



Systematic Review

From Black Box to Clarity: A Systematic Review of Explainability Methods in Deep Convolutional Neural Networks

Zina Tayari * and Mourad Zaied *

Research Team in Intelligent Machines, National Engineering School of Gabes, University of Gabes,
Gabes 6029, Tunisia

* Correspondence: zina.tayari@univgb.tn (Z.T.); mourad.zaied@univgb.tn (M.Z.)

Abstract

Deep neural networks (DNNs) have significantly advanced machine perception and reasoning; however, their lack of transparency in decision-making continues to pose a major challenge, particularly in high-stakes domains such as healthcare, finance, and law. This is especially concerning with the black-box nature of convolutional neural networks (CNNs), where the rationale for making a decision can be as important as the decision itself. This paper is driven by a question that is easier to ask than to answer: how can CNNs be made to explain themselves? To answer the question, we wrote a PRISMA-compliant systematic review of 154 studies published between 2017 and 2025. These studies were selected from 4421 studies retrieved through Web of Science, Scopus, IEEE Xplore, and ACM Digital Library. CNN-specific taxonomy was developed. This taxonomy organizes explainable artificial intelligence (XAI) methods on four axes: explanation timing, model dependency, output type, and target component. We found that there is a huge bias in the field regarding post hoc visual methods. Grad-CAM is the most widely cited visual explanation methodology, and within the model-agnostic framework, LIME and SHAP prevail. This research was also the first to analyze standard assessment methods. It was found that out of the 154 studies in the review, 98 used objective methods to evaluate fidelity, stability, or sensitivity. Conversely, fewer than ten of them used human-centered methods to evaluate how tasks were performed, how the users trusted the method, or how the users were prepared to interact with the system. We argue for a dual-reporting convention under which metrics should be reported together at least once, as per the family of metrics. The third contribution is an evidence-based challenge map, where we outline four issues: absence of standardized benchmarks, post hoc mechanism scalability limitations, vulnerability to adversarial perturbations, and the persistent gap between the technical descriptions and human understanding. For each challenge, we propose concrete directions: integrating causal reasoning, adopting participatory evaluation design, and building hybrid transparent architectures. We offer this review as a practical roadmap for researchers and practitioners working toward more explainable deep neural networks.



Academic Editor: Jianlong Zhou

Received: 28 May 2026

Revised: 28 June 2026

Accepted: 14 July 2026

Published: 23 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: explainable AI; deep neural networks; convolutional neural networks; post hoc explainability; systematic review

1. Introduction

These days, artificial intelligence (AI) has demonstrated its potential to revolutionize technological advances, societal changes, and industrial advancements, meeting or exceeding the performance level of humans in terms of accuracy. AI is a multidisciplinary field of

