International Journal of Computers and Informatics

Journal Homepage: https://www.ijci.zu.edu.eg

Int. j. Comp. Info. Vol. 1 (2023) 38–52

Paper Type: Review Article

Navigating the Depths of Explainable AI (XAI): Methods, Applications, and Challenges in Neurological Diseases

Nabil M. AbdelAziz [1], Mohamed M. AbdelHafeez [1,*], Mohamed M. Hassan [1] and Asmaa H. Ali [1]

[1] Department of Information Systems, Faculty of Computers and Informatics, Zagazig University, Zagazig 44519, Egypt.
Emails: nmabedelaziz@fci.zu.edu.eg; mmabdelhafiz@fci.zu.edu.eg; monirhm2002@yahoo.com; asmaa_289@zu.edu.eg.

Received: 05 Sep 2023 **Revised**: 22 Nov 2023 **Accepted**: 19 Dec 2023 **Published**: 21 Dec 2023

Abstract

Artificial intelligence (AI) systems have been constructed as black boxes that cover their internal logic and learning approach from humans, and this has led to several unanswered questions regarding the process and rationale behind AI decisions. Explainable Artificial Intelligence (XAI) is a developing branch of AI that focuses on creating various methods and tools to unbox the inner workings of black-box AI systems. It aims to generate explanations for AI decisions that are easily understood by humans, providing insights and transparency. This paper presented a taxonomy that allows comprehensive categorization of XAI studies. The study aims to illuminate the similarities and differences among various algorithms used in XAI and highlight the characteristics, benefits, and limitations of these algorithms.

Keywords: Explainable Artificial Intelligence; XAI; Black-box; Deep Learning.

1 | Introduction

Artificial Intelligence (AI) is a subdivision of computer science that has transformed how individuals carry out their daily activities through the use of machines that require minimal human involvement, thereby enabling automated and intelligent actions. AI is considered an incredible prospect for resolving neurology disease issues, generating additional perspectives, and enhancing the quality of decision support. AI and Machine Learning (ML) are already revolutionizing several medical systems, with further advancements expected in the future [1]. The majority of AI algorithms have been referred to as 'black boxes' by researchers due to their intricate and virtual nature, making them challenging to explain and justify to individuals. A black box concept is one where the inputs and outputs are known, but you are unable to determine how the outputs are produced from the inputs. Developers are also unable to explain why the model has reached a particular conclusion or which factors were taken into consideration when making a decision. This is due to the models' intricate internal structure, and the poor offer of interpretability. The consideration of complicated models' ambiguous nature has limited their potential use in making key decisions, such as those involving medical procedures that could endanger people's lives and health. Users can accept or reject forecasts and recommendations based on the justification behind the predictions made by interpretable ML systems [2]. The existence of this particular obscurity has led

Corresponding Author: mmabdelhafiz@fci.zu.edu.eg

to a demand for algorithms in the field of Explainable Artificial Intelligence (XAI) [3]. XAI has been designed to explain its purpose, perception, and process decision in terms that the common person can understand. The concept behind XAI is that AI algorithms and systems shouldn't be "mysterious models" that are incomprehensible to humanity [4]. The development of interpretable models, methodologies, and interfaces can help to explain the behavior of ML algorithms in a form that is understandable to humans. This is part of having fairness, accountability, and transparency for the logic behind its predictions. Several XAI techniques deal with the problem of the lack of interpretability and transparency in black-box ML algorithms. Thus, the challenges of interpretability and explainability of AI algorithms have become urgent. This study aims to highlight the similarities and differences among various algorithms used for XAI and explain the characteristics, benefits, and limitations of these algorithms.

The following are the contributions of this study:

- First, this study presents and explains the fundamental concept of the XAI toward neurology diseases.
- A detailed taxonomy is presented to classify the current XAI solutions based on different categorization criteria, including stage, scope, applicability, and visualization.
- A comparative analysis of XAI approaches is designed to interpret decisions made by AI systems operating on visual explanation, textual explanation, and more.
- Finally, this study introduces some of XAI applications for neurology diseases.

The remainder of this work is arranged in the following manner. Section 2 discusses the background of XAI, especially toward neurology diseases. Section 3 discusses the taxonomy of XAI techniques. Section 4 compares explanation techniques that can be applied to neurology diseases. Section 5 introduces XAI applications that are applied in neurology diseases. Section 6 presents the pros and cons of XAI methods in the medical field. Finally, this study is concluded in section 7.

2 | Background

Transparency in ML and DL algorithms involves explaining their outcomes and decisions, which can be achieved by developing interpretable models, methods, and interfaces to provide human-understandable explanations for their behavior [5]. Interpretable ML systems provide users with explanations for accepting or rejecting predictions and recommendations, thereby enabling them to understand the reasoning behind these outcomes [6]. XAI is a new field of research in ML that examines how AI systems react to black-box decisions. The creation of recommendation systems for healthcare systems is also possible using XAI. The main cause of the rare successful integration and adoption of AI tools into clinical practice is the lack of acceptable explainability and transparency in the majority of the present AI systems. So, XAI is becoming more and more critical for DL-based driven applications, mainly in medical and clinical studies. [7].

The most recent XAI systems that are related to the neurology diseases field are presented in this section. According to the authors in [8], various XAI-enabled methods for medical diseases and XAI applications were described, along with recent and current trends in medical diagnosis and application using XAI based on findings from various research platforms. Finally, the research directions and challenges achieved were discussed. In [9], authors discussed XAI in healthcare in a multidisciplinary way to examine its importance from a legal, medical, patient, and technological perspective. The importance of XAI in the clinical system from an ethical and personal perspective is concluded by the authors after deducing a set of results for the applicability of views. In [10], the authors introduced an overview of current XAI developments and recent advancements in healthcare applications. Through the use of two descriptive clinical-level case studies, the authors demonstrate how XAI makes use of multi-modal and multi-center data fusion. According to [11], ML algorithms are interpretable and explainable, but the authors identify open challenges and opportunities in the medical context by analyzing their interpretability and explainability into two categories: perceptive interpretation and mathematical structural

interpretation. The study in [12] discusses already-developed AI methods, such as ML/DL, and expands the survey to discuss the implications of XAI in biomedical and future medical applications.

