

## Explainable AI: Techniques for Transparent Decision-Making in Neural Networks

Madhusudhan M<sup>1</sup>, Dr. Pushpa Mamoria<sup>2</sup>

<sup>1,\*</sup>Research Scholar, Department of Computer Science and Engineering, Chhatrapati Shahu Ji Maharaj University, Kanpur (Formerly Kanpur University) ORCID ID: 0009-0003-1093-8772, Email ID: madhusim04@gmail.com

<sup>2</sup>Department of Computer Science and Engineering, Chhatrapati Shahu Ji Maharaj University, Kanpur (Formerly Kanpur University), Email ID: pushpamamori76@csjmu.ac.in

### Abstract

Artificial Intelligence (AI), particularly deep neural networks, has achieved strong predictive performance across healthcare, finance, autonomous systems, cybersecurity, and natural language processing, yet its increasing complexity has intensified concerns about transparency and interpretability. This review examines Explainable Artificial Intelligence (XAI) techniques for improving transparent decision-making in neural-network-based systems. It synthesizes major categories of explainability, including intrinsic and post-hoc approaches, and critically reviews LIME, SHAP, Grad-CAM, Integrated Gradients, saliency methods, Layer-wise Relevance Propagation, attention-based explanations, and counterfactual explanations. The study compares these methods in terms of model dependency, explanation scope, computational requirements, architectural compatibility, fidelity, robustness, and interpretability. It also evaluates current approaches using criteria such as faithfulness, sensitivity, stability, human comprehensibility, scalability, and fairness. Applications across healthcare, finance, autonomous systems, cybersecurity, and natural language processing are discussed to demonstrate the practical relevance of explainable neural networks. Key limitations include explanation fragility, adversarial manipulation, computational cost, bias, privacy concerns, and the absence of standardized evaluation frameworks. Finally, the review identifies emerging directions involving XAI 2.0, concept-based explanations, counterfactual recourse, multimodal and generative AI, and human-centered explainability, emphasizing the need for reliable, actionable, and trustworthy neural-network explanations. These directions are essential for responsible deployment and effective human oversight in high-stakes decision environments.

**Keywords:** Explainable Artificial Intelligence (XAI); Neural Networks; Model Interpretability; Transparent Decision-Making; Explainability Evaluation.

### Introduction

Artificial Intelligence (AI) is a major application of modern computing systems with the ability to predict, classify, recognize and make decisions automatically in various areas. More specifically, deep learning has made a significant breakthrough in the ability to learn highly complex and non-linear representations from big and diverse data sets. Neural network architectures like convolutional neural networks, recurrent neural networks and the newer deep learning models have had high predictive values in fields ranging from healthcare to finance, autonomous systems, cyber security and natural language processing. But, as the models become more complex, there has been a fundamental interpretability issue. Neural networks with a large number of parameters can make good predictions, but can't explain what they learned about the relationships between its inputs and how specific internal representations led to the decision. Hence the need for careful assessment of interpretability has emerged to see if the explanation actually conveys some useful information about the behavior of the model or simply provides a nice-looking answer [1].

In response to this opacity, a new research field has been developed, called Explainable Artificial Intelligence (XAI). XAI involves techniques, principles and frameworks designed to improve the explainability of AI decision making to developers, domain experts, decision makers, and the users of the AI systems. In contexts where neural-network

outputs can affect high-stakes decisions, such as in regulatory compliance, safety, reliability, fairness and accountability become especially important, as these factors may become concerns when outputs are not explained. Through XAI, the company will extend beyond the visualization of models' outputs to cover broader needs and requirements of responsible AI, such as human oversight, trustworthiness, traceability and interpretability. Existing research therefore has created taxonomies that distinguish transparent models from post-hoc explanations, and that categorize the different types of explanation techniques based on the characteristics of the models they used, on how much information can be conveyed to the user, and on the information itself [2].

Despite the black-box nature of neural networks, they still pose significant challenges for reliable human–AI interaction. A model can yield a technically correct output without users being able to tell if the model output was influenced by any evidence, spurious correlations, biased features or undesirable shortcuts. Explainability is therefore vital to boost user trust, for model debugging, error analysis, bias detection, validation, and to support for informed decision making. Most of the previous work on XAI has focused on the need to use multiple explanations in order to open a black-box system, since a single technique will not meet all interpretability needs for a particular system, user, or application [3].

In this context, the most important approaches to transparent decision making in neural networks are investigated such as LIME, SHAP, saliency maps, Integrated Gradients, Layer-wise Relevance Propagation, Grad-CAM, attention-based explanations and counterfactual explanations. A comparative analysis of these approaches is carried out from the perspective of interpretability, model dependence, explanation fidelity, computational demands, robustness and scope of application for various types of neural network architectures. This also brings together evaluation strategies and application areas, current research gaps, and research directions. The article seeks to explain the strength and weakness of existing XAI techniques and draw out the key agenda points for the future of neural network decision-making that is more reliable, transparent and human-centered.