computer science dedicated to creating systems and algorithms that can perform tasks typically associated with human intelligence, such as reasoning, learning, and decision-making. Within AI, deep learning (DL) uses architectures, namely DNNs, to process data and generate predictions. However, DNNs often operate as 'black boxes,' making it difficult to understand the processes behind specific decisions. This lack of transparency underscores the critical need for explainability in AI, given that these technologies are increasingly deployed in high-stakes applications. By enhancing the interpretability of DNNs, we aim to bridge the gap between complex model behavior and user comprehension, fostering trust and accountability in AI systems. These models are designed to carry out operations, such as speech recognition [1], language translation [2], and decision-making, that have historically required human intellect. AI techniques are created using algorithms and models that allow them to learn from inputs. One of the principal goals of AI is to replicate human intelligence through systems capable of adapting and enhancing their output without requiring specific programming. The capacity to learn from experience and modify behavior accordingly is a pivotal goal of AI systems, making them adaptable and proficient in addressing several tasks in multiple domains [3]. AI is fundamentally considered an attempt to develop machines capable of thinking, learning, and acting with intelligence, transforming industries and altering how we engage with technology in everyday routines. This technological advance is partly due to the appearance of a new subset of machine learning (ML), termed deep learning (DL), where numerous neuronal parameters are trained to perform a particular function in a generalized manner proficiently [4]. DL relies on a layered structure of algorithms named artificial neural networks, permitting processing unstructured data, such as images and text, with little human intervention. The incorporation of ML methods, such as DL, helped AI systems to deliver extraordinary results in healthcare [5,6], speech and audio processing [7,8] and image processing [9,10]. Deep networks are considered complex and undeniably challenging to interpret because of their extensive parameter count [11]. Despite their high validation accuracy or other performance evaluation parameters, DL models may inherently succeed or fail in learning representations from the data that a human might consider significant. As the complexity of these methods increases, we see more emphasis on transparency/interpretability in developing AI models. Understanding how deep neural network (DNN) explainability accounts for data scientists and industrial stakeholders [12] has been a preoccupation for researchers [13] in the last few years. Therefore, building trust in AI models remains an open challenge. Indeed, we need to explain the AI model not only to ML experts but also to domain experts who require a human-understandable explanation. Convolutional neural networks (CNNs) have outperformed traditional ML models in tasks such as object recognition and image classification. Indeed, these networks have a sophisticated architecture identified as a black box. Their internal mechanism lacks transparency. Therefore, understanding the rationale behind certain decisions can be challenging for humans [14]. Offering explanations in CNNs can help AI stakeholders to determine whether a model succeeds or fails. Therefore, explainable artificial intelligence (XAI) was introduced to enhance explainability in these neural networks. We refer to these efforts as explainable CNNs. Building explainable neural networks can deliver significant insights to end users. Because humans prefer visual data, explaining CNNs can be simpler than other models. Meanwhile, some difficulties regarding XAI models, such as the necessity to determine the level of explanation supplied by CNNs and the practicality of integrating models such as decision trees and linear regression in neural networks, have been discussed in the literature [15]. Furthermore, offering different levels of explanations to stakeholders is crucial [16]. For example, describing specific technical aspects for stakeholders can be insurmountable. Nevertheless, offering abstract explanations may reduce the transparency of

the model and influence the end user's trust. There are two primary ways in which abstract explanations may affect trust. First, highly technical explanations may lead non-expert users to simply dismiss an otherwise correct explanation and ultimately trust the system less. Second, in the presence of abstraction, users may place trust in a system beyond what is warranted; for example, users may not question a recommendation if they cannot identify the features that led the model to "attend to the lesion area". The right balance often depends on the user and the situation. Clinicians, for example, prefer the use of heatmaps, which provide clear localization to the feature. In contrast, stakeholders prefer more quantitative explanations that incorporate metrics [14]. To demonstrate various levels of explanation, we explore the explanations in tasks involving classifying images and tables. In image classification, XAI models supply class activation maps for the interpretation of CNN decisions. These activation maps are valuable to developers as well as end users. XAI is universally applicable throughout the ML life cycle as seen in Figure 1. This figure illustrates the sub-factors essential for understanding the ML model lifecycle. It highlights the various components involved, such as data collection, model training, evaluation, and deployment, while emphasizing the critical role of stakeholders, including end users, in interacting with these elements. We recognize that explaining complex ML models to stakeholders who may not possess a robust statistical background can be challenging. Encouraging collaboration between technical teams and stakeholders throughout the ML lifecycle can improve transparency and ensure that the models align with user needs and expectations. This interaction not only helps to clarify how the models work but also promotes trust and accountability in their application. To address this issue, data scientists often use visualizations and simplified explanations [17] to enable stakeholders to understand the predictions and accuracy of the models. This approach goes beyond the promise to understand the overall decision performance of a model. Therefore, different stakeholders [18] gain the ability to explore models that match real-world applications in several business and economic environments.

