IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 May 2024, accepted 5 May 2024, date of publication 8 May 2024, date of current version 17 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3398203

SURVEY

Unraveling the Black Box: A Review of Explainable Deep Learning Healthcare Techniques

NAFEESA YOUSUF MURAD[1], MOHD HILMI HASAN[1], MUHAMMAD HAMZA AZAM[2], NADIA YOUSUF[3], AND JAMEEL SHEHU YALLI[1], (Graduate Student Member, IEEE)

[1]Department of Computer and Information Sciences, Universiti Teknologi PETRONAS, Seri Iskandar, Perak 32610, Malaysia
[2]Centre for Research in Data Science, Universiti Teknologi PETRONAS, Seri Iskandar, Perak 32610, Malaysia
[3]Department of Computer Science, Benazir Bhutto Shaheed University, Karachi 75660, Pakistan

Corresponding author: Nafeesa Yousuf Murad (nafeesa_22009985@utp.edu.my)

The work presented in this paper is part of ongoing research under the Fundamental Research Grant Scheme (FRGS) funded by the Ministry of Higher Education Malaysia (Ref: FRGS/1/2022/ICT02/UTP/02/1)(015MA0-160).

ABSTRACT The integration of deep learning in healthcare has propelled advancements in diagnostics and decision support. However, the inherent opacity of deep neural networks (DNNs) poses challenges to their acceptance and trust in clinical settings. This survey paper delves into the landscape of explainable deep learning techniques within the healthcare domain, offering a thorough examination of deep learning explainability techniques. Recognizing the pressing need for nuanced interpretability, we extend our focus to include the integration of fuzzy logic as a novel and vital category. The survey begins by categorizing and critically analyzing existing intrinsic, visualization, and distillation techniques, shedding light on their strengths and limitations in healthcare applications. Building upon this foundation, we introduce fuzzy logic as a distinct category, emphasizing its capacity to address uncertainties inherent in medical data, thus contributing to the interpretability of DNNs. Fuzzy logic, traditionally applied in decision-making contexts, offers a unique perspective on unraveling the black box of DNNs, providing a structured framework for capturing and explaining complex decision processes. Through a comprehensive exploration of techniques, we showcase the effectiveness of fuzzy logic as an additional layer of interpretability, complementing intrinsic, visualization, and distillation methods. Our survey contributes to a holistic understanding of explainable deep learning in healthcare, facilitating the seamless integration of DNNs into clinical workflows. By combining traditional methods with the novel inclusion of fuzzy logic, we aim to provide a nuanced and comprehensive view of interpretability techniques, advancing the transparency and trustworthiness of deep learning models in the healthcare landscape.

INDEX TERMS Artificial intelligence, deep learning, explainability, XAI.

I. INTRODUCTION

Deep learning, a subset of machine learning distinguished by intricate neural network structures, has emerged as a revolutionary force, bringing cutting-edge solutions across a wide range of industries [1]. Its success is attributable to technological advances that have exponentially expanded processing capacity, allowing for the training of sophisticated neural networks [2]. This increase in processing capacity, combined with lower hardware prices, has democratized deep learning, accelerating its application across areas such as healthcare [3], agrifood [4], finance, and transportation [5].

Although deep learning, notably models such as Convolutional Neural Networks (CNNs), has transformed fields such as computer vision, the inherent black-box (in figure 1) nature of these models makes comprehending their decision-making processes difficult. This lack of interpretability is especially important in applications such as autonomous driving [6]

The associate editor coordinating the review of this manuscript and approving it for publication was Wai-Keung Fung.

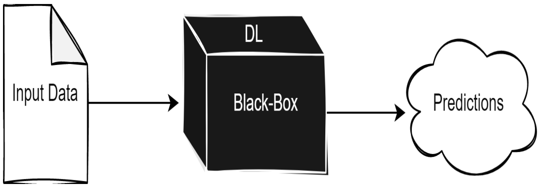


FIGURE 1. **Deep learning as black-box.**

and medical image processing [7], where transparency is vital. The necessity for interpretable deep learning models in healthcare is emphasized by the requirement for transparency in medical diagnosis and treatment recommendations, as well as maintaining patient safety and establishing trust among medical practitioners [8].

The interpretability of deep learning models is critical not just for medical practitioners, but also for regulatory approval and ethical issues [9]. Regulatory organizations demand transparency and responsibility in healthcare Artificial Intelligence models, highlighting the necessity for interpretable deep learning algorithms that offer reasons for predictions. This not only speeds up the regulatory approval process but also raises the ethical status of AI applications in the medical arena. In healthcare, achieving interpretable deep learning models is essential for regulatory compliance, ethical considerations, and establishing trust among stakeholders. While high accuracy is desirable for many applications, ensuring that models are interpretable is crucial for regulatory compliance, ethical considerations, and user trust, particularly in safety-critical domains like healthcare. Interpretability enables stakeholders to understand how AI systems arrive at their decisions, providing transparency and accountability in AI-driven systems. Ethically, interpretable models empower users to validate AI recommendations and mitigate potential biases or errors. In safety-critical domains like healthcare, the ability to explain AI decisions is imperative for ensuring patient safety and clinical acceptance [10].

Efforts to improve deep learning model interpretability are critical for broader acceptance and ethical adoption, notably in healthcare applications [11]. When AI algorithms deliver unexpected or contradictory findings, the need for interpretability becomes even more pressing [9]. Interpretability enables researchers to discover the sources of bias or inaccuracy, allowing for better model performance and, as a result, more trustworthy and accurate predictions. Reference [12], a rising field of research has evolved to produce transparent and interpretable AI. In [13], the Defense Advanced Research Projects Agency (DARPA) identifies various techniques that are actively being studied to build XAI. Many studies have proposed various metrics and frameworks to capture interpretability in artificial intelligence (AI), often referred to as Explainable AI. Reference [14] has emerged as a hot topic in the DL research field.

An explanation, as illustrated in the figure 2, is employed to validate the decision made by a machine learning agent or algorithm. In the example of a tumor detection model that employs microimages, an explanation could take the form of a pixel map that shows the input pixels that influence the model's output, as cited by [15]. Similarly, a voice identification model explanation could give power range information at a certain time, highlighting its importance to the current output decision.

A. MOTIVATION

The integration of modern technology, notably deep learning, holds the promise of altering diagnoses and treatment procedures in the fast-expanding healthcare scene [16]. However, this transformation is fraught with difficulty due to the black-box nature of deep learning models. Transparency and comprehension in the decision-making processes of these models are critical in healthcare because judgments directly affect patient well-being. The stakes are enormous, and as we enter an era of AI-assisted healthcare [17], the need for explainable deep learning approaches becomes critical.

The patient-practitioner connection is built on trust, which is dependent on understandable and justifiable AI-driven judgments. Deep learning models' opacity restricts the creation of trust, causing problems among healthcare professionals and patients alike [18]. Transparency must be incorporated into the fabric of these advanced technologies to create confidence, allowing medical practitioners to validate and comprehend the rationale behind deep learning models' recommendations.

Traditional categories of explainability, such as intrinsic, visualization, and distillation procedures, have emerged as essential foundations for this goal of transparency. Intrinsic explainability focuses on constructing intrinsically interpretable models, giving a platform for understanding without the need for additional procedures [19]. Visualization methods use graphical tools to show and explain the decision-making process, providing insights into the characteristics and patterns that influence model predictions [20]. Distillation techniques reduce complex models to more interpretable forms, allowing for a better grasp of the underlying mechanisms [21].

The motivation to improve explainability in healthcare, however, extends beyond these established strategies. Recognizing the diverse nature of medical data and the complexities of healthcare decision-making, we introduce the justification for including fuzzy logic. The complexities inherent in medical data analysis are aligned with fuzzy logic [22], which is known for its ability to handle ambiguity and imprecision. The use of fuzzy logic aims to improve the interpretability of deep learning models by offering a formal framework for navigating uncertainty in medical information [23].