3 | Taxonomic of XAI Approaches

In this section, as observed in

Figure 1, the XAI Approaches will be classified according to four different criteria: Stage, Scope, Applicability, and Visualization. The subsequent paragraphs will investigate an examination of the four mentioned criteria.

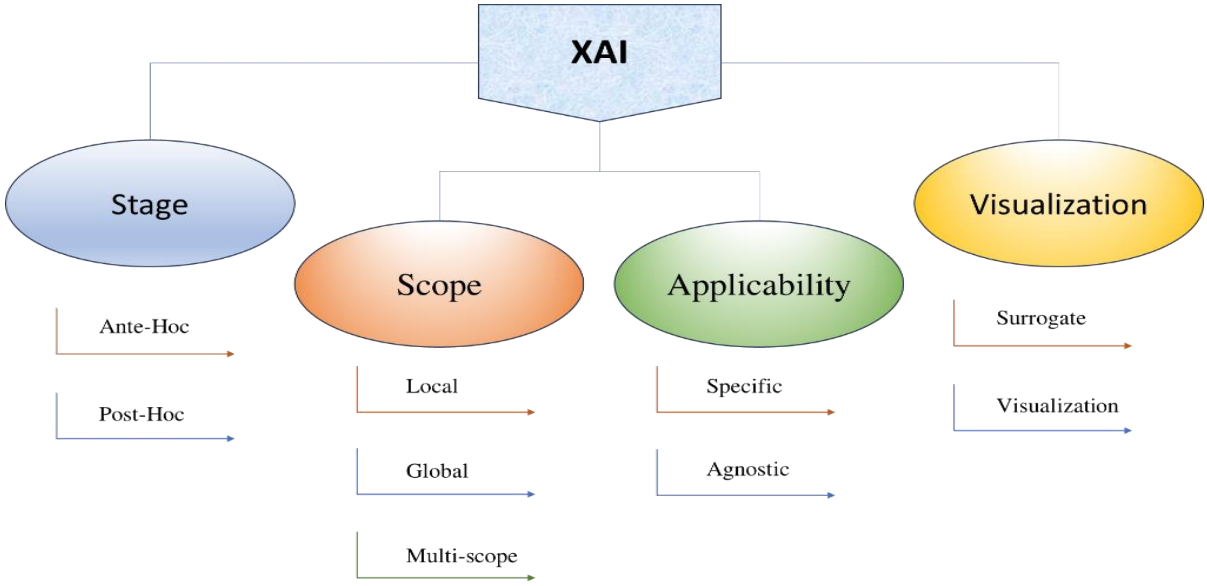

Figure 1. Taxonomy of XAI approaches.

3.1 | Stage

The ante-hoc explanation also called intrinsic explanation is achieved by putting limits on how complicated the AI methodology can be (intrinsic), where intrinsic refers to the model's ability to be interpreted at its essence. On the other hand, post-hoc explanation refers to the utilization of methods that analyze AI models after training. With this method, models can be trained, and explainability is only evaluated during testing [13].

3.1.1 | Ante-hoc explanation

Ante-hoc explanation refers to the traditional ML models such as rule-based models, decision trees, fuzzy inference systems, regression models, k-nearest neighbors, naive Bayes, and simple association rules to be understood by human experts, but more sophisticated to fit a relationship between input and output well [14].

3.1.2 | Post-hoc explanation

Post-hoc explanation attempts to provide local models for a single prediction and make it repeatable rather than explaining the behavior of the entire system. Post-hoc methods have increased in popularity in the applications of deep learning because algorithmic transparency is regarded to be unachievable for these systems [15]. Compared to an ante-hoc explanation, which trains a DNN and then tries to interpret the behavior of the resulting black box network, a post hoc explanation causes the DNN to be explicable [16].

3.2 | Applicability

The applicability of model-specific techniques is limited to specific model classes due to their reliance on the inner mechanisms of a particular model. Model-specific approaches incorporate interpretability constraints into the basic framework and training procedures of algorithm models [17]. On the other hand, model-agnostic methods use the inputs and predictions from black box models to produce explanations. They may be used with any AI model and are used once training is complete [17].

3.2.1 | Model-specific explanation

Model-specific explanation techniques are built around the individual model's parameters. The GNNExplainer is a type of model-specific explainability technique that is specifically designed to handle the intricate nature of data representation, necessitating the use of GNNs in particular [18].

3.2.2 |Model agnostic explanation

Model Agnostic approaches are not restricted to certain model architectures and are typically applied in the post-hoc investigation. The structural parameters or internal model weights cannot be accessed directly by these methods [19]. Model-agnostic explanation depends only on the input and output of the neural network. The user can analyze how the neural network's output has changed by modifying the input. Also, this explains which regions are responsible for the output [16].

3.3 |Scope

An explanation's scope determines whether it involves the whole model (global explanation) or just one single output (local explanation). Therefore, the scope of explanation is categorized into three types: 1) local explanation 2) global explanation 3) multiple-scope explanation.

3.3.1 |Local explanation

Local explanation methods apply to single-input interpretability. This can be accomplished by developing techniques that can explain a certain prediction or result [20]. Local explanation in the context of XAI refers to the interpretation of a model's decision to proceed for a particular instance or expectation. By prioritizing the key aspects or features that affected the model's decision, it attempts to provide insights into how the model arrived at a specific output for a given input [21]. The generation of local explanations has been approached using a variety of methodologies, such as LIME, SHAP, feature importance, gradient-based approaches, and counterfactual analysis [22].

