Viewpoint

Framework for Classifying Explainable Artificial Intelligence (XAI) Algorithms in Clinical Medicine

Thomas Gniadek[1*], MD, PhD; Jason Kang[1*], MD; Talent Theparee[1*], MD; Jacob Krive[2,3,4,5*], PhD

[1]Department of Pathology and Laboratory Medicine, NorthShore University Health System, Evanston, IL, United States

[2]Department of Biomedical and Health Information Sciences, University of Illinois at Chicago, Chicago, IL, United States

[3]Department of Health Information Technology, NorthShore University Health System, Evanston, IL, United States

[4]Department of Health Informatics, Dr Kiran C Patel School of Osteopathic Medicine, Nova Southeastern University, Fort Lauderdale, FL, United States

[5]Pritzker School of Medicine, University of Chicago, Chicago, IL, United States

[*]all authors contributed equally

Corresponding Author:
Jacob Krive, PhD
Department of Biomedical and Health Information Sciences
University of Illinois at Chicago
1919 W Taylor St 233 AHSB
MC-530
Chicago, IL, 60612
United States
Phone: 1 312 996 1445
Email: krive@uic.edu

Abstract

Artificial intelligence (AI) applied to medicine offers immense promise, in addition to safety and regulatory concerns. Traditional AI produces a core algorithm result, typically without a measure of statistical confidence or an explanation of its biological-theoretical basis. Efforts are underway to develop explainable AI (XAI) algorithms that not only produce a result but also an explanation to support that result. Here we present a framework for classifying XAI algorithms applied to clinical medicine: An algorithm's clinical scope is defined by whether the core algorithm output leads to observations (eg, tests, imaging, clinical evaluation), interventions (eg, procedures, medications), diagnoses, and prognostication. Explanations are classified by whether they provide empiric statistical information, association with a historical population or populations, or association with an established disease mechanism or mechanisms. XAI implementations can be classified based on whether algorithm training and validation took into account the actions of health care providers in response to the insights and explanations provided or whether training was performed using only the core algorithm output as the end point. Finally, communication modalities used to convey an XAI explanation can be used to classify algorithms and may affect clinical outcomes. This framework can be used when designing, evaluating, and comparing XAI algorithms applied to medicine.

(Online J Public Health Inform 2023;15:e50934) doi: 10.2196/50934

KEYWORDS

explainable artificial intelligence; XAI; artificial intelligence; AI; AI medicine; pathology informatics; radiology informatics

Introduction

Algorithmic classifiers like artificial neural networks were first implemented many years ago [1]. Recently, unsupervised neural networks have allowed context-agnostic training and deployment. Without the need to embed a priori knowledge of the real-world system being studied, the use of these applications has expanded rapidly, and there has been much excitement about artificial intelligence (AI) algorithms in nearly every industry, including medicine.

Meanwhile, government policy that incentivizes the use of electronic medical record systems expanded the availability of digital health care information [2]. This created an environment where data analysis, predictive analytics, and ultimately AI can readily influence the interpretation of patient data and potentially prevent errors in real time during the course of clinical care [3]. Along these lines, radiologists, and to a lesser extent pathologists, are increasingly using image analysis algorithms as an assistive technology for image interpretation [4-6]. These technologies, rather than feeding into misconceptions about threats and capabilities of AI, could potentially put radiologists and pathologists at the forefront of purposeful AI innovation [7].

Initially, AI may seem like a threat to health care jobs, removing providers from the decision-making process by introducing algorithms that function as a "black box" [8]. With this perceived threat are concerns about patient safety, some stemming from comparisons to non–health care applications of AI. Like any system, AI is not infallible. For example, early versions of self-driving automobile algorithms may have caused accidents [9].

The practice of clinical medicine remains an "art" where decisions of licensed providers are relied upon to ensure patient safety. Unfortunately, in contrast to transparent, rule-based systems, a trained AI model is not transparent to a clinician [10]. Therefore, there are currently efforts to find a middle ground that combines human involvement and AI in a complementary manner [11]. For example, AI might be used to generate insights not always or easily identified by a human, but a human would still determine their significance [12,13]. In this way, AI becomes a tool used by a clinician.

