Artificial intelligence in radiology – beyond the black box

Künstliche Intelligenz in der Radiologie – jenseits der Black-Box

Authors

Luisa Gallée[1], Hannah Kniesel[2], Timo Ropinski[2], Michael Götz[1, 3]

Affiliations

1 Division of Experimental Radiology, Department for Diagnostic and Interventional Radiology, University Ulm Medical Centre, Ulm, Germany
2 Visual Computing, University of Ulm, Germany
3 Medical Image Computing, DKFZ, Heidelberg, Germany

Key words

Artificial Intelligence, Explainable AI, Machine Learning, Black Box, Deep Learning, Medical Image Processing

received 22.12.2022
accepted 22.03.2023
published online 09.05.2023

Bibliography

Fortschr Röntgenstr 2023; 195: 797–803
DOI 10.1055/a-2076-6736
ISSN 1438-9029

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

Correspondence

Prof. Michael Götz
Division for Experimental Radiology, University Ulm Medical Centre, Albert-Einstein-Allee 23, 89081 Ulm, Germany
Tel.: +49/7 31 50 06 11 91
michael.goetz@uni-ulm.de

ABSTRACT

Background Artificial intelligence is playing an increasingly important role in radiology. However, more and more often it is no longer possible to reconstruct decisions, especially in the case of new and powerful methods from the field of deep learning. The resulting models fulfill their function without the users being able to understand the internal processes and are used as so-called black boxes. Especially in sensitive areas such as medicine, the explainability of decisions is of paramount importance in order to verify their correctness and to be able to evaluate alternatives. For this reason, there is active research going on to elucidate these black boxes.

Method This review paper presents different approaches for explainable artificial intelligence with their advantages and disadvantages. Examples are used to illustrate the introduced methods. This study is intended to enable the reader to better assess the limitations of the corresponding explanations when meeting them in practice and strengthen the integration of such solutions in new research projects.

Results and Conclusion Besides methods to analyze black-box models for explainability, interpretable models offer an interesting alternative. Here, explainability is part of the process and the learned model knowledge can be verified with expert knowledge.

Key Points:

- The use of artificial intelligence in radiology offers many possibilities to provide safer and more efficient medical care. This includes, but is not limited to support during image acquisition and processing or for diagnosis.
- Complex models can achieve high accuracy, but make it difficult to understand data processing.
- If the explainability is already taken into account during the planning of the model, methods can be developed that are powerful and interpretable at the same time.

Citation Format

- Gallée L, Kniesel H, Ropinski T et al. Artificial intelligence in radiology – beyond the black box. Fortschr Röntgenstr 2023; 195: 797–803

ZUSAMMENFASSUNG

Hintergrund Die Bedeutung von Künstlicher Intelligenz nimmt in der Radiologie stetig zu. Doch gerade bei neuen und leistungsfähigen Verfahren, vor allem aus dem Bereich des Deep Learnings, ist das Nachvollziehen von Entscheidungen oft nicht mehr möglich. Die resultierenden Modelle erfüllen ihre Funktion, ohne dass die Nutzer die internen Abläufe nachvollziehen können und werden als sogenannte Black-Box eingesetzt. Gerade in sensiblen Bereichen wie der Medizin ist die Erklärbarkeit von Ergebnissen von herausragender Bedeutung, um deren Korrektheit zu verifizieren und Alternativen abwägen zu können. Aus diesem Grund wird aktiv an der Durchleuchtung dieser Black-Boxen gearbeitet.

Methode Dieser Übersichtsartikel stellt unterschiedliche Ansätze für erklärbare Künstliche Intelligenz mit ihren Vor- und Nachteilen vor. Anhand von Beispielen werden die vorgestellten Verfahren veranschaulicht. Die Arbeit soll es dem Leser erlauben, die Grenzen der entsprechenden Erklärungen in der Praxis besser abzuschätzen und die Einbindung solcher Lösungen in neue Forschungsvorhaben stärken.