As more healthcare organizations recognize the value of deep learning applications, the desire to make these models more transparent and interpretable becomes a driving factor. We hope to address the crucial requirement for trustworthy, intelligible, and ethically sound AI-driven judgments in the complicated area of healthcare by embracing not

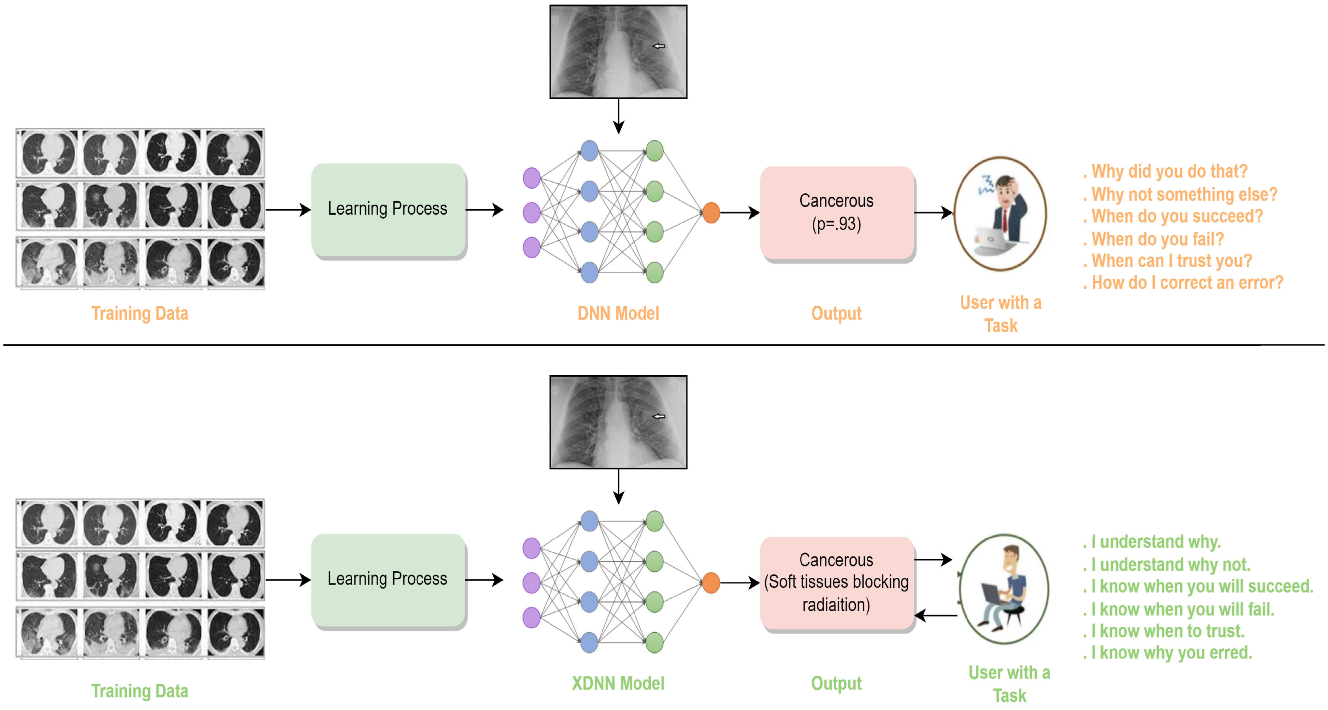


FIGURE 2. XAI concept.

only established categories of explainability but also the distinctive contributions of fuzzy logic. Finally, we believe that explainable deep learning approaches will not only increase AI acceptance in healthcare [23], but will also lead to better patient outcomes and more seamless integration of new technology into the medical decision-making process.

B. OBJECTIVES

This study aims to achieve a diverse collection of goals targeted at enhancing the knowledge and application of explainable deep learning methods in the context of healthcare.

First and foremost, the primary purpose is to perform a thorough examination of three fundamental types of explainability methods: intrinsic, visualization, and distillation procedures. We hope to provide knowledge of the strengths, limits, and uses of these traditional methodologies by carefully exploring current literature and landmark works. This paper lays the groundwork for building a comprehensive assessment of the current landscape of explainability in the area of deep learning, with a particular emphasis on its implications and significance in healthcare settings.

Secondly, by introducing fuzzy logic as an extra category of explainable approaches, the research hopes to offer a new level of healthcare interpretability. Recognizing the inherent ambiguity and imprecision in medical data, fuzzy logic provides a novel approach to resolving these issues. Our goal is to explain the principles and applications of fuzzy logic in the context of healthcare, demonstrating how it may improve interpretability and transparency in deep learning models. We hope to illustrate the complementing nature of fuzzy logic and its role in navigating the intricacies of medical decision-making through a comparison with traditional methodologies.

The rest of the paper is organized as follows: section II reviews prior related surveys, section III describes the explainable deep learning, section IV discusses the types of explainable AI techniques, section V does the comparative analysis of existing techniques and existing surveys, and section VI summarizes our conclusions.

II. RELATED SURVEYS

In this section, we discuss in detail the existing related surveys.

Salahuddin et al. [24] provided a thorough examination of the significance of interpretability in DNNs for medical imaging. It addressed the issues posed by DNNs' "black-box" nature, emphasizing the importance of openness, robustness, and accountability in AI systems, notably in healthcare. The review investigated several interpretability methodologies such as post-hoc interpretability methods (feature visualization, saliency mapping, and attribution methods), used to explicate DNNs decision-making processes, highlighting the importance of intelligible explanations for AI model predictions. It went over the technical aspects, limitations, and uses of interpretability methodologies, emphasizing the significance of both quantitative and qualitative evaluations to assure the reliability of the explanations given. Similarly,

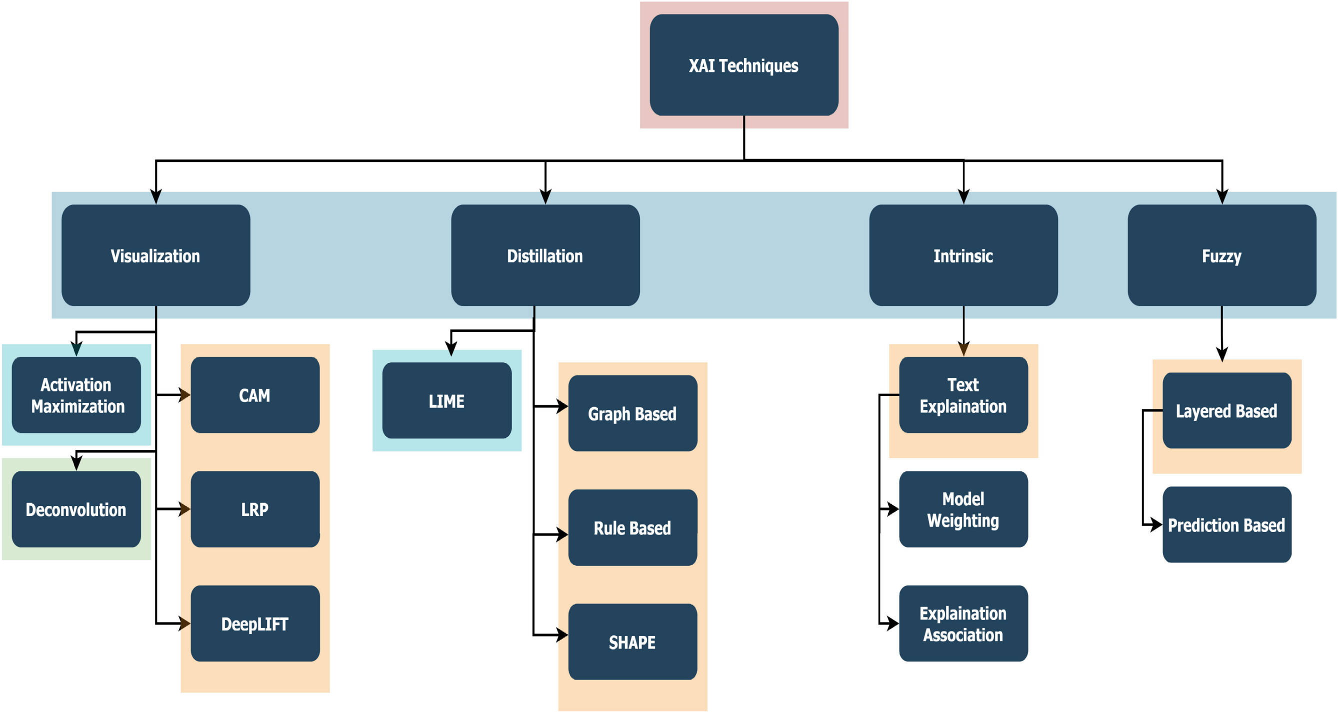


FIGURE 3. XAI taxonomy.

Tjoa and Guan [11] discussed numerous strategies and techniques for attaining interpretability in machine learning models, as well as the obstacles and opportunities that exist in this domain. The survey covered using visuals, probability distributions, and textual reasons to improve the interpretability of deep learning systems. It also emphasized the significance of dependability and responsibility in machine judgments, especially in the medical field. The study provided insights on how XAI could improve the transparency and trustworthiness of AI systems. Furthermore, Chakraborty et al. [25] gave a complete assessment of the interpretability of deep learning models, emphasizing the importance of providing human-understandable reasons for model outputs to foster confidence. It defined several characteristics of interpretability, including model functioning, model transparency, and local explanation, and categorized previous work based on these dimensions.

