Guest Editorial
Explainable AI: Towards Fairness, Accountability, Transparency and Trust in Healthcare

RECENT advances in artificial intelligence, precision health, and medicine [1] have paved the way for the accelerated adaptation and use of intelligent tools and systems in decision-making processes across the healthcare spectrum. Insights and knowledge derived from complex analytics are used to implement diagnostic and therapeutic solutions and targeted interventions in individuals and communities across the globe. Given the complexity of the current multi-dimensional clinical and public health data landscape, providing explainability in the context of socio-environmental [2] and technical systems is a key to revealing pathways from socio-economic disadvantages to health disparities [3] and implementing equitable interventions.

As the complexity of the underlying data sets and AI-based algorithms increases, the explainability and justifiability of the insights generated decrease. Humans need to understand the underlying mechanism behind these insights to know whether they are sound, correct, trustable, and justifiable to make informed decisions. Lack of understandability and explainability in the biomedical domain often leads to poor transparency and accountability and ultimately lower quality of care and suboptimal and unfair health policies. Explainability is considered one of the prerequisites [4] for deep medicine, where AI is meant "to provide composite, panoramic views of individuals' medical data; to improve decision-making; to avoid errors such as misdiagnosis and unnecessary procedures; to help in the ordering and interpretation of appropriate tests, and to recommend treatment" [5].

Explainable Artificial Intelligence (XAI) [6], a relatively new field in AI, aims to provide justification, transparency, and traceability of often black-box machine learning methods as well as testability of causal assumptions [7], [8]. The evaluation of causality is especially instrumental in healthcare to justify why a decision is made, and why an intervention or a treatment option is favored over other available options [9]. XAI is a step towards the realization of the FATE (Fairness, Accountability, Transparency, and Ethics) [10] principles in AI.

XAI has been used in a broad range of applications such as mental health surveillance [11], COVID-19 surveillance [12], prediction of acute critical illness and early detection of sepsis [13], [14], cancer diagnosis [15], and genomics [16]. This special issue includes the latest achievements in research and development in artificial intelligence with an emphasis on explainable models in public health and medicine.

Towards this end, Shi *et al* [Appendix, item 1] propose an explainable attention-transfer classification model based on the knowledge distillation network structure to address challenges of automatically differentiating COVID-19 and community-acquired pneumonia from healthy lungs in radiographic imaging. Jeon *et al*. [Appendix, item 2] emphasize the importance of explainability in medical image processing and present an interpretable and lightweight 3D deep neural network model that diagnoses anterior cruciate ligament (ACL) tears from a knee MRI exam. Ivaturi *et al*. [Appendix, item 3] describe a post-hoc explainability framework for deep learning models applied to quasi-periodic biomedical time-series classification. The authors explained their methods through a case scenario on atrial fibrillation (AF) detection from electrocardiography signals.

Beebe-Wang *et al*. [Appendix, item 4] apply a machine learning model to an aging cohort study with several longitudinal clinical variables to detect and screen individuals at risk of dementia with higher accuracy than standard rudimentary approaches. The presented method also provides individualized prediction explanations that retain non-linear feature effects present in the data. Alexandre *et al*. [Appendix, item 5] propose a structured view on why, when, and how to apply biclustering to mine discriminative patterns of post-surgical risk in the oncological domain. These patterns offer a comprehensive view of how the patient profile, cancer histopathology, and entailed surgical procedures determine post-surgical complications patient survival, and hospitalization needs.

Wickstrom *et al*. [Appendix, item 6] address the lack of uncertainty in deep-learning-based decision support systems, by proposing a deep ensemble approach for explainable convolutional neural networks (CNNs). Xu *et al*. [Appendix, item 7] propose a framework for predicting the long-term recurrence risk in patients with ischemic cerebrovascular events after discharge from hospitals based on process mining and transfer learning. Chaddad *et al*. [Appendix, item 8] demonstrate a new radiomic model to formalize 3D CNN features using the Gaussian mixture models. The authors show how their proposed method is used as a prognostic biomarker for Pancreatic ductal adenocarcinoma (PDAC) patients. This model helps clinicians improve their treatment plans. The radiomic features explained in the paper showed a high predictive value compared to other popular radiomic features and clinical markers using a random forest classifier.