## **2. Literature Review**

### **2.1 Artificial Intelligence and Neural Networks**

Neural networks are one of the most important types of AI models that can be used for learning nonlinear relationships from high-dimensional and complex data. They have a multi-layered computational architecture that can be used to learn increasingly abstract representations from the input information, which enables, among others, classification, prediction, recognition, and automated decision making. The relationship between an input and the prediction becomes more complicated to follow, however, as the model's complexity increases. In this respect, this is especially noticeable when the user gets a result from a black-box predictive system without having an easy-to-understand presentation of the reasoning process which produced the result. Guidotti et al. listed some approaches for explaining black-box models, and highlighted that explanatory methods can vary significantly in terms of their type and whether they can provide local or global understanding, model-specific or model-agnostic, and in terms of the type of explanatory information they produce [4].

This interpretability problem gets worse as the depth of neural network architectures increases and internal representation becomes more and more complex. CNNs are widely used for visual feature extraction, and RNNs for sequential dependencies. In more recent architectures, such as Transformer models, the capacities are huge and the representations are complex. These architectures can give the user very strong predictive powers but their inner calculations are not always easy to explain to the end user. This has led to the study of how these models make decisions to be a central issue in the field of explainable AI.

### **2.2 Concept of Explainable AI**

Explainable AI (XAI) is a broad term used for methods and principles that aim to make the actions, forecasts, and decision making of AI systems comprehensible to humans. Explainability is related to interpretability, but not

necessarily synonymous. Interpretability typically is about the degree of understanding of the internal reasoning or relations of a model, while explainability is about an information produced that communicates the reasons for a given prediction or decision. Vilone and Longo demonstrated that there are several concepts in the XAI literature of what it means to be explained and being interpretable, and that the assessment of an explanation is influenced by its purpose, audience and context [5].

An interesting difference is also between inherent transparency models and methods for post hoc explanations. The former, in the form of transparent systems, enable one to examine parts of the reasoning of their system in a straightforward manner, and the latter are designed to analyze the behaviour of a well-trained system without changing its architecture. In the context of neural networks, post-hoc explanations are especially crucial as the complexity of learned representations can make them hard to understand by direct inspection. Thus, good XAI should not just generate pretty outputs, it should deliver informative, understandable and relevant information to a target audience.

### **2.3 Need for Explainability in Neural Networks**

Especially important for deep neural networks is the need for explainability: They are extremely powerful in predicting an output, but also very complex in structure and representation. Saleem et al. reviewed different methods of interpreting deep neural networks and emphasized the need for global interpretations to better understand the behavior of the models beyond making predictions [6]. This interpretation can be useful to uncover significant components, learned representations, and general decision making processes in neural models.

Explainability also meets practical needs like trust, accountability, model validation, debugging, identification of biases, and human oversight. However, in safety-critical or high impact situations, a correct prediction could be inconclusive if the user doesn't know if the model was based on adequate evidence. Clear explanations may then help the experts to locate incorrect reasoning, unhelpful relationships and to determine if the model's behavior matches domain expectations. The requirements are also applicable to ethical and regulatory aspects since AI systems have an increasing impact on consequential decisions. As a result, explainability is now entering the equation of responsible deployment of neural networks – not just as a visual layer, but as a necessary layer.

## **3. Taxonomy of Explainable AI Techniques**

Explainable AI (XAI) refers to a wide range of techniques that aim to improve the interpretability of the actions of sophisticated predictive systems. There are several goals, mechanisms and levels of transparency in explanation methods, hence a structured taxonomy is required to categorize them. In neural networks, XAI techniques are generally categorized into three groups: intrinsic or interpretable-by-design, post-hoc model-agnostic, and post-hoc model-specific. XAI techniques for neural networks can be divided into three main categories: intrinsic or interpretable-by-design, post-hoc model-agnostic, and post-hoc model-specific. Such categories vary depending on whether interpretability is embedded within the model or applied subsequently to the trained model, whether it is based on a specific model architecture, and the various ways in which the interpretability is communicated, such as features, concepts, visual evidence, examples, and alternative outcomes.

