Can XAI foster the deployment of AI in a medical setting? A case study on contrastive visual explanations for X-ray scans

Anonymous Authors[1]

Abstract

Collaboration between AI and health-care professionals needs to be grounded in a particularly trustful relationship. Building on requirements for decision-support systems in a medical setting, we conduct a case study to assess the potential and limitations of an XAI system to foster the deployment of AI in diagnosis based on medical imaging techniques. To this end, we supplement an X-ray scan classifier to detect pneumonia with contrastive visual explanations automatically generated by an ExplainGAN. While results for a real-world dataset show that the quality of explanations is not yet sufficient for application, our findings underscore the great potential of XAI systems in a medical setting and outline directions for further research.

1. Introduction

Pneumonia is one of the most widespread diseases and remains a leading cause of death. Indeed, it was the most common cause of death of children under five years in 2008 (Black et al., 2010; Rudan et al., 2008), killing more children than AIDS/HIV, malaria, and measles combined (Adegbola, 2012). It also came to the fore due to the spread of the coronavirus 2019-nCov, as pneumonia is one of the central symptoms of COVID-19 patients (Zhou et al., 2020). The most commonly used diagnostic tool for lung diseases like pneumonia is chest radiography (Qin et al., 2018), which exhibits an excellent cost-benefit ratio, easy operation, and availability even in underdeveloped regions (Li et al., 2012; Qin et al., 2018). However, the appearance of pneumonia in X-ray scans can be vague or resemble that of benign abnormalities (Rajpurkar et al., 2017). Further, the overlapping of different tissue structures makes the correct interpretation of the scans very challenging for radiologists (Qin et al., 2018).

Help can be found in the application of AI, which is increasingly used for decision support in various contexts (Crawford et al., 2019). In the context of image recognition, deep neural networks have been found to detect abnormalities with high accuracy and efficiency (Qin et al., 2018). Indeed, AI systems like CheXNet have been found to outperform human experts in the detection of pneumonia and other diseases from chest X-rays (Rajpurkar et al., 2017).

The best diagnostic results can be achieved by combining AI predictions with the expertise of human specialists (Wang et al., 2016). To this end, cooperation between human specialist and AI predictions needs to be grounded in a trustful relationship. Currently, despite AI's potential, its application in a medical setting is often hindered by the "black-box" character of AI systems, i.e., their internal logic stays hidden from the user (Guidotti et al., 2018). If potential users of an AI system do not trust the system, they won't use it (Ribeiro et al., 2016). For instance, IBM's AI system Watson for Oncology was ignored by the doctors it was meant to assist (Bloomberg, 2018). Against this background, Explainable Artificial Intelligence (XAI) aims at providing explanations along with the AI's decisions, i.e., human understandable lines of reasoning (Guidotti et al., 2018). Indeed, if trust can be established this way, the cooperation of human experts and AI systems has been shown to yield great benefits (Wang et al., 2016).

To the best of our knowledge, no supplementary XAI system has been established to support the adoption of AI in diagnosis based on medical images. In this case study, we investigate the potential and current limitations of an XAI system supplementing an AI system that diagnoses pneumonia based on X-ray scans. We implement and instantiate the promising ExplainGAN approach (Samangouei et al., 2018) on a real-world dataset and discuss its efficacy and practical applicability. Our results pave the way for further steps exploiting the potential of XAI to leverage deployment of AI in a clinical setting.

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

2. Theoretical background

2.1. Requirements for the application of AI methods in clinical settings

To evaluate the effects of adding an XAI supplement to an AI system deployed in a clinical setting, in the following, we derive requirements for decision-support systems in this context.

Initial research sheds light on the fact that the deployment of AI systems in clinical settings requires an evaluation beyond system accuracy. Both socio-environmental factors and the AI systems' overall ability to improve patient-care need to be considered (Beede et al., 2020). Against this background, we rely on prior literature investigating the deployment of decision support systems in real world clinical settings to carve out requirements for AI systems in this context. First, the integration of an AI system into an existing process of medical care should not impede the underlying process (Beede et al., 2020; Pontefract et al., 2018). For instance, one study found a detrimental effect of the implementation of computerized physician order entry and clinical decision support systems as it reduced the ability to amend prescriptions and thus increased overall communication load (Pontefract et al., 2018). Another study highlighted the large amount of time that the implementation of Electronic Health Record required physicians to spend with the tool itself (**?**). Thus, the assessment of a decision support system should reflect the effect on the overall clinical workflow with a specific focus on the time medical professionals are required to spend with the system.