Multiple countries have passed or proposed regulations on the use of algorithms in clinical medicine. Under the US Food, Drug, and Cosmetic Act, an algorithm can be classified as a "nonregulated medical device" if it meets certain criteria; otherwise, it may represent a regulated medical device. One of the key criteria is whether the algorithm is "intended for the purpose of enabling such health care professional to independently review the basis for such recommendations that such software presents so that it is not the intent that such health care professional rely primarily on any of such recommendations to make a clinical diagnosis or treatment decision regarding an individual patient" [14]. It remains to be seen how the FDA enforces this criterion on a case-by-case basis, and regulations may change over time. Similarly, the UK Department of Health and Social Care has issued robust guidance for best practices in digital health care innovation [15]. One of the key elements of this guidance is transparency about algorithm limitations, algorithm type, and evidence of effectiveness. Because of these regulatory frameworks, concerns about medical malpractice issues, and the general awareness that algorithm predictions are not always correct, there is a growing recognition that AI algorithms should allow health care providers to independently review some form of explanation of their core results [16].

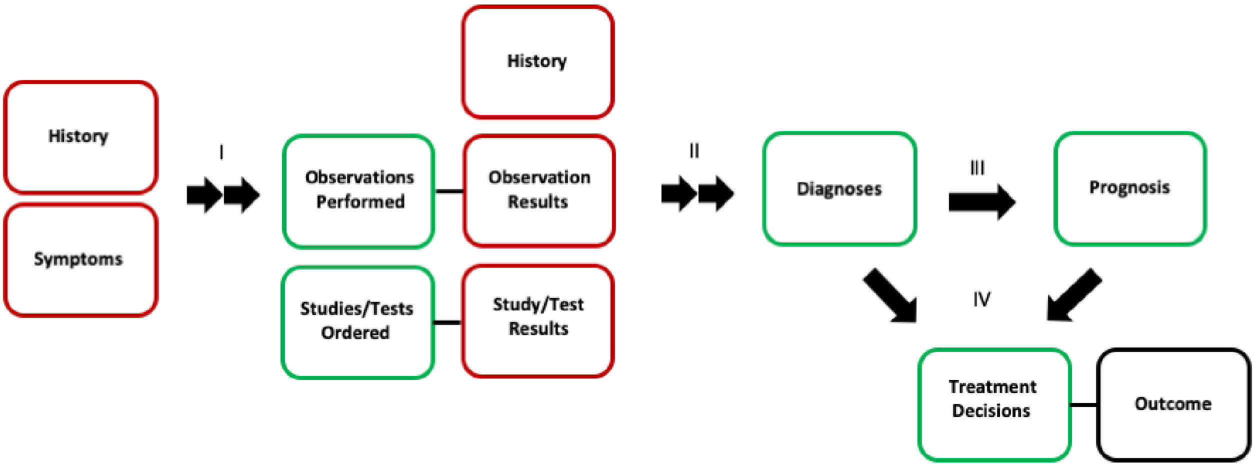
Recently, efforts began to build AI algorithms that allow humans to evaluate the significance of their results, with the goal of better integration and communication between the two. Most notably, the US Defense Advanced Research Projects Agency (DARPA) has called for further development of "explainable artificial intelligence" (XAI) [17]. The core algorithmic result or prediction is provided to the user along with an explanation that is intended to convey insight into the confidence of the core prediction, increase a user's understanding of the real-world process being studied, or both [18].

With its many benefits, XAI also brings added complexity in the form of process-specific outputs and integration with a subject matter expert end user. Not only does this elevate the importance of partnerships between clinicians and AI developers, it also raises the somewhat paradoxical possibility that algorithms with inferior core predictive power may perform better if the explanations provided result in superior outcomes overall. Furthermore, the clinical decision points supported by XAI as well as the manner in which explanations are provided to the user may differ greatly between algorithms and influence their efficacy. Here, we propose a framework for classifying XAI algorithms in clinical medicine in order to simplify this additional complexity and allow for performance evaluation of XAI in clinical practice.