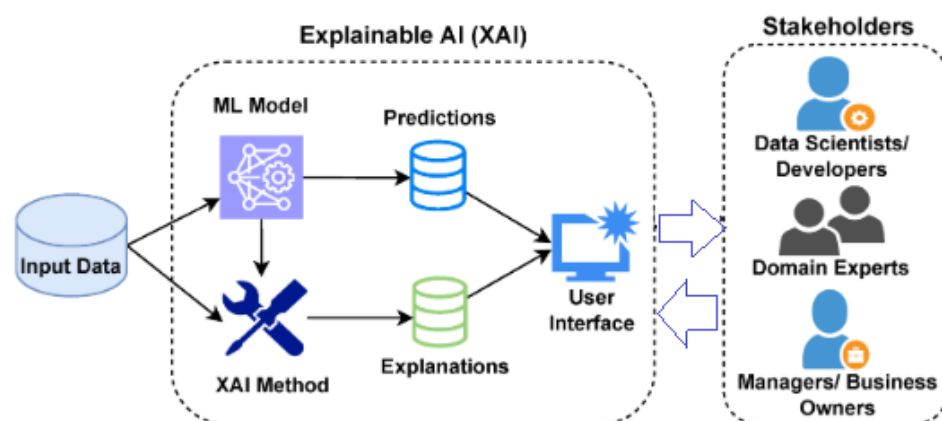


Figure 1. Explainable machine learning life cycle [19].

Many studies have covered reviews of XAI for machine learning. Even so, there are two major gaps in the current literature. First, these papers have not developed a classification of XAI for CNNs that either incorporates architecture or provides an architecture-aware classification of XAI by intrinsic or post hoc, agnostic or specific, and architecture modification, simplification, feature relevance, or visual XAI. Second, they have not compared the various human-centered and objective evaluation metrics in a systematic way.

In order to bridge these two gaps, the current study offers the following original contributions based on a systematic review of 154 peer-reviewed publications from 2017–2025, which have been organized under the PRISMA framework:

1. A CNN-specific taxonomy classifies XAI methods along four analytical dimensions: **Explanation timing** differentiates intrinsic approaches, which use models whose structures are transparent by design, such as decision trees or linear regression, where the model itself is the explanation, from post hoc approaches, which are applied after training to a black-box model (e.g., LIME, SHAP, Grad-CAM) to generate insights such as heatmaps, feature attributions, or rule sets without altering the model's architecture. **Model dependency** differentiates model-agnostic methods, applicable across architectures, from model-specific methods designed for CNNs in particular. **Output type** categorizes explanations as visual, rule-based, or feature attribution outputs. **Target component** identifies which part of the CNN pipeline the method addresses: the architecture, internal representations, feature relevance, or visual output. This taxonomy is systematically applied across the 154 reviewed studies to offer a clear and comparable view of the XAI landscape for CNNs.
2. A systematic analysis of XAI evaluation metrics that identifies objective metrics such as fidelity, stability, and sensitivity, and human-centered metrics such as task performance, subjective trust, and user calibration. We provide evidence that these two groups often provide assessment results that do not agree with one another, and we support a dual-reporting requirement that dictates the concurrent reporting of at least one metric from each group. This enables stringent comparison of studies.
3. An evidence-based challenge map identifying a number of present challenges and future research directions for CNN-based systems in XAI. It highlights the integration of heterogeneous explanation methods, the weak robustness of model-centered adversarial attacks, the susceptibility of XAI strategies to adversarial attacks, and the incorporation of causal reasoning in explanation systems. Each is provided with a number of research avenues, including, but not limited to, hybrid modeling and automation of the explanation pipeline.

2. Review of Explainable AI in the Industrial Domain

AI has become incredibly popular, and ML models are being used in a wide range of industries, such as social media, healthcare, finance, and agriculture. Because of this rapid growth, developers and professionals now urgently need to comprehend and have trust in the choices that these models make. As a result, several XAI techniques have been created to improve reliability and interpretability in the industrial domain.

2.1. Explainability in Vision-Based Applications

In computer vision, significant progress has been made in improving the interpretability of AI models. For instance, an Inception-v3 model was designed to identify malicious software using a color image display of Android applications [20]. Similarly, an XAI approach was proposed to explain fundamental AI concepts from stakeholders' perspectives, enhancing interpretability in practical applications [21]. Furthermore, visual analytics techniques have been employed to elucidate the workings of neural networks, enabling a deeper understanding of these models [22].

2.2. Financial Applications

In the financial sector, the integration of XAI has proven essential in fostering trust and ensuring accountability in predictive models. A hybrid approach combining traditional methods with human-interpretable explanations was implemented to improve the explainability of deep learning (DL) predictions, thereby aiding business decision-making [23]. Another XAI model, applied in financial applications, has demonstrated the ability to support transparency in complex AI-driven analyses [24].