Ergebnisse und Schlussfolgerung Neben Methoden, Black-Box-Modelle auf Erklärbarkeit zu untersuchen, bieten interpretierbare Modelle eine interessante Alternative. Die Erklär-

barkeit ist hier Teil des Verfahrens und das erlernte Modellwissen kann mit Fachwissen überprüft werden.

Kernaussagen:

- Der Einsatz von Künstlicher Intelligenz in der Radiologie bietet viele Möglichkeiten, etwa zur Unterstützung während der Bildaufnahme und -verarbeitung oder zur Diagnosestellung.
- Komplexe Modelle können eine hohe Genauigkeit erreichen, erschweren allerdings die Nachvollziehbarkeit der Datenverarbeitung.
- Wird die Erklärbarkeit bereits bei der Planung des Modells berücksichtigt, können leistungsfähige und zugleich interpretierbare Verfahren entwickelt werden.

Introduction

Algorithms in artificial intelligence (AI) make it possible to effectively process large quantities of data and address various questions. In the initial training phase, already known or previously hidden relationships in sample data are identified and represented in a model. With AI models trained in this way, identified correlations can be applied directly to new data so that it can be processed quickly and easily. Particularly in radiology, due to the high degree of digitalization [1] and the openness to technical progress, this approach has proven to be a very powerful tool for effectively processing the continuously increasing amount of image data [2] in spite of the skilled labor shortage [3].

The spectrum of applications ranges from efficient image acquisition and optimized workflows to automatic diagnostic support. For example, AI algorithms make it possible to reduce measurement time and radiation exposure while maintaining the same image quality due to improved image reconstruction [4–6]. A further application in the daily routine is the preselection of image data to decrease the unnecessary interpretation of unremarkable images. Particularly in screening programs like mammography, the workload can be significantly reduced [7–9]. In addition, AI methods allow faster and better diagnosis, e. g., as a result of the automatic annotation of organs and pathologies [10–12] and new quantitative and image-based markers as are currently being intensively researched in radiomics [13–15].

Advances with respect to AI methods are based on improved methods [16, 17], larger amounts of data [18], and increased computing capacity [19] allowing the generation of increasingly complex models. However, one challenge when using such complex AI methods is that it is often difficult to understand the reasoning behind decision processes [20, 21]. Particularly in the clinical routine, it must be possible to understand the reasoning behind decisions, including those made with the help of AI algorithms [22]. The reasons for this include acceptance by patients and the possibility to evaluate the model decision.

When training an AI method, the knowledge is implicitly acquired from the training data and is applied to new tasks. However, this process results in some uncertainties. Was all relevant information used or was it missing during training? Can the identified correlations be generalized? Is there a causal relationship for the identified correlations or are they random? To ensure the reliability of an AI system, it must be shown that the system learned the underlying properties and the decisions are not based on irrelevant correlations between input and output values in the training dataset.

Weaknesses can be reduced but not ruled out by carefully selecting the model architecture and the training algorithm of an AI method. Additional information helps to minimize the effect of interference factors, and validation of algorithms on external datasets allows evaluation of the generalizability and is being explicitly examined and promoted in data-driven areas like radiomics research [23, 24]. However, errors are possible even when being careful as shown by practical examples. For example, researchers at the Mount Sinai hospital developed an AI method for evaluating pneumonia risk based on radiographs. However, the method achieved significantly lower accuracy outside of that particular hospital. As it turned out, the approach used information about the imaging devices and detected high-risk patients based on devices used in the intensive care unit [25]. This example clearly shows how important it is to be able to understand an AI system so that such false correlations can be discovered not just by accident but systematically.

There are major differences between the individual AI methods not only with respect to performance but also regarding the ability to understand generated models (see ▶ **Table 1**). If the models can't be interpreted, the image of a closed black box is often used (see ▶ **Fig. 1**) This refers to models whose modes of operation cannot be interpreted and only the input and output values are understandable. To understand how a black box works, explanation models are consequently needed for the actual model. In contrast, interpretable models are referred to as white boxes. An intermediate stage between the two extremes is the gray box. This refers to models that allow certain insight into internal data processing. It must be taken into consideration that in practice a method cannot always be clearly classified as a white, gray, or black box method.