Clinical Scope

The ultimate scope of clinical medicine is to prolong and improve the quality of human life. Within this, there are many decisions and actions that can be evaluated independently (eg, ordering a test, prescribing a medication, performing a surgery). XAI algorithms can be classified based on which step(s) in the clinical care pathway they support (see Figure 1). A single algorithm may provide outputs that encompass multiple areas of clinical scope. Defining clinical scope is critical for XAI, because it will determine which individuals on clinical care teams will be best suited to interact with the algorithm and evaluate the explanations provided. Furthermore, the ultimate impact of XAI on clinical outcomes will be limited by the potential impact of the process steps that an algorithm supports.

Figure 1. Clinical scope for XAI algorithms. XAI algorithms can be classified based on which steps in the clinical decision-making process they support. A simplified process flow map divides clinical decision-making into information (boxes) and information processing (arrows). Information processing steps (I-IV) can involve both human cognitive processing and computerized algorithms. Disease process evolution introduces biologic time dependency (red boxes), leading to a requirement for repeated information processing over time (double arrows). Some recorded information more directly reflects underlying disease (red boxes), while some is mainly the result of information processing (green boxes). Clinical outcome reflects underlying biology, the performance of the entire process, and the effectiveness of treatments. XAI fundamentally influences the information processing steps (I-IV) in partnership with clinicians. XAI performance can be evaluated at each information processing step or studied in the context of overall outcome. Performance of tests and treatments (black lines) are assumed to be static; however, they can be incorporated as inputs into a decision process. XAI: explainable artificial intelligence.



Clinical Insight

Explanations provided by XAI algorithms should aim to provide evidence and ultimately insight to the end user. In the case of pathology, generation of insight to assist clinicians can assist with formation of differential diagnoses, quantitative classification of features, risk prediction, and identification of features imperceptible to the human observer [19]. Both the content of the information and its delivery will determine effectiveness. Evidence can be presented in the form of empiric assessments of statistical confidence, such as a *P* value. Alternatively, an algorithm could provide an assessment of the degree of association between the current patient's data and historical groups of patients or established disease mechanisms (see Table 1).

Clinical providers evaluate empiric assessments of confidence differently than associative power, and the existence of a high degree of uncertainty in any patient-specific medical prediction necessitates a continued role for the "art of medicine" in the form of decision-making by end users. This is due to an incomplete accounting for biological factors that influence disease processes, incomplete documentation of observable factors in the electronic medical record, and the importance of the doctor-patient relationship in clinical care [20]. As a result, associative explanations may be more powerful in certain situations, since an association may support a nonquantifiable opinion held by provider or patient.

Table 1. Classifying explainable artificial intelligence explanations by type. The explanations produced by an explainable artificial intelligence algorithm can provide addition information to a clinician in 3 general ways.

XAI[a] explanation type	XAI explanation output	Primary task for clinician	Benefit to clinician
Empiric	Statistical confidence based on historical sample data	Weigh the degree of confidence provided with risks, benefits, and training data used	Assess the validity of the prediction
Population associative	Association between signs and symptoms of a patient with historical groups of patients	Assess the validity of associating this patient with historical groups of patients	Consider alternative options processed by the algorithm
Mechanism associative	Association with known pathologic mechanism(s)	Assess the validity of the pathologic mechanism(s) and diagnoses proposed	Assess validity of the prediction and consider alternatives using established medical paradigms

[a] XAI: explainable artificial intelligence.

