2021

Explainable Artificial Intelligence in the Medical Domain: A Systematic Review

Shuvro Chakrobartty

Omar El-Gayar
Dakota State University

Aug 9th, 12:00 AM

Explainable Artificial Intelligence in the Medical Domain: A Systematic Review

Shuvro Chakrobartty
Dakota State University, shuvro.chakrobartty@trojans.dsu.edu

Omar El-Gayar
Dakota State University, omar.el-gayar@dsu.edu

Explainable Artificial Intelligence in the Medical Domain: A Systematic Review

Completed Research

Shuvro Chakrobartty
Dakota State University
Shuvro.Chakrobartty@trojans.dsu.edu

Omar El-Gayar
Dakota State University
Omar.El-Gayar@dsu.edu

Abstract

The applications of Artificial Intelligence (AI) and Machine Learning (ML) techniques in different medical fields is rapidly growing. AI holds great promise in terms of beneficial, accurate and effective preventive and curative interventions. At the same time, there is also concerns regarding potential risks, harm and trust issues arising from the opacity of some AI algorithms because of their un-explainability. Overall, how can the decisions from these AI-based systems be trusted if the decision-making logic cannot be properly explained? Explainable Artificial Intelligence (XAI) tries to shed light to these questions. We study the recent development on this topic within the medical domain. The objective of this study is to provide a systematic review of the methods and techniques of explainable AI within the medical domain as observed within the literature while identifying future research opportunities.

Keywords

Artificial intelligence, Explainability, XAI.

Introduction

While the very first AI systems were easily interpretable, recent years have witnessed the rise of opaque (black-box) decision systems such as Deep Neural Networks (DNNs) (Barredo Arrieta et al. 2020). Black-box approaches does not foster trust and acceptance of ML among end-users (Holzinger et al. 2017). The opposite of black-box-ness is transparency, i.e., a direct understanding of the mechanism by which a model works as it makes a decision (Barredo Arrieta et al. 2020).

There are various multi-level artificial neural network-based architectures for solving decision or classification problems in different domains. For our discussion we use DNN as an umbrella term to represent all these architectures. The empirical success of Deep Learning (DL) models such as DNNs stems from a combination of efficient learning algorithms and their huge parametric space. This space comprises hundreds of layers and millions of parameters, making DNNs very complex black-box models (Bohanec et al. 2017; London 2019). For example, ResNet50, a relatively small DNN, has 25 million parameters (You et al. 2018). There are also other non-DL ML models that are not transparent i.e., black box. For example, Support Vector Machine (SVM), random forests, boosted trees and tree bagging (tree ensembles) etc (Guidotti et al. 2018). In addition to compromising trust, growing legal and privacy aspects, such as the European General Data Protection Regulations (GDPR), make black-box ML approaches difficult to use. The models often are not able to explain how a decision has been made, e.g., why two objects are similar (Holzinger et al. 2017). Moreover, black-box approaches are not suitable for safety-critical domains such as the medical domain because of the lack of transparency (Holzinger et al. 2017). Further, the lack of explainability of the decisions amplified ethical concerns (Beil et al. 2019; Cabitza et al. 2017; London 2019).

Accordingly, this paper aims to provide a systematic review of existing works on explainable AI methods and techniques within medical domain by showing how the various models were attempted for different use cases for human-understandable explanations using a proposed XAI methods and techniques classification framework. The rest of the paper is organized as follows: first we elaborate on the notion of XAI and prior work, then we provide a conceptual framework for a classification. Later, we describe the review protocol applied in the study. Following that we analyze the outcomes and present its results. Finally, we discuss some future research directions and conclude the paper.

Background and Related Works

Explainable AI is not a new field since, in expert systems of the 1980s, there were reasoning architectures to support an explanation function for complex AI systems (Holzinger et al. 2018). In expert system-based AI, human knowledge is first codified, then an inference engine is used to provide an expert decision to a non-expert user through an interface (London 2019). By design, this is an explainable system since the inference engine follows specific rules to make the decision. Explainability is an essential aspect of trust since trust would depend on the visibility that a human has into the working of the AI system. Therefore, DNN and other algorithms should provide human-understandable justifications for its output, leading to insights into the AI system's inner workings. Interpretable models are able to explain why a certain prediction was made for a specific patient, by showing characteristics that led to the prediction. Lack of interpretability therefore limits the use of otherwise powerful deep and ensemble learning models in medical decision support (Lundberg et al. 2018).

In the literature, we find various terminology for explainability, such as understandability or intelligibility, comprehensibility, interpretability, transparency, and sometimes they are used synonymously in XAI (Holzinger et al. 2019). There isn't a single definition of explainability and there is the interchangeable use of interpretability and explainability in the literature (Arrieta et al. 2020). There are different purposes of explainability in ML models sought by different audience profiles while identifying two principal goals that cuts across them, 1) need for model understanding, and 2) regulatory compliance (Doshi-Velez and Kim 2017). A Taxonomy of Interpretability Evaluation has been presented by (Doshi-Velez and Kim 2017), where the authors lay out an analogous taxonomy of evaluation approaches for interpretability such as application-grounded, human-grounded, and functionally-grounded. In our review we focused on the aspect of the human understandability, where the human is a medical professional. Schoenborn and Althoff (2019) describe that an explainable artificial intelligence enables a user to learn a transparent, relevant, and justified information at the right time using an appropriate size.