Second, on a technical level, an AI systems should cope with imperfect and resource constraint environments (Beede et al., 2020). For instance, one study found that integrating a deep learning system for the detection of diabetic eye disease in clinics increased the time needed for patients to receive care, as the system only supported high-quality images that in turn required longer time to be produced due to, e.g., poor lighting conditions. Further, delays occurred as a result of network connectivity issues (Beede et al., 2020).

Third, the design of a novel decision support system should be informed by the diagnostic decision-making strategy. While decision support systems oftentimes focus on one particular data source (e.g., X-ray images), in practice a diagnosis is generally achieved by consulting a variety of different data sources (Hartswood et al., 2003). Thus, prior research suggests that systems should include multiple explanations and additional data sources (e.g., raw data) to best support medical professionals, rather than merely providing a particular decision (Wang et al., 2019). Further, research identified that medical professionals required transparency of the model properties (e.g., subjective point-of-view, overall design-objective) to be able to incorporate the system's suggestion in the decision-making (**?**).

2.2. The ExplainGAN approach

The need to establish trust can be an obstacle in the deployment of powerful, but highly complex AI systems like deep neural networks. Although they achieve human-expert level results on numerous classification tasks for medical images (McKinney et al., 2020; Rajpurkar et al., 2017), their black-box character can lead to a lack of trust among users (Ribeiro et al., 2016). Against this background, XAI systems aim to provide explanations for the output or inner workings of AI systems (Abdul et al., 2018). One promising approach for binary image classifiers is the ExplainGAN introduced by Samangouei et al. (2018). Given any black-box image classifier that distinguishes between two classes, the ExplainGAN produces a decision-boundary crossing transformation of the original image. I.e., for an image classified as class 0, it creates a modified image that the classifier would predict to belong to the opposite class 1 and vice versa.

The ExplainGAN itself is a system of deep neural networks based on the DCGAN architecture (Radford et al., 2016). Contrary to regular GANs as introduced by Goodfellow et al. (2014), where images are generated from a random seed, the output generated by the ExplainGAN is based on the given image whose classification is to be explained. To this end, the original image is initially fed into an encoder network specific to its predicted class. The resulting encoded vector of reduced dimensionality is then passed as an input to a generator network. A pair of discriminator networks, one for each of the two classes, forms the adversarial part of the ExplainGAN structure that guarantees that the transformations resemble the images from the training dataset.

Besides the transformation, the ExplainGAN generates two further images: a mask that indicates where the differences occur and a composite image created by combining the original image and the transformation according to the mask. This way, only the learned local changes are applied to the original image, which leads to a composite image that stays closer to the original than the transformation. Comparing the transformation and/or composite image to the original image constitutes a contrastive visual explanation for the decision of the classifier.

The ExplainGAN is trained to optimize a four-part loss function (Samangouei et al., 2018):

$$\mathcal{L} = \mathcal{L}_{\text{GAN}} + \mathcal{L}_{\text{classifier}} + \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{prior}}. \tag{1}$$

Here, $\mathcal{L}_{\text{GAN}}$ is a conventional GAN loss function, $\mathcal{L}_{\text{classifier}}$ ensures that the composite image is classified with high confidence, and $\mathcal{L}_{\text{recon}}$ is the reconstruction loss for the encoder networks. $\mathcal{L}_{\text{prior}}$ encourages various properties of the resulting image, e.g., a local concentration and a desired maximum ratio of pixels that are changed. During training, the ExplainGAN is provided with a set of images compatible with the classifier. This can be the same dataset that was used to train the classifier before, but can also be comprised of unlabeled images, since the ExplainGAN only requires the classifier's predictions and no ground truth labels of the images. During training, each image is first fed into the pre-trained classifier, whose output subsequently determines which of the encoder networks of the ExplainGAN is used.