Training and Validation

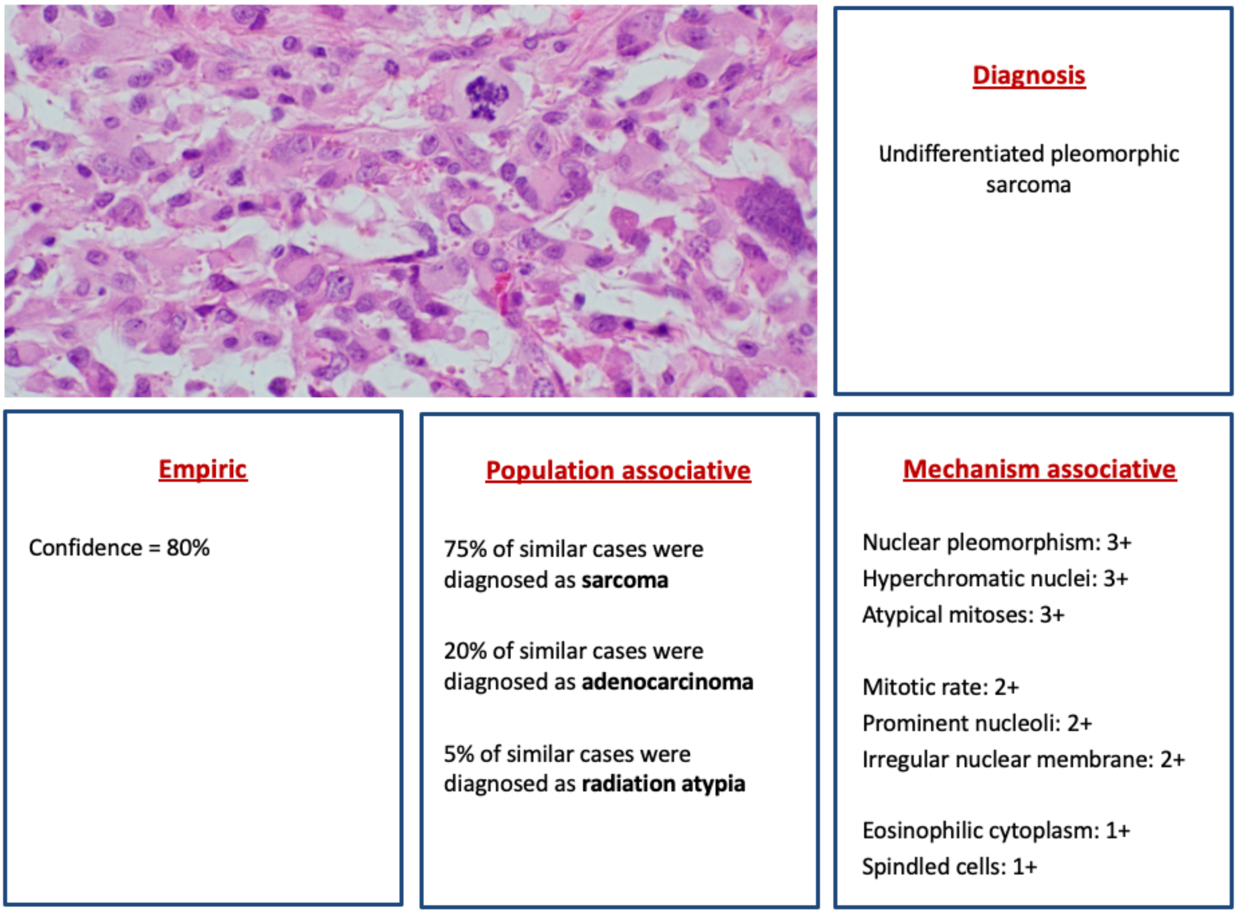
The loss of context and end user agnostic efficiency of traditional AI algorithms remains a great challenge to the initial design and implementation of XAI. In fact, the meaning of model validation in medicine differs from the traditional validation process typically undertaken in technology fields in that it refers to validation relative to patient outcomes and evidence-based medicine principles—not just whether outcomes are technically correct, match a reference method, or agree with expectations [21]. Ultimately, only patient outcomes can confirm whether the model is valid and whether AI investment is or was a worthy investment. Therefore, XAI takes special meaning in such evidence-based validation processes, since explainable analytics will help support outcomes or facilitate corrections and adjustments. Likely, the development of context-specific XAI will evolve from traditional AI in phases, each supposing a core algorithm output in addition to some form of explanation: phase I will involve traditional AI training and validation; phase II will involve traditional AI training and XAI validation, taking into account end-user actions; and phase III will involve XAI training and validation, both taking into account end-user actions.

During the final phase of XAI development as described above, the algorithm will train not to maximize the predictive power of the core algorithm output but to maximize the outcome of the combined effects of core output, explanation, and end-user actions. It is during this phase of development that XAI implementations may regain some degree of the context-agnostic advantages of traditional AI, since the behavior of the end-user context expert can be studied by the algorithm during validation.

