Explainable Artificial Intelligence Methods in Combating Pandemics: A Systematic Review

Felipe Giuste, *Student Member, IEEE*, Wenqi Shi, *Student Member, IEEE*, Yuanda Zhu, *Student Member, IEEE*, Tarun Naren, Monica Isgut, Ying Sha, Li Tong, Mitali Gupte, and May D. Wang, *Fellow, IEEE*

(Methodological Review)

***Abstract*—Despite the myriad peer-reviewed papers demonstrating novel Artificial Intelligence (AI)-based solutions to COVID-19 challenges during the pandemic, few have made a significant clinical impact, especially in diagnosis and disease precision staging. One major cause for such low impact is the lack of model transparency, significantly limiting the AI adoption in real clinical practice. To solve this problem, AI models need to be explained to users. Thus, we have conducted a comprehensive study of Explainable Artificial Intelligence (XAI) using PRISMA technology. Our findings suggest that XAI can improve model performance, instill trust in the users, and assist users in decision-making. In this systematic review, we introduce common XAI techniques and their utility with specific examples of their application. We discuss the evaluation of XAI results because it is an important step for maximizing the value of AI-based clinical decision support systems. Additionally, we present the traditional, modern, and advanced XAI models to demonstrate the evolution of novel techniques. Finally, we provide a best practice guideline that developers can refer to during the model experimentation. We also offer potential solutions with specific examples for common challenges in AI model experimentation. This comprehensive review, hopefully, can promote AI adoption in biomedicine and healthcare.**

***Index Terms*—COVID-19, electronic health records, explainable artificial intelligence, explanation evaluation, explanation generation, explanation representation, medical imaging.**

Manuscript received 18 December 2021; revised 4 April 2022; accepted 29 May 2022. Date of publication 23 June 2022; date of current version 6 January 2023. This work was supported in part by Wallace H. Coulter Distinguished Faculty Fellowship (M. D. Wang), in part by Petit Institute Faculty Fellowship (M. D. Wang), and in part by Microsoft Research. *(Felipe Giuste and Wenqi Shi contributed equally to this work.) (Corresponding author: May D. Wang.)*

Felipe Giuste, Ying Sha, Li Tong, Mitali Gupte, and May D. Wang are with the Wallace H. Coulter School of Biomedical Engineering, Georgia Institute of Technology, Emory University, Atlanta, GA 30322 USA (e-mail: fgiuste@gatech.edu; ysha8@gatech.edu; ltong9@gatech.edu; mitali.gupte@gatech.edu; maywang@gatech.edu).

Wenqi Shi and Yuanda Zhu are with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: wshi83@gatech.edu; yzhu94@gatech.edu).

Monica Isgut is with the School of Biology, Georgia Institute of Technology, Atlanta, GA 30322 USA (e-mail: misgut@gatech.edu).

Tarun Naren is with the Department of Nuclear and Radiological Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: tnaren3@gatech.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/RBME.2022.3185953, provided by the authors.

Digital Object Identifier 10.1109/RBME.2022.3185953

I. Introduction

CORONAVIRUS disease 2019 (COVID-19) has become a worldwide phenomenon with over 545 million cases and claiming over six million lives [1]. Medical imaging, such as X-ray and computed tomography (CT), and electronic health records have been used in addition to molecular tests for diagnosing and precision staging patients potentially infected with COVID-19 [2]–[4]. The need for fast COVID-19 detection has led to a massive number of state-of-art Artificial Intelligence (AI) solutions to alleviate this clinical burden [3], [5]. Unfortunately, very few have succeeded in making a real impact [6]. As the world transitions from disease detection and containment to maximizing patient care outcomes, AI solutions must also improve. In preparing for the future pandemic, we must get lessons learned from this pandemic process. One such big lesson is why many successful models published before have failed to make a meaningful clinical impact. The low AI adoption in clinical decision support is due to the lack of transparency in AI model development, and the lack of interpretability of their results. Thus, physicians and other healthcare practitioners are often reluctant to adopt high-performing yet black-box AI systems. For AI developers, without explaining AI, there exists a high risk of generating models relying on noise instead of real, clinically-meaningful signals [7]–[9], and the ability of researchers and model developers to identify potential pitfalls and avenues for improvement is also very limited.

Explainable artificial intelligence (XAI) is a collection of processes and methods that enables human users to comprehend and trust machine learning algorithms' results [10]. XAI techniques improve the transparency of AI models, which leads to more clinical decision-making confidence and more real-world adoption of AI. Clinicians benefit from XAI by gaining insight into how the AI models reach solutions from clinical data, as shown in Fig. 1.

In general, the goals of AI-based solutions for clinical settings are: to achieve high performance, to instill user trust, and

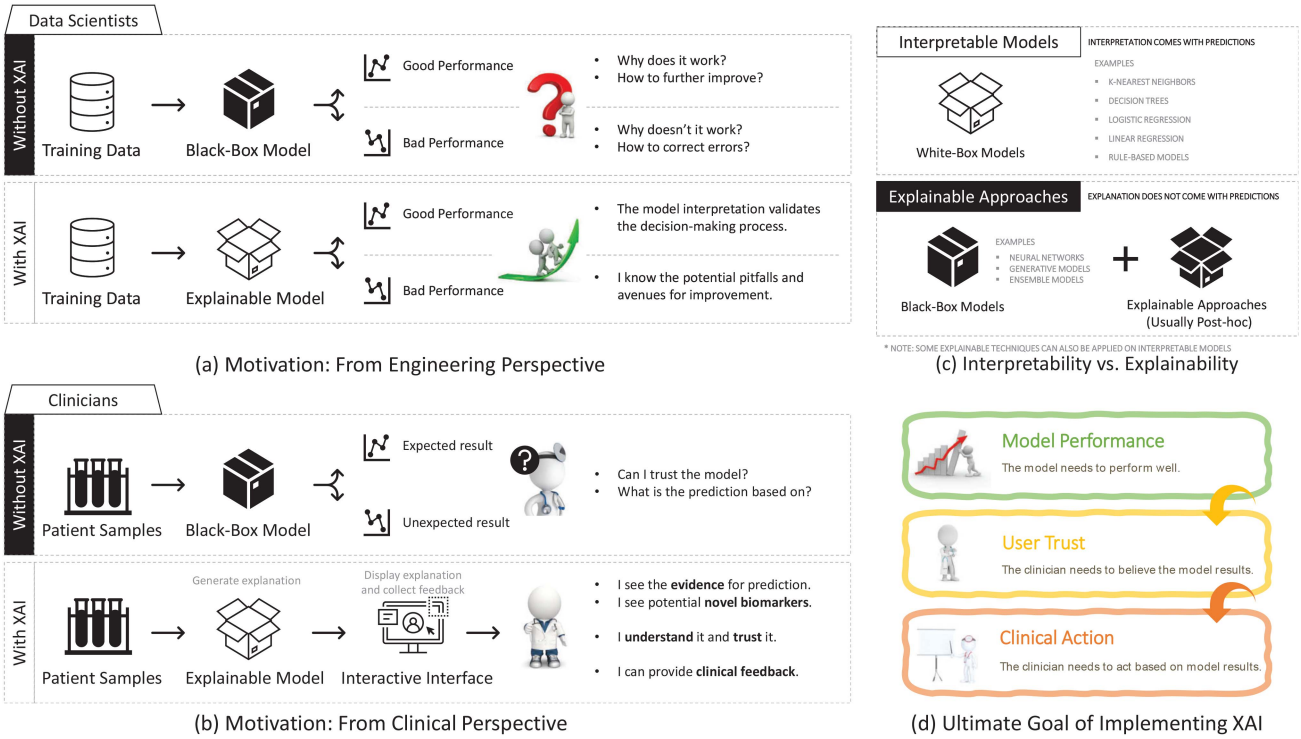

Fig. 1. Problem statement and motivation of XAI in clinical applications. (a) From a model development perspective, XAI techniques enhance the transparency of AI models, allowing for more confident clinical decision-making and increasing the real-world utility of AI approaches. (b) From a clinical perspective, clinicians can benefit from XAI by gaining insight into how the AI models reach solutions from clinical data. (c) The term "interpretability" refers to a property of AI systems in which the process by which they arrive at a conclusion is easily understood. K-nearest neighbors, decision trees, logistic regression, linear models, and rule-based models are all popular interpretable machine learning methods. Explainable AI is frequently used to refer to methods (usually post-hoc) for enhancing comprehension of black-box models such as neutral networks and ensemble models. Explainable AI methods attempt to summarize the rationale for a model's behavior or to generate insights into the underlying cause of a model's decision. Both interpretability and explainability are frequently used interchangeably, and both seek to shed light on the model's credibility. In this review, we will focus on XAI methods used in clinical settings. (d) AI-based clinical solutions should meet three criteria: achieve high performance, instill user trust, and generate user response, all of which demonstrate the importance of XAI in clinical applications.

to reflect user response. Specifically, the model should have achieved sufficient performance at their task on a real-world dataset not used during the training process in order to be considered for real-world use. Guidelines for establishing and reporting real-world clinical trial performance can be found in the SPIRIT-AI [11] and CONSORT-AI [12] guidelines. Trust in the AI solution may be established with XAI, especially when visual feedback is provided to the user on important metrics used to obtain the model prediction. Finally, no solution is effective if it does not result in a change in user response. This response may include a change in the treatment plan, patient prioritization, or diagnosis. This response must be consistent with clinical expertise and evidence-based protocols.

To address low AI adoption, we will mainly focus on XAI solutions to improve end-user trust. Model performance and user interfaces are also mentioned where appropriate. XAI can allow for validation of extracted features, confirm heuristics, identify patient subgroups, and generate novel biomarkers [13]. In addition, XAI can also support research conclusions and guide research field advancement by identifying avenues of model performance improvement. We hope to contribute a unique resource for biomedical engineers working on healthcare-related challenges so that their AI models have a better potential for a positive clinical impact.

In this systematic review, we describe XAI utility during COVID-19. We illustrate how the XAI-based studies applicable to COVID-19 were selected using the Preferred Reporting Items on Systematic Reviews and Meta-analysis (PRISMA) model [14] and exclusivity criteria (see Fig. 2). Upon review of the current literature leveraging AI for COVID-19 detection and risk assessment, XAI is strongly needed for clinical adoption. The remainder of this paper is structured as follows: Section II provides a comprehensive overview of the XAI approaches used to support AI-enabled clinical decision support systems during COVID-19 pandemic; Section III describes the existing evaluation pipelines of XAI methods; Section IV and V summarize the contribution of this paper, provide a schema of the integration of explainable AI module in both model development and clinical practice, and discuss potential challenges and future work of XAI. We have presented a comprehensive review of current efforts in solving existing and future pandemic challenges with XAI approaches.

II. XAI Methods in COVID-19 Applications

In this section, we introduce XAI approaches used to support AI-enabled clinical decision support systems. We categorize

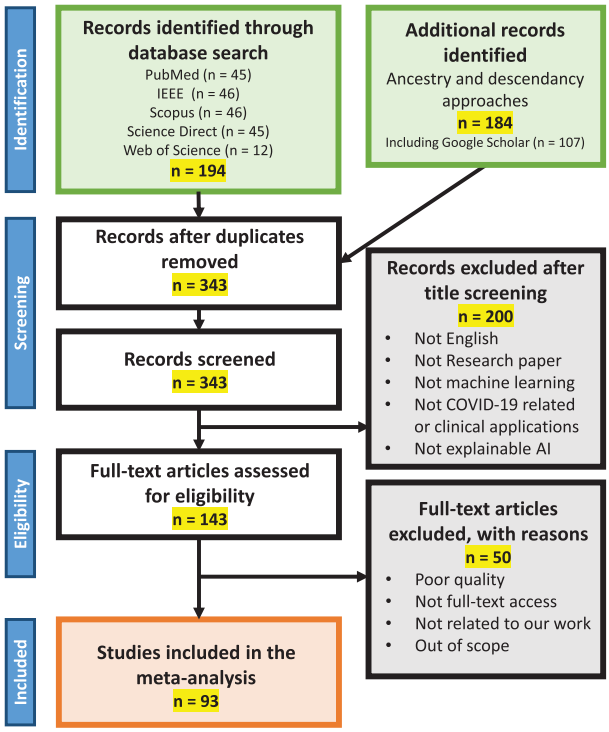

Fig. 2. PRISMA chart for systematic paper selection and quality assessment.

them as follows: data augmentation, outcome prediction, unsupervised clustering, and image segmentation. Moreover, we organized XAI methods according to the underlying theory within each task, as shown in Fig. 3. Additional technical details and clinical applications will be discussed below.