White box AI

The entire data processing chain is ideally able to be understood – these methods are referred to as white box methods. In particular, methods from the areas of classic machine learning and static learning that provide transparent information processing of input values, e. g. patient data, lab values, or image data, and output values, e. g., a diagnosis, should be mentioned here. One example is **linear regression**, which calculates a linear combination from various numeric features. These methods are used, for example to determine a radiomics signature and to weigh the individual features of structure, form, and texture. The influence of every feature is determined by an individual weight and can be easily read out and interpreted [26]. The situation is similar for other

methods like **Naive-Bayes classification** [27], which predicts class based on relative probabilities of occurrence. By using probability distribution, the Naive-Bayes classifier allows simple interpretation of the influence of an input value on the model output.

However, transparency is not synonymous with interpretability. Therefore, interpretability can also be limited in white box methods. This is clearly seen in the case of **decision trees** and **Random Forests**, which are often also used in radiomics [28–31]. Decision trees model a structured series of conditions in a tree structure. If the decision tree is complex or if Random Forests with multiple trees are used, decisions are transparent and theoretically able to be understood but this is no longer the case in practice because of the complexity [32].

Black box AI

If the decisions of a method can no longer be understood, for example due to their complexity, these models are referred to as black box models. Deep learning-based methods (DL), which often exceed the performance of classic methods, are typical here. They are the foundation for leading methods in a broad spectrum of complex tasks including medical image analysis and are increasingly used in radiology. Deep learning is based on the structure and function of the brain and uses a dense network of millions of artificial neurons that are series-connected in multiple layers. The interconnection of the neurons allows flexible adjustment to the particular task at hand with the input images being processed within the neural network to create visual features and segmentation or classifications being performed. The artificial neurons in which the model knowledge is stored are defined by learnable parameters.

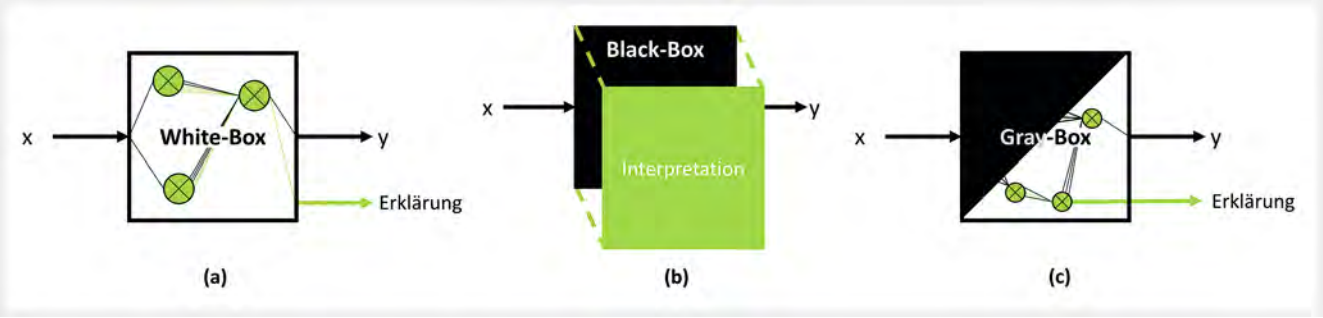
Due to the high number of parameters, deep learning models are de-facto no longer able to be understood [33] and new methods are needed to make it possible to understand the decision process. To make the black box of deep learning more transparent, methods that attempt to explain the unclear functionalities and interconnections of the neural networks in a targeted manner are therefore being developed. Many of these methods can be applied to current DL methods from general image processing. However, the value and the contribution to the interpretability of the methods vary. If the limitations of these methods are not taken into consideration, there is a risk of seeming explainability and the deduction of incorrect conclusions.

Most image-based DL architectures are based on Convolutional Neural Networks (CNNs), which are extracted with image feature filters. The **visualization** of these filters (see ▶ **Fig. 2**) can provide information about the extracted properties of the image data. For example, filters in the early layers of the network extract line or circle patterns. However, filters from deeper layers are difficult to interpret. The visualization of filters primarily contributed to the better understanding and verification of the functioning of CNNs. Due to the high degree of abstraction of filter visualization, this technique is not helpful for explaining model output in an individual application case.