Example 1: Anatomic Pathology

Anatomic pathologists interpret microscopic tissue morphology based on architectural and cytomorphologic criteria shown to correlate with pathologic diagnoses such as cancer. Criteria may include features such as hyperchromatic nuclei, high mitotic rate, and irregular nuclear membrane contours. Unfortunately, none of these features are 100 percent specific for a particular diagnosis like cancer, since nonneoplastic conditions may produce similar cellular features. Additionally, noninvasive premalignant conditions such as carcinoma in situ can contain individual cells that appear morphologically identical to cells within an invasive cancer. Incorporating concepts of XAI into digital anatomic pathology workflows will aid pathologists not only in making the correct diagnosis, but also in considering alternative diagnoses and recognizing potential diagnostic pitfalls (see Figure 2). Potentially, XAI systems can also incorporate ancillary information, such as clinical history, immunohistochemistry staining, and genomic testing, to aid the pathologist.

Figure 2. Illustrative example of 3 types of XAI output applied to anatomic pathology. XAI core algorithm output is shown as a diagnosis. Several forms of output explanation are succinctly outlined beneath the image, enabling a physician to make a visual interpretation in conjunction with immediate access to an explanation under multiple categories. “Empiric” information provides overall accuracy expressed as a single number; “population associative” provides a more detailed glimpse into the “black box” result; “diagnosis” relates to other cases an algorithm has access to; “mechanism associative” maps the AI process onto clinically relevant features found in the image (scored based on degree of association, 1 to 3+). XAI: explainable artificial intelligence.



Diagnosis

Undifferentiated pleomorphic sarcoma

Empiric

Confidence = 80%

Population associative

75% of similar cases were diagnosed as **sarcoma**

20% of similar cases were diagnosed as **adenocarcinoma**

5% of similar cases were diagnosed as **radiation atypia**

Mechanism associative

Nuclear pleomorphism: 3+
Hyperchromatic nuclei: 3+
Atypical mitoses: 3+

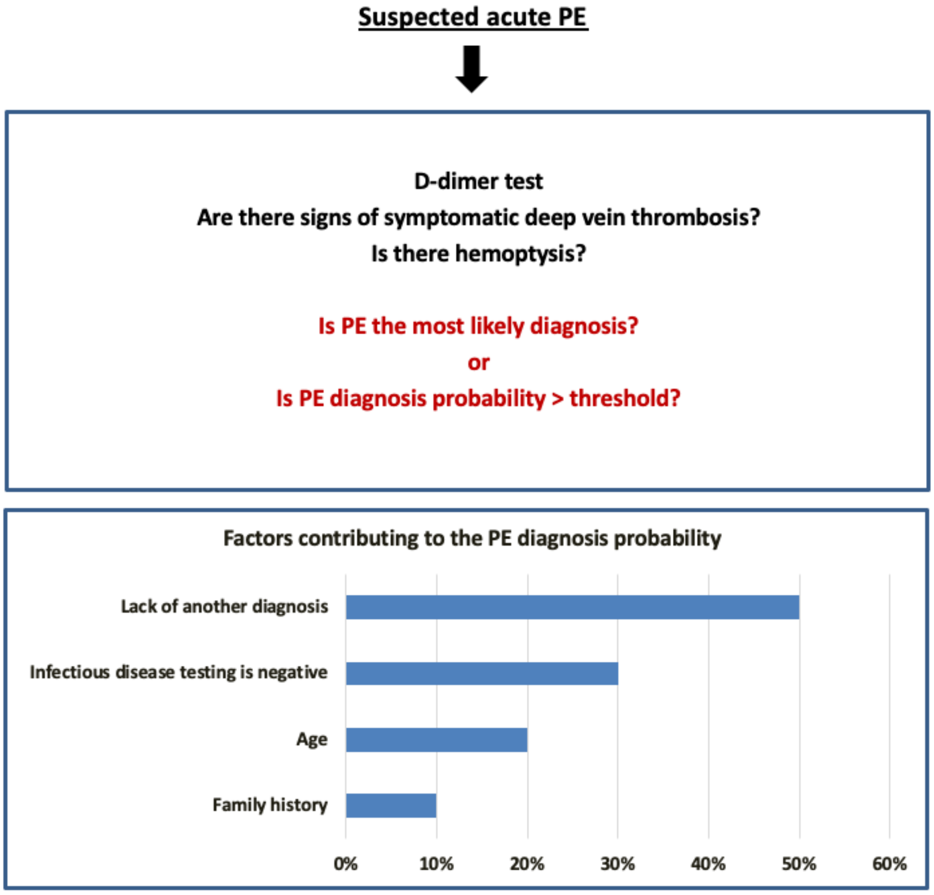
Mitotic rate: 2+
Prominent nucleoli: 2+
Irregular nuclear membrane: 2+