A. Data Augmentation

The need for labeled data for model training was highlighted in the early stages of the COVID-19 pandemic. This was also a point in time where AI-based solutions could have made the most impact by supplementing scarce public datasets. Future pandemics will likely result in the same urgency for labeled data, and AI-solutions would greatly benefit from synthetic data augmentation. Generative Adversarial Networks (GAN) are used to supplement available labeled COVID radiology data with synthetic images and labels. This allows for improved model training with limited labeled datasets by increasing the number of labeled images available for training. Example of classical and modern data augmentation approaches with model interpretation is shown in Fig. 4.

Singh et al. tested a wide variety of GAN models to generate synthetic X-ray images while training a COVID-19 detection deep learning model named COVIDscreen [15]. They compared the quality of four different GAN-based X-ray image generators including Wasserstein GAN (WGAN), least squares GAN (LSGAN), auxiliary classifier GAN (ACGAN), and deep convolutional GAN. They visualized the resulting synthetic X-ray images and showed that WGAN produces visibly higher quality images than the tested alternatives. To the best of our knowledge, this was the first publication to show successful X-ray image generation for COVID-19 data augmentation. A significant limitation of this study was that, although they generated realistic X-ray images using WGAN, they did not leverage this additional data to improve their classifier performance. This is likely due to the lack of label generation during image synthesis which prevents the use of their synthetic images for supervised learning approaches. Despite this limitation, their success in generating synthetic clinical images from a limited COVID-19 dataset illustrated the feasibility of this approach for future work.

Waheed et al. [16] train an Auxiliary Classifier Generative Adversarial Network (ACGAN) to generate synthetic X-ray images. ACGANS take both a label and noise as input to generate new images with known labels. Using COVID-19 status as the label, the proposed model CovidGAN is able to generate normal and COVID-19 images. They train a convolutional neural network (CNN) COVID-19 classifier and compare its performance when trained on a real labeled dataset and a dataset augmented with synthetic images from CovidGAN. They demonstrate that augmentation of their labeled dataset with synthetic images improves classifier performance from 85% to 95% classification accuracy.

Loey et al. [17] trained four CNN classifiers to detect COVID-19 within chest CT images. Synthetic CT images were generated with a conditional GAN (CGAN). They compared the performance of each classifier when trained with four different datasets. Training datasets include: the original dataset alone, the original with morphological augmentation, the original with synthetic images, and the original with morphological augmented combined with synthetic images. They demonstrated that the best classifier ResNet50 was trained on the original dataset with morphological enhancement and balanced accuracy of 82.64%.

Although GANs are widely used for clinical image generation, XAI techniques are not commonly used to understand how they generate the final images from the latent space. Without XAI, it is difficult to detect potential biases in generated images. This is especially important when models are trained on small clinical datasets and subject to a wide range of confounding variables, such as hospital-specific signal properties associated with COVID-19 diagnosis. The following novel XAI techniques allow for the interpretation of the GAN latent space in order to understand how sampling of the latent space affects the final image.

Voynov and Babenko [18] created a GAN learning scheme to maximize the interpretability of the GAN latent space. This approach allows the latent space to describe a set of independent image transformations. They showed that this latent space can be visually interpreted and manipulated to generate synthetic images with specific properties (e.g., object rotation, background blur, zoom, etc.). Their method produced synthetic images with interpretable latent space sampling effects across a wide range of datasets, including MNIST, AnimeFaces, CelebA-HQ, and BigGAN. They show that their interpretation of the latent space can be used to create images with specific properties including zoom, background blur, hair type, skin type, glasses, and many others. These properties were specific to the dataset the GANs were trained on.

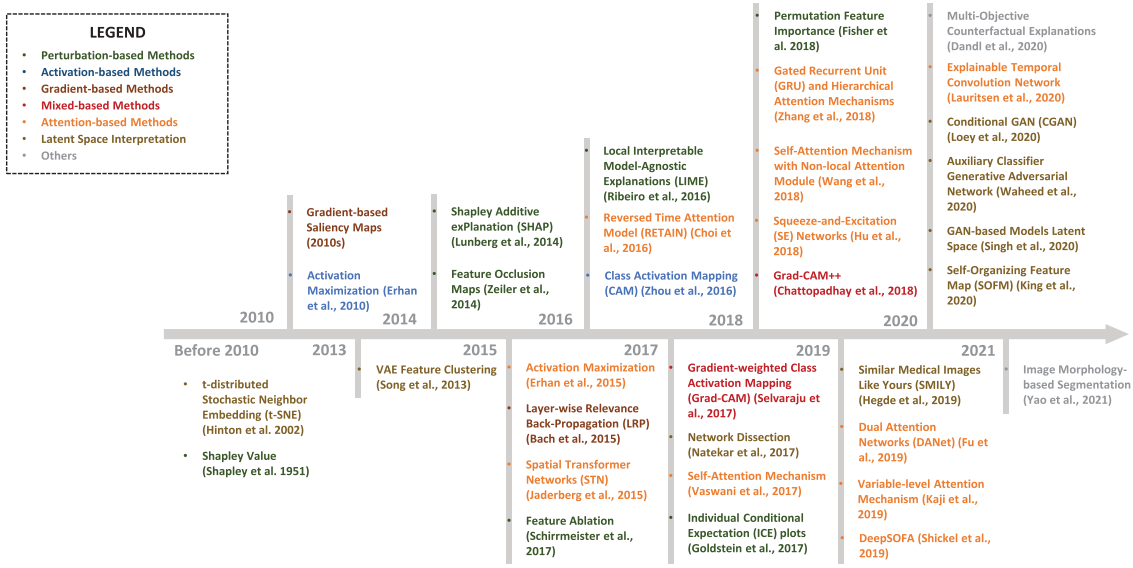

Fig. 3. A brief summary of significant milestones in the development of XAI methods. According to their underlying theory, we classified these popular XAI methods into six categories: perturbation-based, activation-based, gradient-based, mixed-based, attention-based, and latent space interpretation. In the early stages of XAI development, perturbation-based, activation-based, and gradient-based methods are critical for model interpretation and generation of explanations. Recent years have seen significant advancements in mixed-based methods (combination of activation- and gradient-based methods), attention mechanisms, and latent space interpretations, all of which have played a significant role in medical XAI.

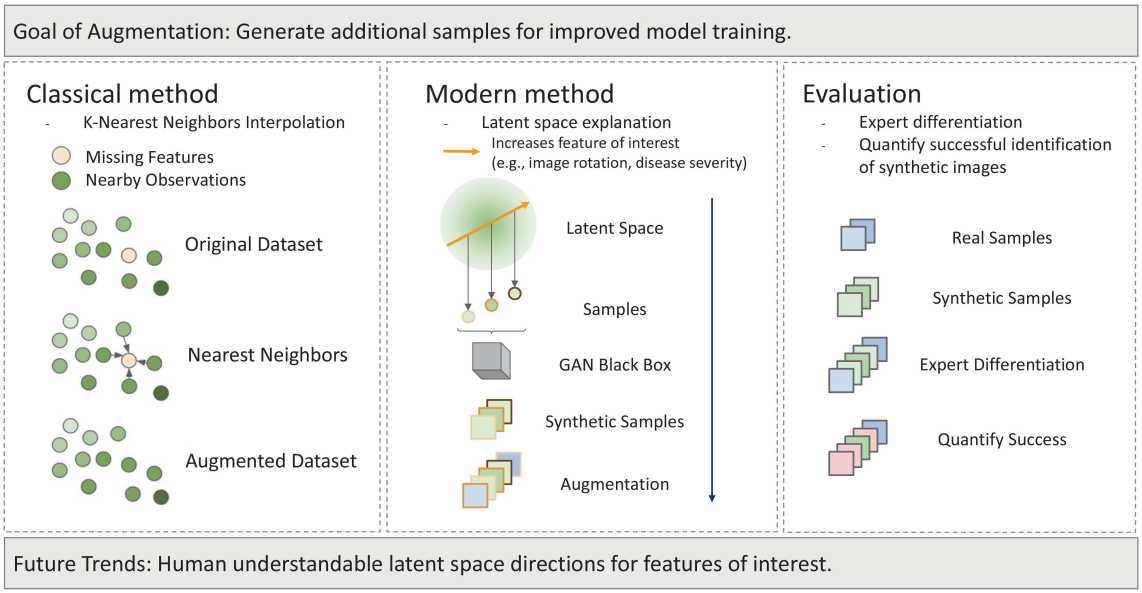

Fig. 4. Examples of classical and modern XAI approaches in clinical data augmentation tasks. It is common to see K-Nearest Neighbors (KNN) [22] interpolation and other classic approaches used in place of more complex modern solutions [18] when performing data augmentation due to a dearth of real-world examples of their successes and failures. The modern approach may result in data bias that is difficult to comprehend without more real-world examples. The trend in data augmentation has been to increase the number of features and the complexity of data transformations in order to more accurately model the underlying distribution of real data.

Härkönen et al. [19] also sought to utilize the GAN latent space for image synthesis with specific properties. Instead of re-training models to isolate latent space axis of greatest interpretation, they take existing GANs and identify explainable latent space axes. This allowed them to modify an image's properties, such as converting concrete to grass and changing the color of an object. Principal component analysis (PCA), which requires no additional model training, was used to extract the interpretable latent space axes. This technique could also be used to modify image properties such as adding wrinkles and gray hair to a person while retaining the original image's label. This methodology allowed the synthesis of additional labeled images containing known object properties.

GANs used to generate additional radiology images can be interpreted to determine the most interpretable directions. This would enable users to deduce which image properties the

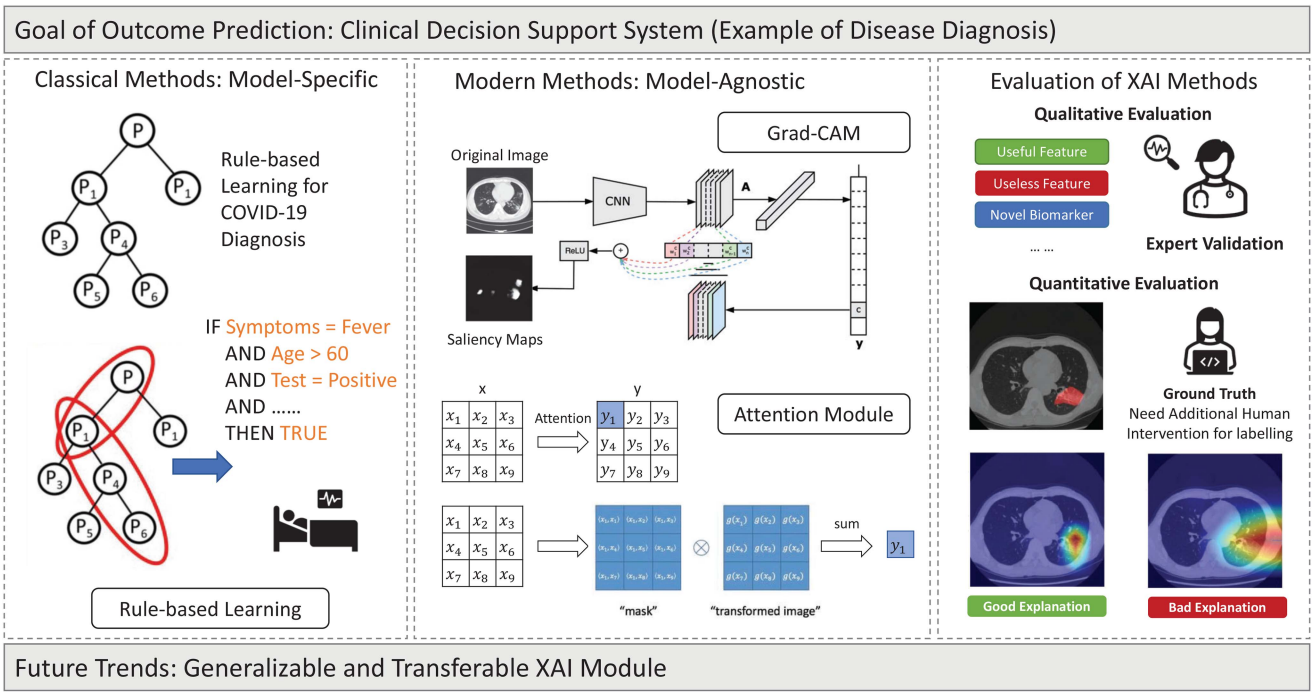

Fig. 5. Examples of classical and modern XAI approaches in clinical decision support system. For the task of disease diagnosis, the trend has been to create visualizations of input importance that can be used with a wide variety of popular deep learning models (model-agnostic) [23], [24]. This is in contrast to early XAI approaches [25], which emphasized model-specific solutions in order to improve interpretability.

generator has been trained to reproduce. There is also the potential to identify latent space directions that are significantly associated with the presence of COVID-19 infection. Examining these vectors may aid in the development of a more complete understanding of COVID-19 disease pathology. Non-COVID-19 directions can be used to alter labeled images without affecting their class labels, allowing for the addition of interpretable noise to datasets. This augmentation method improves classifier performance by training on a larger variety of images, thereby reducing the likelihood of overfitting.