Although it's not the focus of this review, data bias is a major challenge in the field of AI and ML. ML algorithms operate by learning models from historical data and generalizing them to unseen new data (Suresh and Guttag 2020). Tommasi et al. (2017) looked at the dataset bias from an image processing perspective, they analyze Convolutional Neural Networks (CNN) as it has emerged as a reliable and robust image processing ML model. It should be noted that even with an explainable model, a ML model can produce undesirable results suffering from biased decision with fairness issues, which is another very active area of AI and ML research (Barredo Arrieta et al. 2020; Bhatt et al. 2020; Gill et al. 2020; Gilpin et al. 2018; Kaul 2018).

For the purpose of this review we adopt the definition of XAI as eluded by Gunning and Aha (2019), which denotes that XAI's goal is "to create a suite of new or modified ML techniques that produce explainable models that, when combined with effective explanation techniques, enable end users to understand, appropriately trust, and effectively manage the emerging generation of AI systems" (Gunning and Aha 2019). We choose this definition because it clearly identifies the objective of the explainability from an end user's perspective.

Prior Reviews

There have been prior survey studies conducted on XAI related to the general explainability and interpretability of AI focusing on grouping of methods for explainability, encouraging the need for organizing the XAI literature as well as on the identification of the dimensions that are useful for model interpretability, while categorizing prior work along those dimensions (Adadi and Berrada 2018; Chakraborty et al. 2017; Došilović et al. 2018; Gilpin et al. 2018). Others, such as Zhang and Zhu (2018), did a survey study on visual interpretability for DL, focusing on Convolutional Neural Networks (CNNs) methods by diagnosing representations of pre-trained CNNs on model interpretability. On the medical domain there has been a few reviews as well (Beil et al. 2019; Cabitza et al. 2017; Fellous et al. 2019; London 2019; Tjoa and Guan 2019). However, only E. Tjoa and C. Guan (2019) provides a review of interpretability suggesting different research works and categorizing them. Overall, in previous review studies while the authors have focused on XAI methods in more general ML domain and use cases, in this study we focus entirely on the medical domain as a representative of high-stake mission critical and knowledge intensive application domains to identify, understand and categorize XAI methods.

Conceptual Classification Framework for XAI

At a high level there are two ways we can categories of XAI methods. Transparent explainability is where models interpretable by design (transparent models) and Post-hoc explainability is when it's explained employing external XAI techniques. Transparent can further be divided into pre-modeling and during-modeling explainability. As mentioned by Elshawi et al. (2019), the goal of pre-modeling explainability is to understand and describe data used to develop models, whereas the goal for during-modeling explainability is to develop inherently more explainable models. Another way to further classify Post-hoc explainability methods is based on whether the interpretation of the model is global or local (Arya et al. 2019; Barredo Arrieta et al. 2020; Elshawi et al. 2019). A global explanation describes the behavior of the entire model by understanding the overall logic of a black-box model. In contrast, the local explanation is for a single prediction to find correlations between feature values and the outcome.

Further, model-specific interpretation methods are limited to specific types of models (Arya et al. 2019; Barredo Arrieta et al. 2020; Du et al. 2020). For example, the regression weights in a linear model are a model-specific interpretation and do not work for any other model. On the other hand, model-agnostic interpretation methods are more general; they treat a model as a black-box and do not inspect internal model parameters; therefore, it can be applied to any ML model. For example, Black Box Explanations through Transparent Approximations (BETA) is a post-hoc model agnostic XAI method. Moreover, as model-agnostic feature importance is broadly applicable to various ML models, they are usually post-hoc.

Accordingly, we propose a graphical conceptual classification framework for the available literature on the applications of XAI methods and techniques for ML models (Figure 1). The classification framework is based on a literature review of existing knowledge on the methods and techniques on explainable AI research. The classification of XAI methods comprises two levels, as shown in Figure 1. The first, transparent systems could be based on various methods like REverse Time AttentIoN model (RETAIN), Generalized Additive Model (GAM) and Bayesian Deep Learning (BDL) etc., and the second, post-hoc systems comprising of the global and local explanation categories. As shown in the diagram below to help visualize the categorical placement of the various method and techniques, the post-hoc systems could be based on various methods like Local Interpretable Model agnostic Explanations (LIME), Shapley Additive Explanations (SHAP), Layer-wise Relevance Propagation (LRP) etc. We incorporate the XAI methods and techniques based on the literature review for which the result is summarized in the table 1. We find that supervised ML paradigm is used mostly and have been applied to decision-making systems.