Eosinophilic cytoplasm: 1+
Spindled cells: 1+

Example 2: Diagnostic Management

One of the most difficult tasks for a clinician is to identify which patients should undergo screening tests and which should not [22]. This is particularly difficult when the condition screened for has a high mortality rate if not recognized, but the screening test is expensive and not without risks. Such a situation exists in deciding whether to screen for pulmonary embolism using computed tomography pulmonary angiography [23]. As a result, algorithms have been developed to aid clinical decision-making, but a clinician's assessment of whether pulmonary embolism is the most likely diagnosis plays a large role in determining a patient's score and management. Scenarios like this represent an opportunity for XAI to contribute toward more accurate assessments of pretest diagnostic likelihood (see Figure 3).

Figure 3. Diagnostic management. Possible modification to the YEARS algorithm for decisions on screening for PE by computed tomography. Rather than relying on clinician assessment of whether PE is the most likely pretest diagnosis, simple scoring algorithms can use an explainable artificial intelligence core algorithm output to assess pretest probability in the context of well-defined historical patient populations. Furthermore, the contribution of factors contributing to the core probability assessment can be displayed. Users can then assess whether each factor is valid, which may influence their assessment of the core algorithm output. For example, factors may be considered invalid if the electronic medical record is recognized as being incomplete or inaccurate. PE: pulmonary embolism.


Suspected acute PE
D-dimer test
Are there signs of symptomatic deep vein thrombosis?
Is there hemoptysis?
Is PE the most likely diagnosis?
or
Is PE diagnosis probability > threshold?
Factors contributing to the PE diagnosis probability
Lack of another diagnosis
Infectious disease testing is negative
Age
Family history
0%
10%
20%
30%
40%
50%
60%

Conclusions

The 2 recognized advantages of XAI over traditional AI can be summarized as insight into the statistical significance of a core algorithm output and mechanistic insight into the process being studied. It has been suggested that forcing AI to provide mechanistic understanding could decrease the predictive power of the algorithm itself. This may be true in a situation where algorithm inputs include all data relevant to the real-world process; however, clinical medicine remains an area where digitized information is incomplete relative to the totality of factors influencing human disease. Therefore, humans will likely remain the ultimate "trusted" decision-makers during critical, high-risk decisions in clinical care for the foreseeable future. In this framework, even clinical algorithms that are approved as regulated medical devices will remain ancillary to the human practice of medicine. XAI offers the potential to improve not the predictive power of black box algorithms but rather their usefulness as a tool for clinical providers, offering the opportunity to classify and categorize data [24], as well as ensure meaningful feedback that fits clinical workflows [25]. Information should include identification of tasks, the nature and purpose of the tasks, their outcome, and methods applied to produce the outcome [26].

Medical leaders have discussed the need for a "learning health care system" for many years. The development of XAI offers the potential to build algorithms that learn with clinical care providers. To realize the potential of XAI, we must understand how each type of algorithm might fit into the real-world process of care delivery and the minds of medical decision-makers. At least initially, this will challenge algorithm developers to understand clinical information and clinicians to efficiently integrate algorithms into their workflow.

Acknowledgments

The authors acknowledge the support of the NorthShore University Health System in providing the time, encouragement, and article processing fee (APF) funding to complete this manuscript. The authors also thank Clara Baumgarten in the Biomedical Visualization department of the University of Illinois at Chicago for producing a custom table of contents image in support of displaying the manuscript on the JMIR website.

