Highlights

A Survey on XAI methods in Neuroimaging

S Falciglia

- Limited transparency and explainability of deep learning models have prevented their adoption across the sciences. This has resulted in simpler models being used, which are easier to interpret but have less predictive power.
- We categorize post-hoc XAI methods into two categories: visualization and explanation by example. Visualization techniques are further classified as perturbation-based and gradient/backpropagation-based.
- Patients may be more likely to use and trust a system if they can comprehend the logic behind a model as perceived through their own chain of reasoning.

A Survey on XAI methods in Neuroimaging

S Falciglia[a,1]

[a]Department of Computer, Control and Management Engineering, Via Ariosto 25, Rome, 00185, Italy,

ARTICLE INFO

Keywords:
neuroimaging
explainable AI
visualization
explanation by example

ABSTRACT

Medicine needs Deep Learning (DL) models to enhance diagnosis accuracy. Despite the improvement they offer in diagnosis, deep neural networks have limited transparency and explainability, which makes society wary of relying solely on their black-box predictions to make decisions about human health and life. To overcome the next bottlenecks in medicine and identify particular structures in biomedical images that are not discernible to the human eye or simple machine learning models, Explainable AI (XAI) becomes critical. New applications of DL models are emerging in the field of neuroimaging, where techniques such as magnetic resonance imaging (MRI), functional MRI (fMRI), computerized tomography (CT), and ultrasound are attracting XAI researchers. This work aims to conduct a comprehensive survey of the state-of-the-art methods of Explainable AI (XAI) in neuroimaging. Some critical examples of their usage will be presented to demonstrate the usefulness of these methods in diagnosis.

1. Introduction

Artificial intelligence (AI) can improve biomedical imaging by revealing additional and more detailed information about the region of interest (RoI) than, for instance, human radiologists [1]. However, for the regulated healthcare domain, it is paramount to understand, justify and explain the predictions of AI models for the wider adoption of automated diagnosis. DARPA Explainable AI (XAI) program, launched in May 2017, defines explainable AI as “AI systems that can explain their rationale to a human user, characterize their strengths and weakness, and convey an understanding of how they will behave in the future” [2]. Moreover, “Explanation refers to the ability to accurately describe the mechanism, or implementation, that led to an algorithm’s output, often so that the algorithm can be improved in some way” [3].

Biomedical imaging covers a wide range of image modalities, such as conventional radiography and ultrasound. Datasets on medical image acquisition, DNN models and XAI techniques cover a broad range of disease studies and include specialties such as radiology, where there is a high demand for the application of XAI techniques [4]. Remarkably, there is a growing demand across healthcare and medicine to develop AI approaches that perform well and guarantee transparency and interpretability for medical experts [5, 6], whose knowledge in interpreting the machine learning (ML) results – i.e., the human-in-the-loop concept – would be critical, leading eventually to user trust [7, 8]. In particular, with the emerging field of neuroimaging and its extensive use in deep learning (DL) studies, techniques such as magnetic resonance imaging (MRI), functional MRI (fMRI), computerized tomography (CT), and ultrasound have attracted considerable interest from XAI researchers [9, 10].

This work aims to provide a comprehensive survey on XAI methods in neurology, focusing on post-hoc methods for practitioners and researchers working with deep neural networks and neuroimaging techniques. The rest of the paper is organized as follows. Section 2 introduces the concept of explainability in biomedical imaging. In Section 3, we categorize modern explainability techniques into classes and highlight their application in the field of neurology. In Section 4 we focus on recent case studies applying some of the state-of-the art methods described previously. Finally, Section 5 draws the conclusions of this study.

2. Explainability in Biomedical Imaging

AI systems have been categorized as (1) opaque, where the input-output mappings are invisible to the user, (2) interpretable, where a user can both see and understand how the inputs are mapped to the output, and (3) comprehensible, where some additional output such as visualizations, text, etc. is provided [11].

Deep neural networks (DNNs) are opaque systems that outperform standard AI systems in terms of performance and accuracy but are not interpretable. The limited transparency and explainability of such non-linear methods have prevented their adoption across the sciences, resulting in greater popularity of simpler models (e.g. shallow decision trees, linear regression, or non-negative matrix factorization) with higher interpretability than complex models in many applications, including bioinformatics and neuroscience, despite the fact that these choices often reduce predictive power [12].