The latent space of COVID-19 GANs are not being examined enough for interpretable features. This is a missed opportunity to identify novel COVID-specific image properties. Using XAI to understand latent space effects on image generation would also allow generation of images with desired properties. XAI also allows examination of image transformation "directions" such as object rotation and zoom to ensure that they are independent, and uncorrelated with potential sources of confounding, such as scanner model, hospital source, and technician bias. In future pandemics, reliable and explainable synthetic data augmentation approaches may facilitate the training of high-performing AI models to help in the clinical arena.

In addition to data augmentation, synthetic examples can be used to improve model robustness to outliers. Rahman et al. showed that many COVID-19 diagnostic models are vulnerable to attacks by adversarial examples [20]. Palatnik de Sousa et al. [21] also demonstrated the utility of adding random colored artifacts to CT images to identify model architecture which are most robust to such perturbation. This illustrates the importance of robust validation of models prior to their integration within clinical settings. XAI may also be used to verify the validity of models' approach to guard against such unexpected, and potentially harmful, results.

B. Outcome Prediction

Due to their rapid acquisition times and accessibility, imaging modalities such as X-rays and CT scans have aided clinicians tremendously in diagnosing COVID-19. Radiographic signs, such as airspace opacity, ground-glass opacity, and subsequent consolidation, aid in the diagnosis of COVID-19. However, medical images contain hundreds of slices making diagnosis difficult for clinicians. COVID-19 also exhibits similarities to a variety of other types of pneumonia, posing an additional challenge for clinicians. Although AI-based clinical decision support systems outperform conventional models that have been adapted for clinical use, clinicians frequently lack trust in or understanding of them due to unknown risks, posing a significant barrier to widespread adoption. In the context of outcome prediction, we define conventional models as SVMs, tree-based approaches, and logistic regression. Thus, XAI-assisted diagnosis via radiological imaging is highly desirable, as it can be viewed as an explainable image classification task for distinguishing COVID-19 from other pneumonia and healthy subjects, as shown in Fig. 5. Another important clinical application in outcome predication task is risk prediction. Clinicians and researchers use Electronic Health Records (EHRs) to predict risk of adverse clinical events, such as mortality or ICU readmission, and to identify top-ranking clinical features to mitigate negative consequences.

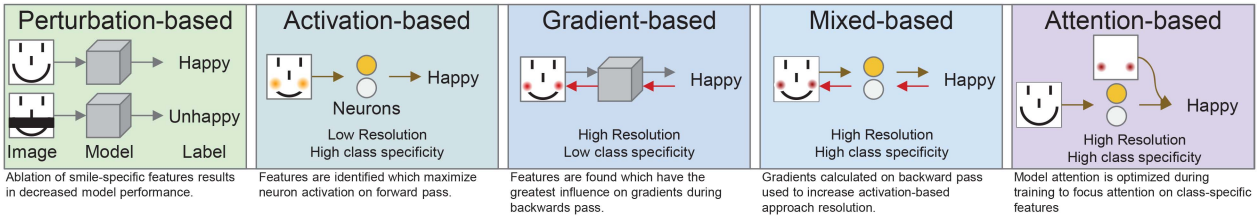

Fig. 6. Overview of Outcome Prediction XAI Approaches: Perturbation, Activation, Gradient, Mixed, and Attention-based. A) Permutation-based approach compares the model outcome between an original image, and the image with a region masked (e.g., with a black rectangle). If ablation of a region results in a change in model output, then the regions are considered to be important for correct image labelling. B) Activation-based approach identifies the regions of the input image which result in the highest neuron activation for producing a specified model label. C) Gradient-based approach back-propagates the final label onto the input image to identify important image regions for each label. D) Mixed-based approach combines activation and gradient-based approaches to improve the resolution of the activation-based region importance by weighing it with the gradients calculated as in the gradient-based approach. E) Attention-based approach learns important image regions during model training and uses this attention map to improve the final model prediction.

Interpretation by feature scoring, also known as saliency, relevance, or feature attribution, is the most common XAI strategy in outcome prediction. Interpretation by feature scoring finds evidence supporting individual predictions by calculating importance scores associated with each feature of the input. Specifically, given an input, we need to find a vector of importance scores that is the same size as the input. In general, feature scoring can be grouped into five categories: perturbation-based, activation-based, gradient-based, mixed-based (combination of activation-based and gradient-based), and attention-based approaches, as shown in Fig. 6.

1) Perturbation-Based Approach: Perturbation is the simplest way to analyze the effect of changing the input features on the output of an AI model. This can be implemented by removing, masking, or modifying certain input features, running the forward pass, and then measuring the difference from the original output. The input features affecting the output the most are ranked as the most important features.

Permutation- or occlusion-based methods measure the importance of a feature by calculating the increase in the model's prediction error after permuting the feature. A feature occlusion study was performed by [26] to show the influence of occluding regions of the input image to the confidence score predicted by the CNN model. The occlusion map was computed by replacing a small area of the image with a pure white patch and generating a prediction on the occluded image. While systematically sliding the white patch across the whole image, the prediction score on the occluded image was recorded as an individual pixel of the corresponding occlusion map. In biomedical application, Tang et al. [27] utilized occlusion mapping to demonstrate that networks learn patterns agreeing with accepted pathological features in Alzheimer's disease. Similarly, Hamon et al. [28] also implemented the occlusion map to indicate decision relevant regions in the chest X-ray image from a pneumonia use case scenario. In multi-modal integration studies [29]–[31], feature permutation and occlusion techniques also played a crucial role in determining the relative importance of different biomedical data modalities for the final prediction.

In the COVID-19 imaging applications, Gomes et al. [32] presented an interpretable method for extracting semantic features (e.g., 'Consolidation', 'Alveolar', 'Effusion', etc.) from X-ray images that correlate to severity from a dataset with patient ICU admission labels. A decision tree was implemented to analyze extracted features and provide interpretable results. Researchers mitigate the effects of overfitting through pruning mechanisms, which could limit the number of samples in tree leaves to reduce model complexity and dimensionality. The interpretable findings indicated that patients with features 'Bi-lateral' and 'Peripheral' showed a higher chance to be admitted into the ICU. Casiraghi et al. [33] calculated COVID-19 patient risk for significant complications from radiographic features extracted using deep learning and non-imaging features. A novel feature selection method was performed by combining the Boruta algorithm [34] and a permutation-based feature selector embedded in Random Forests via a five-fold cross-validation strategy. The most important features were then used to train a final random forest model to predict risk. In order to maximize final model interpretability, they generated a sequence of steps to generate an association decision tree from the final random forest model. The final association tree is easily interpretable by experts and could be used in emergency departments to provide rapid and accurate risk prediction for COVID-19 patients.

Another perturbation-based approach is Shapley value sampling [35], which estimates input feature importance via sampling and re-running the model. Calculating these Shapley feature importance values is computationally expensive as the network has to be run for each sample and feature (sample $\times$ number of features) times. Lunberg et al. [36] proposed a fast implementation for tree-based models named SHapley Additive exPlanation (SHAP) to boost the calculation process. By formulating the data features as players in a coalition game, Shapley values can be computed to learn to distribute the payout fairly. SHAP has demonstrated its efficacy in the medical domain to explain clinical decision-making both from image [37] and non-image [38] inputs and has also been well explored for COVID-19 cases [39]–[43].

Similarly, Local Interpretable Model-Agnostic Explanations (LIME) [44] is a procedure that enables an understanding of how the input features of a deep learning model affect its predictions. For instance, LIME determines the set of super-pixels (a patch of pixels) that have the most grounded relationship with a prediction label when used for image classification. LIME performs clarifications by creating a new dataset of random perturbations (each with its own forecast) around the occasion and then fitting

a weighted neighborhood proxy model. Typically, this neighborhood model is a simpler one with natural interpretability, such as a linear regression model. LIME generates perturbations by turning on and off a subset of the super-pixels in the image. To derive a representation that is understandable by humans, LIME tries to find the importance of contiguous superpixels in a source image towards the output class. It has been widely implemented in COVID-19 diagnosis tasks [41], [45]–[48] to further explain the process of feature extraction, which contributes to a better understanding of what features in CT/X-ray images characterize the onset of COVID-19. Ahsan et al. [45] implemented LIME to interpret top features in COVID-19 X-ray imaging and build trust in an AI framework to distinguish between patients with COVID-19 symptoms with other patients. Similarly, Ong et al. [41] implemented both SHAP and LIME to expound and interpret how Squeezenet performs COVID-19 classification and highlight the area of interest where they can help to increase the transparency and the interpretability of the deep model.

2) Activation-Based Approach: Interpreting layer-wise feature importance of a CNN is simpler in the first layer which generally learns the high-level textures and edges. However, as we move deeper into the CNN, the importance of specific layers towards a particular prediction is hard to summarize and visualize since parameters of subsequent layers are influenced by that of the previous layers. Hence, preliminary research tried to understand the neuronal activations to input instances as well as individual filters of specific layers.

Activation-based approaches identify important regions in a forward pass by obtaining or approximating the activations of intermediate variables in a deep learning model. Because extracted features within deep layers are closer to the classification layer, they capture more class-discriminative information than those in bottom layers. Erhan et al. [49] focused on input patterns which maximize a given hidden unit activation called Activation Maximization to express feature importance of deep learning models. Zhou et al. [50] proposed Class Activation Maps (CAM), which used global average pooling to calculate the spatial average of feature maps in the last convolutional layer of a CNN. Han et al. [51] proposed an attention-based deep 3D multiple instance learning (AD3D-MIL) to semantically generate deep 3D instance following the potential infection regions. Additionally, AD3D-MIL used an attention-based pooling to gain insight into each instance's contribution over a broader spectrum, allowing for more in-depth analysis. In comparison to conventional CAM, AD3D-MIL was capable of precisely detecting COVID-19 infection regions via key instances in 3D models. It achieved an accurate and interpretable COVID-19 screening that has the potential to be generalized to large-scale screening in clinical practice.

3) Gradient-Based Approach: Gradient-based approaches identify important features by evaluating gradients of an input through back-propagation. The intuition behind this idea is that input features with large gradients have the largest effects on predictions. Simonyan et al. [52] constructed the importance map of input features by calculating the absolute value of partial derivatives of class score with respect to the input through back-propagation. However, feature importance calculated above could be noisy because of the saturation problems caused by the existence of non-linear operations such as rectified linear units (ReLU). That is, changes in gradients could be removed in a backward pass if the input to ReLU are negative. To address this issue, several modifications to the way ReLU is handled have been proposed. Springenberg et al. [53] proposed guided back-propagation by combining standard back-propagation with the "deconvnet" approach: gradients are retained only when both the bottom input and top gradients are positive. The 'deconvnet' function inverts the data flow of a CNN given a high-level feature map, going from neuron activations in the given layer to an image. Thus, guided back-propagation can sharpen feature importance scores when compared to back-propagation using vanilla gradients.