Data Availability

All data generated or analyzed during this study are included in this published article.

Authors' Contributions

TG contributed clinical expertise and major ideas for the manuscript, wrote and influenced several sections of the paper, helped edit the paper, and compiled materials and visualizations. J Kang contributed clinical expertise for the manuscript, leadership for the project, and perspectives into applications of artificial intelligence in pathology. TT contributed to visualizations in the manuscript, added ideas for application of technology in clinical pathology, and edited the manuscript. J Krive contributed major ideas, literature review, provided clinical informatics and artificial intelligence expertise, compiled materials and supporting visualizations, helped edit the paper, and oversaw development of the manuscript.

Conflicts of Interest

J Kang is employed by Abbott Laboratories in their Tranafusion Medicine business unit. TG is employed by Fenwal, a Fresenius Kabi company. The knowledge shared in the manuscript is not influenced by any of these companies. All the other authors declare no conflicts of interest.

References

1. Farley B, Clark W. Simulation of self-organizing systems by digital computer. Trans IRE Prof Gr Inf Theory. 1954 Sep;4(4):76-84 [doi: 10.1109/TIT.1954.1057468]
2. Friedman DJ, Parrish RG, Ross DA. Electronic health records and US public health: current realities and future promise. Am J Public Health. 2013 Sep;103(9):1560-1567 [doi: 10.2105/AJPH.2013.301220] [Medline: 23865646]
3. Islam M, Hasan M, Wang X, Germack H, Noor-E-Alam M. A systematic review on healthcare analytics: application and theoretical perspective of data mining. Healthcare (Basel). 2018 May 23;6(2):54 [FREE Full text] [doi: 10.3390/healthcare6020054] [Medline: 29882866]
4. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017 Dec;2(4):230-243 [FREE Full text] [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
5. Yasaka K, Abe O. Deep learning and artificial intelligence in radiology: Current applications and future directions. PLoS Med. 2018 Nov;15(11):e1002707 [FREE Full text] [doi: 10.1371/journal.pmed.1002707] [Medline: 30500815]
6. Chang HY, Jung CK, Woo JI, Lee S, Cho J, Kim SW, et al. Artificial intelligence in pathology. J Pathol Transl Med. 2019 Jan;53(1):1-12 [FREE Full text] [doi: 10.4132/jptm.2018.12.16] [Medline: 30599506]
7. Pesapane F, Codari M, Sardanelli F. Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. Eur Radiol Exp. 2018 Oct 24;2(1):35 [FREE Full text] [doi: 10.1186/s41747-018-0061-6] [Medline: 30353365]
8. Goebel R, Chander A, Holzinger K. Explainable AI: the new 42? In: CD-MAKE 2018: Machine Learning and Knowledge Extraction. 2018 Presented at: International Cross-Domain Conference for Machine Learning and Knowledge Extraction; August 27-30, 2018; Hamburg, Germany p. 295-303 URL: https://openaccess.city.ac.uk/id/eprint/20659/1 [doi: 10.1007/978-3-319-99740-7_21]
9. Yu K, Kohane I. Framing the challenges of artificial intelligence in medicine. BMJ Qual Saf. 2019 Mar;28(3):238-241 [doi: 10.1136/bmjqs-2018-008551] [Medline: 30291179]
10. Holzinger A. From machine learning to explainable AI. In: DISA 2018 - IEEE World Symposium on Digital Intelligence for Systems and Machines, Proceedings. In. DISA 2018 - IEEE World Symposium on Digital Intelligence for Systems and Machines, Proceedings. IEEE; 2018 Presented at: 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA); August 23-25, 2018; Košice, Slovakia p. 55-66 URL: https://ieeexplore.ieee.org/abstract/document/8490530 [doi: 10.1109/disa.2018.8490530]
11. Tacchella A, Romano S, Ferraldeschi M, Salvetti M, Zaccaria A, Crisanti A, et al. Collaboration between a human group and artificial intelligence can improve prediction of multiple sclerosis course: a proof-of-principle study. F1000Res. 2017;6:2172 [FREE Full text] [doi: 10.12688/f1000research.13114.2] [Medline: 29904574]
12. Miller D, Brown E. Artificial intelligence in medical practice: The question to the answer? Am J Med. 2018 Feb;131(2):129-133 [FREE Full text] [doi: 10.1016/j.amjmed.2017.10.035] [Medline: 29126825]