Adadi and Berrada [26] gave a detailed assessment of Explainable Artificial Intelligence, highlighting the field's multidisciplinary character and the importance of a strong literature background. The survey provided an orderly overview of extant XAI methodologies based on a literature review of 381 papers, addressing prior surveys' weaknesses by focusing on holism and clarity. The authors recognized the difficulty of comprehensively collecting all XAI studies and instead concentrated on a sample of prominent papers, including non-academic sources.

Arrieta et al. [27] presented a detailed introduction to XAI, including motives, methodology, applications, and problems. It also explained how XAI relates to other AI concepts, including justice, privacy, and accountability. The research proposed two taxonomies for categorizing different XAI techniques: one based on general ML models and the other on the specialized area of deep learning. In addition, the publication provided various examples of XAI image visualization approaches. Similarly, Das and Rad [28] reviewed current studies on XAI for deep learning models. It focused on the mathematical models and algorithms underlying various XAI methods. The research provided a taxonomy for categorizing XAI methods according to their nature, techniques, including perturbation-based and gradient-based, and outcomes. The research created and evaluated explanation maps for eight distinct XAI algorithms, emphasizing their advantages and disadvantages. The study discussed different XAI techniques, including saliency maps, Grad-CAM, LIME, SHAP, LRP, Axiomatic Attribution, and intrinsic methods. It also addressed the obstacles and future directions of XAI research.

Patrício et al. [29] examined explainable deep learning approaches for medical picture classification, with an emphasis on three forms of explanation: feature attribution, text, and concept attribution. The study discussed saliency maps, CAM, Grad-CAM, LIME, SHAP, and other methods for highlighting the key areas or regions in the image being used that contribute to the model decision. The article examined various strategies for generating textual descriptions to explain model decisions, including picture captioning, transcription with visual explanation, and idea attribution.

The study also discussed methods for identifying high-level ideas or semantic aspects linked with model decisions, such as TCAV, ACE, and Concept Bottleneck Models. Further, Speith [30] examined eleven current taxonomies of XAI methods, highlighting their benefits and challenges. It distinguished four methods for building taxonomies: model-specific, model-agnostic, mixed, and meta. It examined the differences between the groups and classifications of several taxonomies. The research discovered that present taxonomies share some features, such as ante-hoc and post-hoc explainability, but also differ in some ways, such as misleading nomenclature, inconsistent classification, and a lack of consistency. It also highlighted certain specific limits for each strategy. Consequently, the paper presented a new taxonomy that seeks to address some of the issues while providing a full and consistent review of XAI techniques.

Gulum et al. [31] work covered explainable deep learning methods for cancer detection utilizing medical imaging, namely MRI scans. The research contended that deep learning models are black-box techniques that lack explainability, limiting their therapeutic applicability and reliability. The research performed a literature assessment of current techniques for explainable deep learning, including architecture modification, after-the-fact explanations, and the use of visualizations and saliency maps. The paper analyzed the gap between what clinicians require and what current technologies deliver, and it proposed some future paths for improving explainability, such as combining domain knowledge, leveraging multimodal data, and reviewing explanations.

Cao et al. [32] provided a comprehensive overview of explainable AI techniques applied in medical diagnosis. It highlighted the use of Fuzzy Inference Systems with interpretable fuzzy rules to enhance explainability in disease diagnosis. Comparative analysis included fuzzy rules alongside SHAP and heat maps across diverse medical data types such as sequence signals, medical images, and tabular data. Moreover, Loh et al. [33] provided a retrospective review of XAI applications in healthcare over the past decade, focusing on areas requiring more attention from the research community. Employing PRISMA guidelines, it systematically analyzed 99 articles from Q1 journals, investigating diverse XAI techniques like SHAP, LIME, GradCAM, and LRP in healthcare contexts. The review emphasized the need for further research on XAI tailored for 1D biosignals and clinical notes, proposing the development of a comprehensive cloud-based system for smart cities.

Futhermore, Wang et al. [34] focused on improving the intelligibility of AI applications in healthcare, particularly within the context of hospital recommendations. It introduced a cross-domain tools and techniques under the umbrella of XAI to enhance AI's transparency, including universal expression, color coordination, and segmented proximity diagrams. The proposed methodology advocated for applying these tools to existing AI technologies without increasing complexity or altering the target user group. Similarly, Wang et al [35] explored the application of XAI tools to enhance AI in healthcare, focusing on diabetes diagnosis. A systematic approach is outlined, employing seven XAI techniques—smart technologies, common expression, color management, LIME, CART, and donut charts different aspects of AI applications. Using type 2 diabetes diagnosis as an example, the methodology demonstrates how an artificial neural network can be approximated to a Classification and Regression Tree (CART) using LIME. Results indicate that this XAI methodology improves transparency, comprehensibility, interpretability, and understandability of AI applications in healthcare.

When comparing the XAI review studies listed above, they mostly focus on approaches such as visualization, intrinsic, and distillation, excluding talks of fuzzy methods. The majority of studies and taxonomies focus on these techniques, leaving fuzzy methods as an underexplored area in the current body of research. Our review stands out because it includes fuzzy-type elements, which have not been thoroughly explored in previous evaluations.

III. EXPLAINABLE DEEP LEARNING

In this section, we provide a comprehensive discussion on the concept of explainable deep learning, including definitions of key terms such as interpretability, transparency and explainability.

The field of XAI focuses on developing AI tools and techniques that enhance transparency and interpretability in machine and deep learning models. Interpretability refers to the ability to understand and explain how a model makes decisions, while transparency involves making the decision-making process accessible and understandable to stakeholders [24], [25]. Within the realm of deep learning, achieving explainability is crucial given the inherent complexity and black-box nature of DNNs [36]. As highlighted by DARPA and other research initiatives, various methodologies and frameworks are being explored to enable XAI [13], ensuring that AI systems can provide meaningful explanations for their predictions and decisions.

The first strategy is to make deep learning algorithms more interpretable. This method is known as deep explanations [13] and saliency mapping is frequently used for these models. Saliency mapping is accomplished by evaluating a network frequently to determine whether parts of the input impact the outcome [37]. Some of the approaches that employ saliency mapping to offer deep explanations include LRP [37], DeepLIFT [38], CAM [39], and others. The difficulty with these approaches is that the models only indicate the link between the inputs and the outputs; however, the interconnections between the inputs that produce the intermediate layers are more difficult to evaluate and often require the assistance of specialists in these techniques. The second technique recommended for achieving XAI is to leverage model induction using model-agnostic approaches. LIME [40] is one of the model-independent techniques. LIME is used to investigate the model's actions by causing perturbations on the inputs, and this data is subsequently utilized to determine what attributes contribute to the results to generate a locally trustworthy linear model, i.e., the model faithfully reproduces the output of the original model for a specific input. The issue with this method is that an external model was employed, which implies that the explanations supplied are not always correct. The third technique is to employ pre-existing comprehensible and causal models, such as decision trees, Bayesian rules, hidden markov models, fuzzy logic, and so on. These models, however, may be less accurate than matching black box models, and they may also become opaque for high-dimensional inputs [41].

Gabbay et al. [42] used LIME to predict the severity of illness in COVID-19 patients. They used LIME to explain the forecast using ANN MPL with 80% accuracy. The success of LIME, on the other hand, is highly dependent on the perturbation strategy and size. Small changes to the perturbation procedure can lead to alternate interpretations that may or may not adequately characterize the sophisticated model's underlying decision constraints [43]. Lundberg and Lee [44] proposed a similar approach called Shapley Additive Explanation (SHAP) [45] to highlight the relevance of the individual portion of input data while explaining the prediction.

SHAP values are calculated by taking into account all possible feature subsets and their combinations, which can result in significant computational complexity, especially for models with numerous features. As a result, SHAP calculation can become slow and resource-intensive, limiting its relevance to real-time or resource-constrained contexts [45]. Layer-wise relevance propagation (LRP) and sensitivity analysis (SA) are visualization approaches used to explain deep learning model predictions in terms of input variables [46]. Arras et al. [47] utilized deep Taylor decomposition to communicate explanations from DNN outputs to the contributions of its inputs. This explanatory procedure yields a heat-map that displays the importance of each pixel in the forecast. When dealing with a more intricate or hierarchical system, the LRP approach falls short. Zhu et al. [48] created a DNN that recognizes image content and generates subtitles. For each word in the description, an explanation in the form of highlighted relevant sections of the input image is provided. Other approaches include explanations in several media, such as visualization, text, and examples.