Therefore, XAI systems are designed to depend on knowledge that is machine-processable and oriented towards human language, i.e. symbolic systems. An XAI system derives its conclusions (decisions/diagnoses) by employing scientific reasoning methods [13]. Specifically, the decision steps are directly provided by ’symbolic’ machine learning methods. These methods use a set of hierarchically or non-hierarchically organized rules to assign a class to a given case [14, 15]. However, for a system to be comprehensible to a domain expert, it must use simple information, a quality

✉ falciglia.2015426@studenti.uniroma1.it (S. Falciglia)
ORCID(s): 0009-0009-4871-2553 (S. Falciglia)

that is often lacking in machine learning symbolic systems. For instance, explanations that include decision trees with hundreds of conditions and sub-trees lie beyond human understanding. According to Miller's Law, the standard limit of the human capacity for processing information is 7 ± 2 elements [16]. Thus, XAI explanations should be as simple as they can be (i.e. Occam's Razor) and rely on abstractions (generalizations) from example situations.

As AI models and their intended purposes vary widely, it is unlikely a single explainability strategy would work effectively for different stakeholders involved in the systems [17]. XAI has major stakeholders that include (1) AI experts and developers, (2) regulators, (3) medical practitioners, and (4) patients [18]. AI experts would require much more detailed information compared to other stakeholders to improve the model's performance [19]. Regulators are primarily interested in the overall performance of the model, along with a combination of model explanations, rather than specific cases. They also have a keen interest in XAI systems to set standards and validate and certify models [20]. Physicians wish to understand the model better by asking why a particular decision was made, and how it can validate their own diagnosis. Users are more likely to use and trust a system if they can comprehend the logic behind a model's prediction [19]. Patients may be concerned about the validity of a model's decision-making process for their condition and may require assurance of its trustworthiness. They would also be interested in comprehending the logic behind it as perceived through their own chain of reasoning.

The so-called post-hoc explainable AI methods [21] are useful for researchers and practitioners working with deep neural networks and neuroimaging techniques. These methods extract information about the model input-decisions mappings from a fitted and trained model without affecting its performance. In contrast to post-hoc approaches, model-based approaches modify the model to allow for mechanistic (functional) or archetypal explanations. Our focus is on post-hoc approaches as the interpretation of weight vectors has historically been the standard practice when applying encoding and decoding models to functional imaging and neuroimaging data [22]. Post-hoc procedures provide a novel strategy that can allow the application of predictive techniques and the generation of scientific and/or neuroscientific knowledge during the interpretation step.

3. XAI Methods in Neuroimaging

This section presents various state-of-the-art techniques developed to comprehend the model classification executed by a DNN at the pixel, superpixel, or feature level. These techniques are divided into two categories: visualization and explanation by example.

The focus of visualization techniques is to present the model's internal mechanisms as data visualizations, such as a heat map that displays the hidden information about the significance of features using colors. These techniques can be further classified as perturbation or gradient/back-propagation-based approaches.

Explanations by examples help to explain the model by extracting examples similar to a prediction, generating additional features such as text, or using further details.

3.1. Visualization by Perturbation

Perturbation-based visualization approaches alter the input to a DNN model to determine the importance of the features, pixels, or image areas for the prediction.

Local Interpretable Model-agnostic Explanation (LIME) [23] has been proposed as an interpretable and faithful explanation technique for the predictions of any classifier. This is accomplished by locally developing an interpretable model around the prediction. To ensure interpretability and local consistency, the objective is to minimize a loss function comprising two components: a measure of the model's complexity (which influences its interpretability), and a measure of the extent to which the model fails to approximate its own explanation function around a sample (which influences the faithfulness of the explanation). Furthermore, as the explainer is model-agnostic, no assumptions about the explanation function need to be made when minimizing the loss function. Therefore, the final optimization problem is solved by sampling instances around the given input of the model. This is achieved by randomly drawing non-zero elements of the input with uniform probability, with the number of draws also uniformly sampled. This method is quite resilient to sampling noise because the samples are weighted by a locality function that defines the considered neighborhood of the input. In neuroimaging, LIME has been utilized to provide a plausible explanation for the precise classifications of a given DaTSCAN[1] as Parkinson's disease (PD) or non-PD [25]. This is accomplished by using visual superpixels on the input images.