3.3.2 |Global explanation

Global explanation, also known as a dataset-level explanation focuses on the whole model by applying the overall knowledge of the model and the related data [14]. It usually also explains the model's behavior. It aims to offer a thorough comprehension of the behavior and performance of the model by recognizing the key features or factors that produce an effect on the model's predictions [21].

3.3.3 |Multiple-scope explanation

In [23], the authors introduced a gradient-based method for determining the main local and global properties of the model. Multiple-scope explanation refers to the explanation of a model's decision-making process across multiple levels of abstraction, ranging from low-level features to high-level concepts or decisions [24].

3.4 |Visualization

Surrogate methods involve training a simpler, more interpretable model that approximates the behavior of a more complex, black-box model. The decision-making process of the more complex model can then be explained using the simpler model. Surrogate approaches are helpful when the underlying black-box model is too sophisticated to be understood directly or when the training data is too private or sensitive to be made public [25]. On the other hand, visualization techniques make use of visual tools like graphs, charts, and heatmaps to illustrate the connection between the model's inputs and outputs. In addition to helping users find patterns, trends, and anomalies in the data that may be important to the decision-making process of the model, visualization techniques are useful for creating both local and global explanations [26].

4. |XAI Techniques

As shown in Figure 2, The XAI techniques will be classified into four separate types, which will be more detailed in the subsequent subsections.

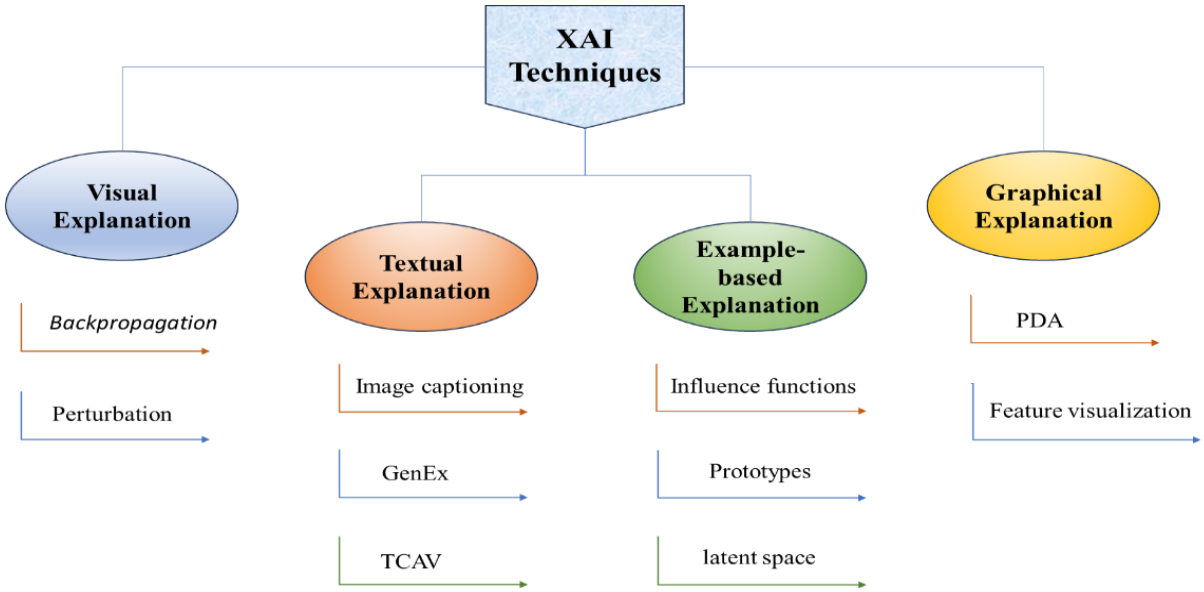

Figure 2. Four types of XAI techniques.

4.1 | Visual explanation

This section will present and explore XAI approaches designed to interpret decisions made by AI systems operating on visual data. Table 1 presents a comparison of visual explanation methods.

4.1.1 | Backpropagation-based approaches

(Guided) backpropagation and deconvolution: Guided backpropagation and deconvolution can be combined to provide more accurate and interpretable visualizations. In [27], the authors offered one such example of this combination. Grad-CAM is a technique presented by the authors that applies guided backpropagation and global average pooling to provide visualizations that emphasize the image domains essential to a specific neural network's decisions.

Class activation mapping (CAM): CAM is a visual explanation approach that generates visualizations of picture regions that are significant for a neural network's classification decision. In [34], the authors suggested an approach for generating class activation maps from the last convolutional layer of a CNN. The CAM technique highlights the regions of an image that are important for a particular class, allowing for better interpretability of the network's decision. Specifically, this approach utilizes the pooling operation before the ultimate decision layer to determine the importance of input regions. It accomplishes this by propagating the final layer's parameters in reverse through the convolutional maps of each layer within the model [28]. Consequently, the computation of the localization map is defined as follows:

$$L_{CAM}^{(c)}(x,y) = ReLU\left(\sum_n W_n^{(c)} \sum_{x,y} f_n(x,y)\right), \quad (1)$$

The weight equivalent to class c for unit n is denoted as $W_n^{(C)}$, while f_n (x, y) represents the output of unit n at the decision layer of the network.

***Gradient-weighted class activation mapping (Grad-CAM*):** Grad-CAM is an extension of the CAM method that uses gradients to compute the importance of each feature map in a CNN approach. Grad-CAM was proposed by Selvaraju et al. in 2017 [27]. The authors proposed a technique to create class activation maps using gradients, which allows for more accurate visualizations than traditional CAM. Grad-CAM has become a popular technique for visualizing the regions of an image that are important for a CNN classification decision. Grad-CAM has been used in several applications, such as medical image analysis, object detection[29], and natural language processing [30]. In object detection, Grad-CAM has been used to create heatmaps that highlight the regions of an image that contain objects[29]. In natural language processing, Grad-CAM has been used to visualize the attention of a model on different words in a sentence [30].