Similarly, Lima et al. [23] proposed the Explainable Fuzzy-Based Deep Learning (EFBDL) method, which strives for both accuracy and comprehension. This system consists of two primary components: The first section uses a Deep network technique to accurately classify images using an Inception V4-based CNNs. It uses transfer learning and a feature extraction technique based on neuron disruption to increase its performance. The second segment lays the groundwork for Soft Computing. It makes use of a Fuzzy Rule-Based System (FRBS) to capture nuanced correlations, a Granular Linguistic Model of a Phenomenon (GLMP) for natural language production, and a technique called Highly Interpretable Linguistic Knowledge (HILK), which combines human insights with linguistic rules. The EFBDL system generates a natural English explanation of the neural network's decisions. This strategy focuses especially on skin cancer prevention, with the potential to allow governments to adopt proactive prevention plans and dramatically reduce overall disease treatment costs.

IV. TYPES OF XAI TECHNIQUES

This section focuses on numerous Explainable Artificial Intelligence strategies, each with a distinct approach to improving transparency and interpretability. These strategies are critical in demystifying sophisticated models' complicated decision-making processes, encouraging trust, and promoting a deeper knowledge of AI systems. Various approaches to categorizing XAI techniques have been investigated. This thorough classification is depicted visually in a figure, which encapsulates the various techniques used to improve transparency and interpretability in artificial intelligence systems. The picture serves as a visual reference, displaying the varied tactics and approaches used in various branches of XAI, allowing for a more in-depth knowledge of the areas of explainability in machine learning models.

A. VISUALIZATION

Visualization methods relate a DNNs assessment of input properties to its decision, also known as attribution. This link is visually represented by saliency maps, which are widely used as a form of explanation [49], [50]. These translucent-colored heatmaps, which are typically overlaid on the original input image, emphasize salient input characteristics, indicating those with the most influence on the model's output by evoking a significant response or stimulation.

1) ACTIVATION MAXIMIZATION

Erhan et al. [51] developed deep architecture visualization by providing the activation maximization approach for identifying important characteristics in any layer of a deep network. This method improves the input, X, to maximize the activation, a, of a selected unit i in a layer j:

$$\text{argmax}_X \, a_{i,j}(X, \theta) \tag{1}$$

During activation maximization, the parameters of a pre-trained network remain constant. The optimal X is found by computing the gradient of $a_{i,j}(X, \theta)$ and updating X in the direction of the gradient. Practitioners can customize hyperparameters such as learning rate and iteration count. The optimal X serves as a representation inside the input space, emphasizing qualities that enhance the activation of a single or several units in an identified network layer. By visualizing these internal representations, practitioners can assess the human interpretability of the topics they have acquired. The quality of these concepts can provide insights into model universality and help determine whether additional labeled data is required. While activation maximization offers useful insights into model training and generalization, it is not intended to explain single model predictions.

2) CAM

Class Activation Maps (CAM) [39], [52] are visualization techniques used in CNNs to highlight critical parts of an input image that contribute to a specific class prediction. The activation maps of the final convolutional layer, shortly before the fully connected (FC) output layer, are subjected to global average pooling (GAP). The final setup is GAP(Conv) FC softmax. In terms of mathematics, CAM incorporates the activations A from the convolutional layer (Conv) with K filters and the weights, $w\ k,c$ from the FC layer, where k denotes the filter index and c represents the class index. The relevance score map map_c for a particular class is calculated as follows:

$$\mathrm{map}_c = \sum_{k=1}^{K} w_{k,c} A_k \tag{2}$$

The relevance score map emphasizes portions of the input image that have a strong influence on the prediction of the specific class c. CAM illustrates where the model spends its attention to produce a categorization conclusion.

3) DECONVOLUTION

Deconvolution is a technique for unsupervised image feature learning that can visualize higher-layer features in a working space [53]. It is based on CNNs with consecutive layers of convolution and activation of Rectified Linear Units (ReLU). Convolution is represented in the original CNN as

$$A' = \mathrm{maxpool}(\mathrm{ReLU}(A_{\mathrm{prev}} * K + b)) \tag{3}$$

where A is the present layer output, A_{prev} is the output of the previous layer, K is the previously learned filter, b is the bias, and indicates convolution. When max-pooling is employed, indices are saved for later unpooling. The CNN is reversed in a deconvolutional neural network (DeCNN), which uses reversed convolutions for deconvolutional layers and unpooling for max-pooling layers. The deconvolution procedure can be described mathematically as

$$A_{\mathrm{prev}} = \mathrm{ReLU}(A' * K'^{\top} + b') \tag{4}$$

where $K\ T$ is the transposed filter and b is the bias. A deconvolution is a useful tool for analyzing neural network features since it allows DeCNN to recreate the input from the CNN output in a top-down way.

4) LRP

Layer-wise Relevance Propagation (LRP) is an explanatory technique for neural network models that handles inputs like images, videos, and text [54]. LRP operates by backward propagating the prediction $f(x)$ across the neural network using specially developed local propagation rules. The propagation follows a conservation property, which is similar to Kirchhoff's rules in electrical circuits. For neurons j and k in successive layers, relevance scores R_k in a particular layer propagate onto neurons in the lower layer according to the rule:

$$R_j = \frac{\sum_k z_{jk}}{z_{jk}} \cdot R_k \tag{5}$$

where z_{jk} indicates the role of neuron j to the significance of neuron k. When the input features are reached, the process comes to a halt.

5) DEEPLIFT

DeepLIFT assigns relevance ratings to input features based on the difference between the activation of a neuron and the activation of a reference neuron. It aids in identifying crucial aspects of the input data that have a major impact on model predictions. DeepLIFT is used to determine the significance of individual characteristics or properties in the input data, providing a more detailed understanding of the model's decision logic [38].

B. INTRINSIC

In the context of model design and training, intrinsic explanation refers to the ideal scenario in which models seamlessly include explainations for their decisions within the model output or allow uncomplicated calculation of explanations from the underlying architecture. This includes making explanation creation an integral component of the model process of creation. The emphasis on intrinsic model expressiveness stems from the belief that models purposefully designed with explainability in mind not only have the ability to produce accurate outputs based on given inputs, but also can produce outputs that serve as optimal explanations for the network's actions, meeting a specified standard of explanatory accuracy [19].

1) TEXT EXPLANATION

By incorporating an explanation-generating component into the original design and undertaking joint training, text explanation approaches achieve interpretability in DNNs. Explanations might be generated word for word or anticipated given a set of options [55]. Joint training enables practitioners to employ cutting-edge models to personalize explanations to the needs of consumers. Obtaining a suitably labeled dataset for the explanation-generating component, on the other hand, is difficult, and joint training adds extra complexity. Inconsistencies in the generated explanations may further undermine confidence in the model's offered explanations [55].

2) MODEL WEIGHTING

Model weighting involves analyzing attention mechanisms in neural networks during a forward pass to understand how different input features are weighted at various phases of model inference. Attention mechanisms assign learned weights to individual phrases in tasks like language translation or sentiment analysis, allowing downstream modules to focus on relevant aspects [56]. Heatmaps illustrating the size and sign of each weight value provide a simple explanation of attention weights. Attention processes promote feature alignment and fusion across multiple feature spaces in multi-modal [57] interaction tasks such as picture captioning or visual question responding, improving model interpretability and providing insights into the model's decision-making process.

3) EXPLAINATION ASSOCIATION

Explanation Association methods provide intrinsic model explanations by linking input items or item properties to human-understandable concepts or objects [58]. In computer vision tasks, these strategies associate input features or latent activations with semantic ideas, connect model predictions to input elements, or map explanations to object saliency maps using regularization terms and model architecture tweaks. Regardless of format or technological differences, these techniques intrinsically connect difficult-to-interpret aspects with easily understood components during joint training, improving the model's interpretability [59].

C. DISTILLATION

Model distillation is a subsequent training explanation technique that condenses the insights engrained in a trained DNNs into a form suitable for user-friendly explanations. This technique entails transforming the complicated knowledge embedded inside the DNNs into a more interpretable representation, allowing users seeking insights into the model's decision-making process to gain a better understanding and transparency [60].