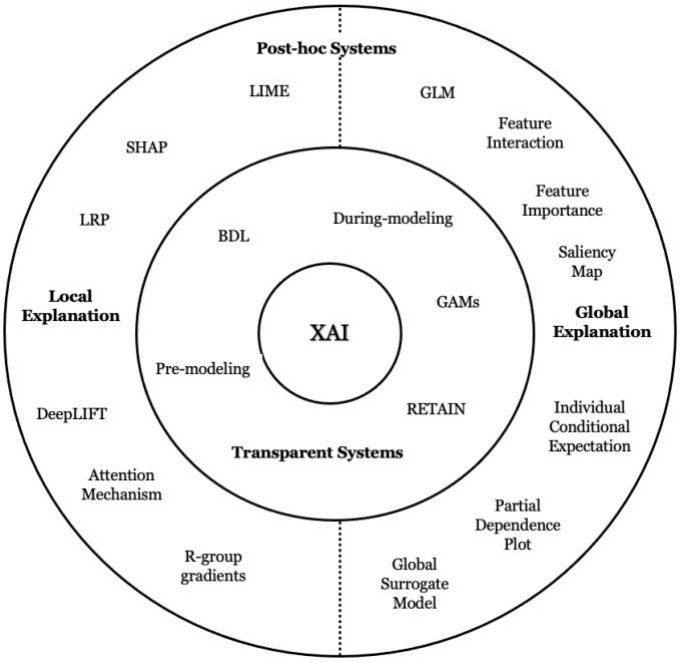

Figure 1. Explainable methods and techniques categories

Methodology

This research's methodological framework can be divided into three essential phases: research definition, research methodology, and research analysis, as depicted in Figure 2. In the research definition phase, the research area is academic research on explainable AI methods and techniques for medical domain. The research goal is to provide a general overview of existing works on explainable AI methods and techniques by showing how the various models were attempted for multiple different use cases for human-understandable explanations in the medical domain with their corresponding classification, as well as to suggest directions for future research. The research scope is the literature on the methods and techniques of explainable AI. As the research on this topic is relatively recent, the scope of this investigation is limited to the time frame of 2008 to 2020. In the research methodology phase, and given the emphasis on the medical domain, PubMed was searched to provide an ample listing of journal articles focused on the biomedical sciences. This literature search was based on the descriptors "XAI", "explainable", "explainability", "interpretability", "trustworthiness, "artificial intelligence", "machine learning", "deep learning". We used Boolean expressions to apply these terms to a search of online PubMed database. The review and classification process were carefully verified, and only articles related to explainable AI methods and techniques that met the following selection criteria were included: First, the articles must have been published in academic journals for which the full-text versions are available. Duplicate articles, conference articles, master or doctoral dissertations, textbooks, survey articles, and unpublished working papers were excluded. Second, the articles had to have been published between 2008 and 2020 inclusive. Third, the articles had to present explainable AI methods and techniques and discuss their application to explain AI models for a specific use case in the medical domain.

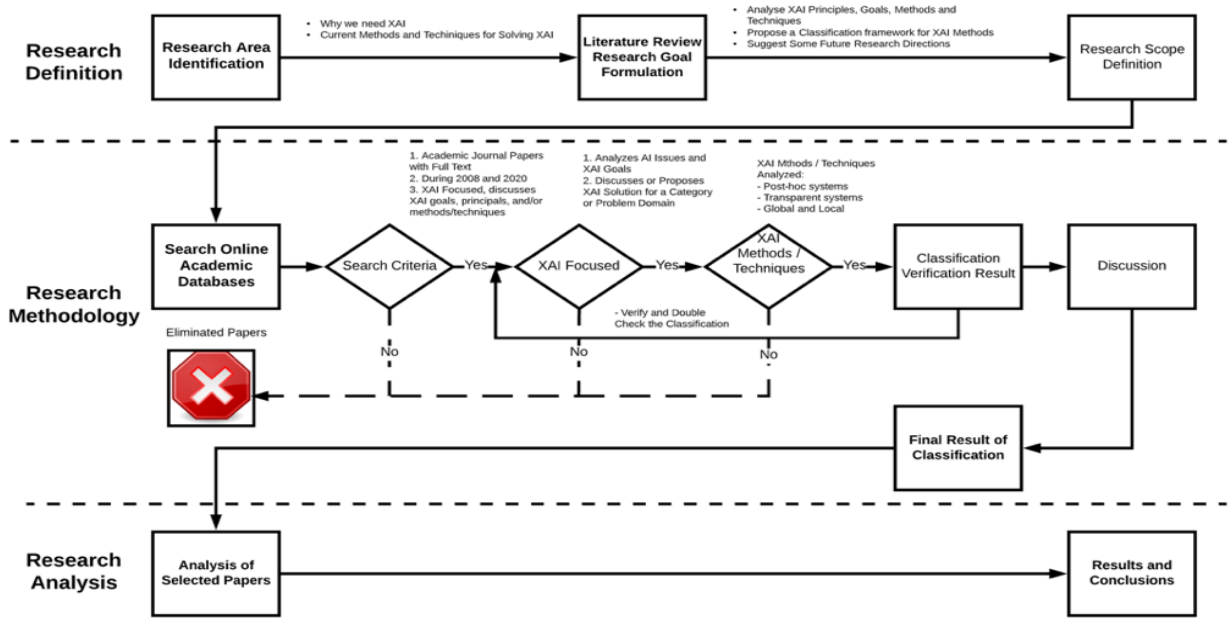

Figure 2. Methodological framework for research

Result

As shown in Figure 3, the search initially produced 66 articles. Upon applying the exclusion and inclusion criteria we included 22 articles for analysis. Categorizing the XAI method proposed in the article we see that 6 articles used the transparent model and 16 articles used post-hoc models. We notice that research on XAI has been getting much attention from the AI community as evident from the fact that the number of XAI publications has been increasing over the years. In Figure 3, we describe the flow diagram of article selection and filtration for the study according to the methodology. In Table 1, we classify those 22 journal articles based on the XAI method and techniques used.

Discussion

The results demonstrate that post-hoc and local interpretability with the visualization-based interpretation technique are more popular. The popularity of post-hoc models results from the need for more accurate models. We also see that the post-hoc interpretability is used for use cases that depends on some variation of the DL models. Local interpretability techniques give explanations at the level of specific instances. But the global interpretability techniques have the benefit that it can generalize over the entire population, and method to use would mostly depend on the explanation need for the application (Elshawi et al. 2019).