Similarly, to generate a contrastive visual explanation, the original image is first classified and then fed into the fully-trained ExplainGAN. Starting from the encoder network associated with the predicted class, transformation, mask, and composite image are created by a single forward pass. This way, the ExplainGAN provides contrastive visual explanations for classifiers that do not provide any explanation with their output, which is generally the case for highly advanced classifiers like CheXNet (Rajpurkar et al., 2017).

3. Use case: Application of ExplainGAN to a chest X-ray classifier predicting pneumonia

To assess the applicability of the ExplainGAN approach to X-ray classification, we instantiated an ExplainGAN on a real-world dataset of chest X-rays. The dataset (Kermany et al., 2018a) comprises 5,863 chest X-rays from pediatric patients of one to five years old. The labels for the images, stating whether the patient suffers from pneumonia or not, were cleared by two expert physicians (Kermany et al., 2018b). The training set consists of 3,883 images characterized as depicting pneumonia and 1,349 normal images, i.e., images not depicting pneumonia.

As the classifier, we trained a deep CNN based on a VGG-16 architecture (Simonyan & Zisserman, 2015) pre-trained on ImageNet (accuracy of 86.7% on the X-ray test data set). In order to achieve compatibility with the classifier, the images in the dataset are stretched or compressed to square shape, resized to 224×224 pixels, and transformed from a grayscale to an RGB color channel format.

Given an X-ray scan of a lung that is classified as exhibiting signs of pneumonia, the ExplainGAN creates a transformation to an image that is classified as healthy, and vice versa. As a starting point for the application to the chest X-ray dataset, we used an ExplainGAN implementation for the MNIST dataset of handwritten digits (LeCun et al., 1998). In order to accommodate for the bigger input size (MNIST images are 28×28 pixels in size, whereas the already resized X-ray scans are still 224×224 pixels), we expanded the dimensionality of the encoded vector from 128 to 3,136 entries and increased the depth of each component network[1]. All component networks of the ExplainGAN are trained simultaneously with respect to a weighted sum of several loss functions (eq. (1)). Our ExplainGAN was trained on Google Colaboratory, where one epoch of training on the whole training portion of the pneumonia data set with a Tesla P100 as GPU support typically took around ten minutes. We trained instances of the ExplainGAN for up to 200 epochs.

The individual network components worked satisfactorily on their own: The encoders and generators were able to produce reconstructions that were hardly distinguishable from the respective original images with the naked eye. Further, the discriminators could learn to accurately tell apart real X-rays from fake images created by fixed, partially trained generators. However, the transformations produced in our experiments either looked almost exactly like the original images or showed drastic and widespread differences to the degree that they only vaguely resembled real X-ray scans. Figure 1 illustrates both cases and compares the results to the ones obtained with an ExplainGAN for MNIST.

With more computational resources and time, it might be possible to achieve both a resemblance of the transformations to the original images and sparse differences that lead to opposite predictions of the classifier. In spite of extensive experimentation with the associated weight coefficients, generally one single loss function from eq. (1) ended up dominating the training process, resulting in the aforementioned extreme results. The use of deeper neural networks or an increased number of filters in the convolutional layers might increase the ExplainGAN's capability to simultaneously factor in the conflicting incentives during training and better capture the more complex nature of the X-ray images compared to MNIST.

However, the instantiation for significantly down-scaled X-ray images constitutes only a small step towards the application of an ExplainGAN to supplement X-ray scan classifiers in clinical practice. The low-resolution contrastive visual explanations would hardly be useful for diagnosis. Further, the ExplainGAN's restriction to a binary classification task constitutes another simplification compared to actual medical diagnosis processes. Hence, the architecture would need to be amended, further increasing the demand for computational resources and increasing the instability of the training process. The transfer and adjustment of an ExplainGAN instance to different X-ray scanners poses yet another challenge. In this light, it is questionable whether the ExplainGAN approach, despite its elegant architecture and reliance on established GAN components, constitutes a suitable candidate for practical application in medical settings.