Hierarchical Occlusion (Hiho) [26] is an explanation technique that expands the standard neighborhood concept on a pixel-wise basis. Hiho assesses occlusion sensitivity in a hierarchical manner, reasoning that crucial features are usually localized within an image, and that feature aggregates affecting model predictions exist on several scales. By doing so, regions within images, not significant for assessing feature importance, can be rapidly eliminated, resulting in faster identification of essential feature sets. The algorithm creates an array of all occluded copies of the input image and processes it with a classification model. The result is a score (feature importance) stored in the occluded region of each image, which is calculated based on the difference between the confidence levels of the input image's classification and the occluded one's. The array is aggregated into a single image to implement the algorithm on a smaller region of

[1] DaTSCAN. Here used improperly, it refers to the input image. DaTSCAN is a diagnostic radiopharmaceutical used in nuclear medicine imaging to assess the integrity and function of dopamine transporters in the brain. It is administered to a patient intravenously, in order to be taken up by the dopamine transporters in the brain, such that it is possible to detect its distribution using a gamma camera or SPECT imaging [24].

interest. The iteration takes place until the feature importance within that region does not change anymore. Hiho has been demonstrated to be effective in visualizing the most salient features of brainstem images in neuroimaging, based on a cohort of 151 patients from the Parkinson's Progression Markers Initiative (PPMI) DW-MRI[2] study []. It produces results over 1000 times faster than the traditional occlusion sensitivity pixel-wise algorithm and 20 times faster than the Grad-CAM visualization method (see Section 3.2).

3.2. Visualization by Gradient/Backpropagation

Saliency map-based post-hoc explanation methods propose easily interpretable maps for the input images, which are actually heatmaps usually superimposed on the input to highlight the pixels more important for the prediction [27]. In neuroimaging, various saliency map explanation methods have been used for several applications: fetal head circumference estimation [28], Parkinson's disease detection and classification [29], tumor grading and identification of the tumor location [30, 31], automatic classification of dopamine transporter DAT-SPECT[3] images [32], ictal[4] prediction [33], brain aging [34].

Image-specific class saliency, based on the class score derivative [27], is a technique that computes the vector derivative through a single back-propagation pass and rearranges the elements of the resulting vector. For an RGB image, a single class saliency value for each pixel is derived by taking the maximum magnitude of the vector derivative across all color channels.

The Class Activation Map (CAM) approach [35] generates the output maps by using global average pooling, which is also known as the GAP layer. A class activation map for a particular category shows the discriminative image regions used by the CNN to identify that category. It is the weighted sum of the spatial averages of feature maps for each unit at the last convolutional layer.

Grad-CAM [36] is a technique that generalizes CAM and is applicable to a considerably wider variety of CNN model families. The approach makes use of gradient information coming into the last convolutional layer of the CNN to determine the value of each neuron in making an important choice. This is achieved by assigning a neuron importance weight to each one of them. Subsequently, a weighted combination of forward activation maps is executed, which is followed by a ReLU to acquire a coarse heat map of the same dimensions as the convolutional feature maps. By applying ReLU to the linear combination of maps, we may capture just those characteristics that have a positive effect on the class of interest. These features correspond to the pixels whose intensity needs to be amplified in order to augment the prediction's likelihood.

The layer-wise relevance propagation (LRP) technique [37] aims to determine the contribution of a single pixel in an image to the prediction of a classifier using a set of constraints that the solution must meet. In a feed-forward network, the total relevance of the input is preserved from one layer to another, and the total node relevance equals the sum of all relevance messages sent to the same node. By considering the pixels with positive relevance as critical for output classification, it is possible to map the relevance of each input pixel to a color space and obtain a conventional heatmap.

Deep Learning Important Features (DeepLIFT) [38] is another technique that assigns a significant score to input variables. These scores can be efficiently computed in a single backward pass, by tracking the change in the output layer relative to the input layer and comparing the activation of each neuron to its reference activation, thus assigning contribution scores based on the difference. Considering positive and negative contributions, interesting dependencies can be revealed.

The Deconvnet [39] utilizes a multi-layered Deconvolutional Network to project feature activations back to the input pixel space. In contrast to the utilization proposed in the original paper [40], a Deconvnet is attached to each layer of an already trained convnet and used as a probe. Examining a particular convnet activation involves setting all non-target activations in the layer to zero and passing the input feature maps to the linked Deconvnet layer. Then, we proceed to unpool, rectify and filter successively to reconstruct the activity in the lower layer that created the selected activation. This method only enables the visualization of a single activation rather than the collective activities of a layer. The visuals produced, on the other hand, are precise representations of the input pattern that stimulates the specific feature map in the model.