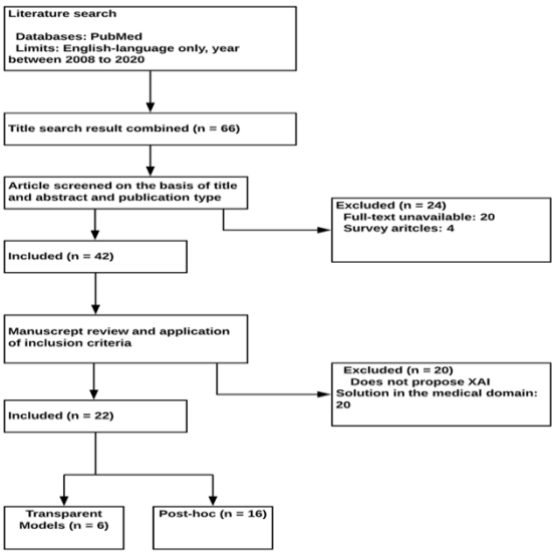

Figure 3. Flow diagram of study selection

Visualization turns out to be a good interface to communicate the explanation of the model prediction to the end users in the medical domain. We observe the popular use of visualization technique to convey the interpretability results. For example, Warman et al. (2020) used all labeled features in an interpretable decision tree to suggest a diagnosis for radiological review and supported its proposal by presenting visuals with radiological evidence highlighted. Qiu et al. (2020) used a color map overlay on the traditional imaging modality to create a simple presentation of interpretable disease risk. Also, Liao et al. (2020) proposed a novel clinical interpretable ConvNet architecture (EAMNet) that provides a visual region of interest to support the reliability of their diagnosis decision. Hao et al. (2020) used effective visualizations of their explanations for anesthesiologists with details that highlights the relevant contributing features.

We also see the use of the attention mechanism either as an added feature or as part of an existing XAI technique. For example, Shickel et al. (2019) used self-attention approach to highlight particular time steps of the input time series for real-time prediction in a novel acuity score framework (DeepSOFA). Karimi et al. (2019) used attention models between RNN encoders and 1D convolution layers to pinpoint Secondary Structure Elements (SSEs) in proteins and atoms in compounds with high attention scores. Since, in the context of medical domain explanation is very important for the decision maker, we see that sometimes more emphasis is given toward interpretability while sacrificing performance if necessary.

We observed the use of data augmentation technique such as Generative Adversarial Networks (GAN) to compensate for the limited amount of real training data set. Xiang and Wang (2019) used the GAN technique to train their model with mode data as it is important to train these models with more data for robustness. Moreover, Fiosina et al. (2019) shows how ML based prediction of metadata (data augmentation) can considerably improve the quality of expression data annotation by systematically benchmarking DL and random forest based metadata augmentation of tissue, age, and sex using small RNA (sRNA) expression profiles. Another observation is that there is obvious tradeoff between the accuracy of prediction and the transparency of algorithms. Kanda et al. (2020) achieved a balance by developing a blended system, which was easily applicable to clinical settings and was useful for identifying patients at a high risk of death. Despite the increase in XAI research in the medical domain, there still remains opportunities for further research as presented in the following paragraphs.

Human-in-the-loop XAI ML: Within the medical domain there are many scenarios of human in the loop and real-time decision making (Caywood et al. 2016). In (Kanda et al. 2020; Liao et al. 2020; Lundberg et al. 2018; Qiu et al. 2020; Warman et al. 2020) we have seen ML decisions are being used by a domain expert to make final decision on a specific scenario. In real-time applications providing explanations are critical for any human-in-the-loop ML system (Ribeiro et al. 2016). To provide accurate, real-time explanations that are critical, we need predictable computational behavior for the XAI techniques. Therefore, more research is required to understand theoretical properties such as the appropriate number of samples and computational resource optimizations for deploying XAI systems for real-time human in the loop applications. Without proper understanding and testing of these limits deploying XAI systems for real-time human in the loop applications will remain challenging.