1) LIME

LIME (Local Interpretable Model-agnostic Explanations) is a black-box deep learning model explanation technique [40]. It builds an interpretable model, designated as g, from a class of inherently interpretable models in order to approximate the complex model f locally around a given input x. The interpretable model works with an interpretable representation of the input data, x_0. LIME includes a complexity metric, *(g)*, to assure the model's interpretability. The optimization goal is to find the model g that minimizes a loss function L while taking approximation accuracy and model simplicity into account. The sampling procedure entails perturbing the input space in order to generate a dataset *(Z)* for training the local interpretable model. The strength of LIME is its model-independence and ability to deliver interpretable insights into complex models [61]. The optimization equation is as follows:

$$argmin_g \{L(f, g, \Pi_x) + \Omega(g)\} \tag{6}$$

2) GRAPH BASED

The graph-based distillation involves developing an explanatory graph for a pre-trained CNNs to clarify model predictions. It analyzes semantic patterns in the data and systematically constructs a graph for explanatory purposes. The nodes in this graph reflect various part patterns, and the edges represent the integration or spatial relationship between these patterns [41]. The resulting explanatory graph reveals the model's knowledge hierarchy by displaying activated nodes and their related geographical locations in feature maps. This distillation method uses graphs to collect relational data in a transparent manner, giving intuitive visualizations such as heatmaps or interlinked graphs that emphasize interpretable image elements and their impact on model predictions [62].

3) RULE BASED

Rule-based [63], [64] distillation entails defining semantics across a wide set of "concepts," which can range from clusters of neuron activations to labeled semantic concepts. The method begins with the extraction of concepts from intermediate representations using an autoencoder. These concepts are then used to build a graphical Bayesian causal model, which establishes relationships between model inputs, extracted concepts, and outputs. The causal model makes it easier to identify input features that are considerably causally related to a given categorization result. This method allows practitioners to connect model predictions to previously learned ideas, providing an alternative to activation minimization and deconvolution. In contrast to other methodologies, rule-based distillation represents concepts as heatmaps overlayed on the input, resulting in better interpretable explanations.

4) SHAP

SHAP computes Shapley values for input feature sets to provide explanations [44]. An incomplete altered input is supplied to the model in this method, which is similar to perturbation-based techniques, and the effects of perturbation are measured. SHAP, on the other hand, examines the contribution of adding a feature, as opposed to perturbation-based approaches that focus on feature removal. Each feature is considered a member of a group, and the approach computes each member's contribution to the group. SHAP's core function provides contribution values to individual input features and represents these values as coefficients in a linear model [82].

D. FUZZY

Fuzzy logic is a computer paradigm that allows the depiction of uncertainty and imprecision in decision-making processes within the field of XAI [83]. Unlike traditional binary logic, which rigorously sticks to true or false values, fuzzy logic offers a spectrum of truth degrees, allowing for more nuanced and flexible information interpretation [23]. Fuzzy logic serves as an explanation approach in the context of XAI by providing a structured and interpretable framework for expressing complex interactions between input variables and model outputs. Fuzzy logic systems generate human-understandable explanations using language variables

TABLE 1. Summary of explainability techniques with different metrics.

Ref No	Model Type	Explanability Technique	Problem Type	Data Type	Performance Metrics
[65]	VGG16, ResNet50, AlexNet	Saliency Map & DeepLIFT	Classification	Images	Accuracy/Heatmap
[66]	VGG16 , ResNet50, Xception	Fuzzy Logic	Classification	Images	Accuracy/Mathew's corr & Cohen's kappa
[67]	CNN	Post-hoc	Classification	Signals	Accuracy /Feature Visualization
[68]	CNN	SHAP	Regression	Tabular	AUC ROC / Feature Importance
[69]	Autoencoders	Distillation	Timeseries	Tabular	AUC /Feature Importance
[70]	Random Forest	LIME and SHAP	Classification	Tabular	Accuracy /Feature Importance
[71]	RNN	Attention Mechanism	Regression	Tabular	Attention Specificity
[72]	EAMNet	Saliency Map	Classification	Images	AUC/ ROI
[73]	Likelihood Ratio	Rule Based	Classification	Tabular	Relative Feature's Weight
[74]	DNN	Fuzzy Logic	Classification	Tabular	Accuracy/Feature Importance
[75]	DNN	DeepLIFT	FeatureSelection	Tabular	Ranking Quality Score
[76]	DNN	Visualization	Classification	Images	Heatmap
[77]	VGGNet& DenseNet	Saliency Maps & GradCAM	Classification	Images	Accuracy
[78]	COVID-Net	Fuzzy Logic	Classification	Images	Accuracy
[35]	ANN-DNN	LIME	Regression	Tabular	Probability
[79]	VGGNet & DenseNet	Saliency Maps & GradCAM	Classification	Images	Accuracy
[80]	Transer Learning & DNN	LIME & GradCAM	Classification	Images	Precision/Recall
[81]	DeepFCM	Fuzzy Logic	Classification	Text & Images	Accuracy /Specificity/Sensitivity

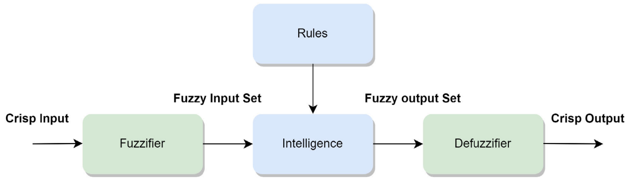


FIGURE 4. Fuzzy logic architecture.

and rules, allowing for a clear grasp of the decision-making process within artificial intelligence algorithms. Because of its inherent interpretability, fuzzy logic is a significant tool for improving the explainability and trustworthiness of machine learning systems, particularly in fields where clear and intuitive insights into decision reasons are critical [84]. Additionally, figure 4 shows how fuzzy logic transforms crisp input data into linguistic variables, applies fuzzy rules to make decisions, and produces crisp output based on fuzzy inference and defuzzification. This process enables handling of uncertainty and imprecision in decision-making, making fuzzy logic a valuable tool in AI and control systems.

1) LAYER-BASED

The implementation of fuzzy logic in CNNs improves their interpretability [85]. Fuzzy clustering groups features, providing more clarity into learned patterns. Fuzzy inference and rule-based systems create language rules by connecting raw features to meaningful concepts. Using fuzzy classifiers increases decision boundary flexibility while accepting uncertainty [86]. Membership functions describe model uncertainty by quantifying feature importance. Human-readable explanations are provided by rule-based systems, which promote transparency. Fuzzy clustering recognizes cluster medoids, which aids interpretation [87].

2) PREDICTION-BASED

Fuzzy logic, used after CNNs prediction, improves explainability by converting model outputs into interpretable insights [88]. Fuzzy inference methods label predictions with language labels and degrees of membership, offering a human-readable context. This linguistic interpretation aids in conveying the model's uncertainty and logic. Fuzzy logic allows for the creation of clear, rule-based explanations that connect individual attributes to the ultimate forecast. By introducing unclear reasoning into the post-prediction stage, the CNN's decision-making procedure becomes more accessible and understandable [89].

V. COMPARATIVE ANALYSIS

There are many reviews on the topic of the explainability of deep learning in healthcare. Most of them focus on general methods of XAI [27], [37], [90], [91], [92]. This review contributes significantly in two key areas. For starters, it focuses on deep learning explanations, with a particular emphasis on applications in the healthcare industry. Second, it pioneers the use of fuzzy logic as a novel technique for improving explainability in machine and deep learning systems. In our paper evaluation procedure, we focused on recent publications within the last five years. However, we noticed constraints, particularly in the availability of articles on healthcare and XAI. As a result, we conducted a review of 18 implementation papers that investigated the practical application of AI models with a focus on explainability. In addition, to fill a vacuum in current research, we thoroughly analyzed 13 survey publications. These survey studies provide in-depth insights and analyses on the present state of research in our subject.

The table 1 presents a complete review of several methodologies in current works on Explainable Artificial Intelligence. Several topics are examined, including model selection, explainability methodologies, problem and data kinds, and performance indicators. VGG16, ResNet50, DenseNet and CNNs are common selections throughout studies, demonstrating their versatility and effectiveness in tackling a variety of tasks. Saliency maps, DeepLIFT, SHAP,

TABLE 2. Overview of the XAI techniques used in various review papers.

Ref No	Visualization	Intrinsic	Distillation	Fuzzy
[24]	✓	×	×	×
[11]	✓	✓	✓	×
[25]	✓	×	✓	×
[26]	✓	×	✓	×
[27]	✓	✓	×	×
[28]	✓	✓	✓	×
[29]	✓	×	✓	×
[30]	✓	×	✓	×
[31]	✓	×	×	×
[32]	✓	×	×	✓
[34]	✓	×	×	×
[33]	✓	✓	✓	×
[93]	✓	×	×	✓

GrandCAM and Fuzzy Logic are popular explainability techniques that highlight the need for interpretable methods in understanding model decisions and establishing trust in AI systems.