### **A. Intrinsic/Interpretable Models**

Intrinsic interpretability tries to 'build it in' to the architecture or representation of the model itself, rather than waiting until after prediction to provide explanations. Concept-based methods are especially significant because they make an attempt to capture the behavior of neural networks in terms of higher-level concepts that are comprehensible to humans. Kim et al. proposed the Testing with Concept Activation Vectors (TCAV) method to check whether the model predictions are sensitive to the user-defined semantic concept rather than single features [7]. This provides an example of how explainability can go beyond the simple feature attribution and instead provide concept-level reasoning that is more in line with human understanding. In addition to such intrinsic approaches, there are also other explainable-by-design neural architectures, such as attention

mechanisms, which aim at revealing meaningful internal relations.

Another class of interpretable-by-design learning models are prototype-based neural models. In their model, Chen et al. explain image predictions by similarity to previously learned prototypical image regions, enabling the model to reason about its selection using an "if this looks like that then that must be the case" type of logic [8]. These methods give a meaningful explanation as an integral part of the prediction process, diminishing the need for additional explanation layers after the fact.

#### **B. Post-hoc Model-Agnostic Methods**

Post-hoc model agnostic approaches that yield explanations once a model is trained and usually do not require any knowledge of the inner workings of the model. Some of the most important examples are LIME, SHAP, partial dependence techniques and surrogate models. The following methods are typically used to approximate the behavior of a model, quantify feature contributions or describe input–output relationships. They offer the great flexibility that they can be used in various predictive systems, but the explanation they produce can be only an approximation of the system's internal workings and not the full explanation of the model's workings.

#### **C. Post-hoc Model-Specific Methods**

Model-specific explanations take advantage of the properties of a specific architecture of a neural network. Among them are saliency maps, Integrated Gradients, Layer-wise Relevance Propagation, and Grad-CAM. These methods typically rely on gradients, activations, or relevance propagation to determine what the most important information in the input was to a prediction. While more model-specific than model-agnostic methods, they often have the advantage of being more easily accessible in terms of internal computations, and are typically less transferable across vastly different architectures.

#### **D. Example-Based Explanations**

Example-based methods explain predictions by means of representative, contrasting or influential examples. In counterfactual explanations, the change in an input required to get a different prediction is determined; prototype-based explanations give examples that represent the decisions made by the model. In addition, influence based methods can detect training observations that have a significant impact on predictions.

Multiple dimensions are thus necessary for a comprehensive taxonomy of XAI, as opposed to a single classification criterion. They structured the wider XAI space, covering a methodological space, an explanation target space, a model space, and an interpretability space, showing the diversity and overlap of explanation families [9]. This taxonomy is a conceptual framework that allows for the systematic comparison of the major XAI techniques discussed in the next section.

## **4. Major XAI Techniques**

Explainable Artificial Intelligence offers a range of complementary ways of explaining the output of neural-network and other complex machine learning models. These approaches vary in their explanatory scope, model architecture reliance, computational mechanism and the mode of communicating explanations to the user. A few methods approximate models locally, while others rely on gradients, activations, feature attribution, relevance propagation, attention mechanisms, and other approaches to uncover why a specific prediction was made. The key methods that are applicable to transparent neural-network decision making are presented below.

### **4.1 Local Interpretable Model-Agnostic Explanations (LIME)**

LIME is a post-hoc, model agnostic explanation method designed to explain individual predictions of complex classifiers. Whereas global models of the entire distribution are not interpretable, LIME builds a simpler interpretable model in the local neighbourhood of a chosen example. This idea was proposed by Ribeiro et al. (2014), who introduced a perturbation in the input, made predictions with the original black-box model and fit a locally weighted interpretable surrogate that mimics the model's behavior on the instance to be explained [10]. The resulting explanation describes features that are most likely to back up or contradict a specific prediction.

LIME's model agnostic design allows it to be used with various data types and predictive models. Its explanations, however, are local approximations and may be impacted by the construction of the neighborhood, the perturbation strategy and surrogate-model selection.

#### **4.2 Grad-CAM**

Gradient-weighted Class Activation Mapping (Grad-CAM) was initially developed to give visual explanations for convolutional neural networks. Selvaraju et al. proposed a technique called Grad-CAM which employs gradients that cascade into the last convolutional layer to highlight the significance of each feature map to a given prediction [11]. These importance weights are multiplied with activation maps to form a coarse localization map that indicates the areas of the image that are important to the model's decision. In computer vision, Grad-CAM is very useful because one can directly map predictions to spatial information in the input image. It has been helpful for exploring CNN behavior thanks to its intuitive heatmap visualisation, but its applicability is naturally more limited to CNNs with appropriate convolutional representations.

#### **4.3 Integrated Gradients**

Integrated Gradients is a method for attribute features of deep neural networks. Sundararajan et al. suggested the method based on axiomatic requirements for attribution (sensitivity and implementation invariance) [12]. Integrated Gradients calculates gradients along a path from a baseline to the input, rather than just at the input. This results in an attribution score that shows the percentage attributed of one's prediction to each input dimension. It is easy to compute, and works well for differentiable neural networks. A key practical question is, however, baseline selection, as this can impact the attribution.