Layer-wise Relevance Propagation (LRP) proposed by Bach et al. [54] is also used to find relevance scores for individual features in the input data by decomposing the output predictions of the DL models. The relevance score for each input feature is calculated by back-propagating the class scores of an output class node towards the input layer. The propagation follows a strict conservation property whereby a equal redistribution of relevance received by a neuron must be enforced. In COVID-19 X-ray imaging, LRP was implemented in DL models to provide explanations of diagnosis predictions and identify the critical regions on patients chest [55], [56].

Saliency map generation in deep neural networks were first introduced by Simonyan et al. [52] as a way of computing the gradient of the output class category with respect to an input image. By visualizing the gradients, a fair summary of pixel importance can be achieved by studying which positive gradients had more of an influence on the output. Shamout et al. [57] proposed a a data-driven approach for automatic prediction of deterioration risk using a deep neural network that learns from chest X-ray images and a gradient boosting model that learns from routine clinical variables. To illustrate the interpretability of proposed model, they performed the saliency maps for all time windows (24, 48, 72, and 96 h) to highlight regions that contain visual patterns such as airspace opacities and consolidation, which are correlated with clinical deterioration. These saliency maps could be used to guide the extraction of six regions of interest patches from the entire image, each of which is then assigned a score indicating its relevance to the prediction task. Similarly, [58]–[62] also include saliency maps as an explainable deliverable to interpret deep models and find potential infection regions in COVID-19 diagnosis and detection.

4) Mixed-Based Approach: Both activation-based and gradient-based methods have their own set of benefits and drawbacks. Specifically, activation-based methods generate feature scores that are more class discriminative, but they suffer from the coarse resolution of importance scores. On the other hand, although gradient-based methods produce fine resolution of feature scores, they tend not to show ability to differentiate between classes. Gradient-based and activation-based approaches could be combined to produce both fine and discriminative features importance scores.

Gradient-weighted Class Activation Mapping (Grad-CAM) [23] proposed by Selvaraju et al. uses the gradients

flowing down to the last convolutional layer to multiply CAM from a forward pass. The resolution is enhanced by multiplying Grad-CAM with guided-backpropagated gradients. It allows class-specific query of an input image as well as counterfactual explanations which highlights regions in the image which negatively contribute to a particular model output. Grad-CAM++ [63] replaces the globally averaged gradients in Grad-CAM with a weighted average of the pixel-wise gradients since the weights of pixels contribute to the final prediction, which leads to better visual explanations of CNN model predictions. It addresses the shortcomings of Grad-CAM, especially multiple occurrences of a class in an image and poor object localization.

Due to the vanishing non-linearity of classifiers, CAM is often unsuitable for interpreting deep learning models in COVID-19 image classification tasks. Grad-CAM and Grad-CAM++ improved the CAM operation for deeper CNNs and better visualizations and are usually considered the most popular interpretation strategy in COVID-19 automatic diagnosis on radiographic imaging [64]–[70]. Additionally, Oh et al. [71] proposed patch-wise deep learning architecture to investigate potential biomarkers in X-ray images and find the globally distributed localized intensity variation, which can be a discriminatory feature for COVID-19. They extended the idea of Grad-CAM to a novel probabilistic Grad-CAM that took patch-wise disease probability into account, resulting in more precise interpretable saliency maps that are strongly correlated with radiological findings.

5) Attention-Based Approach: Attention mechanism is a critical component of human perception, as it enables humans to selectively focus on critical portions of an image rather than processing the entire scene. Simulating the human visual system's selective attention mechanism is also critical for comprehending the mechanisms underlying black-box neural networks. Attention mechanism has been widely applied to computer vision applications [24], endowing the model with several new characteristics: 1) determine which portion of the inputs to focus on; 2) allocate limited computing resources to more critical components.

The efficacy of attention mechanism has been demonstrated in a variety of medical image analysis tasks. Specifically, several state-of-the-art methods have been proposed to leverage attention mechanisms in order to improve the discriminative capability of classification models for both X-ray and CT image analysis tasks [72]–[77]. In COVID-19 diagnosis, Shi et al. [76] proposed an explainable attention-transfer classification model based on a knowledge distillation network structure to address the difficulties associated with automatically differentiating COVID-19 and community-acquired pneumonia from healthy lungs in radiographic imaging. Extensive experiments on public radiographic datasets demonstrated the explainability of the proposed attention module in diagnosing COVID-19. Similarly, Zhang et al. [77] developed an end-to-end multiple-input deep convolutional attention network (MIDCAN) by leveraging the effectiveness of the convolutional block attention module [78] to generate model explanation as well as improve model performance.

In addition to medical imaging, the attention mechanism is also useful in other feature interpretation setting, such as unstructured clinical notes with natural language processing (NLP) [79]–[81]. Diagnostic coding of clinical notes is a task that aims to provide patients with a coded summary of their disease-related information. Recently, Dong et al. [80] proposed a novel Hierarchical Label-wise Attention Network (HLAN) to automate a medical coding process and to interpret model prediction results by evaluating the attention weights at word and sentence level. The label-wise attention scores in the proposed HLAN model provide comprehensive and robust explanation to support the prediction. Zhang et al. [81] proposed Patient2Vec to learn interpretable deep representations and predict risk of hospitalization on EHR data. The backbones of the model are gated recurrent units (GRU) and a hierarchical attention mechanism that learn and interpret the importance of clinical events on individual patients.

However, the attention mechanism continues to struggle when confronted with missing codes, rare labels, or clinical notes containing subtle errors. Additionally, clinical notes in real-world clinical practice frequently contain multiple sentences, and it is unknown how well the attention mechanism would function when interpreting multiple sentences. Additionally, external domain knowledge in the medical field is required to verify interpretation results. In general, the attention mechanism has enormous potential for emphasizing critical features and fostering trust in clinical practice.

C. Unsupervised Clustering

Development of an AI-based diagnosis system for COVID-19 was different from traditional epidemiological challenges: in the early stage of a new disease there is limited amounts of available data, especially diagnostic information [82]. The major downside of traditional deep learning methods is that they largely rely on the availability of labeled data, while COVID-19 datasets often contain incomplete or inaccurate labels. In biomedical applications, unsupervised learning has the benefit of not needing labeled data to train, extract features, and cluster data, which makes it a great candidate for COVID-19 diagnosis (see Fig. 7).

The application of unsupervised learning approaches, especially clustering techniques, represents a powerful means of data exploration. Discovering underlying data characteristics, grouping similar measurements together, and identifying patterns of interest are some of the applications which can be tackled through clustering. Being unsupervised, clustering does not always provide clear and precise insight into the produced output, especially when the input data structure and distribution are complex and unlabeled. Applying XAI can allow researchers to understand the reasons leading to a particular decision under clinical scenarios and suggest an explanation to the clustering results for the end-users.

Recent advances in Auto-Encoders (AEs) have shown their ability to learn strong feature representations for image clustering [83]–[85]. By designing the constraint of the distance between data and cluster centers well, Song et al. [83] artificially

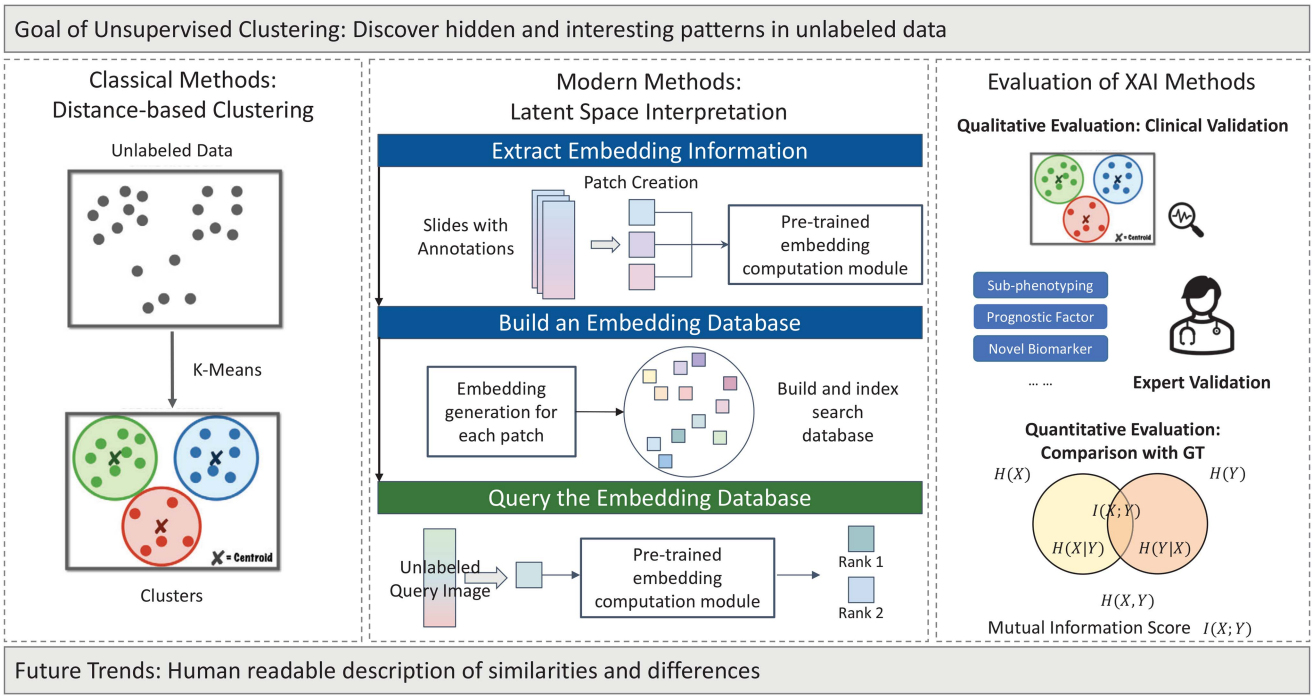

Fig. 7. Examples of classical and modern XAI approaches in unsupervised clustering task. Unsupervised clustering has benefited from the use of latent spaces generated by deep learning models to generate sample similarities. This shift from the conventional approach of calculating input feature distances enables the use of custom transformations to optimize the space in which similarity is measured. This can result in improved sample disentanglement.

re-aligned each point in the latent space of an AE to its nearest class neighbors during training to obtain a stable and compact representation suitable for clustering. Lim et al. [84] generalize Song's approach by introducing a Bayesian Gaussian mixture model for clustering in the latent space and replacing the input points with probability distributions which can better capture more hidden variables and hyperparameters. Prasad et al. [85] introduced a Gaussian Mixture prior to help clustering based on Variational Auto-Encoders to efficiently learn data distribution and discriminate between different clusters in a latent space.

In addition to guided feature representation achieved by AEs, King et al. [86] applied chest X-ray images of COVID-19 patients to a Self-Organizing Feature Map (SOFM) and found a distinct classification between COVID-19 and healthy patients. SOFM was first proposed to provide data visualization to cluster unlabeled X-ray images as well as reducing the dimensions of data to a map to understand high dimensional data. SOFM applied competitive learning to selectively tune the output neurons to the classes of the input patterns and then cluster their weights in locations respective to each other based off the feature similarities. They demonstrate that image clustering methods, specifically with SOFM networks, can cluster COVID-19 chest X-ray images and extract their features successfully to generate explainable results.

Yadav [87] proposed a deep unsupervised framework called Lung-GANs to learn interpretable representations of lung disease images using only unlabeled data and classify COVID-19 from chest CT and X-ray images. They extracted the lung features learned by the model to train a support vector machine (SVM) and a stacking classifier and demonstrated the performance of proposed unsupervised models in lung disease classification. They visualized the features learned by Lung-GANs to interpret deep models and empirically evaluate its effectiveness in classifying lung diseases.

Singh et al. [88] used image embedding generated from a prototypical part network (ProtoPNet) inspired network to calculate similarities and differences of X-ray image patches to known examples of pathology and healthy patches. This metric was then used to classify subjects into COVID-19 positive, pneumonia, or healthy classes.

The task of image clustering in COVID-19 and other clinical scenarios naturally requires good feature representation to capture the distribution of the data and subsequently differentiate one category from one another. In general, unsupervised clustering is an XAI technique which can be implemented to validate that images cluster in meaningful groups and facilitate expert annotation by extrapolating labels within samples belonging to the same cluster, when labels need be estimated.