[1] Our encoder networks are comprised of five convolutional layers, the generator of six transposed convolutional layers, and the discriminators of four convolutional layers. The number of filters varies between layers, with a maximum of 128 filters

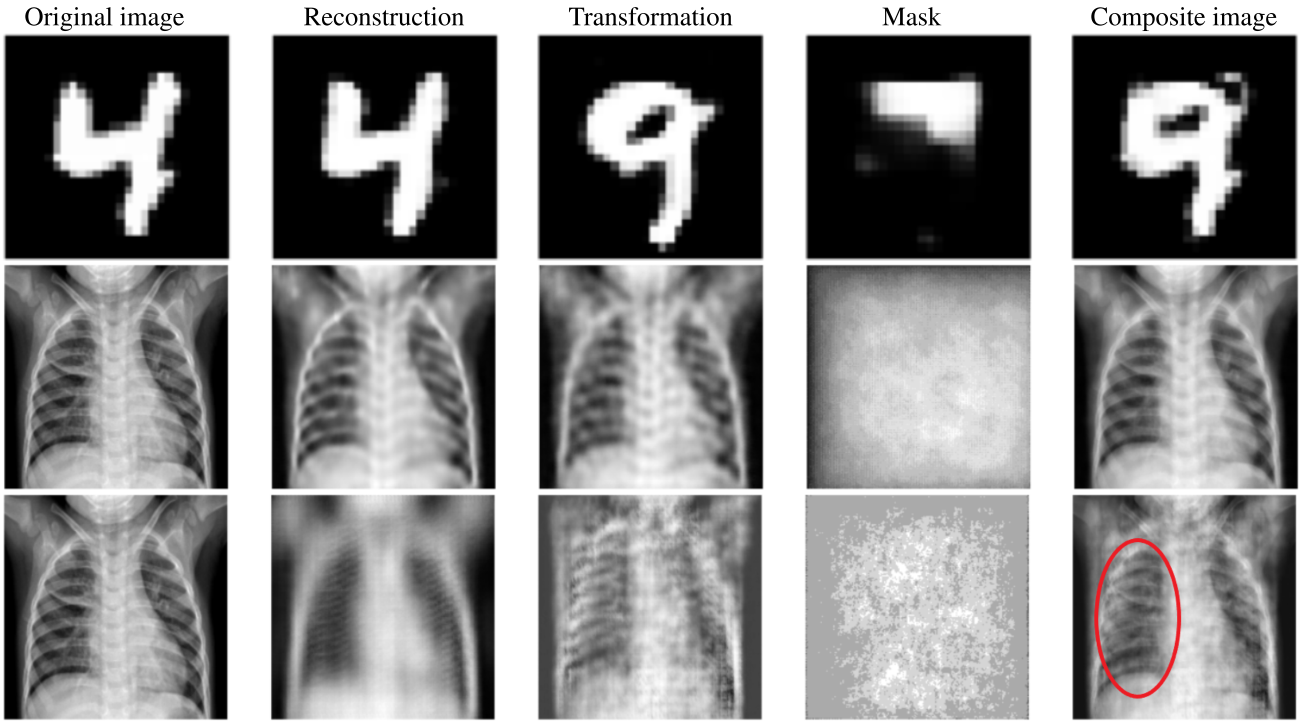

Figure 1. Example of images generated by ExplainGANs. The first row shows images produced by an ExplainGAN instantiated for a classifier for handwritten 4's and 9's from the MNIST data set. The original image is correctly classified to depict the digit 4. Therefore, a reconstruction of class 4 (used for training purposes) and a transformation to the opposite class 9 are created. The fourth column shows a mask highlighting which parts of the composite image in the last column are based on the original image (mask black) and the transformation (mask white), respectively. The second and third row show two sets of images generated by two instances of the ExplainGAN resulting from different training setups and parameters. The first column shows the original chest X-ray scan of a healthy patient. In the middle row, reconstruction and transformation both look real and similar to the original, but the transformation does not show any easily perceptible differences to the original image. In the third row, the transformation differs from the original image so much that it is difficult to recognize it as a chest X-ray scan. One observation indicating that the ExplainGAN correctly identified how the classifier distinguishes healthy from pneumonia patients is the general increase in brightness in the lung area (best visible in the red circle in the composite image). Bright patches correspond to lobar consolidations typical for pneumonia. However, the changes are not sparse and localized and both the transformation and composite image are blurry and of little explanatory value. Experimentation with the weight of $\mathcal{L}_{\text{prior}}$, which controls the mask properties, indicates that an increase does not solve this problem, but instead leads to fully black or white masks. For both depicted ExplainGAN samples, the mask does not fulfill the desired properties of being binary and localized.