A significant number of studies address classification problems, notably in picture and tabular data, which corresponds to frequent machine and deep learning use cases where interpretability is critical for decision-making. These studies primarily use image and tabular data, which reflect their ubiquity in real-world healthcare applications, and the data type used frequently corresponds to the scope of the problem being addressed. Accuracy is a commonly used statistic in research, stressing the significance of model accuracy. Additional measures, like AUC, feature significance, and correlation coefficients, demonstrate the varied evaluation methodologies used.

Techniques like as SHAP, LIME, Fuzzy Logic, and Likelihood Ratio are frequently used when dealing with tabular data, emphasizing the need to comprehend model decisions in situations where structured data is critical. Visual interpretability in image classification is demonstrated by the consistent usage of saliency maps, heatmap analysis, and feature visualization. These visual interpretability techniques help not only grasp model predictions but also acquire insight into the significance of various parts in an image.

Several studies use unique methodologies, such as attention specificity, relative feature weights, and a ranking quality score. These developments illustrate ongoing attempts to improve and broaden the toolkit of interpretable deep learning. Overall, the table depicts a diverse set of techniques in interpretable AI, underscoring the need for transparency and comprehension in AI models. The similarities in model selection, explainability methodologies, and performance indicators highlight common problems and best practices in the development of interpretable and trustworthy AI systems.

The table 2 presents a quick overview of the use of numerous technologies in different review papers. it also shows a clear emphasis on visualization, intrinsic, and distillation technologies, indicating their main role in the examined studies. Notably, Fuzzy logic is utilized minimally in the context of explainable XAI approaches, indicating an underexplored aspect. The limited adoption of Fuzzy logic in these studies may be attributed to a relative scarcity of research on its applications for interpretability compared to other techniques.

Visualization technology is the most prevalent, as it is used in all submissions. This frequency emphasizes how important it is in the context of the cited studies, stressing the importance of illustrations in the offered research. Almost all of the entries make use of intrinsic technology, suggesting that its importance is well recognized. Its widespread application indicates that intrinsic visualization approaches are highly valued for their capacity to portray inherent qualities or properties. Distillation technology is also widely used, highlighting its widespread use in the surveyed studies. This indicates a widespread acknowledgment of the necessity for distillation processes, maybe to simplify complex information or improve interpretability.

On the other side, Fuzzy technology seems to be the least used of the listed items, appearing in only a few studies. This indicates a more selective use of fuzzy logic in the research, potentially due to its particular applicability for certain sorts of data or situations.

In our review, we add to this landscape by filling the gap in research on Fuzzy as an explainable XAI technique. Recognizing its underrepresentation, we hope to shed light on Fuzzy logic's usefulness and relevance in improving interpretability. By looking into this underexplored aspect, we want to provide significant insights into the nuanced benefits and prospective applications of fuzzy logic in the field of explainable artificial intelligence. This contribution aims to increase awareness of available approaches and provide a more comprehensive toolkit for researchers and practitioners working on explainable AI.

VI. CONCLUSION

In conclusion, our investigation thoroughly investigates the context of explainable deep learning in healthcare, acknowledging the limitations faced by the complexity of DNNs. By categorizing and critically examining intrinsic, visualization, and distillation methodologies, we were able to reveal their benefits and drawbacks in healthcare applications. One distinguishing feature of the study is the inclusion of fuzzy logic as a critical category, which addresses uncertainty present in medical records and contributes to DNN interpretability. Fuzzy logic, which has traditionally been used in decision-making contexts, provides a unique perspective on breaking down the black box of DNNs and a systematic framework for describing complex decision processes. Furthermore, our work stands out by comparing previous evaluations in terms of the use of fuzzy logic as an explainable technique. We also look at numerous studies that have used explainable methodologies in healthcare, highlighting their applicability and significance. Moving forward, a possible avenue for future research is actual experimentation and comparison of all four techniques: intrinsic, visualization, distillation, and fuzzy logic.

REFERENCES

[1] P. Angelov and E. Soares, "Towards explainable deep neural networks (xDNN)," *Neural Netw.*, vol. 130, pp. 185–194, Oct. 2020.

[2] S. Dong, P. Wang, and K. Abbas, "A survey on deep learning and its applications," *Comput. Sci. Rev.*, vol. 40, May 2021, Art. no. 100379, doi: 10.1016/j.cosrev.2021.100379.

[3] T. T. Pham and Y. Shen, "A deep causal inference approach to measuring the effects of forming group loans in online non-profit microfinance platform," 2017, *arXiv:1706.02795*.

[4] N. Y. Murad, T. Mahmood, A. R. M. Forkan, A. Morshed, P. P. Jayaraman, and M. S. Siddiqui, "Weed detection using deep learning: A systematic literature review," *Sensors*, vol. 23, no. 7, p. 3670, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/7/3670

[5] L. Nie, M. Wang, L. Zhang, S. Yan, B. Zhang, and T.-S. Chua, "Disease inference from health-related questions via sparse deep learning," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 8, pp. 2107–2119, Aug. 2015.

[6] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *J. Field Robot.*, vol. 37, no. 3, pp. 362–386, 2020.

[7] H.-P. Chan, R. K. Samala, L. M. Hadjiiski, and C. Zhou, "Deep learning in medical image analysis," in *Deep Learning in Medical Image Analysis: Challenges and Applications* (Advances in Experimental Medicine and Biology), vol. 1213, G. Lee and H. Fujita, Eds. Cham, Switzerland: Springer, 2020, ch. 1.

[8] M. I. Razzak, S. Naz, and A. Zaib, "Deep learning for medical image processing: Overview, challenges and the future," in *Classification in BioApps: Automation of Decision Making* (Lecture Notes in Computational Vision and Biomechanics), vol. 26, N. Dey, A. Ashour, and S. Borra, Eds. Cham, Switzerland: Springer, 2018, ch. 12.

[9] A. Fourcade and R. H. Khonsari, "Deep learning in medical image analysis: A third eye for doctors," *J. Stomatol., Oral Maxillofacial Surg.*, vol. 120, no. 4, pp. 279–288, 2019.

[10] X. Li, H. Xiong, X. Li, X. Wu, X. Zhang, J. Liu, J. Bian, and D. Dou, "Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond," *Knowl. Inf. Syst.*, vol. 64, no. 12, pp. 3197–3234, Sep. 2022, doi: 10.1007/s10115-022-01756-8.

[11] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021.

[12] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Vaughan, and H. Wallach, "Manipulating and measuring model interpretability," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2018, pp. 1–52.

[13] D. Gunning and D. W. Aha, "DARPA's explainable artificial intelligence program," *AI Mag.*, vol. 40, no. 2, pp. 44–58, Jun. 2019.

[14] J. Schneider, C. Meske, and M. Vlachos, "Deceptive XAI: Typology, creation and detection," *Social Netw. Comput. Sci.*, vol. 5, no. 1, p. 81, 2023.

[15] S. Lapuschkin, S. Waldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking Clever Hans predictors and assessing what machines really learn," *Nature Commun.*, vol. 10, no. 1, p. 1096, 2019.

[16] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *J. Imag.*, vol. 6, no. 6, p. 52, Jun. 2020, doi: 10.3390/jimaging6060052.

[17] T. Lysaght, H. Y. Lim, V. Xafis, and K. Y. Ngiam, "AI-assisted decision-making in healthcare: The application of an ethics framework for big data in health and research," *Asian Bioethics Rev.*, vol. 11, pp. 299–314, Sep. 2019.

[18] W.-C. Juang, M.-H. Hsu, Z.-X. Cai, and C.-M. Chen, "Developing an AI-assisted clinical decision support system to enhance in-patient holistic health care," *PLoS ONE*, vol. 17, no. 10, 2022, Art. no. e0276501.

[19] G. Ras, M. van Gerven, and P. Haselager, "Explanation methods in deep learning: Users, values, concerns and challenges," in *Explainable and Interpretable Models in Computer Vision and Machine Learning* (The Springer Series on Challenges in Machine Learning), H. Escalante et al., Eds. Cham, Switzerland: Springer, 2018.

[20] D. T. Huff, A. J. Weisman, and R. Jeraj, "Interpretation and visualization techniques for deep learning models in medical imaging," *Phys. Med. Biol.*, vol. 66, no. 4, 2021, Art. no. 04TR01.

[21] X. Liu, X. Wang, and S. Matwin, "Improving the interpretability of deep neural networks with knowledge distillation," in *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2018, pp. 905–912.

[22] R. Chimatapu, H. Hagras, A. Starkey, and G. Owusu, "Explainable AI and fuzzy logic systems," in *Proc. 7th Int. Conf. Theory Pract. Natural Comput. (TPNC)*. Dublin, Ireland: Springer, 2018, pp. 3–20.