Article	ML Model	XAI Method and Technique
A biochemically-interpretable machine learning classifier for microbial GWAS (Kavvas et al. 2020)	Metabolic Allele Classifier (MAC)	Post-hoc, Visual
Application of explainable ensemble artificial intelligence model to categorization of hemodialysis-patient and treatment using nationwide-real-world data in Japan (Kanda et al. 2020)	K-Means and SVM Ensemble	Post-hoc, Visual, Signature gradients and R-group gradients
Beyond the scope of Free-Wilson analysis: building interpretable QSAR models with machine learning algorithms. (Chen et al. 2013)	Support Vector Machine (SVM)	Post-hoc, Visual
Clinical Interpretable Deep Learning Model for Glaucoma Diagnosis (Liao et al. 2020)	EAMNet based on Convolutional Neural Network (CNN)	Post-hoc, Visual
DeepAffinity: interpretable deep learning of compound-protein affinity through unified recurrent and convolutional neural networks. (Karimi et al. 2019)	Recurrent Neural Network (RNN)-CNN	Post-hoc, Attention mechanism, Visual
DeepSOFA: A Continuous Acuity Score for Critically Ill Patients using Clinically Interpretable Deep Learning. (Shickel et al. 2019)	(RNN) with Gated Recurrent Units (GRU)	Post-hoc, Attention mechanism, Visual
Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. (Qiu et al. 2020)	Fully Convolutional Network (FCN), Multilayer Perceptron (MLP)	Post-hoc, Visual
eXplainable Artificial Intelligence (XAI) for the identification of biologically relevant gene expression patterns in longitudinal human studies, insights from obesity research. (Anguita-Ruiz et al. 2020)	Sequential Rule Mapping (SRM)	Transparent, Visual
Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. (Lamy et al. 2019)	Weighted K-nearest Neighbor (WkNN)	Transparent, Visualization
Explainable Deep Learning for Augmentation of Small RNA Expression Profiles (Fiosina et al. 2019)	Neural Network (NN)	Post-hoc, DeepLIFT, Visualization
Explainable machine-learning predictions for the prevention of hypoxaemia during surgery (Lundberg et al. 2018)	XGBoost	Post-hoc, Local, SHAP, Visualization
Gaussian Process Regression for Predictive But Interpretable Machine Learning Models: An Example of Predicting Mental Workload across Tasks (Caywood et al. 2016)	Gaussian Process Regression (GPR)	Transparent, Global
Interpretable Artificial Intelligence for COVID-19 Diagnosis from Chest CT Reveals Specificity of Ground-Glass Opacities (Warman et al. 2020)	Decision Tree (DT)	Transparent, Visual
Interpretable Machine Learning Techniques for Causal Inference Using Balancing Scores as Meta-features. (Nohara et al. 2018)	Generalized Linear Regression Model (GLM)	Transparent, Rule based
ISeeU: Visually interpretable deep learning for mortality prediction inside the ICU (Caicedo-Torres and Gutierrez 2019)	Multi-scale CNN	Post-hoc, Visual
On the interpretability of machine learning- based model for predicting hypertension (Elshawi et al. 2019)	Random Forests Ensemble	Post-hoc, global-feature importance, Local-LIME, SHAP
PAGE-Net: Interpretable and Integrative Deep Learning for Survival Analysis Using Histopathological Images and Genomic Data. (Hao et al. 2020)	Pathway-based sparse deep neural network (Cox-PASNet)	Post-hoc
Plant disease identification using explainable 3D deep learning on hyperspectral images (Nagasubramanian et al. 2019)	Deep CNN (DCNN)	Post-hoc, Class saliency map-based visualization
Predicting adverse drug reactions through interpretable deep learning framework (Dey et al. 2018)	DNN	Post-hoc, Attention map with feature importance
Predicting diabetic retinopathy and identifying interpretable biomedical features using machine learning algorithms (Tsao et al. 2018)	SVM	Post-hoc, Decision Tree
The Virtual Operative Assistant: An explainable artificial intelligence tool for simulation-based training in surgery and medicine (Mirchi et al. 2020)	Linear SVM	Transparent
Towards Interpretable Skin Lesion Classification with Deep Learning Models. (Xiang and Wang 2019)	Auxiliary Classifier Generative Adversarial Network (AC-GAN), CNN	Post-hoc, Local, LIME

Table 1. List of papers with application of XAI methods and techniques

Accuracy vs. Explainability: There is obvious tradeoff between the accuracy of prediction and the transparency of algorithms. Gunning and Aha (2019) identified this inherent tension between ML

performance such as predictive accuracy and explainability. Often the highest-performing methods such as DL are the least explainable, and the most explainable method such as decision trees are the least accurate (Gunning and Aha 2019). In medical domain with critical decision-making process, one cannot take precedence over another. We see researchers try to achieve a balance by developing a blended system to optimize for both of these characteristics in identifying patients at a high risk of death (Kanda et al. 2020). There needs to be more research on what it means to balance between these two characteristics for a specific class of ML model. There are opportunities to map class of problem to specific ML techniques with identifying data sample representation requirements as well as model measurement metrics that can bring a balance between these two requirements.

Robustness of DL with Data Augmentation: Data Augmentation (DA) works as a solution to the problem of limited dataset for training a ML model. Therefore, DA has become a popular technique in building DL models because it improves the models generalization ability (Cook 2000). DA involves a suite of techniques that increases the size and improves the quality of training datasets so that robust ML models can be built with the dataset (Shorten and Khoshgoftaar 2019). In medical domain for many of the use case the availability of dataset can be small and DA techniques can be of help. For example, in (Fiosina et al. 2019) and (Mirchi et al. 2020) the researchers used GAN-based DA technique to improve robustness of their models. However, Shorten and Khoshgoftaar (2019) identifies that it is difficult to interpret the representations learned by neural networks for GAN-based augmentation, so human level explainable interface for convolutional networks features could greatly help guide the DA process for image dataset.

Moreover, DL models are generally vulnerable to adversarial attacks, where Adversarial Examples (AEs) are maliciously generated to mislead the model to output wrong predictions. AEs are often used to attack a DL model (Choo and Liu 2018), where modified samples force one or multiple DNN models outputs with wrong results (Zeng et al. 2020). Zeng et al. (2020) suggested a DA-based defense method against adversarial attacks in neural network based models. However, based on an empirical assessment of the effect of DA, regarding both classification accuracy and adversarial vulnerability, Cook (2000) suggests that general-purpose DA that do not take into account the characteristics of the data and the task, must be applied with care. Therefore, there exists research opportunity for DA and its application to various dataset and task models to better understand the effect on explainability and weakness toward adversarial vulnerability. Overall, how to maintain the robustness of a DL model is critical in real-world applications and especially within the medical domain.

Conclusion

We conducted a systematic literature review (SLR) focusing on the XAI methods and techniques used in the ML systems used in medical domain. Further, and to better organize and discuss extant literature we proposed and presented a conceptual framework for the classification of XAI methods and techniques. Accompanying the use of transparent and post-hoc systems, the literature appears to emphasize the models' interpretability as a requirement for ML models while using post-hoc methods for greater accuracy. While the literature tends to emphasize the balance between interpretability and accuracy, there is evidence of research that emphasized interpretability over accuracy. We also identify opportunities for XAI research most notably in the areas of human-in-the-loop XAI ML, the robustness of DL with data augmentation, and the tradeoff between accuracy and explainability. Moreover, research opportunity remains related to various specific use cases, methods, and techniques within the XAI domain. Some limitations of our study are as follows: 1) we used a short list of query terms, and 2) we limited the literature search to PubMed. While such limitations are commensurate with the scope of the review, namely providing a synopsis of XAI methods in the medical domain, a comprehensive review will entail extending the review to other databases, expanding the search query, and including snowballing (tracking forward and backward citations). Opportunities exist to potentially enhance the conceptual classification framework to describe methods and techniques at a finer level of abstraction.