2.3. Agriculture and Environmental Monitoring

XAI methods have also found applications in agriculture. For example, a deep learning algorithm was enhanced with explainability techniques to detect leaf disease classification [25]. Additionally, interpretable neural network models were developed to control water flux and assess water quality, showcasing the applicability of XAI in environmental monitoring [26–28].

2.4. Healthcare Applications

The healthcare domain, where trust and reliability are paramount, has benefited significantly from XAI advancements. For instance, explainable techniques have been utilized to study non-melanoma skin cancers in patients [29]. In another study, XAI was applied to X-ray imaging to improve the transparency of diagnostic processes [30]. An explainable approach was also introduced for predicting Alzheimer's disease [31].

2.5. Signal Processing and Forecasting

In signal processing, an XAI method using optimal data augmentation was proposed to evaluate the quality of vehicle sounds [32,33]. In forecasting, a modified XAI model architecture incorporating semantic and AI tools was developed to address real-world demand prediction challenges [34].

2.6. Social Media and Audio Analysis

The application of XAI in social media includes the detection of hate speech using deep learning mechanisms [35]. Additionally, the LIME method was employed to identify deepfake audio, contributing to better understanding and mitigation of such threats [36].

Although the previous examples demonstrate a wide range of industrial XAI applications, this study notes that the coverage is not all-inclusive. A number of advanced XAI models are not well covered, particularly those that make use of sophisticated methods like Shapley values, causal inference, and counterfactual explanations. These advanced techniques will be incorporated later in this evaluation to offer a more thorough analysis. The use of XAI in speech processing has increased and researchers have begun looking into the existence of interpretable representations within deep models. Recently, studies have used layer-wise relevance propagation (LRP) to determine which input feature time points in waveforms or time–frequency elements in spectrograms are most relevant for spoken digit and speaker sex classification. Consonants and vowels were shown to be relevant for distinguishing speaker traits in research that utilized integrated gradients to explain how phonetic variations impacted speaker verification systems. To understand the internal functioning of automatic speech recognition systems based on end-to-end models (E2E ASR), different model-explainability techniques have been employed. Shallow E2E ASR layers were shown to align with prominent syllables of speech and form rhythmic elements, suggesting that E2E ASR systems may be similar to human speech processing. Finally, due to the growth of multi-modal emotion recognition, XAI has used gradient SHAP to determine relevant features in audio, video, and text, thereby creating interpretable automated expert systems.

3. Research Methodology

This study uses a systematic literature review (SLR) [37] to closely follow the PRISMA [38] guidelines to mitigate selection bias while also ensuring that future studies can be replicated and that a clear audit trail is maintained. The completed PRISMA 2020 checklist, documenting compliance with each of the 27 reporting items, is provided as Supplementary Materials (Table S1). This systematic review was retrospectively registered

on the Open Science Framework (OSF). The registration number is 39bd2 (available at <https://osf.io/39bd2>, accessed on 5 June 2026). The review entails publications from January 2017, the year in which LIME, SHAP, and Grad-CAM emerged and the Defense Advanced Research Projects Agency (DARPA) XAI programme [39] began, until December 2025. It incorporates the entire span of modern XAI research in deep learning. Considering the SLR process, our methodology is divided into the following three stages: planning, conducting, and reporting.

3.1. Planning the Review

3.1.1. Research Questions

The review's scope is limited to answering four specific research questions (RQ1–RQ4) in Table 1, which, in turn, correspond to the review's analytical sections. Thus, this framework is constructed from evidence, not as a hypothesis.

Table 1. Research questions governing the systematic review: scope, scientific objective, and location.