Table 1 Comparisons of visual explanation methods. Ex= Explainability; L= Local; G= Global; H= High; M= Medium.

XAI Method	Description	Ex-Type	Ex-Level	Agnostic?	Scalability	Data Types	Pros	Cons
LIME [22]	Produces any model's locally accurate explanations by modifying the data input.	L	H	Yes	H	Tabular/text/images	Local explanations that are simple to comprehend and interpret are provided.	Potentially fails to represent the model's global behavior.
SHAP [21]	Give each characteristic a value that represents how much it contributed to the prediction.	L/G	H	Yes	M	Tabular/text/images	Offers coherent and scientifically supported explanations for both local and global explanations.	If the data is high dimensional, it could be computationally costly.
Deep SHAP [31]	Deep SHAP is a technique that uses deep neural networks to apply Shapley values. By examining the interactions between several characteristics, it calculates the contribution of each feature to the model's output.	G	H	Yes	M	Tabular/text/images	Allows for global interpretability and can capture complicated feature relationships.	Computation-intensive and maybe not scalable to huge models or datasets.
CAM [27, 28]	CAM is a technique that creates a heatmap of class activation using the last convolutional layer of a DNN.	L	H	No	H	Images	Offers global interpretability and the ability to record complicated feature relationships.	It might not be realistic for complicated pictures.
GRAD-CAM [32, 33]	Creates visual explanations by emphasizing the areas of a picture that are crucial for the predictions.	L	H	No	H	Images	Provides visual explanations that are easy to understand and interpret.	Applied just to data in images.
Integrated Gradients [34]	By integrating the gradients from a baseline input to the actual input, this method calculates the contribution of each feature to the prediction.	L	H	Yes	H	Tabular/text/images	Consistent, theoretically supported local explanations are offered.	If the data is high dimensional, it could be computationally costly.
(Guided) Backpropagation and Deconvolution [27]	These gradient-based techniques show how each input characteristic affects the final result of the model.	L	M	No	H	Tabular/text/images	Easy to use, effective in terms of calculations.	Can't capture intricate feature relationships and might not generalize to other models or datasets.
Occlusion-based [35].	Occlusion-based techniques include masking or occluding specific input characteristics and evaluating how the outcome prediction is affected.	L/G	M	Yes	H	Tabular/text	May offer local and global explanations, and it can be used with many models.	Only discrete data can be interpreted, and high-dimensional data might be computationally costly.
LRP [36].	Calculates each neuron's contribution to the prediction by backpropagating the prediction across the network and giving each neuron a relevance score.	L/G	H	Yes	H	Tabular/text/images	Explain in detail at the neuronal level.	Depending on the rule selected for determining relevance ratings.

Grad-CAM is computed as the following:

$$L_{Grad-CAM}^{(c)}(x,y) = ReLU\left(\sum_n W_n^{(c)} \sum_{x,y} f_n(x,y)\right), \quad (2)$$

$$W_k^{(c)} = \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{j=1}^{W} \frac{\partial Y^{(c)}}{\partial f_n(i,j)}, \quad (3)$$

Layer-wise relevance propagation (LRP): The LRP method is utilized to explain the predictions made by DNNs. This methodology historically propagates the prediction score from the output layer back through the network's layers. The LRP method identifies the individual contribution of each input feature towards the final output [36]. In [37], the authors proposed a technique for explaining the predictions of DNNs that are based on the conservation of relevance principle, which aims to attribute relevance to the input features based on their contribution to the output prediction. LRP has become a popular technique for visualizing the contribution of input features to the prediction of a neural network. LRP has also been extended and improved upon in various ways. One such extension is Deep Taylor Decomposition (DTD) [36], which decomposes the relevance of each neuron into contributions from its input features. Another extension is relevance Propagation By Reconstruction (RPR) [37], which generates a reconstruction of the input features based on their relevance scores. The basic rule of the LRP method can be defined as the following.

$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k, \quad (4)$$

The variables j and k are used to denote two neurons that are members of consecutive layers. The above formula is applied in a recursive way to calculate the value of R for every neuron in the preceding layer. "a" represents the activation of a specific neuron, whereas "w" represents the weight associated with the connection between two neurons.

Deep SHaply additive explanation (Deep SHAP): Deep SHaply additive explanation (Deep SHAP) is a method for interpreting the predictions of DNNs by assigning importance values to each input feature. It's based on the Shapley value, a cooperative game theory notion that values each player's contribution to the game's outcome. Deep SHAP was proposed by Lundberg and Lee in 2017 [21]. The authors proposed a technique for explaining the predictions of DNNs that are based on the Shapley value. Deep SHAP computes the contribution of each input feature to the model's output by comparing the model's prediction with and without the feature. This allows for the identification of important features and the creation of a summary plot that shows the contribution of each feature to the model's prediction.

Trainable attention: Trainable attention is a technique in DL that involves learning a set of attention weights to weigh the importance of different input features during the model's computation. Trainable attention was published by [38]. The authors introduced a neural network architecture that uses a set of trainable attention weights to dynamically select a subset of input features that are most relevant to the current task. The attention weights are learned along with the other parameters of the network using backpropagation. Trainable attention has been extended and improved upon in various ways. One such extension is self-attention [39], which uses the input features themselves to compute the attention weights, rather than learning them from scratch. Another extension is multi-head attention [40], which uses multiple sets of attention weights to attend to different parts of the input features simultaneously.

4.1.2 | Perturbation-based approaches

Perturbation-based approaches are a class of explainability methods that involve introducing small changes, or perturbations, to input data to observe changes in model outputs. These methodologies are especially valuable in understanding the decision-making mechanism of intricate models, such as deep neural networks, where the connection between input and output is frequently complex.