D. Image Segmentation

Segmentation algorithms make pixel-level classifications of images and the overall segmentation produced provide insight into the decisions of the model. In the realm of XAI, image segmentation itself can be considered highly interpretable. Therefore, explanations of the segmentation process are currently not widely explored for medical image analysis. In the current climate, segmentation algorithms function as useful tools

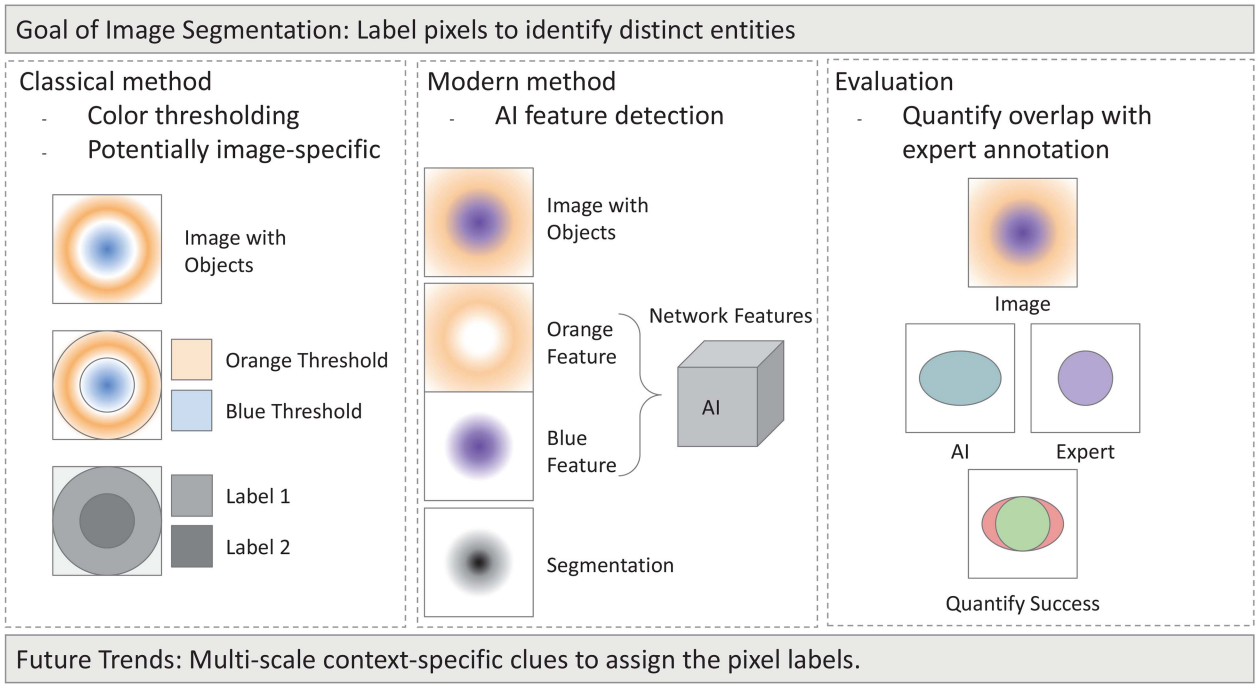

Fig. 8. Examples of classical and modern XAI approaches in image segmentation task. Segmentation models have progressed from being highly interpretable (when simple color thresholds are used) to requiring numerous nonlinear transformations to generate the final segmentation. Although XAI approaches to image segmentation are not widely used, recent techniques have used the model activation maps generated by deep layers to identify significant associations with the final segmentation.

for isolating regions significant to COVID-19 diagnosis or for determining infection severity. Application of explainable AI techniques to segmentation techniques could provide valuable information to improve COVID-19 segmentation approaches, as shown in Fig. 8.

Current COVID-19 segmentation approaches often use convolutional neural networks to delineate the regions of interest. One example of this model was developed by Saeedizadeh et al. for segmenting CT images of COVID-19 patients based on U-Net, which they call TV-UNet [89]. The framework was trained to detect ground glass regions on the pixel level, which are indicative of infected regions, and to segment them from normal tissue. TV-UNet differs from regular U-Net by the addition of an explicit regularization term in the training loss function which the authors report improves connectivity for predicted segmentations. Their model was trained on a COVID CT segmentation dataset with three different types of ground truth masks and reported an average DICE coefficient score of 0.864 and an average precision of 0.94. However, the results of the segmentation algorithm do not provide any intuition on why the model made the decisions it did. Part of this is due to the black box nature of U-Net. The residual connections between layers are inherently obscure to human intuition which makes it difficult to understand how U-Net decided to apply the labels. Application of a technique that explains the model's decision-making process could provide information on possible biases in the model and ways to improve it. Pennisi et al. [90] achieved sensitivity and specificity of COVID-19 lesion categorization of over 90% using a combination of lung lobe segmentation followed by lesion classification. In addition, they also created a clinician-facing user interface to visualize model prediction. This expert oversight was leveraged to improve future prediction by integrating clinician feedback through the same user interface (expert in the loop). Wang et al. [91] proposed an interpretable DeepSC-COVID designed with 3 subnets: a cross-task feature subnet for feature extraction, a 3D lesion subnet for lesion segmentation, and a classification subnet for disease diagnosis. Different from the single-scale self-attention constrained mechanism [24], they implemented multi-scale attention constraint to generate more fine-grained visualization maps for potential infections.

Image morphology-based segmentation approaches are not as common within the context of COVID-19 image segmentation, but they do exist. An example from [92] demonstrates the successful use of an maximum entropy threshold segmentation-based method along with fundamental image processing techniques, such as erosion and dilation, to isolate a final lung-only binary mask. These lung masks can also be used to generate bounding boxes to limit classification to regions surrounding, and including, lung tissue [93]. In addition to lesion segmentation, some approaches first segment lung tissue prior to classification or further segmentation of clinically-relevant lesions. Jadhav et al. utilized this approach to allow radiologists to use a user-interface to view the two and three-dimensional CT regions used for the classification task with a saliency map overlay [94]. This combination of XAI approaches sought to

increase radiologist trust of classification predictions by gaining multiple visual insights of the automated workflow.

Natekar et al. described one such method of explaining segmentation algorithms known as network dissection [95]. Their focus was on explaining segmentations done on MR images of brain tumors with U-Net, but the techniques could be applicable to COVID-19 segmentation. They explain network dissection as follows: for a single filter in a single layer, collect the activation maps of all input images and determine the pixel-level distribution over the entire dataset. In CNNs, individual filters can focus on learning specific areas or features in an image, however, it is not clear from the outside which filter does which. Dissecting the network would make the purpose of each filter clearer and allow for better understanding of the decisions made by the model. Application to COVID-19 algorithms such as TV-UNet could allow for visualization of specific features that the model looks for to make a segmentation decision, thereby increasing user confidence in the model.

Another COVID-19 segmentation approach is the joint classification and segmentation diagnosis system developed by Wu et al. [58]. In their framework, they include an explainable classification model and segmentation model that work together to provide diagnosis prediction for COVID-19. Their segmentation is done via an encoder-decoder architecture based on VGG-16, plus the addition of an Enhanced Feature Module to the encoder which the authors proposed to improve the extracted feature maps. They trained and tested their model on a private COVID dataset and reported a DICE coefficient score of 0.783. Typically, image segmentation tasks are used to help explain classification decisions but the authors of this paper extend this idea by having the classification also help explain the segmentation. The segmentation algorithm references information from the classifier by merging their feature maps together to improve its decisions but this also helps indicate the reasoning behind the decisions made when producing segmentation. Utilizing classification information to help train and explain segmentation is an avenue which merits further exploration.

III. Evaluation of Explainable AI Methods

Qualitative visualization plays an important role in evaluating XAI methods. For biomedical applications, qualitative evaluation focuses on whether visualization can align with established knowledge. For instance, expert radiologists can assess how well the generated attention map identifies image regions of high diagnostic relevance [96]. Based on the previous work [97], a guideline for evaluating XAI methods from both model behavior and human understanding perspectives is proposed and illustrated in Fig. 9.

Although qualitative evaluation is important, quantitative evaluation of interpretation is still desirable, which can be obtained through either user study or automatic approaches. When conducting user studies, target users (e.g., physicians for medical applications) perform specific tasks with and without the assistance of visual interpretation in order to quantify the efficacy of model explanations. For example, clinicians will be asked to differentiate between cases involving original images and those involving images with visual interpretation. Then, improvement in performance is measured with the assistance of visual interpretation. User studies could be considered as one of the most reliable approaches for evaluating interpretability, if they are designed to resemble real application scenarios. However, conducting such user studies are expensive and time-consuming, especially for biomedical applications.

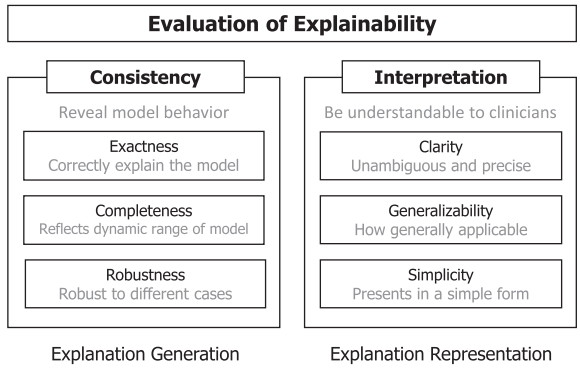

Fig. 9. Evaluation of model explainability. To assess the generation of explanations and the revealing of model behaviors, it is critical to consider their correctness, completeness, and robustness. When evaluating explanation representations, it is critical to consider their clarity of presentation, their generalizability, and their simplicity of form.

When evaluating the generation of explanations, an alternative approach is to use automatic evaluation, which acts as a proxy for user research without involving real users. Zeilar and Fergus first introduced the idea of the occlusion experiment [26], in which portions of input images were systematically occluded by a grey square for monitoring the performance of deep learning models. Samek et al. [98] further formalized the occlusion experiments by introducing a procedure called "pixel flipping", which destroys data points ordered by their feature importance scores and compares the decrease in classification metrics among multiple interpretation methods. A larger decrease in the metrics suggests a better interpretation method. Because occlusion experiments are model agnostic, they can be used as an objective measure for interpretation methods. On the other hand, the occlusion experiments can not serve as objective evaluation for perturbation-based feature scoring methods, such as Randomized Input Sampling for Explanation (RISE) [99], that perturb input directly to identify important features.

Apart from occlusion studies, metrics such as R-squared for the global surrogate model could also be used to evaluate the model interpretation. A global surrogate model is an interpretable model (e.g., linear models, decision tree, etc.) that is trained to approximate the predictions of a black-box model [100]. We can draw conclusions about the black-box model by examining how well the surrogate model can mimic the behavior of the original black-box model. The R-squared measure is one way to determine how well the surrogate model replicates the black-box model, which can be interpreted as the percentage of variance captured by the surrogate model. If it is close to 1, the interpretable model closely approximates the black-box model's behavior, which indicates

TABLE I
SUMMARY OF XAI METHODS IN CLINICAL APPLICATIONS

Strategy	Category	Technique
Data Augmentation	Latent Space Interpretation	Intrinsic latent space guidance [15]–[18], [20], [21]
		Post-hoc PCA-based [19]
Outcome Prediction	Perturbation-based	Feature occlusion and ablation [26]–[28], [32], [33]
		SHAP feature importance [35]–[41]
		Local interpretable model-agnostic explanations (LIME) [41], [44]–[48]
	Activation-based	Activation maximization [49]
		Class activation maps (CAM) [50], [51]
	Gradient-based	Gradient-based class score [52]
		Deconvnet [53]
		Layer-wise relevance propagation (LRP) [54]–[56]
		Gradient-based saliency analysis [52], [57]–[62]
	Mixed-based	Grad-CAM and Grad-CAM++ [23], [63]–[71]
	Attention-based	Self-attention mechanism [24], [72]–[78]
		Hierarchical attention mechanism [80], [81]
Unsupervised Clustering	Guided Embedding	VAE-feature clustering [83]–[85]
	Feature Extraction	Self-organizing feature map (SOFM) clustering [86]
		Similarity calculation [88]
		Latent space interpretation [87]
Image Segmentation	Morphology-based	Maximum entropy threshold [92]–[94]
	Context-based	Mult-scale attention [91]
		Saliency Analysis [58]
		Network dissection [95]

the feasibility of replacing the complex model with the interpretable model. If the R-squared is close to zero, the interpretable model does not adequately explain the black-box model.