#### **4.4 SHapley Additive exPlanations (SHAP)**

SHAP gives explanations for predictions based on principles from the cooperative game theory. Lundberg and Lee proposed a unified framework where each feature is used to make the prediction and is allocated an attribution value that represents the marginal value with respect to the possible combinations of features [13]. SHAP values give an additive explanation of the effects of features on a prediction compared to a reference or expected model prediction. A key benefit is that it can offer local explanations to individual predictions as well as aggregated interpretations of model behaviour. SHAP also has desirable theoretical properties with respect to feature attribution. However, it is computationally costly to calculate the exact Shapley values, especially for complex models and high dimensional data sets, so approximation techniques are often used in practice.

#### **4.5 Saliency Maps and SmoothGrad**

Saliency maps are used to explain neural-network outputs by quantifying the change in output based on changes in input, either visually or as a feature. Gradient-based saliency methods typically calculate the gradient of the prediction to be made on the target with respect to the input features; the higher the gradient, the more influential the feature. However, the traditional saliency maps can have a lot of visual noise. In order to increase the quality of gradient-based explanation, Smilkov et al. proposed to generate multiple noisy versions of an input, run saliency maps on them, and average the resulting gradients to get SmoothGrad [14]. The method yields smoother and more often visually interpretable attribution maps without the need of adapting the neural network. Even with the better visualization, saliency-based explanations are subjective to the properties used during the computation of the gradient, and do not ensure that highlighted areas capture all aspects of model reasoning.

#### **4.6 Layer-Wise Relevance Propagation**

The aim of Layer-wise Relevance Propagation (LRP) is to understand the predictions made by neural networks by propagating the relevance of an output back through successive layers of the network. Instead of merely using gradient information from the lower level neurons, the relevance scores are propagated up through the layers to lower level neurons and then to the original input features. LRP was analyzed in the context of general understanding of deep neural networks in [15] and its ability to determine the parts of the network that depended on the input component for making a decision was described. For image-based tasks, these scores can be

represented as relevance maps on which the regions with positive scores indicate regions that are helpful to the prediction and regions with negative scores indicate regions that are considered unhelpful to the prediction. The method gives a detailed explanation of the models but could be specific to the network architecture and propagation rules set for the various layers.

#### **4.7 Attention-Based Explanations**

Attention mechanisms are used to weight different parts of an input differently, and are frequently viewed as a way to gain insight into the information that a neural model attended to. This has resulted in the use of attention weights as possible explanations, especially in natural language processing and architectures that depend on sequences such as neural networks. But all the connections between attention and actual interpretability are not agreed. Jain and Wallace showed that attention distributions do not necessarily reflect other metrics of feature importance and that significantly different patterns of attention can lead to the same predictions [16]. In their results, they conclude that the attention weights are not necessarily considered as faithful explanations of an attention-based model's decision process.

To make sure that attention weights have interpretable evidence, Serrano and Smith also explored what happens when the highly attended inputs are modified or deleted from the model [17]. In their conclusions, they argue that useful information about model behavior might be present in attention, but may not be reliable in all models and instances. So attention-based visualization is useful for interpretation, but attention weights should be used in conjunction with other explanatory evidence, and should not be viewed as explanatory evidence.

#### **4.8 Counterfactual Explanations**

Unlike feature-attribution approaches, counterfactual explanations explain how an input's attributes would change to produce a different prediction. Wachter et al. [18] formalized counterfactual explanations for automated decision systems as a way to give meaningful information about the decision made by the model without the user having to explore the full internal structure of the underlying black box. For instance, a counterfactual explanation might suggest, not only which variables lead to the denial of a financial application, but also the minimum changes needed for the financial application to be approved. This is "what if" thinking that is especially useful in decision support, as it helps to convert the model behavior into something potentially actionable.

However, sometimes a single counterfactual isn't enough flexible for users as there might be a number of different changes that produce the same desired effect. This is why Mothilal et al. suggested producing multiple counterfactuals to allow users to consider multiple possible cases to view different plausible options than the one presented to them [19]. Diversity, proximity, feasibility, and actionability are therefore key attributes of good counterfactual explanations. Even then, generated counterfactuals need to be carefully examined to ensure that the changes that are proposed are realistic, valid and practically achievable.