Occlusion-based methods: Occlusion-based methods involve systematically masking or occluding parts of the input data to observe changes in the model's output. By comparing the model's output with and without the occlusion, these methods can identify which input features are most important for the model's decision [35].

Local interpretable model-agnostic explanations (LIME): LIME is a popular approach for producing local explanations of a model's behavior by perturbing input data. LIME generates a surrogate model that approximates the original model's behavior in a local region around the input of interest. The surrogate model is subsequently used to explain the decision-making procedure of the primary model by highlighting the input elements that exerted the most significant impact on the decision [22].

Integrated Gradients: Integrated Gradients is a perturbation-based approach for generating feature importance scores that are based on the model's sensitivity to changes in the input data. Integrated Gradients compute the gradient of the model's output concerning the input data and integrate it along a path from a baseline input to the input of interest. This integration process generates a feature importance score for each input feature that reflects its contribution to the model's decision [34]. To determine the gradients of points, Integrated Gradients constructs a path from the reference input $\tilde{x}$ to x. Points are generated in images by intersecting $\tilde{x}$ and x and gradually changing the transparency of x, and integrated gradients are obtained by collecting gradient data. In mathematical notation, the reference $\tilde{x}$ input x integrated gradient is as follows:

$$\mathrm{i-th\ gradient} = \frac{\partial B(x)}{x_i}, \tag{5}$$

Relevant scores are determined by the following formula:

$$e_i(x) = (x_i - \tilde{x}_i) \int_{\propto=0}^{1} \frac{\partial B(\tilde{x}+((x-\tilde{x})))}{\partial x_i} . d \propto, \tag{6}$$

4.2 | Textual explanation

Textual explanation is an important aspect of XAI that aims to provide understandable and interpretable explanations of ML models to humans. It helps users understand the decision-making process of these models, which is essential for building trust and accountability. Table 2 presents a comparison of textual explanation methods.

4.2.1 | Image captioning

Image captioning is a popular application of XAI that generates textual explanations for images. Image captioning models use DL techniques to automatically generate natural language descriptions of the content in an image. The generated captions can provide textual explanations of what the image contains and what is happening in it. In [41], the authors proposed a model called bottom-up and top-down Attention for image captioning. The model uses a bottom-up approach to generate visual features of the image and a top-down approach to generate the caption. The model learns to attend to the most relevant regions of the image and generate a caption that describes the content of those regions.

4.2.2 | Image captioning with visual explanation

Image captioning with visual explanation is an XAI technique that combines image captioning with visual aids to provide more informative and interpretable explanations of image content. In [42], the authors proposed a model called Generating Visual Explanations (GenEx) that generates image captions along with spatial attention maps that highlight the most important regions of the image that contributed to the caption. The model uses the DNN to generate captions and attention maps that are trained jointly to ensure that they are aligned with each other. In addition to providing more informative and interpretable explanations, image captioning with visual explanation techniques can also be used for applications such as image retrieval and content-based image retrieval. In [43], the authors proposed a model that generates captions and attention maps for images in a large dataset and uses them to retrieve images based on their content.

4.2.3 | Testing with concept activation vectors (TCAV)

Testing with Concept Activation Vectors (TCAV) is an XAI technique used to explain the predictions of DNNs by analyzing their internal representations. TCAV uses a set of user-defined concepts (such as "striped" or "floral" for image classification models) and compares the activations of different layers in the neural network for each concept [44].

Table 2 Comparison of textual explanation methods. Ex= Explainability; L= Local; G= Global; H= High; M= Medium; W= Weak

XAI Method	Description	Ex-Type	Ex-Level	Agnostic?	Scalability	Data Types	Pros	Cons
Image Captioning [45, 46]	Uses deep learning models to provide textual descriptions of pictures that let people comprehend what's being shown in the picture.	L	W	No	H	Images	Provides explanations of the images that are understandable to humans.	Possibly not going to announce how the model predicts things.
Image Captioning with Visual Explanation [42]	Extends image captioning by producing not just verbal but also visual explanations that emphasize the parts of the picture that the model used to make a judgment.	G	M	No	M	Images	Provide more thorough explanations of how the model generates its predictions than only the image descriptions do.	To create visual explanations, more processing power can be needed.
Testing with Concept Activation Vectors (TCAV)[47]	Tests the model's sensitivity to changes in the activation of high-level concepts to determine how important those concepts are to the model's decision-making process.	G	H	Yes	H	Tabular/text/images	Sheds light on the broad ideas that a model believes are essential for its predictions.	Choosing concepts to test requires domain expertise.

4.3 |Example-based explanation

An example-based explanation is a type of explanation that uses specific examples to illustrate a broader concept or idea. It involves breaking down complex concepts or processes into smaller, more manageable pieces that can be understood through concrete examples. An example-based explanation is useful because it helps make abstract or complex concepts more tangible and relatable, allowing the listener to better understand and retain the information.

4.3.1 |Influence functions

Influence functions are a class of methods used in the example-based explanation, which aim to identify which examples in a dataset have the most influence on a specific prediction or model outcome. The basic idea is to perturb each example in the dataset and observe the resulting change in the prediction, to estimate the influence of that example on the model [48]. One of the most commonly used influence functions is the Leave-One-Out (LOO) method, which involves retraining the model with all but one example and then measuring the difference in prediction for the excluded example. The LOO influence function is useful for a variety of tasks, including feature selection, outlier detection, and model diagnosis [49]. Another popular influence function is the influence function based on derivatives, which involves computing the gradient of the model's loss function concerning each training example and then using the gradient to estimate the influence of each example on the model. This method has been used in a variety of applications, including robust optimization, model pruning, and adversarial example detection [50].

4.3.2 |Prototypes

Prototypes are a class of methods used in example-based explanation, which aim to identify representative examples in a dataset that can help interpret the behavior of a model. The basic idea is to identify examples that are typical or representative of a particular class or decision boundary and use these examples to provide insight into how the model is making predictions [51]. Prototype-based explanation methods have been applied to a wide range of tasks, including time series analysis, image classification, and natural language processing. They are effective in providing insights into the behavior of complex models and can be used to improve the interpretability and trustworthiness of ML systems [51].