Besides correctness and completeness, the evaluation for robustness of model interpretability remains challenging without human intervention. Lin et al. [101] proposed an adversarial attack to evaluate the robustness of interpretability in XAI methods by checking whether they can detect backdoor triggers present in the input. Researchers employed data poisoning to create trojaned models and generated saliency maps that will highlight the trigger to evaluates the saliency map output using three quantitative evaluation metrics (IoU, recovery rate, and recovering difference).

Quantitative evaluation of data synthesis is still in its infancy. DeVries et al. [102] designed an evaluation metric, named Fréchet Joint Distance (FJD), for the quality of images generated by conditional GAN based on visual quality, intra-conditioning diversity, and conditional consistency. Assuming the joint distribution of hidden space and labels are Gaussian, they used FJD to compare the mean and variance between real and generated images. Recently, Yang et al. [103] created a ground-truth dataset consisting of mosaic natural images for interpretation methods and tried to unify the evaluation of both feature scoring and data synthesis methods. Their aforementioned methods are early in their developmental stage, even for natural images, and ways to adopt them into biomedical images and other biomedical data modalities remains an ongoing challenge.

IV. DISCUSSION

Upon review of the existing works leveraging XAI to facilitate the interpretation of AI-based COVID-19 solutions to clinical challenges, we have identified key features present in papers which have made substantial impacts in the field. Table I summarizes the XAI techniques used in COVID-19 related clinical applications covered in this work. Furthermore, inspired by [104], we summarize these findings and references to example implementations in a checklist of important considerations during the process of AI-based experimental design, as shown in Table II.

A. Checklist for AI-Enabled Clinical Applications

Using the framework values of performance, user trust, and user response, we noticed the need for incorporating clinical insights throughout the study design process. This includes understanding the factors influencing response variables in the real world, as illustrated in Haimovich et al. [105] when they stated that ICU admission was not an ideal outcome variable due to site-specific and time-dependent patient admission requirements. Clinical input may also be obtained during and after model optimization via real-time expert feedback [90] and during implementation via expert-facing user interfaces [42]. In addition to web-based applications, visualizing sample clusters [67], [87] and feature importance metrics [33], [71], [80] can offer users without expertise in data analysis an option of understanding the decision-making process of otherwise obscure models.

A very common approach to generating easily-interpretable models is to optimize a decision tree approach to define a clear decision-making process using available features [4], [32], [33], [106]. This approach is also similar to widely-used clinical guidelines to generate fast and consistent metrics for patient triage and management [107], [108].

Validating feature importance ranking by using multiple methods, such as tree-based importance metrics and Shapley values, can establish features lists which are consistent between

TABLE II
CHECKLIST FOR AI-ENABLED CLINICAL DECISION SUPPORT SYSTEMS

Suggestions
Incorporate clinical insights [105]
Interactive user interface [42], [90], [94], [105]
Clinical feedback [90], [90]
Visualization of feature importance [33], [71], [80]
Clustering analysis [67], [87]
Decision tree [4], [32], [33]
Use multiple feature importance approaches [13], [41], [74]
Cross validation when comparing models [46], [105], [106], [109]
Use appropriate and robust performance metrics (AUROC, MCC) [109], [110]
Adversarial example testing [20], [21]

Problems	Solutions
Small sample size	Data imputation [32], [59], [113]
Bad data quality	Artifact correction [64], [114]
Imbalanced classes	Data augmentation [62], [65]
Complex disease phenotype	Multi-modality data [57], [115]
Data heterogeneity	Data normalization [59]
Lack of expert annotation	Weakly supervised learning [51], [71], [87]
Unkown sources of signal	Key feature extraction [37], [58], [69]
Explanations unclear	Pre-processing changes [67], [68]
Data leakage	Patient-level split [45], [60], [66], [91]
Training is inefficient	Transfer learning [56], [62], [92]

approaches to prevent spurious rankings [105]. This may be especially important if the list will be used for feature selection or simplified feature visualizations, such as displaying only the odds ratios for the most important features.

Comparison of multiple competing models is often necessary to generate high-performance solutions. We noticed the widespread use of cross validation when authors sought to conduct these comparisons [46], [105], [106], [109]. Cross validation is easier to implement when models are quickly trained and tested, but this approach may also be used with more complex models to ensure robust comparisons.

In the quest for easily interpretable results, it is common to see accuracy being reported as a model performance metric. Although accuracy is understood by model developers and end-users alike, it should be avoided when significant data imbalance is present. Examples of works using appropriate performance metrics include [109], [110]. More robust metrics include Area Under the Receiver Operating Curve (AUROC) and Matthews Correlation Coefficient (MCC), the latter being appropriate even in highly imbalanced binary classification tasks [111].

An often overlooked aspect of model development is the potential for adversarial attack within the final implementation context. Cyber attacks on hospital systems are depressingly common with a notable rise in frequency over time [112]. As research tools make their way into the hospital it may become important to understand the vulnerability of models to potential future attacks. Therefore, we included this component to our checklist alongside a recent illustration of adversarial testing approaches [20].

B. Challenges and Solutions

With any clinical informatics work there will be challenges. Often these will arise due to issues with the dataset being used, especially if it was derived from real-world data. After our review of the literature, we summarized common challenges and potential solutions, including example works which successfully solve the problem (see Table II).

Early in the pandemic, there was a scarcity of reliable data available to the general scientific community. This resulted in a significant need for data imputation in order to fill in missing values to maximize the utility of existing data [59]. Poor data quality also affected model performance and artifact correction techniques were implemented [64]. Imbalanced classes were frequently found within COVID-19 datasets due to the accessibility of normal samples relative to COVID-19 positive cases. Data augmentation was found to alleviate this problem in some cases by generating additional samples of the underrepresented class [65].

Lack of expert annotation of key regions of pathology in imaging data created the need for weakly supervised learning models capable of generalizing small ground truth datasets [51], [71]. Without expert insight, it was often necessary to identify features capable of differentiating between similar phenotypes (e.g., bacterial versus viral pneumonia). This problem was frequently solved via key feature extraction [51], [58], [60], [61], [64], [66], [69], [71]. In the case of complex disease phenotypes, multi-modality data were integrated to leverage data obtained from consistent or complementary sources [57]. Ensuring model generalizability requires robust external validation. Data leakage occurs when testing/external dataset information is used during the model design or optimization process. The likelihood of this occurring can be reduced by isolating the test dataset during hyperparameter selection and model training. Special care must be taken to avoid including data derived from the same patient in both the test and training datasets. There exists significant within-patient correlation of features, even across samples. This data leakage may allow models to learn patient-specific patterns which are not generalizable to other patients, resulting in poor performance in the real-world [45], [60], [64], [66].

In addition, XAI may lead to unclear results, either due to inconsistent feature importance ranking or nonspecific image highlighting. In these cases, it is often a good idea to re-establish the quality of the preprocessing pipeline [67], [68]. When training is inefficient, transfer learning may be used to take advantage of prior parameter optimization on similar problems [56], [70].

Ultimately, we designed this checklist to help both academic researchers in general, and clinical data scientists specifically. We summarize the integration of XAI in both settings, along with its benefits in Fig. 10.

C. Evolution of XAI Methods

XAI techniques have developed quickly in recent years to meet the evolving needs of AI researchers and the end-users of their models. Although it is easy to fall into the trap of believing that more recent models are objectively better than their more classic counterparts, it is important to understand that each model was designed to improve our understanding

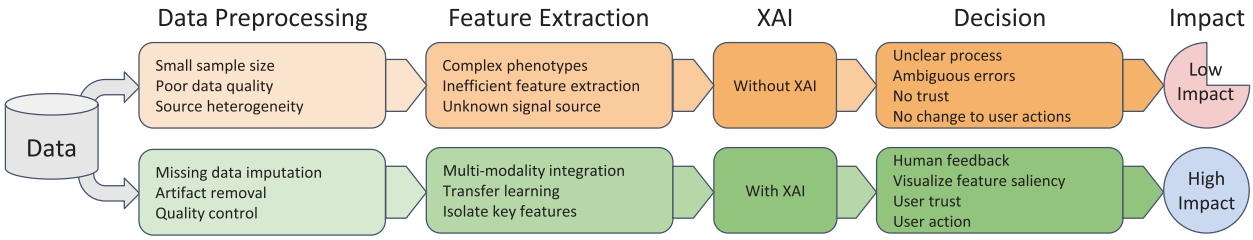

Fig. 10. Summary of insights gained for designing an AI development workflow. We provide a checklist of considerations to make early in the experimental design process in order to avoid common problems. Additionally, we provide a list of common issues encountered when working with clinical data and discuss several common solutions that may assist the reader when working with these data.

of different facets of AI-solutions. For example, in the task of data augmentation, it is common to see K-Nearest Neighbors interpolation and other classic approaches used instead of more complex modern solutions. This is in part because the more classic approach has been around for longer, and its pitfalls have been well established. The modern approach may result in data bias which may be difficult to understand due to the lack of real-world examples of their successes and failures. The trend for data augmentation has been to increase the number of considered factors and complexity of data transformations to better model the underlying data distribution of real-world samples.

Clinical decision support is a very common setting to find AI-solutions in need of explanation. For the task of disease diagnosis, the trend has been to generate input importance visualizations which can be used across a wide range of common deep learning models. This is in contrast to early XAI approaches which relied on model-specific solutions to improve interpretability. XAI in risk prediction for clinical decision support has trended towards generating sample-specific explanations. These may provide the end user with a custom answer to the question of "why did this sample get the score that it did?". This is especially useful in the clinical setting where precision medicine is becoming the standard, and patient-specific explanations for risk scores are vital.

Additionally, depending on the problem scope, XAI methods can be classified into global methods that provide a unified global explanation for the overall model behavior and local methods that provide explanations for each individual instance [100]. In local interpretation methods, XAI methods attempt to accurately describe individual sample predictions as the sum of feature effects; for example, LIME explains individual predictions by replacing a locally interpretable surrogate model for the complex model; Shapley values attempt to fairly assign the prediction to individual features. In contrast to local interpretation methods, global methods such as SHAP feature importance, coefficient of regression models, and permutation-based feature importance are frequently expressed as expected values based on the distribution of the data in order to investigate the knowledge encoded in the model and its effect on predictions [116]. Depending on the scope of the problem, clinicians may consider different levels of interpretability. Local methods show the explanation for specific instances, whereas global methods can generalize over the entire cohort. Global interpretable features, for example, were derived from global interpretation methods to generate a risk score for in-hospital mortality [117], and local explanations were used to investigate COVID-19 progression prediction for individual patients [106].

Unsupervised clustering has benefited from the use of deep learning model latent spaces for their generation of sample similarities. This shift from the classical input feature distance approaches allows custom transformations to optimize the space within which similarity is measured. This can result in better disentanglement of samples [118].

Image segmentation approaches have increased in complexity in recent years due to models such as U-Net and its variants. Models have gone from highly interpretable (if using simple color thresholds) to involving many nonlinear transformations to produce the final segmentation. XAI approaches for image segmentation are still not commonly used, but recent techniques have leveraged the model activation maps produced by deep layers to identify significant associations with final deep learning model output [119]. XAI approaches will continue to adapt as models continue to become better optimized for different tasks. XAI will likely cover a much wider range of approaches to meet the needs of end-users and regulatory agencies.

In future work, with the decrease of COVID-19 incidence and increase of vaccine supply, risk stratification will become vital to determine optimal treatment plan. We also hope our focus of XAI within the ongoing COVID-19 pandemic may increase the relevance of our insights to future disease outbreaks. The framework we provided can be used across common AI-tasks and may improve the clinical implementation of these solutions, especially in the early stages of infection.