The techniques together show that there is no single all-purpose explanation mechanism for transparent decision making by neural networks. Model-agnostic methods offer flexibility in the use of models, attribution methods pinpoint important features, visual approaches go beyond visualizing the data and can reveal spatial evidence, relevance-based tools can uncover neural contributions to predictions, attention-based methods offer potentially informative internal signals, and counterfactual explanations can reveal what the model might decide about alternative decision scenarios. This leads to the need for systematic comparison of their different properties based on criteria of interpretability, fidelity, stability, computational complexity and applicability to architecture.

## **5. Comparative Analysis of XAI Techniques**

The model dependency and scope of explanations, the computational cost and the fidelity of XAI methods vary. Model-specific approaches involve gradients, activations, etc. from within the model, whereas model-agnostic approaches are based primarily on model inputs and outputs.

DeepLIFT computes the difference between the activations of the model relative to a reference input and provides

scores of how much a specific input feature contributes to the overall difference. It can explain local aspects in detail and is restricted to the internal structure of neural network [20]. In contrast to the perturbation-based approaches, which change portions of an input and assess changes in the prediction, these approaches change the input. Fong and Vedaldi demonstrated meaningful perturbations that can help identify “significant” regions in images, but the quality of explanations is affected by the perturbation method and optimization process [21]. Gradient-based attribution methods are in general efficient, as they leverage information provided by backpropagation. Ancona et al. showed that these methods are not the same in estimating the features' contribution and can be influenced by gradient saturation and instability [22]. Unlike that, RISE is model-agnostic and creates visual explanations of the model by randomly masking input areas and monitoring the resulting changes in prediction scores. This makes it easier to change the architecture, but it adds a lot of model evaluations, and hence computational cost [23].

**Table 1. Comparative characteristics of major XAI approaches**

| XAI approach                             | Model dependency       | Explanation scope | Main mechanism                        | Architectural compatibility       | Computational requirement | Key advantage                                       | Key limitation                                         |
|------------------------------------------|------------------------|-------------------|---------------------------------------|-----------------------------------|---------------------------|-----------------------------------------------------|--------------------------------------------------------|
| DeepLIFT / propagation-based attribution | Model-specific         | Local             | Activation-difference propagation     | Differentiable neural networks    | Low–moderate              | Detailed feature contributions                      | Requires internal model access and reference selection |
| Meaningful perturbation                  | Largely model-agnostic | Local             | Input perturbation and optimization   | Primarily image models; adaptable | High                      | Intuitive causal-style importance assessment        | Perturbation design and optimization sensitivity       |
| Gradient-based attribution               | Model-specific         | Local             | Gradients/backpropagation             | Differentiable neural networks    | Low                       | Computationally efficient                           | Gradient saturation and instability                    |
| RISE                                     | Model-agnostic         | Local/visual      | Random input masking                  | Black-box image classifiers       | High                      | Does not require gradients or internal model access | Large number of model evaluations                      |
| Extremal perturbation                    | Largely model-agnostic | Local/visual      | Optimized preservation/deletion masks | Primarily visual neural models    | High                      | Spatially coherent explanations                     | Optimization cost and modality dependence              |

A further benefit of extremal perturbation is that it provides additional interpretation clues by indicating compact regions that, when preserved or removed, have significant impacts on predictions. These explanations may be more spatially coherent than raw gradient maps, but require more computational resources for optimization and aren't as practical for some applications [24].

Model specific approaches tend to be more efficient computationally and more closely related to the neural computations, while model agnostic perturbation methods offer more flexibility with typically more inference operations. Local techniques provide explanations for individual predictions, and global techniques provide summary explanations of the behavior of models. Hence, the selection of an XAI method should be based on the model architecture, data type, intended explanation purpose, computational resources and fidelity.

The comparison illustrates the trade-offs between explanatory fidelity, interpretability, flexibility and computational efficiency when making decisions with neural networks transparently. As such, assessment of XAI should go beyond what the explanation looks like to test for faithfulness, stability, usefulness, and technical appropriateness of the explanation for the underlying model and the decision situation.

## **6. Evaluation of Explainability**

The problem of evaluation is essential, as it is not easy to produce a good, even a believable, explanation that is not faithful, stable, or helpful. Explainability needs to be evaluated in several ways, such as explanation quality, usefulness to humans, technical performance, and ability to provide insight into problematic model behavior.

### **6.1 Explanation Quality**

The quality of explanations is related to the accuracy of the explanation in representing the real reasoning in the model. Fidelity, faithfulness, completeness and consistency are some of the important criteria. Samek et al. highlighted the importance of quantitative assessment of neural-network visualizations, not only through their subjective visual inspection, but also by considering the contribution of features identified in the visualization to model prediction [25]. Benchmarking evaluation also allows for an organization to systematically compare different interpretability techniques. Hooker et al. have shown that attribution techniques can be tested by systematically changing and/or removing certain features deemed important and observing the impact on the model's performance [26]. Yeh et al. also quantified infidelity and sensitivity of explanations, and pointed out the significance of identifying whether attribution scores reflect model behavior and are robust to small variations in the input [27].