In conclusion, XAI is a critical and timely field of research within AI, but more importantly in medical domain. The proliferation and acceptance of AI by the medical community will ultimately depend not only on the efficacy of these approaches but also on the ability to provide a meaningful explanation for clinicians. The proposed classification framework and the findings from the SLR provide a perspective on the current-state-of-the-art and insights for future research for XAI in the medical domain.

REFERENCES

Adadi, A., and Berrada, M. 2018. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access* (6), pp. 52138–52160.

Anguita-Ruiz, A., Segura-Delgado, A., Alcalá, R., Aguilera, C. M., and Alcalá-Fdez, J. 2020. "EXplainable Artificial Intelligence (XAI) for the Identification of Biologically Relevant Gene Expression Patterns in Longitudinal Human Studies, Insights from Obesity Research.," *PLoS Computational Biology* (16:4), p. e1007792. (https://doi.org/10.1371/journal.pcbi.1007792).

Arrieta, A. [Barredo, Díaz-Rodríguez, N., Ser, J. [Del, Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," *Information Fusion* (58), pp. 82–115. (https://doi.org/10.1016/j.inffus.2019.12.012).

Arya, V., Bellamy, R. K. E., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S. C., Houde, S., Liao, Q. V., Luss, R., Mojsilović, A., Mourad, S., Pedemonte, P., Raghavendra, R., Richards, J., Sattigeri, P., Shanmugam, K., Singh, M., Varshney, K. R., Wei, D., and Zhang, Y. 2019. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques.

Barredo Arrieta, A., Diaz-Rodriguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," *Information Fusion* (58), pp. 82–115. (https://doi.org/10.1016/j.inffus.2019.12.012).

Beil, M., Proft, I., van Heerden, D., Sviri, S., and van Heerden, P. V. 2019. "Ethical Considerations about Artificial Intelligence for Prognostication in Intensive Care.," *Intensive Care Medicine Experimental* (7:1), p. 70. (https://doi.org/10.1186/s40635-019-0286-6).

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M. F., and Eckersley, P. 2020. "Explainable Machine Learning in Deployment," in Proceedings of the 2020 *Conference on Fairness, Accountability, and Transparency, FAT* '20*, New York, NY, USA: ACM, pp. 648–657. (https://doi.org/10.1145/3351095.3375624).

Bohanec, M., Robnik-Šikonja, M., and Borštnar, M. K. 2017. "Decision-Making Framework with Double-Loop Learning through Interpretable Black-Box Machine Learning Models," *Industrial Management & Data Systems* (117:7), Wembley: Emerald Group Publishing Limited, pp. 1389–1406. (https://doi.org/10.1108/IMDS-09-2016-0409).

Cabitza, F., Ciucci, D., and Rasoini, R. 2017. A Giant with Feet of Clay: On the Validity of the Data That Feed Machine Learning in Medicine. (http://arxiv.org/abs/1706.06838).

Caicedo-Torres, W., and Gutierrez, J. 2019. "ISeeU: Visually Interpretable Deep Learning for Mortality Prediction inside the ICU.," *Journal of Biomedical Informatics* (98), United States, p. 103269.

Caywood, M. S., Roberts, D. M., Colombe, J. B., Greenwald, H. S., and Weiland, M. Z. 2016. "Gaussian Process Regression for Predictive But Interpretable Machine Learning Models: An Example of Predicting Mental Workload across Tasks.," *Frontiers in Human Neuroscience* (10), p. 647.

Chakraborty, S., Tomsett, R., Raghavendra, R., Harborne, D., Alzantot, M., Cerutti, F., Srivastava, M., Preece, A., Julier, S., Rao, R. M., Kelley, T. D., Braines, D., Sensoy, M., Willis, C. J., and Gurram, P. 2017. "Interpretability of Deep Learning Models: A Survey of Results," in *SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI*, pp. 1–6. (https://doi.org/10.1109/UIC-ATC.2017.8397411).

Chen, H., Carlsson, L., Eriksson, M., Varkonyi, P., Norinder, U., and Nilsson, I. 2013. "Beyond the Scope of Free-Wilson Analysis: Building Interpretable QSAR Models with Machine Learning Algorithms.," *Journal of Chemical Information and Modeling* (53:6), pp. 1324–1336.

Choo, J., and Liu, S. 2018. "Visual Analytics for Explainable Deep Learning," *IEEE Computer Graphics and Applications* (38:4), pp. 84–92. (https://doi.org/10.1109/MCG.2018.042731661).

Cook, R. D. 2000. "ADVERSARIAL ROBUSTNESS IN DATA AUGMENTATION," *Technometrics* (42:1), pp. 65–68. (https://doi.org/10.1080/00401706.2000.10485981).

Dey, S., Luo, H., Fokoue, A., Hu, J., and Zhang, P. 2018. "Predicting Adverse Drug Reactions through Interpretable Deep Learning Framework.," *BMC Bioinformatics* (19:Suppl 21), p. 476.

Doshi-Velez, F., and Kim, B. 2017. "Towards A Rigorous Science of Interpretable Machine Learning," ArXiv:1702.08608 [Cs, Stat]. (http://arxiv.org/abs/1702.08608).