V. Conclusion

The recent confluence of large-scale public healthcare datasets combined with the rapid increase of computing capacity has resulted in a noteworthy increase in AI-based solutions for clinical decision-making. However, making these AI solutions adopted in clinical practice is slow. In this work, we reviewed XAI approaches that can increase AI adoption based on lessons learned from COVID-19 and presented future trends with insights. Clinical informatics is generally risk-averse, which creates the need for AI developers in the field to understand how AI-based decisions are reached. This understanding would provide two key benefits: i) increasing confidence that a deep learning model is unbiased and relies on relevant features to accomplish

desired tasks and ii) detecting biases or discovering new knowledge if the generated explanations elucidate previously-hidden patterns in the data. Meanwhile, clinicians and healthcare practitioners will benefit from the model transparency and result interpretation enabled by XAI to understand the black-box decision-making. This could increase the trustworthiness and accountability of AI solutions and promote their adoption in the clinical workflow. Ultimately, the implementation of XAI techniques will accelerate the translation of data-driven analytic solutions to improve the quality of patient care.

Acknowledgment

The authors would like to thank the Emory Science Librarian, Ms. Kristan Majors, for her support and guidance on search optimization for the PRISMA chart. They like to thank Dr. Siva Bhavani from Emory University for his insights on leveraging artificial intelligence in clinical practice. They also like to thank Mr. Benoit Marteau from Bio-MIBLab for his help on reviewing the manuscript.

References

[1] B. Hu, H. Guo, P. Zhou, and Z.-L. Shi, "Characteristics of SARS-CoV-2 and COVID-19," *Nature Rev. Microbiol.*, vol. 19, no. 3, pp. 141–154, 2021.

[2] G. D. Rubin et al., "The role of chest imaging in patient management during the COVID-19 pandemic: A multinational consensus statement from the fleischner society," *Chest*, vol. 158, pp. 106–116, 2020.

[3] A. A. Soltan et al., "Rapid triage for COVID-19 using routine clinical data for patients attending hospital: Development and prospective validation of an artificial intelligence screening test," *Lancet Digit. Health*, vol. 3, no. 2, pp. e78–e87, 2021.

[4] L. Yan et al., "An interpretable mortality prediction model for COVID-19 patients," *Nature Mach. Intell.*, vol. 2, no. 5, pp. 283–288, 2020.

[5] K. Zhang et al., "Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography," *Cell*, vol. 181, no. 6, pp. 1423–1433, 2020.

[6] W. D. Heaven, "Hundreds of AI tools have been built to catch covid. none of them helped," *MIT Technol. Rev. Retrieved October*, vol. 6, 2021, Art. no. 2021.

[7] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Trans. Evol. Comput.*, vol. 23, no. 5, pp. 828–841, Oct. 2019.

[8] K. Eykholt et al., "Robust physical-world attacks on deep learning visual classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1625–1634.

[9] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "AI for radiographic COVID-19 detection selects shortcuts over signal," *Nature Mach. Intell.*, vol. 3, no. 7, pp. 610–619, 2021.

[10] D. Alvarez-Melis and T. S. Jaakkola, "Towards robust interpretability with self-explaining neural networks," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 7786–7795.

[11] S. Cruz Rivera et al., "Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension," *Lancet Digit Health*, vol. 2, no. 10, pp. e549–e560, Oct. 2020.

[12] X. Liu et al., "Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension," *Lancet Digit. Health*, vol. 2, no. 10, pp. e537–e548, Oct. 2020.

[13] T. Makino et al., "Differences between human and machine perception in medical diagnosis," *Sci. Rep.*, vol. 12, no. 1, pp. 1–13, 2022.

[14] D. Moher et al., "Preferred reporting items for systematic reviews and meta-analyses: The prisma statement," *PLoS Med.*, vol. 6, no. 7, 2009, Art. no. e1000097.

[15] R. K. Singh, R. Pandey, and R. N. Babu, "COVIDScreen: Explainable deep learning framework for differential diagnosis of COVID-19 using chest x-rays," *Neural Comput. Appl.*, vol. 33, no. 14, pp. 8871–8892, Jan. 2021.

[16] A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman, and P. R. Pinheiro, "CovidGAN: Data augmentation using auxiliary classifier GAN for improved COVID-19 detection," *IEEE Access*, vol. 8, pp. 91916–91923, 2020.

[17] M. Loey, G. Manogaran, and N. E. M. Khalifa, "A deep transfer learning model with classical data augmentation and CGAN to detect COVID-19 from chest CT radiography digital images," *Neural Comput. Appl.*, pp. 1–13, 2020. [Online]. Available: https://doi.org/10.1007/s00521-020-05437-x

[18] A. Voynov and A. Babenko, "Unsupervised discovery of interpretable directions in the gan latent space," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 9786–9796.

[19] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, "Ganspace: Discovering interpretable GAN controls," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 9841–9850, 2020.

[20] A. Rahman, M. S. Hossain, N. A. Alrajeh, and F. Alsolami, "Adversarial examples–security threats to COVID-19 deep learning systems in medical IoT devices," *IEEE Internet Things J.*, vol. 8, no. 12, pp. 9603–9610, Jun. 2021.

[21] I. Palatnik de Sousa, M. M. Vellasco, and E. Costa da Silva, "Explainable artificial intelligence for bias detection in COVID CT-scan classifiers," *Sensors*, vol. 21, no. 16, 2021, Art. no. 5657.

[22] E. Fix and J. L. Hodges, "Discriminatory analysis. nonparametric discrimination: Consistency properties," *Int. Stat. Review/Revue Internationale de Statistique*, vol. 57, no. 3, pp. 238–247, 1989.

[23] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[24] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[25] J. R. Quinlan, "Simplifying decision trees," *Int. J. man- Mach. Stud.*, vol. 27, no. 3, pp. 221–234, 1987.

[26] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.

[27] Z. Tang et al., "Interpretable classification of Alzheimer's disease pathologies with a convolutional neural network pipeline," *Nature Commun.*, vol. 10, no. 1, pp. 1–14, 2019.

[28] R. Hamon, H. Junklewitz, G. Malgieri, P. D. Hert, L. Beslay, and I. Sanchez, "Impossible explanations? beyond explainable AI in the GDPR from a COVID-19 use case scenario," in *Proc. ACM Conf. Fairness, Accountability, Transparency*, 2021, pp. 549–559.

[29] J. Venugopalan, L. Tong, H. R. Hassanzadeh, and M. D. Wang, "Multimodal deep learning models for early detection of Alzheimer's disease stage," *Sci. Rep.*, vol. 11, no. 1, pp. 1–13, 2021.

[30] H. R. Hassanzadeh and M. D. Wang, "An integrated deep network for cancer survival prediction using omics data," *Front. Big Data*, 2021, Art. no. 568352.

[31] L. Tong, J. Mitchel, K. Chatlin, and M. D. Wang, "Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis," *BMC Med. Informat. Decis. Mak.*, vol. 20, no. 1, pp. 1–12, 2020.

[32] D. P. Gomes et al., "Features of ICU admission in x-ray images of COVID-19 patients," in *Proc. IEEE Int. Conf. Image Process.*, 2021, pp. 200–204.

[33] E. Casiraghi et al., "Explainable machine learning for early assessment of COVID-19 risk prediction in emergency departments," *IEEE Access*, vol. 8, pp. 196299–196325, 2020.

[34] M. B. Kursa and W. R. Rudnicki, "Feature selection with the boruta package," *J. Stat. Softw.*, vol. 36, pp. 1–13, 2010.

[35] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowl. Inf. Syst.*, vol. 41, no. 3, pp. 647–665, 2014.

[36] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4765–4774.

[37] A. Singh, A. R. Mohammed, J. Zelek, and V. Lakshminarayanan, "Interpretation of deep learning using attributions: Application to ophthalmic diagnosis," *Proc. SPIE*, vol. 11511, 2020, Art. no. 115110A.

[38] S. M. Lundberg et al., "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," *Nature Biomed. Eng.*, vol. 2, no. 10, pp. 749–760, 2018.

[39] Y. Zoabi, S. Deri-Rozov, and N. Shomron, "Machine learning-based prediction of COVID-19 diagnosis based on symptoms," *npj Digit. Med.*, vol. 4, no. 1, pp. 1–5, 2021.

[40] H. Wu et al., "Interpretable machine learning for COVID-19: An empirical study on severity prediction task," *IEEE Trans. Artif. Intell.*, 2021, doi: 10.1109/TAI.2021.3092698.

[41] J. H. Ong, K. M. Goh, and L. L. Lim, "Comparative analysis of explainable artificial intelligence for COVID-19 diagnosis on CXR image," in *Proc. IEEE Int. Conf. Signal Image Process. Appl.*, 2021, pp. 185–190.

[42] B. V. Patel et al., "Natural history, trajectory, and management of mechanically ventilated COVID-19 patients in the United Kingdom," *Intensive Care Med.*, vol. 47, no. 5, pp. 549–565, 2021.

[43] J. Hinns, X. Fan, S. Liu, V. Raghava Reddy Kovvuri, M. O. Yalcin, and M. Roggenbach, "An initial study of machine learning underspecification using feature attribution explainable AI algorithms: A COVID-19 virus transmission case study," in *Proc. Pacific Rim Int. Conf. Artif. Intell.*, 2021, pp. 323–335.

[44] M. T. Ribeiro, S. Singh, and C. Guestrin, "'why should i trust you?' explaining the predictions of any classifierin," *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 1135–1144.

[45] M. M. Ahsan et al., "COVID-19 symptoms detection based on nasnetmobile with explainable AI using various imaging modalities," *Mach. Learn. Knowl. Extraction*, vol. 2, no. 4, pp. 490–504, 2020.

[46] M. M. Ahsan, R. Nazim, Z. Siddique, and P. Huebner, "Detection of COVID-19 patients from CT scan and chest x-ray data using modified mobilenetv2 and lime," in *Healthcare*, vol. 9, no. 9. Multidisciplinary Digital Publishing Institute, 2021, Art. no. 1099.

[47] Q. Ye, J. Xia, and G. Yang, "Explainable AI for COVID-19 CT classifiers: An initial comparison study," in *Proc. IEEE 34th Int. Symp. Comput.-Based Med. Syst.*, 2021, pp. 521–526, ISSN: 2372–9198.

[48] F. Gabbay, S. Bar-Lev, O. Montano, and N. Hadad, "A lime-based explainable machine learning model for predicting the severity level of COVID-19 diagnosed patients," *Appl. Sci.*, vol. 11, no. 21, 2021, Art. no. 10417.

[49] D. Erhan, A. Courville, and Y. Bengio, "Understanding representations learned in deep architectures," Department d'Informatique et Recherche Operationnelle, Univ. Montreal, QC, Canada, Tech. Rep, 1355, 2010.

[50] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.

[51] Z. Han et al., "Accurate screening of COVID-19 using attention-based deep 3D multiple instance learning," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2584–2594, Aug. 2020.

[52] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013, *arXiv:1312.6034*.

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

[54] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS One*, vol. 10, no. 7, 2015, Art. no. e0130140.

[55] M. R. Karim, T. Döhmen, M. Cochez, O. Beyan, D. Rebholz-Schuhmann, and S. Decker, "DeepCOVIDexplainer: Explainable COVID-19 diagnosis from chest x-ray images," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2020, pp. 1034–1037.

[56] P. R. Bassi and R. Attux, "A deep convolutional neural network for COVID-19 detection using chest x-rays," *Res. Biomed. Eng.*, vol. 38, no. 1, pp. 139–148, 2022.

[57] F. E. Shamout et al., "An artificial intelligence system for predicting the deterioration of COVID-19 patients in the emergency department," *NPJ Digit. Med.*, vol. 4, no. 1, pp. 1–11, 2021.

[58] Y. H. Wu et al., "JCS: An explainable COVID-19 diagnosis system by joint classification and segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 3113–3126, 2021.

[59] E. J. M. Barbosa et al., "Machine learning automatically detects COVID-19 using chest CTs in a large multicenter cohort," *Eur. Radiol.*, vol. 31, no. 11, pp. 8775–8785, 2021.

[60] X. Qian et al., "M ^{3} lung-sys: A deep learning system for multi-class lung pneumonia screening from CT imaging," *IEEE J. Biomed. health Informat.*, vol. 24, no. 12, pp. 3539–3550, Dec. 2020.

[61] G. Singh and K.-C. Yow, "An interpretable deep learning model for COVID-19 detection with chest x-ray images," *IEEE Access*, vol. 9, pp. 85198–85208, 2021.