### **6.2 Human-Centered Evaluation**

An explanation that is technically correct is not necessarily useful to human users. Comprehensibility, usefulness, trust and explanations that facilitate informed decision making are therefore aspects of human centered evaluation. Hase and Bansal explored the effect of algorithmic explanations on helping users predict the behavior of a model, and demonstrated that the effectiveness of the explanation must be determined by the measurable increase in human understanding not by the presence of the explanation [28]. This is especially relevant in high impact applications where explanations need to convey relevant information in an understandable way to enable suitable human judgement.

### **6.3 Technical Evaluation**

Technical evaluation is based on robustness, stability, computational efficiency and scalability. OpenXAI was introduced by Agarwal et al. as a transparent benchmarking framework for systematically assessing various explanation techniques based on standardised metrics and datasets [29]. Such frameworks allow for more reproducible comparisons of various XAI methods. Hedström et al. created Quantus, a comprehensive set of quantitative metrics for assessing neural-network explanations on various properties including faithfulness, robustness, complexity, localization, and randomization [30]. These tools enable more systematic assessment and

help to move away from a purely qualitative evaluation.

#### **6.4 Fairness and Bias**

Explainability can also help to uncover adverse dependencies in model decisions. Feature-attribution and visualization techniques could help demonstrate the extent to which predictions are driven by information that is inappropriate and/or potentially discriminatory. But there are two differentiated issues, however: quality of the explanation and fairness of the explanation: An explanation can report on a biased model without being a fair model. Evaluation should thus consider the following questions: Does an explanation reveal a reliance on a biased feature? Do similar cases get similar explanations? Does the mechanism of explanation lead to misleading interpretation across groups?

In general, the best evaluation of XAI should integrate the faithfulness and fidelity evaluation measures, the human-centered assessment, the robustness and stability testing, computational factors, and the fairness-based analysis. Explanation quality is not just one metric, but more multidimensional and standardized assessment is needed for reliable XAI.

## **7. Applications of Explainable Neural Networks**

In areas where automated decisions require to be explainable, audit and trustworthy, such as in applications, explainable neural networks have become more and more prominent. For instance, in NLP, neural models can produce rationales that are human-readable and which clarify the parts of the text that contribute to the prediction, showcasing how explanation mechanisms can enhance understanding of the behavior of the models [31].

### **7.1 Healthcare**

XAI can be used in healthcare for disease diagnosis, medical-image interpretation, and clinical decision support. Explanations can assist clinicians in determining which symptoms, features, or image regions affected a prediction made by a neural network, which can aid in model validation, and decrease the amount of data that cannot be explained.

### **7.2 Finance**

Credit scoring, fraud detection, risk assessment are some of the financial applications. Explainability can uncover the factors behind a prediction, and offer decision makers increased transparency when considering financial outcomes from automation. This is especially relevant when decisions have immediate impact on the lives of people and individuals or organization.

### **7.3 Autonomous Systems**

Deep visual models have been found to be crucial for autonomous systems' ability to recognize objects, understand scenes, and make safety decisions. Bau et al. showed how internal representations of deep visual networks can be analyzed and measured in terms of interpretable visual concepts, which can provide insight into how complex neural models represent objects and scene-level information [32]. This type of interpretation is useful in the context of validating autonomous-system behavior.

### **7.4 Cybersecurity**

In the cybersecurity field, explainable neural networks can aid in intrusion detection, malware classification, and threat analysis. In their role as analysts, XAI can provide insights into the intentions behind specific network traffic or software samples to assist in better investigation, debug of models, and human oversight.

### **7.5 Natural Language Processing**

Examples of NLP applications are text classification, sentiment analysis and question answering. Explainability can highlight influential words, phrases, or semantic concepts and provide insight on the role of textual evidence in neural predictions. More generally, the techniques of explainability have been extended to various architectures of neural networks and various application domains to increase transparency, to find relevant features, or to enable

more reliable human interpretation of the decisions made by the model [33].

In conclusion, the importance of XAI is dependent on the application scenario. Safety and reliability are paramount in healthcare and autonomous systems, transparent and accountable decisions are at the core of finance, interpretable threat evidence is important for cybersecurity, and comprehensible relationships between textual inputs and predictions are crucial for NLP. The applications show the importance of explainability in a practical application of neural-network-based decision-making that is trustworthy.