Došilović, F. K., Brčić, M., and Hlupić, N. 2018. "Explainable Artificial Intelligence: A Survey," in *41st Convention - MIPRO*, May, pp. 0210–0215. (https://doi.org/10.23919/MIPRO.2018.8400040).

Du, M., Liu, N., and Hu, X. 2020. "Techniques for Interpretable Machine Learning," *Communications of the ACM* (63:1), New York: ACM, p. 68. (https://doi.org/10.1145/3359786).

Elshawi, R., Al-Mallah, M. H., and Sakr, S. 2019. "On the Interpretability of Machine Learning-Based Model for Predicting Hypertension.," *BMC Medical Informatics and Decision Making* (19:1), p. 146. (https://doi.org/10.1186/s12911-019-0874-0).

Fellous, J.-M., Sapiro, G., Rossi, A., Mayberg, H., and Ferrante, M. 2019. "Explainable Artificial Intelligence for Neuroscience: Behavioral Neurostimulation.," *Frontiers in Neuroscience* (13), p. 1346. (https://doi.org/10.3389/fnins.2019.01346).

Fiosina, J., Fiosins, M., and Bonn, S. 2019. "Explainable Deep Learning for Augmentation of Small RNA Expression Profiles.," *Journal of Computational Biology* (27:2), pp. 234–247.

Gill, N., Hall, P., Montgomery, K., and Schmidt, N. 2020. "A Responsible Machine Learning Workflow with Focus on Interpretable Models, Post-Hoc Explanation, and Discrimination Testing," *Information* (11:3), p. 137. (https://doi.org/10.3390/info11030137).

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. 2018. "Explaining Explanations: An Overview of Interpretability of Machine Learning," in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, October, pp. 80–89.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. 2018. "A Survey of Methods for Explaining Black Box Models," *ACM Comput. Surv.* (51:5), New York, NY, USA: Association for Computing Machinery. (https://doi.org/10.1145/3236009).

Gunning, D., and Aha, D. 2019. "DARPA's Explainable Artificial Intelligence Program," *AI Magazine* (40:2), La Canada, pp. 44–58. (https://doi.org/10.1609/aimag.v40i2.2850).

Hao, J., Kosaraju, S. C., Tsaku, N. Z., Song, D. H., and Kang, M. 2020. "PAGE-Net: Interpretable and Integrative Deep Learning for Survival Analysis Using Histopathological Images and Genomic Data.," *Pacific Symposium on Biocomputing* (25), United States, pp. 355–366.

Holzinger, A., Kieseberg, P., Weippl, E., and Tjoa, A. M. 2018. Current Advances, Trends and Challenges of Machine Learning and Knowledge Extraction: From Machine Learning to Explainable AI: Second IFIP TC 5, TC 8/WG 8.4, 8.9, TC 12/WG 12.9 *International Cross-Domain Conference*, CD-MAKE 2018, Hamburg, Germany, August 27–30, 2018, Proceedings, pp. 1–8. (https://doi.org/10.1007/978-3-319-99740-7_1).

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., and Müller, H. 2019. "Causability and Explainability of Artificial Intelligence in Medicine.," *Wiley Interdisciplinary Reviews*. DMKD (9:4), p. e1312. (https://doi.org/10.1002/widm.1312).

Holzinger, A., Plass, M., Holzinger, K., Crisan, G. C., Pintea, C.-M., and Palade, V. 2017. "A Glass-Box Interactive Machine Learning Approach for Solving NP-Hard Problems with the Human-in-the-Loop," *CoRR* (abs/1708.01104). (http://arxiv.org/abs/1708.01104).

Kanda, E., Epureanu, B. I., Adachi, T., Tsuruta, Y., Kikuchi, K., Kashihara, N., Abe, M., Masakane, I., and Nitta, K. 2020. "Application of Explainable Ensemble Artificial Intelligence Model to Categorization of Hemodialysis-Patient and Treatment Using Nationwide-Real-World Data in Japan.," *PloS One* (15:5), United States, p. e0233491.

Karimi, M., Wu, D., Wang, Z., and Shen, Y. 2019. "DeepAffinity: Interpretable Deep Learning of Compound-Protein Affinity through Unified Recurrent and Convolutional Neural Networks.," *Bioinformatics* (Oxford, England) (35:18), pp. 3329–3338.

Kaul, S. 2018. "Speed And Accuracy Are Not Enough! Trustworthy Machine Learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, New York, NY, USA: ACM, pp. 372–373. (https://doi.org/10.1145/3278721.3278796).

Kavvas, E. S., Yang, L., Monk, J. M., Heckmann, D., and Palsson, B. O. 2020. "A Biochemically-Interpretable Machine Learning Classifier for Microbial GWAS.," *Nature Communications* (11:1), p. 2580. (https://doi.org/10.1038/s41467-020-16310-9).

Lamy, J.-B., Sekar, B., Guezennec, G., Bouaud, J., and Séroussi, B. 2019. "Explainable Artificial Intelligence for Breast Cancer: A Visual Case-Based Reasoning Approach.," *Artificial Intelligence in Medicine* (94), Netherlands, pp. 42–53. (https://doi.org/10.1016/j.artmed.2019.01.001).

Liao, W., Zou, B., Zhao, R., Chen, Y., He, Z., and Zhou, M. 2020. "Clinical Interpretable Deep Learning Model for Glaucoma Diagnosis," *IEEE Journal of Biomedical and Health Informatics* (24:5), pp. 1405–1412. (https://doi.org/10.1109/JBHI.2019.2949075).