4. Discussion

Our study was motivated by the quest of improving the deployment of AI systems in clinical settings. Thus, we discuss the applicability of ExplainGANs to an X-ray scan classifier not only with regards to the quality of the resulting explanations, but also regarding the requirements for decision support systems in a clinical setting.

On the one hand, our results suggest that the quality of the explanations generated by the ExplainGAN is yet insufficient to support a diagnosis decision. Indeed, to the best of our knowledge, applicability of ExplainGANs had only been demonstrated for a much smaller dataset (Samangouei et al., 2018). A transfer to the real-world problem context of X-ray scans in a clinical setting, i.e., a data set with much larger images, was not possible with restricted resources regarding development time (two months) and computational power (a single Tesla P100 as GPU support), comparable to the resource constraints an ExplainGAN deployment would face in a real-world setting (Beede et al., 2020). Although we were able to successfully train each of the component networks of the ExplainGAN on their own (keeping the other component networks fixed), the simultaneous training of the whole network did not yield a fully functional system. While we identified some potential in the isolated training of the autoencoders, the interaction between generator and discriminator is the centerpiece of GAN training (Goodfellow et al., 2014). As the adversarial networks learn from each other, they cannot be trained independently. Approaches with a higher modularity than GANs might be advantageous for the creation of visual contrastive explanations in real-world settings. Besides the challenging implementation, drawbacks of the ExplainGAN approach include the narrow scope to binary image classification tasks. In the given use case of chest X-rays, there are many diseases other than pneumonia that can manifest themselves on the images and that doctors look for simultaneously. Multi-class classifiers like CheXNet (Rajpurkar et al., 2017) are arguably more useful than a binary pneumonia classifier in such a case. The challenge of adapting ExplainGAN or developing another approach that creates contrastive visual explanations for multi-class classifiers remains open for future research. A new methodical question that arises in this scenario is how to determine the target class for the transformation of a given original image.

On the other hand, we find that XAI systems that produce explanations similar to ExplainGAN bear great potential to enhance the deployment of AI classifying X-ray scans. Indeed, adding contrastive visual explanations to an existing AI system would address the requirements for AI systems in a clinical setting which already have been found to be violated in other studies (Beede et al., 2020). First, an AI classifying X-ray scans combined with contrastive visual explanations fits the underlying decision structure and provides an additional element for the process of ”coherent marshalling of ensembles of evidence” (Hartswood et al., 2003, p. 390): As a decision support system, the ExplainGAN could provide an initial starting point to support a diagnosis decision by providing contrastive explanations along with the classification outcome of an arbitrary AI system. In a situation with several candidate black-box AI systems for a binary classification task, contrastive visual explanations could be used to visualize the internal logic of the different classifiers at one glance and help to choose a candidate with the most convincing explanations. With a trained network stored on a local machine, the creation of these explanations could be done instantly for a new query image since it corresponds to only one forward pass through the neural networks, thus taking into account the time-constrained environment of a clinical setting. Second, contrastive visual explanations are suitable to support the collaborative and interdisciplinary working environment in healthcare, as they provide intuitive visual explanations for the black-box classifier and thus allow easy comprehensibility. Finally, in a running system with one classifier and added contrastive visual explanations, the produced explanations serve as a constant control mechanic of the quality of the classifier AI addressing the requirement of verifiable quality of the AI support. This way, if the quality of the AI classifier is sufficient, the addition of contrastive visual explanations helps to instill trust in the AI system.