Shapley additive explanations (SHAP) [41] is a technique used for understanding model predictions. It assigns an importance value to each feature by combining three other measures of feature importance, derived from cooperative game theory. These measures include Shapley regression values [42], Shapley sampling values [43], and Quantitative Input Influence [44]. Computing the final SHAP values exactly is challenging, and this has led to the development of different approaches to approximate these values, such as model-agnostic (Kernel SHAP) and model-type-specific (Deep SHAP) methods [41]. This approach has been found to be more consistent with human understanding and intuition.

3.3. Explanation by Example

Interactive 'what-if' questions, or counterfactual questions, are gaining importance in the field of Explainable AI (XAI). The aim is to create a representation space of features, encompassing multiple modes, by employing knowledge bases. This can lead to the development of innovative explanation techniques [45].

[2] DW-MRI. Diffusion-Weighted Magnetic Resonance Imaging is a neuroimaging technique.

[3] DAT-SPECT. Dopamine Transporter Single-Photon Emission Computed Tomography isa specific type of nuclear imaging.

[4] Ictal. The ictal phase is the middle stage of a seizure, the time from the first symptom to the end of the seizure activity.

In recent research [45], graph neural networks (GNNs) have been proposed as a preferred method for enabling information fusion for multi-modal causality, quantifying the degree to which an explanation achieves the desired level of causal comprehension for a human expert. GNNs play a critical role in this context as they allow the establishment of causal connections directly between characteristics, by exploiting graph structures. In neuroimaging, it has been observed that constructing a correlation-based Graph Convolutional Network (GCN) model across dissimilar sample groups in the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, and then studying the GNNExplainer's decision-making mechanism [45], has identified unique groups of biomarkers, thereby illuminating the heterogeneity of Alzheimer's Disease progression.

4. Case Studies

This section discusses four different case studies relevant to Parkinson's, Alzheimer's, multiple sclerosis, and fetal brain abnormalities. The data for these studies was collected using various neuroimaging techniques including magnetic resonance imaging (MRI), ultrasound, and nuclear imaging. Deep neural networks (DNNs) have been extensively employed to deepen our understanding of these diseases, while XAI techniques enable the explanation of these models.

Parkinson Parkinson's disease (PD) is a neurodegenerative condition that mainly affects the aging population. PD is mainly characterized by the loss of dopaminergic neurons in the substantia nigra pars compacta (SNc) [46].

In 2019, Shinde et al. [46] conducted a study using Convolutional Neural Networks (CNNs) to create biomarkers for Parkinson's disease from Neuromelanin Sensitive Magnetic Resonance Imaging (NMS-MRI). The study also applied Class Activation Mapping (CAM) to identify the most relevant and discriminative image areas. To obtain discriminative activations, the input image was forward propagated, and the weights at the output layer for the respective class were acquired. To match the resolution of the original input image, the feature maps obtained from the last convolutional layer were upsampled. The class activation map was created by multiplying the weights of the respective class with the corresponding feature maps and adding them together. The resulting map showcases the most distinguishable regions in the image for each subject. Lastly, to assess any contralateral deficits in Parkinson's Disease, as indicated in previous studies, the mean activations were separately computed for the left and right of each correctly classified subject by dividing the input image into two parts. We conducted an analysis of variance (ANOVA) test to compare the average activations on the left and right side of all participants. The results showed that for most patients, the activations were more concentrated on the left SNc, indicating a significant trend with a p-value of 0.09.

In a 2021 study by T. Pianpanit et al. [47], four DNN architectures were examined with six common interpretation methods for SPECT images in detecting Parkinson's disease. The DICE coefficient was used to evaluate the interpretation on the Parkinson's Progression Markers Initiative (PPMI) database. This article aims to be a tutorial demonstrating the procedure for selecting an appropriate interpretation method for the PD recognition model. Evaluating the interpretation methods applied to the four DCNN architectures, presented as an example, shows that the guided backpropagation and SHAP interpretation methods are suitable for PD recognition methods in different ways. The capacity to display fine-grained relevance is particularly obvious in guided backpropagation, as seen by the greatest DICE coefficient and the lowest mean square error. On the other hand, SHAP produces higher quality heatmaps at the uptake depletion site, which surpasses other methods in distinguishing the difference between PD and normal control (NC) subjects.