[23] S. Lima, L. Terán, and E. Portmann, "A proposal for an explainable fuzzy-based deep learning system for skin cancer prediction," in *Proc. 7th Int. Conf. eDemocracy eGovernment (ICEDEG)*, Apr. 2020, pp. 29–35.

[24] Z. Salahuddin, H. C. Woodruff, A. Chatterjee, and P. Lambin, "Transparency of deep neural networks for medical image analysis: A review of interpretability methods," *Comput. Biol. Med.*, vol. 140, Jan. 2022, Art. no. 105111.

[25] S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R. M. Rao, T. D. Kelley, D. Braines, M. Sensoy, C. J. Willis, and P. Gurram, "Interpretability of deep learning models: A survey of results," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov. (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, Aug. 2017, pp. 1–6.

[26] 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.

[27] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," 2019, *arXiv:1910.10045*.

[28] E. Puiutta and E. M. Veith, "Explainable reinforcement learning: A survey," 2020, *arXiv:2005.06247*.

[29] C. Patricio, J. C. Neves, and L. F. Teixeira, "Explainable deep learning methods in medical image classification: A survey," *ACM Comput. Surv.*, vol. 56, no. 4, pp. 1–41, 2023.

[30] T. Speith, "A review of taxonomies of explainable artificial intelligence (XAI) methods," in *Proc. ACM Conf. Fairness, Accountability, Transparency*, 2024, pp. 2239–2248.

[31] M. A. Gulum, C. M. Trombley, and M. Kantardzic, "A review of explainable deep learning cancer detection models in medical imaging," *Appl. Sci.*, vol. 11, no. 10, p. 4573, 2021.

[32] J. Cao, T. Zhou, S. Zhi, S. Lam, G. Ren, Y. Zhang, Y. Wang, Y. Dong, and J. Cai, "Fuzzy inference system with interpretable fuzzy rules: Advancing explainable artificial intelligence for disease diagnosis—A comprehensive review," *Inf. Sci.*, vol. 662, Mar. 2024, Art. no. 120212. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025524001257

[33] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and U. R. Acharya, "Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)," *Comput. Methods Programs Biomed.*, vol. 226, Nov. 2022, Art. no. 107161, doi: 10.1016/j.cmpb.2022.107161.

[34] Y.-C. Wang, T.-C. T. Chen, and M.-C. Chiu, "An improved explainable artificial intelligence tool in healthcare for hospital recommendation," *Healthcare Anal.*, vol. 3, Nov. 2023, Art. no. 100147, doi: 10.1016/j.health.2023.100147.

[35] Y.-C. Wang, T.-C. T. Chen, and M.-C. Chiu, "A systematic approach to enhance the explainability of artificial intelligence in healthcare with application to diagnosis of diabetes," *Healthcare Anal.*, vol. 3, Nov. 2023, Art. no. 100183, doi: 10.1016/j.health.2023.100183.

[36] R. Goebel, A. Chander, K. Holzinger, F. Lecue, Z. Akata, S. Stumpf, P. Kieseberg, and A. Holzinger, "Explainable AI: The new 42?" in *Proc. Mach. Learn. Knowl. Extraction, 2nd IFIP TC 5, TC 8/WG 8.4, 8.9, TC 12/WG 12.9 Int. Cross-Domain Conf.*. Hamburg, Germany: Springer, Aug. 2018, pp. 295–303.

[37] L. H. Gilpin, A. R. Paley, M. A. Alam, S. Spurlock, and K. J. Hammond, "'"Explanation' is not a technical term: The problem of ambiguity in XAI," 2022, *arXiv:2207.00007*.

[38] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3145–3153.

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

[40] M. Tulio Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," 2016, *arXiv:1606.05386*.

[41] Y. Zhang, K. Song, Y. Sun, S. Tan, and M. Udell, "'Why should you trust my explanation?' understanding uncertainty in LIME explanations," 2019, *arXiv:1904.12991*.

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

[43] M. Kinkead, S. Millar, N. McLaughlin, and P. O'Kane, "Towards explainable CNNs for Android malware detection," *Proc. Comput. Sci.*, vol. 184, pp. 959–965, Jan. 2021.

[44] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–10.

[45] P. Meddage, I. Ekanayake, U. S. Perera, H. M. Azamathulla, M. A. M. Said, and U. Rathnayake, "Interpretation of machine-learning-based (black-box) wind pressure predictions for low-rise gable-roofed buildings using Shapley additive explanations (SHAP)," *Buildings*, vol. 12, no. 6, p. 734, 2022.

[46] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," 2017, *arXiv:1708.08296*.

[47] L. Arras, F. Horn, G. Montavon, K.-R. Müller, and W. Samek, "'What is relevant in a text document?': An interpretable machine learning approach," *PLoS ONE*, vol. 12, no. 8, 2017, Art. no. e0181142.

[48] J. Zhu, J. Yao, B. Han, J. Zhang, T. Liu, G. Niu, J. Zhou, J. Xu, and H. Yang, "Reliable adversarial distillation with unreliable teachers," 2021, *arXiv:2106.04928*.

[49] T. Gomez, T. Fréour, and H. Mouchère, "Metrics for saliency map evaluation of deep learning explanation methods," in *Proc. Int. Conf. Pattern Recognit. Artif. Intell.*, Paris, France, 2022, pp. 84–95.

[50] T. N. Mundhenk, B. Y. Chen, and G. Friedland, "Efficient saliency maps for explainable AI," 2019, *arXiv:1911.11293*.

[51] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," *Univ. Montreal*, vol. 1341, no. 3, pp. 1–13, 2009.

[52] M. B. Muhammad and M. Yeasin, "Eigen-CAM: Class activation map using principal components," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–7.

[53] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*. Zurich, Switzerland: Springer, Sep. 2014, pp. 818–833.

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

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

[56] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[57] D. Teney, P. Anderson, X. He, and A. V. D. Hengel, "Tips and tricks for visual question answering: Learnings from the 2017 challenge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4223–4232.

[58] J. Li, W. Monroe, and D. Jurafsky, "Understanding neural networks through representation erasure," 2016, *arXiv:1612.08220*.

[59] D. Alvarez Melis and T. Jaakkola, "Towards robust interpretability with self-explaining neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–10.

[60] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2020.

[61] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 1–9.

[62] Q.-S. Zhang and S.-C. Zhu, "Visual interpretability for deep learning: A survey," *Frontiers Inf. Technol. Electron. Eng.*, vol. 19, no. 1, pp. 27–39, 2018.

[63] M. Harradon, J. Druce, and B. Ruttenberg, "Causal learning and explanation of deep neural networks via autoencoded activations," 2018, *arXiv:1802.00541*.

[64] S. Kabir, M. S. Hossain, and K. Andersson, "An advanced explainable belief rule-based framework to predict the energy consumption of buildings," *Energies*, vol. 17, no. 8, p. 1797, 2024.

[65] L. A. de Souza, R. Mendel, S. Strasser, A. Ebigbo, A. Probst, H. Messmann, J. P. Papa, and C. Palm, "Convolutional neural networks for the evaluation of cancer in barrett's esophagus: Explainable AI to lighten up the black-box," *Comput. Biol. Med.*, vol. 135, Aug. 2021, Art. no. 104578, doi: 10.1016/j.compbiomed.2021.104578.

[66] M. M. Hasan, M. M. Hossain, M. M. Rahman, A. Azad, S. A. Alyami, and M. A. Moni, "FP-CNN: Fuzzy pooling-based convolutional neural network for lung ultrasound image classification with explainable AI," *Comput. Biol. Med.*, vol. 165, Oct. 2023, Art. no. 107407, doi: 10.1016/j.compbiomed.2023.107407.

[67] B. M. Maweu, S. Dakshit, R. Shamsuddin, and B. Prabhakaran, "CEFEs: A CNN explainable framework for ECG signals," *Artif. Intell. Med.*, vol. 115, May 2021, Art. no. 102059, doi: 10.1016/j.artmed.2021.102059.

[68] W. Caicedo-Torres and J. Gutierrez, "ISeeU: Visually interpretable deep learning for mortality prediction inside the ICU," *J. Biomed. Informat.*, vol. 98, Oct. 2019, Art. no. 103269, doi: 10.1016/j.jbi.2019.103269.

[69] Z. Che, S. Purushotham, R. Khemani, and Y. Liu, "Distilling knowledge from deep networks with applications to healthcare domain," 2015, *arXiv:1512.03542*.