4.3.3 |Examples from the latent space

Examples from the latent space are a class of methods used in example-based explanation that aim to identify representative examples in the low-dimensional latent space of a generative model. The basic idea is to identify examples that are typical or representative of a particular class or distribution in the latent space and use these examples to provide insight into the behavior of the generative model. One of the most commonly used latent space-based explanation methods is traversal, which involves exploring the latent space by perturbing the latent code of a specific example and observing the resulting changes in the generated output. This allows us to understand how the generative model maps the latent space to the output space and can provide insights into the underlying structure of the data [52]. Another popular method is latent variable inference, which involves inferring the latent variables that are most likely to have generated a specific example. This allows us to identify the latent factors that are most important for generating the observed output and can provide insights into the generative process of the model [53]. Examples from the latent space are effective in providing insights into the behavior of generative models and can be used to improve the interpretability and controllability of these models [54].

4.4 |Graphical explanation

Graphical explanations are a type of interpretability technique used in XAI to explain how AI models make decisions. In [55], the authors provide a comprehensive survey of different interpretability techniques, including graphical explanations. It covers the advantages and drawbacks of different techniques and provides examples of real-world applications. In [56], the authors introduce a method for visualizing the decision-making process of DNNs called Prediction Difference Analysis (PDA). PDA generates graphical explanations that highlight the features that are most important in making a prediction and compares the predictions of two models to identify areas of disagreement. In [57], the authors provide a conceptual framework for understanding

interpretability techniques, including graphical explanations. It introduces the concept of "feature visualization" techniques that generate graphical representations of features that are most important in making a prediction and provides examples of different types of visualizations.

5|XAI toward Neurology Diseases: Applications

There has been a growing inclination toward employing AI in the medical field, especially in the neurology diseases field. AI algorithms have proven their capacity to analyze vast quantities of medical data and derive valuable insights that may be challenging, if not impossible, for humans to grasp [58]. The key aspect of medical AI lies in obtaining informed consent from patients, ensuring that decision-making is shared between doctors and patients in a manner that prioritizes the patients' final say. Therefore, the implementation of medical AI is contingent upon patients being adequately informed about its essential functionalities beforehand, clearly and understandably. To accomplish this goal, there has been a focus on recent research endeavors aimed at creating XAI systems. These systems aim to provide medical practitioners with interpretations and explanations that can enhance the reasoning and decision-making processes within the neurology diseases field [59]. To achieve this goal, the XAI can be investigated and researched in various medical subdomains such as medical image analysis, medical record analysis, and drug discovery. CNN, 3D CNN, Visual Geometry Group 16 (VGG16), 3D Residual Attention Deep Neural Network (3D ResAttNet), and more are the DL models used in the studies that utilize LRP, GradCAM, and occlusion sensitivity methods as AI explanations.

LRP creates a visual explanation of significant brain areas as heat maps for recognizing brain atrophy. In [60], authors reported that they uncovered comparable significant features by applying composite LRP and multiple propagation rules. Furthermore, the author highlights that damage to the left temporal lobe impairs verbal semantic memory, while injury to the right temporal lobe hinders visual memory. These data are contributed by both authors to aid doctors and radiologists in diagnosing and creating confidence in the system. In [61], the GradCAM method is used to illustrate the predictions of a VGG16 DL model. Nevertheless, the authors demonstrate that CNN models with self-attention outperform VGG16 with GradCAM. The heat maps have been assessed as vastly improved to the baseline model by clinical professionals who found them to be useful. The authors further assert that the approach improves classification performance and interpretability. In [62], The implementation of GradCAM was utilized to visually represent heat maps depicting a four-way classification of AD that was predicted by a Generative Adversarial Network (GAN) model. Differently colored heatmaps generated by the system contribute to the accuracy of predictions regarding the development and severity of dementia. The approach has been shown useful for accurately discriminating between classes and making suitable early predictions. The color-coded heat map in the research, which depicted the advanced characteristics of different stages of dementia, would assist medical practitioners in making decisions. In [63], the authors created a new software technique for the early diagnosis of AD by proposing a simple-to-understand 3D ResAttNet. The researchers claimed that utilization of local, global, and spatial information in the 3D ResAttNet enhances the diagnostic accuracy and explainability of MRI images when employing GradCAM. The research presents a comprehensive end-to-end system for automated disease diagnosis. Furthermore, the methodology employed in this approach elucidates the mechanism by which crucial brain regions, including the hippocampus, lateral ventricle, and a significant portion of the cortex, contribute to the facilitation of transparent decision-making.

In [64], a modified Residual Network (ResNet) architecture was utilized to analyze a dataset from the MR CLEAN Registry, specifically focusing on CT angiography scans. The goal was to demonstrate the superiority of automated DNN over radiological imaging biomarkers in stroke prediction and treatment selection. Grad-CAM++ was used for visualization, and the model was found to be useful for stroke outcome

prediction. In [65], authors used a variety of 3D-CNNs, including 3D GradCAM, for classification and AI explainers. The heat maps are helpful for medical professionals because they highlight the importance of the lateral ventricle and most cortical regions in the diagnosis of AD. In [66], The researchers apply High-Resolution Activation Mapping (HAM) to offer visual explanations with enhanced resolution, integrating values generated from both the final convolutional layer and intermediate data. High-quality heatmaps that show discriminative localization of brain abnormalities outperform earlier studies. The clinical utility of the model was validated by the authors based on its high diagnostic accuracy and transparent explanations. In [67], the occlusion sensitivity approach is used to generate heat maps by occluding a part of the input image with a black patch. From changes in the output probability predictions, the model's brain areas contributing to the classification decision were directly observable. White matter hyperintensity was found and reaffirmed as a neuroimaging biomarker for dementia by the authors. One study utilized LRP to deconstruct the network's output score of 18-Fluoro-Deoxyglucose Positron Emission Tomography (18 FDG-PET) scans into individual contributions while keeping the conservation principle and heat map produced. The study uses saliency maps to build voxel-wise heat maps for each contribution.