Alzheimer Alzheimer's disease (AD) is a neurodegenerative condition that affects approximately 10 million people worldwide annually [48].

In a study conducted in 2021 by S. Kamal et al. [48], multimodal detection of Alzheimer's disease was achieved by combining image and gene expression data. While CNN and SpinalNet were employed for the MRI images, KNN, Support Vector Classifier (SVC), and Xboost methods were used for the microarray gene expression data. Using LIME for explainability, we identified the significant genes. LIME illustrates the gene selection process for a specific AD patient, and determines the most important genes from their gene expression data. The XAI method describes the role of genes, which enables the ranking of genes based on probability values. Improved accuracies resulted from the use of SVC for prediction.

Multiple Sclerosis Multiple sclerosis is a neurological condition that is diagnosed primarily by clinical symptoms and signs of damage to the central nervous system (CNS). An MRI scan is commonly used to detect lesions in the white matter of the brain [49].

In a 2020 study by A. Lopatina et al. [49], a Convolutional Neural Network (CNN) was used for the classification of MRI images, and multiple attribution algorithms were utilized to generate heatmaps that displayed each voxel's contribution to the prediction of the class. To compare the heatmaps generated by the LRP, DeepLIFT, and saliency map methods, perturbation analysis was used. The values of the Area Over Prediction Curve (AOPC) were calculated for all 66 images from the test dataset, which was specifically created for this study. During each perturbation step, ten regions sized at 10x10 voxels were substituted with random values drawn uniformly from a distribution. The heatmap values dictate the order of perturbation, starting from the most positively relevant ones for prediction and finishing with the most negatively relevant ones. Precisely, 34.5% of the image was perturbed by replacing the first 130 regions in 13 steps, assuming that this disturbance has a sufficient impact on the brain region containing critical information for categorization. LRP and DeepLIFT exhibit the most significant AOPC values, with DeepLIFT performing marginally better after a few perturbation steps. As the saliency map's focus is restricted to identifying local relevance, its performance is poorer; however, it still outperforms the random baseline. Furthermore, the selection of the reference input for the DeepLIFT algorithm affects the results. The most promising results were obtained using a blurred reference input image of the original image.

Fetal Brain Abnormalities Fetal brain abnormalities are among the most common congenital malformations and may contribute to mental retardation and neurodevelopmental delay [50].

According to a 2020 study by B. Xie et al. [50], lesions are likely to be associated with the regions of interest (ROIs), so the localization of lesions could be solved by searching the ROIs. The authors identified regions that have a significant impact on the final score by producing a class activation map using Grad-CAM and using that to localize the lesions present in the input ultrasound images. Experts annotated the ground truth boxes, and the predicted boxes were generated by fitting the ROI of the network in abnormal ultrasound images. The researchers compared 729 test abnormal ultrasound images by comparing the predicted lesion localization boxes generated by the network with the ground truth boxes. The study calculated the mean and standard deviation of the intersection over union (IOU) metric. The result indicates the ability of the algorithms to localize lesions. However, there is a need for improvement to more accurately determine the edges of the lesions.

5. Conclusions

This survey presents a discussion on the use of XAI methods in neuroimaging. The concept of explainability for biomedical images is introduced, followed by the presentation of state-of-the-art XAI methods and their applications to enable more reliable diagnosis of several neurodegenerative diseases.

Health stakeholders, such as patients, physicians, pharmaceutical companies, and the government, consider interpretability to be an integral part of selecting the best model. Despite its significance, XAI is still undervalued, particularly in the medical field. In situations where human health and lives are in balance, it is not enough to make a decision based on a "black box" prediction, even if it comes from a superhuman model. Classification alone is insufficient; instead, interpretation is crucial for XAI to provide a comprehensive and thorough description of the voxels comprising a tumor.

XAI has exceptional potential in medicine, as it can provide an explanation supporting the accuracy of a diagnosis.

CRediT authorship contribution statement

S Falciglia: Conceptualization of this study.

References

[1] E. Sorantin, M. G. Grasser, A. Hemmelmayr, S. Tschauner, F. Hrzic, V. Weiss, J. Lacekova, A. Holzinger, The augmented radiologist: artificial intelligence in the practice of radiology, Pediatric Radiology (2021) 1–13.

[2] D. Gunning, D. Aha, Darpa's explainable artificial intelligence (xai) program, AI magazine 40 (2019) 44–58.

