alongside his Computer Science contribution, he is also a Quality Assurance expert in higher education. He has reviewed more than 50 universities in Pakistan for quality assurance on behalf of Higher Education Commission in the role of educational quality reviewer.