6 | Pros and Cons of XAI Approaches toward Medical Field

Explaining the pros and cons of XAI approaches in the medical field can help us understand their potential benefits and limitations.

6.1 | Pros of XAI approaches in the medical field:

1. **Enhanced trust and acceptance:** XAI techniques provide interpretability and transparency, enabling healthcare professionals to understand and trust the decisions made by AI systems. This promotes increased acceptance and adoption of AI in neurology disease applications [68].
2. **Error detection and diagnosis:** XAI methods can help identify errors or biases in the AI models, enhancing their reliability and robustness. These approaches enable healthcare practitioners to detect and rectify potential issues in the decision-making process [68].

6.2 | Cons of XAI approaches in the medical field:

1. **Increased complexity:** Some XAI methods can be computationally expensive and complex, requiring additional resources and expertise to implement and interpret. This may limit their practicality and scalability in certain medical settings [68].
2. **Privacy and security concerns:** The interpretability provided by XAI methods may reveal sensitive patient information, leading to privacy and security risks. Careful consideration must be given to ensure the protection of patient data when implementing XAI approaches [69].

7 | Conclusions

This study presents an overview of the significant topic of XAI techniques and their application in the comprehension of image classification tasks in the field of neurology diseases. XAI combined with DNN models can aid in neurology disease identification and diagnosis and give the doctor new insights about the right diagnosis. This study illuminated a detailed taxonomy that provides an insightful categorization of XAI studies and the similarities and differences among various algorithms used in XAI, thus facilitating further advancements in methodology.

Ethical Approval

Not applicable.

Acknowledgments

The authors express their gratitude to the producers of the dataset for their valuable support.

Author contributions

Conceptualization, Nabil M., Mohamed M. AbdelHafeez, M. M. Hassan and Asmaa H.; Methodology, Nabil M., Mohamed M. AbdelHafeez, M. M. Hassan and Asmaa H.; Investigation Mohamed M. AbdelHafeez; Writing – original draft Mohamed M. AbdelHafeez; Writing – review and editing Nabil M., M. M. Hassan and Asmaa H. All authors have read and agreed to the published version of the manuscript.

Data availability

Not Applicable.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflicts of interest in the research.

References

[1] X. Bai et al., "Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments," Pattern Recognition, vol. 120, p. 108102, 2021.

[2] M. A. Ahmad, C. Eckert, and A. Teredesai, "Interpretable machine learning in healthcare," in Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics, 2018, pp. 559-560.

[3] S. Wachter, B. Mittelstadt, and L. Floridi, "Transparent, explainable, and accountable AI for robotics," Science robotics, vol. 2, no. 6, p. eaan6080, 2017.

[4] U. Pawar, D. O'Shea, S. Rea, and R. O'Reilly, "Incorporating Explainable Artificial Intelligence (XAI) to aid the Understanding of Machine Learning in the Healthcare Domain," in AICS, 2020, pp. 169-180.

[5] A. Abujabal, R. S. Roy, M. Yahya, and G. Weikum, "Quint: Interpretable question answering over knowledge bases," in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2017, pp. 61-66.

[6] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)," IEEE access, vol. 6, pp. 52138-52160, 2018.

[7] D. Saraswat et al., "Explainable AI for healthcare 5.0: opportunities and challenges," IEEE Access, 2022.

[8] Y. Zhang, Y. Weng, and J. Lund, "Applications of Explainable Artificial Intelligence in Diagnosis and Surgery," Diagnostics, vol. 12, no. 2, p. 237, 2022.

[9] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," BMC Medical Informatics and Decision Making, vol. 20, no. 1, pp. 1-9, 2020.

[10] G. Yang, Q. Ye, and J. Xia, "Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond," Information Fusion, vol. 77, pp. 29-52, 2022.

[11] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," IEEE transactions on neural networks and learning systems, vol. 32, no. 11, pp. 4793-4813, 2020.

[12] M. Moradi and M. Samwald, "Deep Learning, Natural Language Processing, and Explainable Artificial Intelligence in the Biomedical Domain," arXiv preprint arXiv:2202.12678, 2022.

[13] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," Electronics, vol. 8, no. 8, p. 832, 2019.

[14] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," Proceedings of the National Academy of Sciences, vol. 116, no. 44, pp. 22071-22080, 2019.

[15] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 9, no. 4, p. e1312, 2019.