[3] D. A. Broniatowski, et al., Psychological foundations of explainability and interpretability in artificial intelligence, NIST, Tech. Rep (2021).

[4] L. Eldridge, What is radiology? understanding diagnostic, interventional, and therapeutic radiology, [Online] (Accessed 9 August 2023) (2021).

[5] P. K. Douglas, S. Harris, A. Yuille, M. S. Cohen, Performance comparison of machine learning algorithms and number of independent components used in fmri decoding of belief vs. disbelief, Neuroimage 56 (2011) 544–553.

[6] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9 (2019) e1312.

[7] A. Holzinger, Interactive machine learning for health informatics: when do we need the human-in-the-loop?, Brain Informatics 3 (2016) 119–131.

[8] L. C. Magister, D. Kazhdan, V. Singh, P. Liò, Gcexplainer: Human-in-the-loop concept-based explanations for graph neural networks, arXiv preprint arXiv:2107.11889 (2021).

[9] G. Zhu, B. Jiang, L. Tong, Y. Xie, G. Zaharchuk, M. Wintermark, Applications of deep learning to neuro-imaging techniques, Frontiers in neurology 10 (2019) 869.

[10] B. H. Van der Velden, H. J. Kuijf, K. G. Gilhuijs, M. A. Viergever, Explainable artificial intelligence (xai) in deep learning-based medical image analysis, Medical Image Analysis 79 (2022) 102470.

[11] D. Doran, S. Schulz, T. R. Besold, What does explainable ai really mean? a new conceptualization of perspectives, arXiv preprint arXiv:1710.00794 (2017).

[12] G. Bologna, Y. Hayashi, Characterization of symbolic rules embedded in deep dimlp networks: a challenge to transparency of deep learning, Journal of Artificial Intelligence and Soft Computing Research 7 (2017) 265–286.

[13] W. Hodges, Classical logic i: First-order logic, The Blackwell Guide to Philosophical Logic (2017) 9–32.

[14] J. R. Quinlan, Induction of decision trees, Machine learning 1 (1986) 81–106.

[15] W.-Y. Loh, Fifty years of classification and regression trees, International Statistical Review 82 (2014) 329–348.

[16] G. A. Miller, The magical number seven, plus or minus two: Some limits on our capacity for processing information., Psychological review 63 (1956) 81.

[17] A. Páez, The pragmatic turn in explainable artificial intelligence (xai), Minds and Machines 29 (2019) 441–459.

[18] S. Nazir, D. M. Dickson, M. U. Akram, Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks, Computers in Biology and Medicine (2023) 106668.

[19] A. Preece, D. Harborne, D. Braines, R. Tomsett, S. Chakraborty, Stakeholders in explainable ai, arXiv preprint arXiv:1810.00184 (2018).

[20] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information fusion 58 (2020) 82–115.

[21] F. V. Farahani, K. Fiok, B. Lahijanian, W. Karwowski, P. K. Douglas, Explainable ai: A review of applications to neuroimaging data, Frontiers in Neuroscience 16 (2022) 906290.

[22] N. Kriegeskorte, P. K. Douglas, Cognitive computational neuroscience, Nature neuroscience 21 (2018) 1148–1160.

[23] M. T. Ribeiro, S. Singh, C. Guestrin, " 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, 2016, pp. 1135–1144.

[24] R. de la Fuente-Fernández, Role of datscan and clinical diagnosis in parkinson disease, Neurology 78 (2012) 696–701.

[25] P. R. Magesh, R. D. Myloth, R. J. Tom, An explainable machine learning model for early detection of parkinson's disease using lime on datscan imagery, Computers in Biology and Medicine 126 (2020) 104041.

[26] W. S. Monroe, F. M. Skidmore, D. G. Odaibo, M. M. Tanik, Hiho: accelerating artificial intelligence interpretability for medical imaging in iot applications using hierarchical occlusion: Opening the black box, Neural Computing and Applications 33 (2021) 6027–6038.

[27] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, arXiv preprint arXiv:1312.6034 (2013).

[28] J. Zhang, C. Petitjean, S. Ainouz, Segmentation-based vs. regression-based biomarker estimation: a case study of fetus head circumference assessment from ultrasound images, Journal of Imaging 8 (2022) 23.