## 8. Challenges and Limitations

Though substantial advances have been made, there are still significant limitations concerning faithfulness, robustness, computational complexity, bias and variability of evaluation in XAI techniques. A major worry is that the visual explanations can be misleading as to the actual reasoning of a neural network. Danesh et al. showed that some of the saliency methods can produce identical explanations even when the model parameters and/or the labels are randomized, introducing doubt about the reliability of the explanations [34].

It is also tough on the robustness. Alvarez-Melis and Jaakkola demonstrated that small changes to the input could significantly change an explanation even if the model's prediction is almost the same [35]. This instability makes the methods of explanation less reliable in certain situations, especially when it comes to safety critical or high-stakes circumstances. Additionally, XAI can be helpful to uncover unexpected model behaviour: for instance, Kauffmann et al. demonstrated that XAI can be used to uncover “Clever Hans” effects, in which the models are using spurious patterns or unintended evidence to achieve their predictions [36].

Another example of fragility of interpretation methods is provided by Ghorbani et al. [37] who showed that while neural-network explanations are fragile, they can be modified by making small changes in the input, which leaves the original prediction unchanged. These findings illustrate how challenging it is to provide explanation faithfulness and consistency. The issues of security and manipulation are also applicable. The authors of slack et al. showed that post-hoc explanations like LIME and SHAP can be deliberately misled, and that this results in noisy explanations, while hiding the biased or unwanted behavior of the model [38].

Additionally, model-agnostic explanation techniques have methodological constraints on the notions of feature dependency, extrapolation, unrealistic perturbations and misjudged associations. All of these drawbacks can be an obstacle for users making faulty conclusions if they do not take into account the assumptions that are used in the explanation as well as the characteristics of the data [39].

Other unaddressed challenges are: trade-off between predictive performance and interpretability, computational cost of using complex explanation methods, privacy concerns, bias in models and explanations, and non-expert users may not fully understand the outputs of the explanations. Additionally, there are no uniform standards and common evaluation protocols in the field to compare XAI techniques for different models and applications. These limitations must be tackled with more powerful, authentic, secure and people-centred explanation frameworks with clear evaluation methods.

## 9. Emerging Trends and Future Research Directions

The future of Explainable Artificial Intelligence is shifting from static post hoc explanations to more trustworthy, interactive and user-centric explanations. Longo et al. refer to this development as XAI 2.0 with emphasis on the importance of interdisciplinary strategies that address human factors, usability, trust, and responsible AI principles alongside with technical explainability [40]. Going forward, the emphasis on explainability in AI will likely grow stronger for more complex architectures like generative AI, Transformers and multimodal neural networks where traditional feature attribution approaches may only give limited insights into model behavior. Interactive XAI, real-time generation of explanations, privacy-preserving interpretation, fair explanations, and more standard evaluation methods are other promising areas of future research into XAI for improved practical application.

Another major research direction is concept-based explanations where model reasoning is described in terms of more abstract and comprehensible concepts, in lieu of individual features. Ghorbani et al. present an automatic method of extracting meaningful concepts from neural network representations and their application to explain predictions [41]. Concept-based XAI may also decrease the reliance on manually defined concepts, enhance semantic interpretability and enable more intuitive explanations in the image, language, and multimodal models.

The focus of counterfactual explanations and algorithmic recourse is also gaining more traction as it shifts the conversation to explainability from interpretation to actionable decision support. Verma et al. note that there are ongoing issues with feasibility, diversity, causality, fairness and actionability in counterfactual explanations [42]. Future research should therefore look to provide explanations that not only reveal the rationale behind a particular decision but also provide realistic and ethically sound recommendations for changes that could result in a different decision.

In general, future studies on XAI will be more focused on human-centric interaction, concept-level reasoning, actionable counterfactuals, explainability of generative and multimodal models, fairness and privacy as well as robust evaluation. Advance in these areas will be critical to creating explainability frameworks that are technically sound, but also easily interpreted, credible, and useful in the making of decisions in real world neural-network computing.

## References

1. F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.
2. A. B. Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.
3. 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.
4. 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*, vol. 51, no. 5, pp. 1–42, 2018.
5. G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Information Fusion*, vol. 76, pp. 89–106, 2021.
6. R. Saleem, B. Yuan, F. Kurugollu, A. Anjum, and L. Liu, "Explaining deep neural networks: A survey on the global interpretation methods," *Neurocomputing*, vol. 513, pp. 165–180, 2022.
7. 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 *Proc. International Conference on Machine Learning*, 2018, pp. 2668–2677.
8. C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, "This looks like that: Deep learning for interpretable image recognition," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
9. G. Schwalbe and B. Finzel, "A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts," *Data Mining and Knowledge Discovery*, vol. 38, no. 5, pp. 3043–3101, 2024.
10. M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.
11. 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 *Proc. IEEE International Conference on Computer Vision*, 2017, pp. 618–626.
12. M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. International*