Shang *et al*. [Appendix, item 9] introduce an electronic health record (EHR) oriented knowledge graph system to utilize non-used information buried in EHRs. Oh *et al*. [Appendix, item 10] demonstrate how trajectories, the order in which diseases manifest throughout life, are predictive of the course of disease progression. The comprehensive set of disease trajectories extracted by their method is shown to explain the observed outcomes substantially better than competing methods.

Feng *et al*. [Appendix, item 11] propose an approach for prediction of patient volumes based on causalities discovered

Digital Object Identifier 10.1109/JBHI.2021.3088832

by Gaussian processes. Their motivations to adopt causal feature selection are to maintain interpretability and reliability of causal discovery. They focus on patients who suffer from allergy, although the proposed method can be generalized to cover other types of patient volumes.

The included articles look at the problem of explainability and interpretability of AI models within biomedical and public health domains, but from different angles. This special issue brings together these different perspectives and serves to highlight the breadth of approaches used to make AI models more usable and trustable in the health domain.

ARASH SHABAN-NEJAD
UTHSC-ORNL Center for Biomedical Informatics, Department of Pediatrics, College of Medicine, University of Tennessee Health Science Center, TN 38103, USA

MARTIN MICHALOWSKI
University of Minnesota - Twin Cities, MN 55455, USA

JOHN S. BROWNSTEIN
Boston Children's Hospital, Harvard University, MA 02115, USA

DAVID L. BUCKERIDGE
McGill University, QC H3A 1A3, Canada

APPENDIX
RELATED WORK

1) W. Shi, L. Tong, Y. Zhu, and M. D. Wang, "COVID-19 automatic diagnosis with radiographic imaging: Explainable attentiontransfer deep neural networks," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 7, Jul. 2021.
2) Y. Jeon *et al.*, "Interpretable and lightweight 3-D deep learning model for automated ACL diagnosis," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 7, Jul. 2021.
3) P. Ivaturi, M. Gadaleta, A. C. Pandey, M. Pazzani, S. R. Steinhubl, and G. Quer, "A comprehensive explanation framework for biomedical time series classification," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 7, Jul. 2021.
4) N. Beebe-Wang, A. Okeson, T. Althoff, and S. I. Lee, "Efficient and explainable risk assessments for imminent dementia in an aging cohort study," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 7, Jul. 2021.
5) L. Alexandre, R. S. Costa, L. L. Santos, and R. Henriques, "Mining pre-surgical patterns able to discriminate post-surgical outcomes in the oncological domain," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 7, Jul. 2021.
6) K. K. Wickstrom, K. OyvindMikalsen, M. Kampffmeyer, A. Revhaug, and R. Jenssen, "Uncertainty-Aware deep ensembles for reliable and explainable predictions of clinical time series," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 7, Jul. 2021.
7) H. Xu, J. Pang, W. Zhang, X. Li, M. Li, and D. Zhao, "Predicting recurrence for patients with ischemic cerebrovascular events based on process discovery and transfer learning," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 7, Jul. 2021.
8) A. Chaddad, P. Sargos, and C. Desrosiers, "Modeling texture in deep 3D CNN for survival analysis," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 7, Jul. 2021.
9) Y. Shang *et al.*, "EHR-oriented knowledge graph system: Toward efficient utilization of Non-used information buried in routine clinical practice," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 7, Jul. 2021.
10) W. Oh, M. S. Steinbach, M. R. Castro, K. A. Peterson, and V. Kumar, "A computational method for learning disease trajectories from partially observable EHR data," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 7, Jul. 2021.
11) G. Feng, K. Yu, Y. Wang, Y. Yuan, and P. M. Djuric, "Exploiting causality for improved prediction of patient volumes by Gaussian processes," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 7, Jul. 2021.