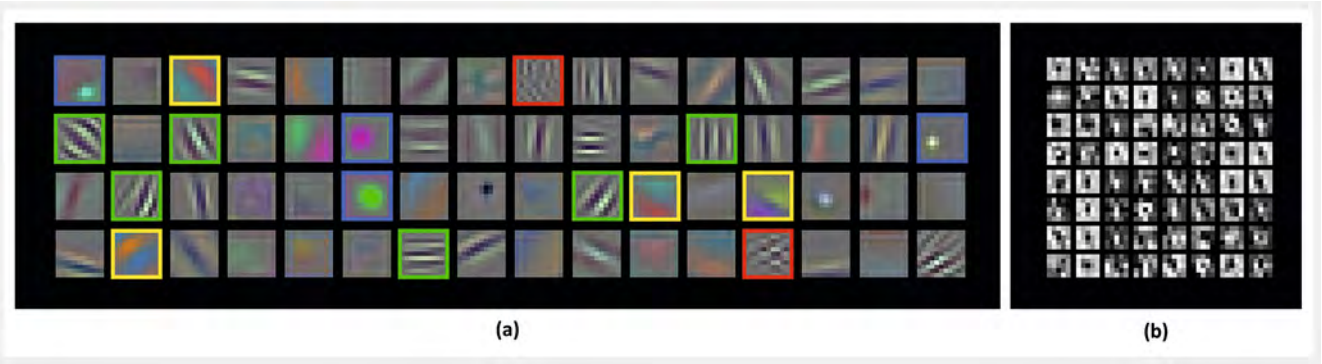
Another approach is to use **optimization** to generate an input image that maximally activates certain neurons [34]. If a neuron is highly activated, this means that an image feature learned by this neuron is present in the input image. Thus, the method converges in images that depict the patterns on which the selected neurons were trained. Either random noise can be optimized as an input image or a search can be performed for images from the training dataset that maximize activation. Initial methods usually deliver only abstract images that can be helpful during model development. The latter provides images that are easier to interpret but limits the specificity when it is not clear which element in the

▶ **Table 1** Comparison of the various levels of performance and explainability of white, black, and gray box methods.

	Performance	Explainability
White box	Only limited model complexity	Direct interpretation of models
Black box	Complex models possible	Subsequent indirect interpretation of individual aspects using explanation models
Gray box	Complex models possible	Interpretation of defined aspects using models, further explanations via black box methods are possible



▶ **Fig. 1** Schematic representation of (**a**) white box, (**b**) black box, and (**c**) gray box methods. Data processing in white box methods is transparent, while only interpretation models, which can be a source of error, can be generated for black box methods. Methods that combine complex information processing with interpretable modules can be referred to as gray box methods.



▶ **Fig. 2** Visualization of the feature filters of a CNN model that can differentiate between 100 different animals. The filters for the first layer (**a**) have an understandable function (green box: line filter, blue box: circle filter, red box: noise filter, yellow box: color filter), while the filters in the second, deeper layer (**b**) can no longer be assigned an understandable function.

input images actually caused the high activation of the neurons. Nonetheless, this approach can be helpful in practice in some cases.

Deconvolution [35, 36] is an approximated inversion of the convolution of a CNN. The regions of the input image that contribute to the activation of individual feature filters are highlighted. Human interpretation of exactly which image features in the image region are highlighted is also needed here. For this reason and as a result of the multitude of filters that are needed for complex image analyses, deconvolution is usually used only during the development of models for supporting analysis.

Regardless of the inner structure of a model, **masking-based saliency** methods examine the model as a true black box only from the outside [37]. With targeted manipulation of the input data and observation of the change in output values, relationships between individual input parameters and results can be established. In the context of image analysis, input is manipulated by masking or manipulating individual image pixels. In the best case, the model output is changed exclusively as a reaction to the masking of relevant areas. Otherwise, incorrectly learned correlations can be concluded. Moreover, spatial significance as well as intensity influences can be examined. However, a comprehensive examination with this method is time-consuming and accuracy cannot be assumed even in the case of a positive result.