5. Conclusion

AI classifiers for medical images have repeatedly been found to outperform human experts in diagnosis (McKinney et al., 2020; Rajpurkar et al., 2017). However, deployment of AI in a medical setting is still hindered by a lack of trust (Bloomberg, 2018). Indeed, cooperation between human specialist and AI systems needs to be grounded in a trustful relationship and deployment of AI systems in clinical settings requires an evaluation beyond system accuracy, considering socio-environmental factors and evaluating the overall ability to improve patient-care (Beede et al., 2020). For our case study, we implemented an ExplainGAN (Samangouei et al., 2018) to supplement an AI classifier diagnosing pneumonia based on chest X-ray scans from a real-world dataset and investigated its potential and current limitations in leveraging the deployment of AI in a clinical setting. Our findings suggest that further research is required to increase the quality of resulting explanations while taking the limitations on resources into account. Based on the evaluation with respect to requirements for application of AI systems in a clinical setting, we argue that contrastive visual explanations show great potential to enhance the deployment of AI systems classifying X-ray scans and assisting pneumonia diagnosis.

We identified three major research avenues to advance explanations for AI systems for diagnosis based on medical images: First, to ensure high-quality explanations as required in a clinical setting, XAI methods that generate contrastive visual explanations need to be capable of efficiently and robustly handling high-resolution images. In particular, the applicability of XAI systems like ExplainGANs would greatly profit from a more robust training process, which could be achieved through a more modular system architecture. Second, development of future XAI systems for contrastive visual explanations should focus on practical applicability. In particular, they should be applicable to multi-class problems, which are frequent in a real-world clinical setting. Third, as both socio-environmental factors and the AI systems' overall ability to improve patient-care need to be considered (Beede et al., 2020) when instantiating an AI system in a medical setting, a crucial next step would be to conduct user-centric evaluation of XAI methods. In particular, adoption rates and users satisfaction of an AI system diagnosing pneumonia based on chest X-ray scans alone might be compared to a system accompanied with contrastive visual explanations. With our paper, we hope to encourage researchers to take the next steps exploiting the potential of XAI to leverage the deployment of AI in a clinical setting.

References

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., and Kankanhalli, M. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2018. doi: 10.1145/3173574.3174156.

Adegbola, R. A. Childhood Pneumonia as a Global Health Priority and the Strategic Interest of The Bill & Melinda Gates Foundation. *Clinical Infectious Diseases*, 54 (suppl_2):S89–S92, 2012. doi: 10.1093/cid/cir1051.

Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., and Vardoulakis, L. M. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2020. doi: 10.1145/3313831.3376718.

Black, R., Cousens, S., Johnson, H., Lawn, J., Rudan, I., Bassani, D., Jha, P., Campbell, H., Walker, C., Cibulskis, R., Eisele, T., Liu, L., and Mathers, C. Global, Regional, and National Causes of Child Mortality in 2008: A Systematic Analysis. *The Lancet*, 375(9730):1969–1987, 2010. doi: 10.1016/S0140-6736(10)60549-1.

Bloomberg, J. Don't Trust Artificial Intelligence? Time To Open The AI 'Black Box', 2018. URL https://www.forbes.com/sites/jasonbloomberg/2018/09/16/dont-trust-artificial-intelligence-time-to-open-the-ai-black-box/. [Online; accessed 16.01.2020].

Crawford, K., Dobbe, R., Drye, T., Fried, G., Green, B., Kaziunas, E., Kak, A., Mathur, V., McElroy, E., Nill Sánchez, A., Raji, D., Ranking, J. L., Richardson, R., Schultz, J., West, S. M., and Whittaker, M. AI Now Report 2019. Technical report, AI Now Institute, New York, NY, 2019.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. Curran Associates, 2014. URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. A Survey of Methods for Explaining Black Box Models. *Association for Computing Machinery Computing Surveys*, 51(5), 2018. doi: 10.1145/3236009.

Hartswood, M., Procter, R., Rouncefield, M., Slack, R., Soutter, J., and Voss, A. 'Repairing' the Machine: A Case Study of the Evaluation of Computer-Aided Detection Tools in Breast Screening. In *Proceedings of the Eighth European Conference on Computer Supported Cooperative Work*, pp. 375–394. Springer Netherlands, 2003. doi: 10.1007/978-94-010-0068-0_20.