REFERENCES

[1] A. Shaban-Nejad, M. Michalowski, N. Peek, J. S. Brownstein, and D. L. Buckeridge, "Seven pillars of precision digital health and medicine," *Artif. Intell. Med.*, vol. 103, Mar 2020, Art. no. 101793.
[2] E. K. Shin, R. Mahajan, O. Akbilgic, and A. Shaban-Nejad, "Sociomarkers and biomarkers: Predictive modeling in identifying pediatric asthma patients at risk of hospital revisits," *NPJ Digit. Med.*, vol. 1, Oct. 2018, Art. no. 50.
[3] E. K. Shin, Y. Kwon, and A. Shaban-Nejad, "Geo-clustered chronic affinity: Pathways from socio-economic disadvantages to health disparities," *JAMIA Open*, vol. 2, no. 3, pp. 317–322, Oct. 2019.
[4] A. Shaban-Nejad, M. Michalowski, and D. L. Buckeridge, "Explainability and interpretability: Keys to deep medicine," in *Explainable AI in Healthcare and Medicine. Studies in Computational Intelligence*, Cham: Springer, vol. 914, 2021. [Online]. Available: https://doi.org/10.1007/978-3-030-53352-6_1
[5] E. Topol, *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Basic Books; 1 st ed. Jul. 2019.
[6] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "XAI-Explainable artificial intelligence," *Sci. Robot.*, vol. 4, no. 37, 2019, Art. no. eaay7120.
[7] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explain-ability of artificial intelligence in medicine," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 9, no. 4, Jul./Aug. 2019, Art. no. e1312.
[8] J. Pearl, "Theoretical impediments to machine learning with seven sparks from the causal revolution," Jan. 2018, *arXiv:1801.04016*.
[9] J. H. Brenas and A. Shaban-Nejad, "Health intervention evaluation using semantic explainability and causal reasoning," *IEEE Access*, vol. 8, pp. 9942–9952, 2020.
[10] FATE: Fairness, Accountability, Transparency, and Ethics in AI. Microsoft Research. Jul. 2021. [Online]. Available: https://www.microsoft.com/en-us/research/theme/fate/
[11] N. Ammar and A. Shaban-Nejad, "Explainable artificial intelligence recommendation system by leveraging the semantics of adverse childhood experiences: Proof-of-concept prototype development," *JMIR Med Inform*, vol. 8, no. 11, Apr. 2020, Art. no. e18752.
[12] M. S. Hossain, G. Muhammad, and N. Guizani, "Explainable AI and mass surveillance system-based healthcare framework to combat COVID-I9 like pandemics," *IEEE Netw.*, vol. 34, no. 4, pp. 126–132, Jul./Aug. 2020.
[13] S. M. Lauritsen *et al.*, "Explainable artificial intelligence model to predict acute critical illness from electronic health records," *Nature Commun.*, vol. 11, no. 1, Jul 31 2020, Art. no. 3852.
[14] K. H. Goh *et al.*, "Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare," *Nature Commun.*, vol. 12, no. 1, Jan. 29, 2021, Art. no. 711.
[15] A. Binder *et al.*, "Morphological and molecular breast cancer profiling through explainable machine learning," *Nature Mach. Intell.*, vol. 3, pp. 355–366, 2021. [Online]. Available: https://doi.org/10.1038/s42256-021-00303-4
[16] A. Anguita-Ruiz, A. Segura-Delgado, R. Alcalá, and C. M. Aguilera, "eXplainable artificial intelligence (XAI) for the identification of biologically relevant gene expression patterns in longitudinal human studies, insights from obesity research," *PLoS Comput. Biol.*, vol. 16, no. 4, Apr. 10, 2020, Art. no. e1007792.