- Conference on Machine Learning, 2017, pp. 3319–3328.
13. S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.
  14. D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, “SmoothGrad: Removing noise by adding noise,” *arXiv preprint arXiv:1706.03825*, 2017.
  15. G. Montavon, W. Samek, and K.-R. Müller, “Methods for interpreting and understanding deep neural networks,” *Digital Signal Processing*, vol. 73, pp. 1–15, 2018.
  16. S. Jain and B. C. Wallace, “Attention is not explanation,” in *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, 2019, pp. 3543–3556.
  17. S. Serrano and N. A. Smith, “Is attention interpretable?” in *Proc. 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2931–2951.
  18. S. Wachter, B. Mittelstadt, and C. Russell, “Counterfactual explanations without opening the black box: Automated decisions and the GDPR,” *Harvard Journal of Law & Technology*, vol. 31, p. 841, 2017.
  19. R. K. Mothilal, A. Sharma, and C. Tan, “Explaining machine learning classifiers through diverse counterfactual explanations,” in *Proc. 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 607–617.
  20. A. Shrikumar, P. Greenside, and A. Kundaje, “Learning important features through propagating activation differences,” in *Proc. International Conference on Machine Learning*, 2017, pp. 3145–3153.
  21. R. C. Fong and A. Vedaldi, “Interpretable explanations of black boxes by meaningful perturbation,” in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3449–3457.
  22. 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.
  23. V. Petsiuk, A. Das, and K. Saenko, “RISE: Randomized input sampling for explanation of black-box models,” *arXiv preprint arXiv:1806.07421*, 2018.
  24. R. Fong, M. Patrick, and A. Vedaldi, “Understanding deep networks via extremal perturbations and smooth masks,” in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 2950–2958.
  25. W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, “Evaluating the visualization of what a deep neural network has learned,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 11, pp. 2660–2673, 2016.
  26. S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, “A benchmark for interpretability methods in deep neural networks,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.
  27. C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D. I. Inouye, and P. K. Ravikumar, “On the (in)fidelity and sensitivity of explanations,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.
  28. P. Hase and M. Bansal, “Evaluating explainable AI: Which algorithmic explanations help users predict model behavior?” in *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 5540–5552.
  29. C. Agarwal et al., “OpenXAI: Towards a transparent evaluation of model explanations,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 15784–15799, 2022.
  30. A. Hedström et al., “Quantus: An explainable AI toolkit for responsible evaluation of neural network explanations and beyond,” *Journal of Machine Learning Research*, vol. 24, no. 34, pp. 1–11, 2023.
  31. T. Lei, R. Barzilay, and T. Jaakkola, “Rationalizing neural predictions,” in *Proc. 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 107–117.
  32. D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, “Network dissection: Quantifying interpretability of deep visual representations,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3319–3327.
  33. W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller, “Explaining deep neural networks and

- beyond: A review of methods and applications,” *Proceedings of the IEEE*, vol. 109, no. 3, pp. 247–278, 2021.
34. J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, “Sanity checks for saliency maps,” *Advances in Neural Information Processing Systems*, vol. 31, 2018.
  35. D. Alvarez-Melis and T. S. Jaakkola, “On the robustness of interpretability methods,” *arXiv preprint arXiv:1806.08049*, 2018.
  36. J. Kauffmann, J. Dippel, L. Ruff, W. Samek, K.-R. Müller, and G. Montavon, “Explainable AI reveals Clever Hans effects in unsupervised learning models,” *Nature Machine Intelligence*, vol. 7, no. 3, pp. 412–422, 2025.
  37. A. Ghorbani, A. Abid, and J. Zou, “Interpretation of neural networks is fragile,” in *Proc. AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, 2019, pp. 3681–3688.
  38. D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, “Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods,” in *Proc. AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 180–186.
  39. C. Molnar, G. König, J. Herbringer, T. Freiesleben, S. Dandl, C. A. Scholbeck, et al., “General pitfalls of model-agnostic interpretation methods for machine learning models,” in *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, Cham, Switzerland: Springer International Publishing, 2020, pp. 39–68.
  40. L. Longo et al., “Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions,” *Information Fusion*, vol. 106, Art. no. 102301, 2024.
  41. A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim, “Towards automatic concept-based explanations,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.
  42. S. Verma, V. Boonsanong, M. Hoang, K. Hines, J. Dickerson, and C. Shah, “Counterfactual explanations and algorithmic recourses for machine learning: A review,” *ACM Computing Surveys*, vol. 56, no. 12, pp. 1–42, 2024.