With **gradient-based saliency** methods, regions in the input image that contribute to the decision regarding a certain output can be highlighted [38, 39]. Using this approach, it is possible to determine whether irrelevant image areas were considered for a decision (see ▶ **Fig. 3**). It turned out that in the detection of COVID-19 pathologies on chest radiographs [40], the focus of the learned AI was also outside of the lungs or even the body, thus reflecting differences in patient position and X-ray projection. Although this field example clearly shows that this method can identify insufficiently generalized deep learning models, care should be taken when introducing these algorithms. Even when the focus is on the correct image region, incorrect image features in this region can be learned and the use of saliency analyses can result in an overestimation of the model.

The **T-CAV** method is a more abstract approach to explaining deep learning models [41]. The goal is to examine the influence of concepts in the input images. A linear classification model is trained to learn different concepts based on the input data. The data can then be examined based on these concepts. A biased model can be detected early, for example during model development. However, the functionality of T-CAV is highly dependent on the trained model and the resulting concepts.

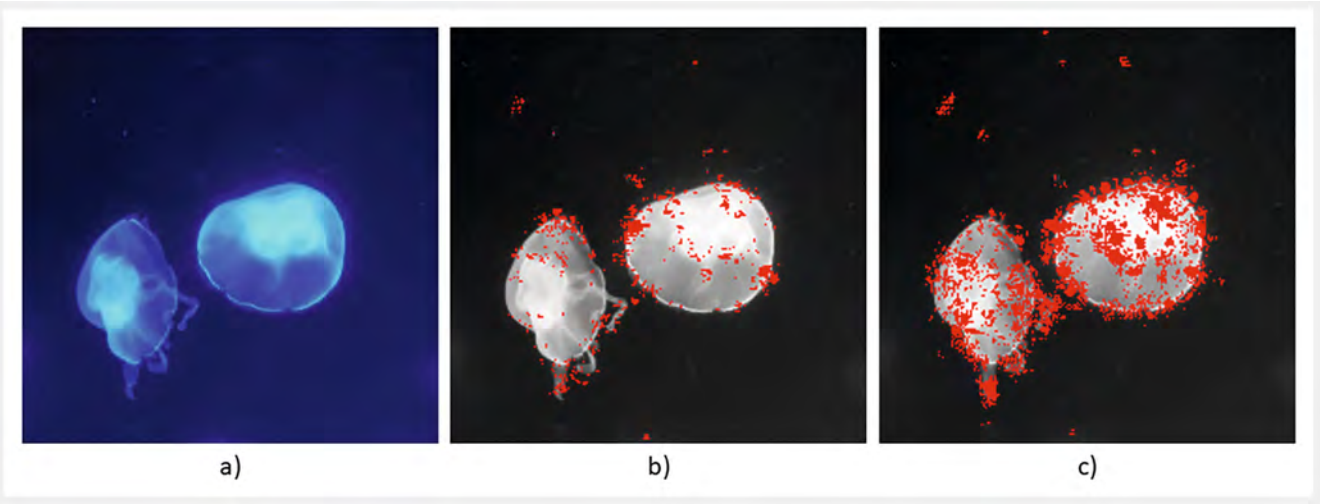
The presented methods show how different the explanation approaches for deep learning networks are. These can make an important contribution to the explanation of black box models but always have systematic limitations. Explanation of complex models always requires a reduction, which is associated with a loss of information and thus provides only partial explanations. In summary, the good applicability for black box models is an advantage of the indicated methods. The restrictions include the limited significance and the associated uncertainties.

Gray box AI

Gray box methods combine the advantages of interpretable white box methods with the powerful performance of black box methods. In this new research field, explainability is taken into consideration in the development of AI methods in order to achieve explanation goals without any notable loss of performance.