[70] M. Panda and S. R. Mahanta, "Explainable artificial intelligence for healthcare applications using random forest classifier with LIME and SHAP," 2023, *arXiv:2311.05665*.

[71] J. Rebane, I. Karlsson, and P. Papapetrou, "An investigation of interpretable deep learning for adverse drug event prediction," in *Proc. IEEE 32nd Int. Symp. Computer-Based Med. Syst. (CBMS)*, Jun. 2019, pp. 337–342.

[72] W. Liao, B. Zou, R. Zhao, Y. Chen, Z. He, and M. Zhou, "Clinical interpretable deep learning model for glaucoma diagnosis," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 5, pp. 1405–1412, May 2020.

[73] M. Ennab and H. Mcheick, "Designing an interpretability-based model to explain the artificial intelligence algorithms in healthcare," *Diag.*, vol. 12, no. 7, p. 1557, 2022.

[74] F. Aghaeipoor, M. Sabokrou, and A. Fernández, "Fuzzy rule-based explainer systems for deep neural networks: From local explainability to global understanding," *IEEE Trans. Fuzzy Syst.*, vol. 31, no. 9, pp. 3069–3080, Sep. 2023.

[75] J. Figueroa Barraza, E. Lopez Droguett, and M. R. Martins, "Towards interpretable deep learning: A feature selection framework for prognostics and health management using deep neural networks," *Sensors*, vol. 21, no. 17, p. 5888, 2021.

[76] A. K. Mondal, A. Bhattacharjee, P. Singla, and A. P. Prathosh, "XViTCOS: Explainable vision transformer based COVID-19 screening using radiography," *IEEE J. Translational Eng. Health Med.*, vol. 10, pp. 1–10, 2022.

[77] T. Mahmud, K. Barua, S. U. Habiba, N. Sharmen, M. S. Hossain, and K. Andersson, "An explainable AI paradigm for Alzheimer's diagnosis using deep transfer learning," *Diagnostics*, vol. 14, no. 3, p. 345, 2024.

[78] Q. Hu, F. N. B. Gois, R. Costa, L. Zhang, L. Yin, N. Magaia, and V. H. C. de Albuquerque, "Explainable artificial intelligence-based edge fuzzy images for COVID-19 detection and identification," *Appl. Soft Comput.*, vol. 123, Jul. 2022, Art. no. 108966. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494622003064

[79] K. Barua, T. Mahmud, A. Barua, N. Sharmen, N. Basnin, D. Islam, M. S. Hossain, K. Andersson, and S. Hossain, "Explainable AI-based humerus fracture detection and classification from X-ray images," in *Proc. 26th Int. Conf. Comput. Inf. Technol. (ICCIT)*, 2023, pp. 1–6.

[80] S. Sarp, M. Kuzlu, E. Wilson, U. Cali, and O. Guler, "The enlightening role of explainable artificial intelligence in chronic wound classification," *Electronics*, vol. 10, no. 12, p. 1406, Jun. 2021, doi: 10.3390/electronics10121406.

[81] A. Feleki, I. D. Apostolopoulos, S. Moustakidis, E. I. Papageorgiou, N. Papathanasiou, D. Apostolopoulos, and N. Papandrianos, "Explainable deep fuzzy cognitive map diagnosis of coronary artery disease: Integrating myocardial perfusion imaging, clinical data, and natural language insights," *Appl. Sci.*, vol. 13, no. 21, p. 11953, Nov. 2023, doi: 10.3390/app132111953.

[82] D. Bowen and L. Ungar, "Generalized SHAP: Generating multiple types of explanations in machine learning," 2020, *arXiv:2006.07155*.

[83] M. Islam, D. T. Anderson, A. J. Pinar, T. C. Havens, G. Scott, and J. M. Keller, "Enabling explainable fusion in deep learning with fuzzy integral neural networks," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 7, pp. 1291–1300, Jul. 2020.

[84] K. Cohen, L. Bokati, M. Ceberio, O. Kosheleva, and V. Kreinovich, "Why fuzzy techniques in explainable AI? Which fuzzy techniques in explainable AI?" in *Proc. Annu. Conf. North Amer. Fuzzy Inf. Process. Soc. (NAFIPS)*. Cham, Switzerland: Springer, 2022, pp. 74–78.

[85] R. Chimatapu, H. Hagras, M. Kern, and G. Owusu, "Hybrid deep learning type-2 fuzzy logic systems for explainable AI," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, 2020, pp. 1–6.

[86] C. Xie, D. Rajan, D. K. Prasad, and C. Quek, "An embedded deep fuzzy association model for learning and explanation," *Appl. Soft Comput.*, vol. 131, Dec. 2022, Art. no. 109738.

[87] M. Yeganejou, S. Dick, and J. Miller, "Interpretable deep convolutional fuzzy classifier," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 7, pp. 1407–1419, Jul. 2020.

[88] X. Gu and X. Cheng, "Distilling a deep neural network into a Takagi–Sugeno–Kang fuzzy inference system," 2020, *arXiv:2010.04974*.

[89] V. Pasquadibisceglie, G. Castellano, A. Appice, and D. Malerba, "FOX: A neuro-fuzzy model for process outcome prediction and explanation," in *Proc. 3rd Int. Conf. Process Mining (ICPM)*, Oct. 2021, pp. 112–119.

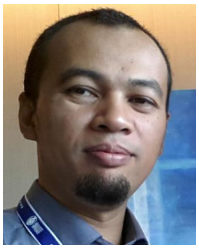
[90] T. Speith, "A review of taxonomies of explainable artificial intelligence (XAI) methods," in *Proc. ACM Conf. Fairness, Accountability, Transparency*, 2022, pp. 2239–2250.

[91] A. M. Antoniadi, Y. Du, Y. Guendouz, L. Wei, C. Mazo, B. A. Becker, and C. Mooney, "Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: A systematic review," *Appl. Sci.*, vol. 11, no. 11, p. 5088, 2021.

[92] V. Chamola, V. Hassija, A. R. Sulthana, D. Ghosh, D. Dhingra, and B. Sikdar, "A review of trustworthy and explainable artificial intelligence (XAI)," *IEEE Access*, vol. 11, pp. 78994–79015, 2023.

[93] M. R. Islam, M. U. Ahmed, S. Barua, and S. Begum, "A systematic review of explainable artificial intelligence in terms of different application domains and tasks," *Appl. Sci.*, vol. 12, no. 3, p. 1353, Jan. 2022, doi: 10.3390/app12031353.

NAFEESA YOUSUF MURAD received the master's degree in data engineering and information management from NEDUET, Pakistan. Currently, she is pursuing the Ph.D. degree in information technology, demonstrating her commitment to advancing her knowledge and expertise in the field. She possesses a strong academic background. In 2023, she coauthored one journal paper and presented two conference publications, showcasing her dedication to research and scholarly contributions.



MOHD HILMI HASAN received the Ph.D. degree in information from Universiti Teknologi PETRONAS. He is an accomplished academician with a proven track record in the higher education sector. Proficient in artificial intelligence, machine learning, fuzzy logic, and data analytics. He brings a wealth of expertise to his field. He is enhancing his credentials as a Distinguished Educator and a Researcher.

MUHAMMAD HAMZA AZAM received the bachelor's degree (Hons.) in computer science from Kohat University of Science and Technology (KUST), Pakistan, in 2017, and the master's degree in information technology from Universiti Teknologi PETRONAS (UTP), in 2021, specializing in artificial intelligence and fuzzy logic. Currently, he is a Research Scientist and a Lecturer with UTP, contributing to the forefront of artificial intelligence focused on fuzzy logic and XAI research.

NADIA YOUSUF received the master's degree in computer science from MAJU, Pakistan. She is currently a Lecturer with Benazir Bhutto Shaheed University, Pakistan, where she imparts knowledge and nurtures the minds of aspiring students. She has solidified her academic foundation. Furthermore, she has contributed as the coauthor to one conference publication, showcasing her dedication to scholarly pursuits and academic collaboration.

JAMEEL SHEHU YALLI (Graduate Student Member, IEEE) received the bachelor's degree in computer science from AI-Qalam University Katsina, Nigeria, in 2010, and the M.Sc. degree in computer and information engineering from International Islamic University Malaysia, in 2015. He is currently pursuing the Ph.D. degree with Universiti Teknologi PETRONAS. From 2015 to 2020, he worked in the industry and later joined Federal University Gusau, as a Lecturer and a Research Assistant with the Department of Computer Science. He has published few articles and coauthored some and still counting. His research interests include the Internet of Things (IoT), security, authentication models, cybercrime, cyber security, cyber and islam, wireless communications, and wireless networks.