[62] D. Singh, V. Kumar, M. Kaur, M. Y. Jabarulla, and H.-N. Lee, "Screening of COVID-19 suspected subjects using multi-crossover genetic algorithm based dense convolutional neural network," *IEEE Access*, vol. 9, pp. 142566–142580, 2021.

[63] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam : Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 839–847.

[64] L. Li et al., "Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: Evaluation of the diagnostic accuracy," *Radiology*, vol. 296, no. 2, pp. E65–E71, 2020.

[65] S.-H. Wang, Y. Zhang, X. Cheng, X. Zhang, and Y.-D. Zhang, "PSSPNN: Patchshuffle stochastic pooling neural network for an explainable diagnosis of COVID-19 with multiple-way data augmentation," *Comput. Math. Methods Med.*, vol. 2021, pp. 1–18, 2021.

[66] Y. Song et al., "Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 6, pp. 2775–2780, Nov./Dec. 2021.

[67] J. D. Arias-Londoño, J. A. Gomez-Garcia, L. Moro-Velázquez, and J. I. Godino-Llorente, "Artificial intelligence applied to chest x-ray images for the automatic detection of COVID-19. A thoughtful evaluation approach," *IEEE Access*, vol. 8, pp. 226811–226827, 2020.

[68] L. Brunese, F. Mercaldo, A. Reginelli, and A. Santone, "Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from x-rays," *Comput. Methods Programs Biomed.*, vol. 196, 2020, Art. no. 105608.

[69] H. Panwar, P. Gupta, M. K. Siddiqui, R. Morales-Menendez, P. Bhardwaj, and V. Singh, "A deep learning and GRAD-CAM based color visualization approach for fast detection of COVID-19 cases using chest x-ray and CT-scan images," *Chaos, Solitons Fractals*, vol. 140, 2020, Art. no. 110190.

[70] H. Alshazly, C. Linse, E. Barth, and T. Martinetz, "Explainable COVID-19 detection using chest CT scans and deep learning," *Sensors*, vol. 21, no. 2, 2021, Art. no. 455.

[71] Y. Oh, S. Park, and J. C. Ye, "Deep learning COVID-19 features on CXR using limited training data sets," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2688–2700, Aug. 2020.

[72] B. Chen, J. Li, G. Lu, and D. Zhang, "Lesion location attention guided network for multi-label thoracic disease classification in chest x-rays," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 7, pp. 2016–2027, Jul. 2020.

[73] R. Xu et al., "Pulmonary textures classification via a multi-scale attention network," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 7, pp. 2041–2052, Jul. 2020.

[74] P. Chikontwe, M. Luna, M. Kang, K. S. Hong, J. H. Ahn, and S. H. Park, "Dual attention multiple instance learning with unsupervised complementary loss for COVID-19 screening," *Med. Image Anal.*, vol. 72, 2021, Art. no. 102105.

[75] W. Shi, L. Tong, Y. Zhuang, Y. Zhu, and M. D. Wang, "Exam: An explainable attention-based model for COVID-19 automatic diagnosis," in *Proc. 11th ACM Int. Conf. Bioinf., Comput. Biol. Health Inform.*, 2020, pp. 1–6.

[76] 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 Informat.*, Jul. 2021.

[77] Y.-D. Zhang, Z. Zhang, X. Zhang, and S.-H. Wang, "Midcan: A multiple input deep convolutional attention network for COVID-19 diagnosis based on chest CT and chest x-ray," *Pattern Recognit. Lett.*, vol. 150, pp. 8–16, 2021.

[78] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[79] Y. Sha and M. D. Wang, "Interpretable predictions of clinical outcomes with an attention-based recurrent neural network," in *Proc. 8th ACM Int. Conf. Bioinf., Comput. Biol., Health Inform.*, 2017, pp. 233–240.

[80] H. Dong, V. Suárez-Paniagua, W. Whiteley, and H. Wu, "Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation," *J. Biomed. Informat.*, vol. 116, 2021, Art. no. 103728.

[81] J. Zhang, K. Kowsari, J. H. Harrison, J. M. Lobo, and L. E. Barnes, "Patient2vec: A personalized interpretable deep representation of the longitudinal electronic health record," *IEEE Access*, vol. 6, pp. 65333–65346, 2018.

[82] F. Shi et al., "Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for COVID-19," *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 4–15, 2021.

[83] C. Song, F. Liu, Y. Huang, L. Wang, and T. Tan, "Auto-encoder based data clustering," in *Iberoamerican Congress on Pattern Recognition*. Berlin, Heidelberg: Springer, 2013, pp. 117–124.

[84] K.-L. Lim, X. Jiang, and C. Yi, "Deep clustering with variational autoencoder," *IEEE Signal Process. Lett.*, vol. 27, pp. 231–235, 2020.

[85] V. Prasad, D. Das, and B. Bhowmick, "Variational clustering: Leveraging variational autoencoders for image clustering," in *Proc. Int. Joint Conf. Neural Netw.*, 2020, pp. 1–10.

[86] B. King, S. Barve, A. Ford, and R. Jha, "Unsupervised clustering of COVID-19 chest x-ray images with a self-organizing feature map," in *Proc. IEEE 63rd Int. Midwest Symp. Circuits Syst.*, 2020, pp. 395–398.

[87] P. Yadav, N. Menon, V. Ravi, and S. Vishvanathan, "Lung-gans: Unsupervised representation learning for lung disease classification using chest CT and x-ray images," *IEEE Trans. Eng. Manag.*, early access, Aug. 30, 2021, doi: 10.1109/TEM.2021.3103334.

[88] G. Singh and K.-C. Yow, "These do not look like those: An interpretable deep learning model for image recognition," *IEEE Access*, vol. 9, pp. 41482–41493, 2021.

[89] N. Saeedizadeh, S. Minaee, R. Kafieh, S. Yazdani, and M. Sonka, "Covid tv-unet: Segmenting COVID-19 chest CT images using connectivity imposed unet," *Comput. Methods Programs Biomed. Update*, vol. 1, 2021, Art. no. 100007.

[90] M. Pennisi et al., "An explainable AI system for automated COVID-19 assessment and lesion categorization from CT-scans," *Artif. Intell. Med.*, vol. 118, 2021, Art. no. 102114. .

[91] X. Wang et al., "Joint learning of 3D lesion segmentation and classification for explainable COVID-19 diagnosis," *IEEE Trans. Med. Imag.*, vol. 40, no. 9, pp. 2463–2476, Sep. 2021.

[92] X.-J. Yao, Z.-Q. Zhu, S.-H. Wang, and Y.-D. Zhang, "Csgbbnet: An explainable deep learning framework for COVID-19 detection," *Diagnostics*, vol. 11, no. 9, 2021, Art. no. 1712.

[93] I. M. Nedumkunnel, L. E. George, K. S. Sowmya, N. A. Rosh, and V. Mayya, "Explainable deep neural models for COVID-19 prediction from chest x-rays with region of interest visualization," in *Proc. 2nd Int. Conf. Secure Cyber Comput. Commun.*, 2021, pp. 96–101.

[94] S. Jadhav, G. Deng, M. Zawin, and A. E. Kaufman, "COVID-view: Diagnosis of COVID-19 using chest CT," *IEEE Trans. Vis. Comput. Graph.*, vol. 28, no. 1, pp. 227–237, Jan. 2022.

[95] P. Natekar, A. Kori, and G. Krishnamurthi, "Demystifying brain tumor segmentation networks: Interpretability and uncertainty analysis," *Front. Comput. Neurosci.*, vol. 14, 2020, pp. 6–6.

[96] N. Tsiknakis et al., "Interpretable artificial intelligence framework for COVID-19 screening on chest x-rays," *Exp. Therapeutic Med.*, vol. 20, no. 2, pp. 727–735, 2020.

[97] J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, "Evaluating the quality of machine learning explanations: A survey on methods and metrics," *Electronics*, vol. 10, no. 5, 2021, Art. no. 593.

[98] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 11, pp. 2660–2673, Nov. 2017.

[99] V. Petsiuk, A. Das, and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models," 2018, *arXiv:1806.07421*.

[100] C. Molnar, *Interpretable Mach. Learn.*. Lulu. com, 2020.

[101] Y.-S. Lin, W.-C. Lee, and Z. B. Celik, "What do you see? evaluation of explainable artificial intelligence (xai) interpretability through neural backdoors," in *Proc. 27th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2021, pp. 1027–1035.

[102] T. DeVries, A. Romero, L. Pineda, G. W. Taylor, and M. Drozdzal, "On the evaluation of conditional GANs," 2019, *arXiv:1907.08175*.

[103] M. Yang and B. Kim, "Benchmarking attribution methods with relative feature importance," 2019, *arXiv:1907.09701*.

[104] W. Hryniewska, P. Bombiński, P. Szatkowski, P. Tomaszewska, A. Przelaskowski, and P. Biecek, "Checklist for responsible deep learning modeling of medical images based on COVID-19 detection studies," *Pattern Recognit.*, vol. 118, 2021, Art. no. 108035.

[105] A. D. Haimovich et al., "Development and validation of the quick COVID-19 severity index: A prognostic tool for early clinical decompensation," *Ann. Emerg. Med.*, vol. 76, no. 4, pp. 442–453, 2020.

[106] P. Pan et al., "Prognostic assessment of COVID-19 in the intensive care unit by machine learning methods: Model development and validation," *J. Med. Internet Res.*, vol. 22, no. 11, 2020, Art. no. e23128.

[107] A. E. Jones, S. Trzeciak, and J. A. Kline, "The sequential organ failure assessment score for predicting outcome in patients with severe sepsis and evidence of hypoperfusion at the time of emergency department presentation," *Crit. Care Med.*, vol. 37, no. 5, 2009, Art. no. 1649.

[108] G. Barlow, D. Nathwani, and P. Davey, "The curb65 pneumonia severity score outperforms generic sepsis and early warning scores in predicting mortality in community-acquired pneumonia," *Thorax*, vol. 62, no. 3, pp. 253–259, 2007.

[109] H. Estiri, Z. H. Strasser, and S. N. Murphy, "Individualized prediction of COVID-19 adverse outcomes with mlho," *Sci. Rep.*, vol. 11, no. 1, pp. 1–9, 2021.

[110] B. Zheng et al., "An interpretable model-based prediction of severity and crucial factors in patients with COVID-19," *BioMed Res. Int.*, vol. 2021, 2021, Art. no. 8840835.

[111] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomic.*, vol. 2 1, no. 1, pp. 1–13, 2020.

[112] S. T. Argaw et al., "Cybersecurity of hospitals: Discussing the challenges and working towards mitigating the risks," *BMC Med. Informat. Decis. Mak.*, vol. 20, no. 1, pp. 1–10, 2020.

[113] A. Mirzazadeh et al., "Improving heart transplant rejection classification training using progressive generative adversarial networks," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Informat.*, 2021, pp. 1–4.

[114] F. Giuste et al., "Automated classification of acute rejection from endomyocardial biopsies," in *Proc. 11th ACM Int. Conf. Bioinf., Comput. Biol. Health Inform.*, 2020, pp. 1–9.

[115] H. Wang et al., "Decoding COVID-19 pneumonia: Comparison of deep learning and radiomics CT image signatures," *Eur. J. Nucl. Med. Mol. Imag.*, vol. 48, no. 5, pp. 1478–1486, 2021.

[116] D. Jin, E. Sergeeva, W.-H. Weng, G. Chauhan, and P. Szolovits, "Explainable deep learning in healthcare: A methodological survey from an attribution view," *WIREs Mechanisms Dis.*, vol. 14, no. 3 2022, Art. no. e1548.

[117] F. Foieni et al., "Derivation and validation of the clinical prediction model for COVID-19," *Intern. Emerg. Med.*, vol. 15, no. 8, pp. 1409–1414, 2020.

[118] M. Kang et al., "Quantitative assessment of chest CT patterns in COVID-19 and bacterial pneumonia patients: A deep learning perspective," *J. Korean Med. Sci.*, vol. 36, no. 5, pp. 1–14, 2021.

[119] D. Bau, J.-Y. Zhu, H. Strobelt, A. Lapedriza, B. Zhou, and A. Torralba, "Understanding the role of individual units in a deep neural network," *Proc. Nat. Acad. Sci.*, vol. 117, no. 48, pp. 30071–30078, 2020.