[29] S. Chakraborty, S. Aich, H.-C. Kim, Detection of parkinson's disease from 3t t1 weighted mri scans using 3d convolutional neural network, Diagnostics 10 (2020) 402.

[30] S. Pereira, R. Meier, V. Alves, M. Reyes, C. A. Silva, Automatic brain tumor grading from mri data using convolutional neural networks and quality assessment, in: Understanding and Interpreting Machine Learning in Medical Image Computing Applications: First International Workshops, MLCN 2018, DLF 2018, and iMIMIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16-20, 2018, Proceedings 1, Springer, 2018, pp. 106–114.

[31] P. Windisch, P. Weber, C. Fürweger, F. Ehret, M. Kufeld, D. Zwahlen, A. Muacevic, Implementation of model explainability for a basic brain tumor detection using convolutional neural networks on mri slices, Neuroradiology 62 (2020) 1515–1518.

[32] M. Nazari, A. Kluge, I. Apostolova, S. Klutmann, S. Kimiaei, M. Schroeder, R. Buchert, Explainable ai to improve acceptance of convolutional neural networks for automatic classification of dopamine transporter spect in the diagnosis of clinically uncertain parkinsonian syndromes, European journal of nuclear medicine and molecular imaging (2022) 1–11.

[33] V. Gabeff, T. Teijeiro, M. Zapater, L. Cammoun, S. Rheims, P. Ryvlin, D. Atienza, Interpreting deep learning models for epileptic seizure detection on eeg signals, Artificial intelligence in medicine 117 (2021) 102084.

[34] A. Lombardi, D. Diacono, N. Amoroso, A. Monaco, J. M. R. Tavares, R. Bellotti, S. Tangaro, Explainable deep learning for personalized age prediction with brain morphology, Frontiers in neuroscience 15 (2021) 578.

[35] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2921–2929.

[36] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[37] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PloS one 10 (2015) e0130140.

[38] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: International conference on machine learning, PMLR, 2017, pp. 3145–3153.

[39] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13, Springer, 2014, pp. 818–833.

[40] M. D. Zeiler, D. Krishnan, G. W. Taylor, R. Fergus, Deconvolutional networks, in: 2010 IEEE Computer Society Conference on computer vision and pattern recognition, IEEE, 2010, pp. 2528–2535.

[41] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).

[42] S. Lipovetsky, M. Conklin, Analysis of regression in game theory approach, Applied Stochastic Models in Business and Industry 17 (2001) 319–330.

[43] E. Štrumbelj, I. Kononenko, Explaining prediction models and individual predictions with feature contributions, Knowledge and information systems 41 (2014) 647–665.

[44] A. Datta, S. Sen, Y. Zick, Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems, in: 2016 IEEE symposium on security and privacy (SP), IEEE, 2016, pp. 598–617.

[45] A. Holzinger, B. Malle, A. Saranti, B. Pfeifer, Towards multi-modal causability with graph neural networks enabling information fusion for explainable ai, Information Fusion 71 (2021) 28–37.

[46] S. Shinde, S. Prasad, Y. Saboo, R. Kaushick, J. Saini, P. K. Pal, M. Ingalhalikar, Predictive markers for parkinson's disease using deep neural nets on neuromelanin sensitive mri, NeuroImage: Clinical 22 (2019) 101748.

[47] T. Pianpanit, S. Lolak, P. Sawangjai, T. Sudhawiyangkul, T. Wilaiprasitporn, Parkinson's disease recognition using spect image and interpretable ai: A tutorial, IEEE Sensors Journal 21 (2021) 22304–22316.

[48] M. S. Kamal, A. Northcote, L. Chowdhury, N. Dey, R. G. Crespo, E. Herrera-Viedma, Alzheimer's patient analysis using image and gene expression data and explainable-ai to present associated genes, IEEE Transactions on Instrumentation and Measurement 70 (2021) 1–7.

[49] A. Lopatina, S. Ropele, R. Sibgatulin, J. R. Reichenbach, D. Güllmar, Investigation of deep-learning-driven identification of multiple sclerosis patients based on susceptibility-weighted images using relevance analysis, Frontiers in neuroscience 14 (2020) 609468.

[50] B. Xie, T. Lei, N. Wang, H. Cai, J. Xian, M. He, L. Zhang, H. Xie, Computer-aided diagnosis for fetal brain ultrasound images using deep convolutional neural networks, International Journal of Computer Assisted Radiology and Surgery 15 (2020) 1303–1312.