One possibility regarding explainability is the use of exemplary examples, referred to as **prototypes**. Based on the human approach to making predictions, decisions are made on the basis of the most similar examples that allow a direct analysis. Either entire images or individual segments can be learned as prototypes. Such systems not only allow the classification of medical images but also show at the same time the most similar images in the training database [42, 43]. The validity of model estimation can thus be classified, thereby inspiring trust on the part of the responsible end user. At the same time, identified prototypes can serve as training material.

Invertible neural networks have an invertible architecture so that the input and output of a model can be inverted. This invertibility can be used to check individual layers of the network. By manipulating relevant features, counterfactual sample images that allow statements like "without feature A the result is..." can be generated. This technique is already used in computer-assisted



▶ **Fig. 3** Visualization of the saliency heatmaps of the CNN model for an input image (**a**). Red pixels of heatmaps (**b**) and (**c**) show image regions that have a large impact on the network output. Heatmap (**b**) shows the focus of the model for the correct "jellyfish" output, which is predominantly on the body of the animal. However, this focusing is almost identical for an incorrect network output, as in (**c**) for the class "Hummingbird".

surgery to determine the degree of uncertainty of perfusion estimation in endoscopy [44]. Even when invertible neural networks limit the possible network structures, they are a good alternative for better understanding AI models.

The advantage of gray box methods is the combination of understandability and high performance which are important properties particularly in sensitive areas like medicine. However, there are currently corresponding methods only for a few application cases. In addition, the explainability is limited to specific elements in these methods. As in all explanation methods, it makes a difference, for example, whether individual cases are taken into consideration or a general statement is to be made. Different explanation approaches must be used depending on this. For this reason, additional research and development in the new field of gray box methods are needed for customized use of these methods in numerous application areas. Only then can the advantages of these methods also be fully utilized.

Summary

Artificial intelligence can make an important contribution to safer and more efficient radiology. However, broad acceptance of such systems in the medical profession as well as among patients requires the ability to understand decisions. Radiologists must be able to understand the models they use in order to be able to continue to fulfill their duty of care, make informed diagnoses, provide patients with comprehensive information, and provide well-founded documentation. To ensure the legal understandability of measures that are taken, the explainability of models is an important requirement for usability. In particular, high-performance systems like deep learning-based algorithms are often too complex to be able to be understood. The need to create interpretable models has already been recognized and is currently being addressed with various approaches particularly by methods that can be retrospectively used on fully trained models. The advances of the last years have resulted in considerable further developments with various levels of transparency and make it possible to answer various questions without limiting the complexity of the models. However, analysis from the outside limits the value of the black box system and the corresponding methods can only provide explanation models of the models. These are necessary reductions of the original models and are therefore also a source of error.

The use of complex but interpretable gray box AI is an interesting alternative here. Since explainability is part of these methods, the intermediate step of creating an explanation model is not needed. The learned features can be analyzed and checked with expert knowledge and offer a decision basis on which the end user can check the reliability of the model results. Since the explanation method is an integral part of AI solutions, this use must be considered early and it must be determined which parts of the AI model should be understandable. Adapted algorithms are necessary here – to provide the correct type of explanation. The close cooperation between medicine and information technology is consequently of essential importance for identifying relevant questions and finding customized solutions.

Funding

University of Ulm
Baustein (L.SBN.0214)

Conflict of Interest

The authors declare that they have no conflict of interest.

References

[1] Hricak H. 2016 new horizons lecture: beyond imaging – radiology of tomorrow. Radiology 2018; 286 (3): 764–775

[2] Bundesamt für Strahlenschutz, Hrsg. Röntgendiagnostik: Häufigkeit und Strahlenexposition für die deutsche Bevölkerung". 14. April 2022. Zugegriffen: 24. Oktober 2022. [Online]. Verfügbar unter: https://www.bfs.de/DE/themen/ion/anwendung-medizin/diagnostik/roentgen/haeufigkeit-exposition.html

[3] Attenberger U, Reiser MF. Future Perspectives: Wie beeinflusst künstliche Intelligenz die Entwicklung unseres Berufsfeldes? Radiol 2022; 62 (3): 267–270