Kermany, D., Zhang, K., and Goldbaum, M. Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification. http://dx.doi.org/10.17632/rscbjbr9sj.2, 2018a.

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M. K., Pei, J., Ting, M. Y., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., Shi, A., Zhang, R., Zheng, L., Hou, R., Shi, W., Fu, X., Duan, Y., Huu, V. A., Wen, C., Zhang, E. D., Zhang, C. L., Li, O., Wang, X., Singer, M. A., Sun, X., Xu, J., Tafreshi, A., Lewis, M. A., Xia, H., and Zhang, K. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*, 172(5):1122 – 1131.e9, 2018b. doi: 10.1016/j.cell.2018.02.010.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-Based Learning Applied to Document Recognition. In *Proceedings of the IEEE*, volume 86, pp. 2278–2324, 1998. doi: 10.1109/5.726791.

Li, F., Engelmann, R., Pesce, L., Armato, S. G., and MacMahon, H. Improved Detection of Focal Pneumonia by Chest Radiography with Bone Suppression Imaging. *European Radiology*, 22(12):2729–2735, 2012. doi: 10.1007/s00330-012-2550-y.

McKinney, S., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C., King, D., and Shetty, S. International Evaluation of an AI System for Breast Cancer Screening. *Nature*, 577:89–94, 2020. doi: 10.1038/s41586-019-1799-6.

Pontefract, S. K., Hodson, J., Slee, A., Shah, S., Girling, A. J., Williams, R., Sheikh, A., and Coleman, J. J. Impact of a Commercial Order Entry System on Prescribing Errors Amenable to Computerised Decision Support in the Hospital Setting: A Prospective Pre-Post Study . *BMJ Quality & Safety*, 27(9):725–736, 2018. doi: 10.1136/bmjqs-2017-007135.

Qin, C., Yao, D., Shi, Y., and Song, Z. Computer-aided detection in chest radiography based on artificial intelligence: a survey. In *Biomedical Engineering Online*, volume 17, 2018. doi: 10.1186/s12938-018-0544-y.

Radford, A., Metz, L., and Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *4th International Conference on Learning Representations Conference Track Proceedings*, 2016. arXiv: 1511.06434.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., and Ng, A. Y. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning, 2017. arXiv: 1711.05225.

Ribeiro, M. T., Singh, S., and Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144. Association for Computing Machinery, 2016. doi: 10.1145/2939672.2939778.

Rudan, I., Boschi Pinto, C., Biloglav, Z., Mulholland, K., and Campbell, H. Epidemiology and Etiology of Childhood Pneumonia. *Bulletin of the World Health Organization*, 86:408–16, 06 2008. doi: 10.1097/INF.0b013e3181950942.

Samangouei, P., Saeedi, A., Nakagawa, L., and Silberman, N. ExplainGAN: Model Explanation via Decision Boundary Crossing Transformations. In *Proceedings of the 15th European Conference on Computer Vision*, pp. 681–696, 2018. doi: 10.1007/978-3-030-01249-6_41.

Simonyan, K. and Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*, 2015. arXiv: 1409.1556v6.

Wang, D., Khosla, A., Gargeya, R., Irshad, H., and Beck, A. H. Deep Learning for Identifying Metastatic Breast Cancer, 2016. arXiv: 1606.05718.

Wang, D., Yang, Q., Abdul, A., and Lim, B. Y. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–15. Association for Computing Machinery, 2019. doi: 10.1145/3290605.3300831.

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L., Chen, H.-D., Chen, J., Luo, Y., Guo, H., Jiang, R.-D., Liu, M.-Q., Chen, Y., Shen, X.-R., Wang, X., Zheng, X.-S., Zhao, K., Chen, Q.-J., Deng, F., Liu, L.-L., Yan, B., Zhan, F.-X., Wang, Y.-Y., Xiao, G.-F., and Shi, Z.-L. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 579:270–273, 2020. doi: 10.1038/s41586-020-2012-7.