London, A. J. 2019. "Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability," *Hastings Center Report* (49:1), pp. 15–21. (https://doi.org/10.1002/hast.973).

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K.-W., Newman, S.-F., Kim, J., and Lee, S.-I. 2018. "Explainable Machine-Learning Predictions for the Prevention of Hypoxaemia during Surgery.," *Nature Biomedical Engineering* (2:10), pp. 749–760. (https://doi.org/10.1038/s41551-018-0304-0).

Mirchi, N., Bissonnette, V., Yilmaz, R., Ledwos, N., Winkler-Schwartz, A., and Del Maestro, R. F. 2020. "The Virtual Operative Assistant: An Explainable Artificial Intelligence Tool for Simulation-Based Training in Surgery and Medicine.," *PloS One* (15:2), p. e0229596.

Nagasubramanian, K., Jones, S., Singh, A. K., Sarkar, S., Singh, A., and Ganapathysubramanian, B. 2019. "Plant Disease Identification Using Explainable 3D Deep Learning on Hyperspectral Images.," *Plant Methods* (15), p. 98. (https://doi.org/10.1186/s13007-019-0479-8).

Nohara, Y., Iihara, K., and Nakashima, N. 2018. "Interpretable Machine Learning Techniques for Causal Inference Using Balancing Scores as Meta-Features.," *IEEE Engineering in Medicine and Biology Society* (2018), United States, pp. 4042–4045. (https://doi.org/10.1109/EMBC.2018.8513026).

Qiu, S., Joshi, P. S., Miller, M. I., Xue, C., Zhou, X., Karjadi, C., Chang, G. H., Joshi, A. S., Dwyer, B., Zhu, S., Kaku, M., Zhou, Y., Alderazi, Y. J., Swaminathan, A., Kedar, S., Saint-Hilaire, M.-H., Auerbach, S. H., Yuan, J., Sartor, E. A., Au, R., and Kolachalama, V. B. 2020. "Development and Validation of an Interpretable Deep Learning Framework for Alzheimer's Disease Classification.," *Brain : A Journal of Neurology, England.* (https://doi.org/10.1093/brain/awaa137).

Ribeiro, M. T., Singh, S., and Guestrin, C. 2016. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," *CoRR* (abs/1602.04938). (http://arxiv.org/abs/1602.04938).

Schoenborn, J., and Althoff, K.-D. 2019. "Recent Trends in XAI: A Broad Overview on Current Approaches, Methodologies and Interactions," *27th International Conference on Case-Based Reasoning*, September 10, 2019.

Shickel, B., Loftus, T. J., Adhikari, L., Ozrazgat-Baslanti, T., Bihorac, A., and Rashidi, P. 2019. "DeepSOFA:A Continuous Acuity Score for Critically Ill Patients Using Clinically Interpretable Deep Learning," *Scientific Reports* (9:1), p. 1879. (https://doi.org/10.1038/s41598-019-38491-0).

Shorten, C., and Khoshgoftaar, T. M. 2019. "A Survey on Image Data Augmentation for Deep Learning," *Journal of Big Data* (6:1), p. 60. (https://doi.org/10.1186/s40537-019-0197-0).

Suresh, H., and Guttag, J. V. 2020. "A Framework for Understanding Unintended Consequences of Machine Learning," ArXiv:1901.10002 [Cs, Stat]. (http://arxiv.org/abs/1901.10002).

Tjoa, E., and Guan, C. 2019. "A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI," *CoRR* (abs/1907.07374). (http://arxiv.org/abs/1907.07374).

Tommasi, T., Patricia, N., Caputo, B., and Tuytelaars, T. 2017. "A Deeper Look at Dataset Bias," in *Domain Adaptation in Computer Vision Applications*, G. Csurka (ed.), Cham: Springer International Publishing, pp. 37–55. (https://doi.org/10.1007/978-3-319-58347-1_2).

Tsao, H.-Y., Chan, P.-Y., and Su, E. C.-Y. 2018. "Predicting Diabetic Retinopathy and Identifying Interpretable Biomedical Features Using Machine Learning Algorithms.," *BMC Bioinformatics* (19:Suppl 9), p. 283. (https://doi.org/10.1186/s12859-018-2277-0).

Warman, A., Warman, P., Sharma, A., Parikh, P., Warman, R., Viswanadhan, N., Chen, L., Mohapatra, Subhra, Mohapatra, Shyam, and Sapiro, G. 2020. "Interpretable Artificial Intelligence for COVID-19 Diagnosis from Chest CT Reveals Specificity of Ground-Glass Opacities," *MedRxiv : The Preprint Server for Health Sciences*. (https://doi.org/10.1101/2020.05.16.20103408).

Xiang, A., and Wang, F. 2019. "Towards Interpretable Skin Lesion Classification with Deep Learning Models," *AMIA Symposium* (2019), pp. 1246–1255.

You, Y., Zhang, Z., Hsieh, C.-J., Demmel, J., and Keutzer, K. 2018. "ImageNet Training in Minutes," in *Proceedings of the 47th International Conference on Parallel Processing*, ICPP 2018, New York, NY, USA: Association for Computing Machinery. (https://doi.org/10.1145/3225058.3225069).

Zeng, Y., Qiu, H., Memmi, G., and Qiu, M. 2020. "A Data Augmentation-Based Defense Method Against Adversarial Attacks in Neural Networks," ArXiv:2007.15290 [Cs]. (http://arxiv.org/abs/2007.15290).

Zhang, Q., and Zhu, S. 2018. "Visual Interpretability for Deep Learning: A Survey," Frontiers of Information *Technology & Electronic Engineering* (19:1), pp. 27–39.