[4] Chen Y et al. AI-Based Reconstruction for Fast MRI – A Systematic Review and Meta-Analysis. Proc. IEEE 2022; 110 (2): 224–245. doi:10.1109/JPROC.2022.3141367

[5] Reader AJ, Corda G, Mehranian A et al. Deep Learning for PET Image Reconstruction. IEEE Trans. Radiat. Plasma Med. Sci 2021; 5 (1): 1–25. doi:10.1109/TRPMS.2020.3014786

[6] Willemink MJ, Noël PB. The evolution of image reconstruction for CT – from filtered back projection to artificial intelligence. Eur. Radiol 2019; 29 (5): 2185–2195. doi:10.1007/s00330-018-5810-7

[7] Rodriguez-Ruiz A et al. Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. Eur. Radiol 2019; 29: 4825–4832

[8] McKinney SM et al. International evaluation of an AI system for breast cancer screening. Nature 2020; 577: 89–94. doi:10.1038/s41586-019-1799-6

[9] Kooi T et al. Large scale deep learning for computer aided detection of mammographic lesions. Med. Image Anal 2017; 35: 303–312. doi:10.1016/j.media.2016.07.007

[10] Gu Z et al. CE-Net: Context Encoder Network for 2D Medical Image Segmentation. IEEE Trans. Med. Imaging 2019; 38 (10): 2281–2292. doi:10.1109/TMI.2019.2903562

[11] Huang H et al. UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation. in ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain 2020. doi:10.1109/ICASSP40776.2020.9053405

[12] Zhou Z, Siddiquee MMR, Tajbakhsh N et al. UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. IEEE Trans. Med. Imaging 2020; 39 (6): 1856–1867. doi:10.1109/TMI.2019.2959609

[13] Bera K, Braman N, Gupta A et al. Predicting cancer outcomes with radiomics and artificial intelligence in radiology. Nat. Rev. Clin. Oncol 2022; 19 (2): 132–146

[14] Shin J et al. MRI radiomics model predicts pathologic complete response of rectal cancer following chemoradiotherapy. Radiology 2022; 303 (2): 351–358

[15] Lisson CS et al. Deep Neural Networks and Machine Learning Radiomics Modelling for Prediction of Relapse in Mantle Cell Lymphoma. Cancers 2022; 14 (8): 2008. doi:10.3390/cancers14082008

[16] Guo Y, Liu Y, Oerlemans A et al. Deep learning for visual understanding: A review. Neurocomputing 2016; 187: 27–48. doi:10.1016/j.neucom.2015.09.116

[17] Feng X, Jiang Y, Yang X et al. Computer vision algorithms and hardware implementations: A survey. Integration 2019; 69: 309–320. doi:10.1016/j.vlsi.2019.07.005

[18] Kiryati N, Landau Y. Dataset Growth in Medical Image Analysis Research. J. Imaging 2021; 7 (8): 155. doi:10.3390/jimaging7080155

[19] Thompson NC, Greenewald K, Lee K et al. The Computational Limits of Deep Learning, MIT INITIATIVE ON THE DIGITAL ECONOMY RESEARCH BRIEF Vol. 4, Sep. 2020.

[20] He J, Baxter SL, Xu J et al. The practical implementation of artificial intelligence technologies in medicine. Nat. Med 2019; 25 (1): 30–36. doi:10.1038/s41591-018-0307-0

[21] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat. Med 2019; 25 (1): 44–56. doi:10.1038/s41591-018-0300-7

[22] Tonekaboni S, Joshi S, McCradden MD et al. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. Proceedings of the 4th Machine Learning for Healthcare Conference 2019; 106: 359–380

[23] Götz M, Maier-Hein KH. Optimal Statistical Incorporation of Independent Feature Stability Information into Radiomics Studies. Sci. Rep 2020; 10 (1): 737. doi:10.1038/s41598-020-57739-8

[24] Zwanenburg A et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology 2020; 295 (2): 328–338. doi:10.1148/radiol.2020191145

[25] Zech JR, Badgeley MA, Liu M et al. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS Med 2018; 15 (11): e1002683. doi:10.1371/journal.pmed.1002683