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

[17] A. Rai, "Explainable AI: From black box to glass box," Journal of the Academy of Marketing Science, vol. 48, pp. 137-141, 2020.
[18] Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "Gnnexplainer: Generating explanations for graph neural networks," Advances in neural information processing systems, vol. 32, 2019.
[19] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," Journal of Imaging, vol. 6, no. 6, p. 52, 2020.
[20] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," Entropy, vol. 23, no. 1, p. 18, 2020.
[21] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," Advances in neural information processing systems, vol. 30, 2017.
[22] M. T. Ribeiro, S. Singh, and 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.
[23] M. Melis, D. Maiorca, B. Biggio, G. Giacinto, and F. Roli, "Explaining black-box android malware detection," in 2018 26th european signal processing conference (EUSIPCO), 2018, pp. 524-528: IEEE.
[24] D. Gunning and D. Aha, "DARPA's explainable artificial intelligence (XAI) program," AI magazine, vol. 40, no. 2, pp. 44-58, 2019.
[25] J. Yuan, B. Barr, K. Overton, and E. Bertini, "Visual exploration of machine learning model behavior with hierarchical surrogate rule sets," IEEE Transactions on Visualization and Computer Graphics, 2022.
[26] G. Alicioglu and B. Sun, "A survey of visual analytics for Explainable Artificial Intelligence methods," Computers & Graphics, vol. 102, pp. 502-520, 2022.
[27] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 618-626.
[28] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2921-2929.
[29] W. Liu et al., "Ssd: Single shot multibox detector," in Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, 2016, pp. 21-37: Springer.
[30] P. Rajpurkar et al., "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," arXiv preprint arXiv:1711.05225, 2017.
[31] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," Nature machine intelligence, vol. 2, no. 1, pp. 56-67, 2020.
[32] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in 2018 IEEE winter conference on applications of computer vision (WACV), 2018, pp. 839-847: IEEE.
[33] J. Adebayo, J. Gilmer, I. Goodfellow, and B. Kim, "Local explanation methods for deep neural networks lack sensitivity to parameter values," arXiv preprint arXiv:1810.03307, 2018.
[34] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in International conference on machine learning, 2017, pp. 3319-3328: PMLR.
[35] M. D. Zeiler and 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, 2014, pp. 818-833: Springer.
[36] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-wise relevance propagation: an overview," Explainable AI: interpreting, explaining and visualizing deep learning, pp. 193-209, 2019.
[37] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," PloS one, vol. 10, no. 7, p. e0130140, 2015.
[38] V. Mnih, N. Heess, and A. Graves, "Recurrent models of visual attention," Advances in neural information processing systems, vol. 27, 2014.
[39] A. Vaswani et al., "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.
[40] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," IEEE transactions on neural networks and learning systems, vol. 32, no. 1, pp. 4-24, 2020.
[41] P. Anderson et al., "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. computer vision and pattern recognition," 2017.
[42] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating visual explanations," in Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14, 2016, pp. 3-19: Springer.
[43] R. Krishna et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," International journal of computer vision, vol. 123, pp. 32-73, 2017.
[44] B. Kim et al., "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV).(2017)," arXiv preprint arXiv:1711.11279, 2017.
[45] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3156-3164.
[46] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3128-3137.

[47] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, and F. Viegas, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in International conference on machine learning, 2018, pp. 2668-2677: PMLR.
[48] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in International conference on machine learning, 2017, pp. 1885-1894: PMLR.
[49] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in International conference on machine learning, 2017, pp. 3145-3153: PMLR.
[50] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," arXiv preprint arXiv:1711.06104, 2017.
[51] F. Di Martino and F. Delmastro, "Explainable AI for clinical and remote health applications: a survey on tabular and time series data," Artificial Intelligence Review, pp. 1-55, 2022.
[52] L. Maaløe, C. K. Sønderby, S. K. Sønderby, and O. Winther, "Auxiliary deep generative models," in International conference on machine learning, 2016, pp. 1445-1453: PMLR.
[53] X. Wang and A. Gupta, "Generative image modeling using style and structure adversarial networks," in Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14, 2016, pp. 318-335: Springer.
[54] J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum, "Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling," Advances in neural information processing systems, vol. 29, 2016.
[55] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," ACM computing surveys (CSUR), vol. 51, no. 5, pp. 1-42, 2018.
[56] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," arXiv preprint arXiv:1702.04595, 2017.
[57] C. Olah et al., "The building blocks of interpretability," Distill, vol. 3, no. 3, p. e10, 2018.
[58] Y. Xie, G. Gao, and X. A. Chen, "Outlining the design space of explainable intelligent systems for medical diagnosis," arXiv preprint arXiv:1902.06019, 2019.
[59] K. Stöger, D. Schneeberger, and A. Holzinger, "Medical artificial intelligence: the European legal perspective," Communications of the ACM, vol. 64, no. 11, pp. 34-36, 2021.
[60] T. Pohl, M. Jakab, W. J. I. J. o. I. S. Benesova, and Technology, "Interpretability of deep neural networks used for the diagnosis of Alzheimer's disease," vol. 32, no. 2, pp. 673-686, 2022.
[61] C. Chunharas et al., "An Explainable Self-Attention Deep Neural Network for Detecting Mild Cognitive Impairment Using Multi-input Digital Drawing Tasks," 2022.
[62] V. Jain, O. Nankar, D. J. Jerrish, S. Gite, S. Patil, and K. J. I. A. Kotecha, "A novel AI-based system for detection and severity prediction of dementia using MRI," vol. 9, pp. 154324-154346, 2021.
[63] X. Zhang, L. Han, W. Zhu, L. Sun, D. J. I. j. o. b. Zhang, and h. informatics, "An explainable 3D residual self-attention deep neural network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI," vol. 26, no. 11, pp. 5289-5297, 2021.
[64] A. Hilbert et al., "Data-efficient deep learning of radiological image data for outcome prediction after endovascular treatment of patients with acute ischemic stroke," vol. 115, p. 103516, 2019.
[65] C. Yang, A. Rangarajan, and S. Ranka, "Visual explanations from deep 3D convolutional neural networks for Alzheimer's disease classification," in AMIA annual symposium proceedings, 2018, vol. 2018, p. 1571: American Medical Informatics Association.
[66] L. Yu, W. Xiang, J. Fang, Y.-P. P. Chen, and R. J. P. R. Zhu, "A novel explainable neural network for Alzheimer's disease diagnosis," vol. 131, p. 108876, 2022.
[67] V. Bordin, D. Coluzzi, M. W. Rivolta, and G. Baselli, "Explainable AI points to white matter hyperintensities for Alzheimer's disease identification: A preliminary study," in 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2022, pp. 484-487: IEEE.
[68] J. Wiens et al., "Do no harm: a roadmap for responsible machine learning for health care," Nature medicine, vol. 25, no. 9, pp. 1337-1340, 2019.
[69] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis," IEEE journal of biomedical and health informatics, vol. 22, no. 5, pp. 1589-1604, 2017.