13. Li D, Kulasegaram K, Hodges B. Why we needn't fear the machines: opportunities for medicine in a machine learning world. Acad Med. 2019 May;94(5):623-625 [doi: 10.1097/ACM.0000000000002661] [Medline: 30768470]
14. 21 U.S.C. 360j - General provisions respecting control of devices intended for human use. U.S. Government Publishing Office. URL: https://tinyurl.com/555t6b4f [accessed 2023-08-18]
15. A guide to good practice for digital and data-driven health technologies. UK Department of Health and Social Care. URL: https://tinyurl.com/32srzb9t [accessed 2023-08-18]
16. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. N Engl J Med. 2018 Mar 15;378(11):981-983 [FREE Full text] [doi: 10.1056/NEJMp1714229] [Medline: 29539284]
17. Gunning D. Explainable artificial intelligence. Defense Advanced Research Projects Agency. URL: https://www.darpa.mil/program/explainable-artificial-intelligence [accessed 2019-03-27]
18. Holzinger A, Biemann C, Pattichis C, Kell D. What do we need to build explainable AI systems for the medical domain? arXiv.. Preprint posted online Dec 28, 2017. [FREE Full text] [doi: 10.1093/oso/9780197529003.003.0015]
19. Holzinger A, Malle B, Kieseberg P. Towards the pathologist: challenges of explainable-AI in digital pathology. arXiv.. Preprint posted online on December 18, 2017. [FREE Full text]
20. Aminololama-Shakeri S, López JE. The doctor-patient relationship with artificial intelligence. AJR Am J Roentgenol. 2019 Feb;212(2):308-310 [doi: 10.2214/AJR.18.20509] [Medline: 30540210]
21. Park S, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology. 2018 Mar;286(3):800-809 [doi: 10.1148/radiol.2017171920] [Medline: 29309734]
22. van der Hulle T, Cheung W, Kooij S, Beenen LFM, van Bemmel T, van Es J, et al. YEARS study group. Simplified diagnostic management of suspected pulmonary embolism (the YEARS study): a prospective, multicentre, cohort study. Lancet. 2017 Jul 15;390(10091):289-297 [doi: 10.1016/S0140-6736(17)30885-1] [Medline: 28549662]
23. Qaseem A, Alguire P, Dallas P, Feinberg LE, Fitzgerald FT, Horwitch C, et al. Appropriate use of screening and diagnostic tests to foster high-value, cost-conscious care. Ann Intern Med. 2012 Jan 17;156(2):147-149 [FREE Full text] [doi: 10.7326/0003-4819-156-2-201201170-00011] [Medline: 22250146]
24. Weng S, Reps J, Kai J, Garibaldi J, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One. 2017;12(4):e0174944 [FREE Full text] [doi: 10.1371/journal.pone.0174944] [Medline: 28376093]
25. Komorowski M, Celi L, Badawi O, Gordon A, Faisal A. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med. 2018 Nov;24(11):1716-1720 [FREE Full text] [doi: 10.1038/s41591-018-0213-5] [Medline: 30349085]
26. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv.. Preprint posted online on February 28, 2017. [FREE Full text]

Abbreviations

AI: artificial intelligence
DARPA: US Defense Advanced Research Projects Agency
XAI: explainable artificial intelligence

Edited by E Mensah; this is a non–peer-reviewed article. Submitted 30.03.20; accepted 27.02.23; published 01.09.23.

Please cite as:
Gniadek T, Kang J, Theparee T, Krive J
Framework for Classifying Explainable Artificial Intelligence (XAI) Algorithms in Clinical Medicine
Online J Public Health Inform 2023;15:e50934
URL: https://ojphi.jmir.org/2023/1/e50934
doi: 10.2196/50934
PMID: 38046562