[26] Nasief H et al. A machine learning based delta-radiomics process for early prediction of treatment response of pancreatic cancer. Npj Precis. Oncol 2019; 3 (1): 25. doi:10.1038/s41698-019-0096-z

[27] Wood A, Shpilrain V, Najarian K et al. Private naive bayes classification of personal biomedical data: Application in cancer data analysis. Comput. Biol. Med 2019; 105: 144–150. doi:10.1016/j.compbiomed.2018.11.018

[28] Masoud Rezaeijo S, Ghorvei M, Alaei M. A machine learning method based on lesion segmentation for quantitative analysis of CT radiomics to detect COVID-19. 6th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS) 2020: 1–5. doi:10.1109/ICSPIS51611.2020.9349605

[29] Chaddad A, Zinn PO, Colen RR. Radiomics texture feature extraction for characterizing GBM phenotypes using GLCM. IEEE 12th International Symposium on Biomedical Imaging (ISBI) 2015: 84–87. doi:10.1109/ISBI.2015.7163822

[30] Haniff NSM, Karim MKBA, Ali NS et al. Magnetic Resonance Imaging Radiomics Analysis for Predicting Hepatocellular Carcinoma. International Congress of Advanced Technology and Engineering (ICOTEN), Taiz, Yemen 2021: 1–5. doi:10.1109/ICOTEN52080.2021.9493533

[31] Wu Q et al. Radiomics analysis of magnetic resonance imaging improves diagnostic performance of lymph node metastasis in patients with cervical cancer. Radiother. Oncol 2019; 138: 141–148. doi:10.1016/j.radonc.2019.04.035

[32] Loyola-Gonzalez O. Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View. IEEE Access 2019; 7: 154096–154113. doi:10.1109/ACCESS.2019.2949286

[33] Leong MC, Prasad DK, Lee YT et al. Semi-CNN Architecture for Effective Spatio-Temporal Learning in Action Recognition. Appl. Sci 2020; 10 (2): 557. doi:10.3390/app10020557

[34] Nguyen A, Dosovitskiy A, Yosinski J et al. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. Adv. Neural Inf. Process. Syst 2016; 29. doi:10.48550/arXiv.1605.09304

[35] Dosovitskiy A, Brox T. Inverting visual representations with convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition 2016: 4829–4837

[36] Zeiler MD, Fergus R. Visualizing and Understanding Convolutional Networks. Computer Vision – ECCV 2014 D. Fleet, T. Pajdla, B. Schiele, und T. Tuytelaars, Hrsg. Cham: Springer International Publishing, 2014; 8689: 818–833. doi:10.1007/978-3-319-10590-1_53

[37] Park SJ, An KH, Lee M. Saliency map model with adaptive masking based on independent component analysis. Neurocomputing 2002; 49 (1/04): 417–422. doi:10.1016/S0925-2312(02)00637-9

[38] Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, in 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings, 2014.

[39] Adebayo J, Gilmer J, Muelly M et al. Sanity checks for saliency maps. Adv. Neural Inf. Process. Syst 2018; 31. doi:10.48550/arXiv.1810.03292

[40] DeGrave AJ, Janizek JD, Lee SI. AI for radiographic COVID-19 detection selects shortcuts over signal. Nat. Mach. Intell 2021; 3 (7): 610–619. doi:10.1038/s42256-021-00338-7

[41] Kim B et al. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). Proceedings of the 35th International Conference on Machine Learning 2018; 80: 2668–2677

[42] Chen C, Li O, Tao D et al. This looks like that: deep learning for interpretable image recognition. Adv. Neural Inf. Process. Syst 2019; 32. doi:10.48550/arXiv.1806.10574

[43] Li O, Liu H, Chen C et al. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. Proceedings of the AAAI Conference on Artificial Intelligence 2018; 32 (1). doi:10.48550/arXiv.1710.04806

[44] Adler TJ et al. Uncertainty-aware performance assessment of optical imaging modalities with invertible neural networks. Int. J. Comput. Assist. Radiol. Surg 2019; 14 (6): 997–1007. doi:10.1007/s11548-019-01939-9