RQ	Research Question	Scientific Objective	Section
RQ ₁	Which intrinsic and post hoc XAI techniques have been proposed for deep neural networks, with particular focus on convolutional architectures?	Establish a comprehensive taxonomy of explainability methods for DNN and CNN architectures, distinguishing model-agnostic from model-specific approaches.	Sections 6–8
RQ ₂	How do model-agnostic and model-specific explainability approaches compare with respect to fidelity, scalability, and applicability to CNN-based models?	Identify the practical trade-offs that govern technique selection across different deep learning tasks and deployment constraints.	Sections 6–9
RQ ₃	What objective and human-centered evaluation metrics are employed to assess the quality of explanations produced by XAI methods in deep learning contexts?	Characterize existing evaluation standards and highlight standardization gaps that impede cross-study comparison.	Section 10
RQ ₄	What are the principal limitations and open challenges hindering the practical deployment of XAI techniques for DNN and CNN architectures?	Synthesize unresolved problems and map actionable future research directions for the XAI community.	Section 11

3.1.2. Search Strategy

Four major databases were used: Web of Science (WoS), Scopus, IEEE Xplore, and ACM Digital Library. These databases provide complementary coverage of computer science, electrical engineering, and human–computer interaction. Table 2 includes the Boolean search string, which structurally combines two thematic fields that are to be searched across titles, abstracts, and keywords, within the years 2017–2025.

Table 2. Boolean search string applied across all four databases.

Cluster 1—XAI and Interpretability	Cluster 2—Deep Learning
("explainable" OR "interpretable" OR "transparent" OR "XAI" OR "post-hoc" OR "saliency" OR "attribution" OR "counterfactual" OR "LIME" OR "SHAP" OR "Grad-CAM" OR "class activation map" OR "attention mechanism" OR "feature importance")	AND ("deep learning" OR "deep neural network" OR "DNN" OR "convolutional neural network" OR "CNN" OR "image classification" OR "object detection" OR "computer vision" OR "representation learning")

3.1.3. Inclusion and Exclusion Criteria

We set and established criteria before the search was conducted. These criteria were established a priori and applied strictly; any record that breached a single exclusion criterion was discarded regardless of its topical relevance (Table 3).

Table 3. Prespecified inclusion and exclusion criteria.

Inclusion Criteria	Exclusion Criteria
Peer-reviewed journal articles or conference papers, 2017–2025.	Book chapters, editorials, technical reports, and opinion pieces.
Proposes, applies, or evaluates an XAI method for DNNs or CNNs.	XAI applied exclusively to shallow models (SVMs, random forests) without DNN integration.
Written in English.	Non-English publications.
Addresses at least one of: explainability technique, evaluation metric, or application domain in a deep learning context.	Studies focused solely on accuracy improvements with no explainability component.
Provides reproducible experimental details (dataset, metric, baseline).	Duplicate records; preprints without subsequent peer review.

3.2. Conducting the Review

We searched the four selected databases using the query string defined in Table 2, retrieving **4421 records**. After removing 1195 duplicates, we independently screened titles and abstracts, excluding records that were clearly out of scope, yielding **2335 records** for full-text review. We then assessed each full text against the criteria in Table 3, excluding 1989 records primarily due to XAI being applied to shallow models only ($n = 658$), no explainability component ($n = 547$), and language or publication-type restrictions ($n = 784$). The remaining **346 articles** underwent a final quality check, retaining only studies that clearly described their method, reported results on a named dataset, and acknowledged their limitations. This yielded a **final corpus of 154 articles**; 93 journal papers and 61 conference contributions spanning 2017–2025, summarized in Figure 2.

3.3. Reporting the Review

The 154 articles were characterized along two descriptive dimensions:

Domain distribution The corpus spans 11 application domains. Computer Vision ($n = 38$) and Healthcare ($n = 29$) attract the largest share of publications, reflecting their early adoption of DNN-based solutions and their stringent transparency requirements. Emerging areas such as Industrial IoT ($n = 10$) and Social Media ($n = 4$) remain underrepresented, signaling open opportunities for future XAI research. Publications in high-stakes domains, such as Healthcare and Finance, appear predominantly in journals, while faster-evolving areas such as Autonomous Vehicles and Cybersecurity are more represented in conference proceedings.

RQ mapping: Each article was tagged to one or more RQs. RQ1 (XAI techniques) attracted the largest share ($n = 89$), followed by RQ3 (evaluation metrics, $n = 61$), RQ4 (challenges, $n = 54$), and RQ2 (comparative analysis, $n = 48$). These counts governed the depth of treatment in each section.

The essential discussion in Sections 7–9 will be more proportional to the level of depth added in a new method or extension offered by a given study in comparison to the level of depth offered to the study, adding a new application domain to an existing method.

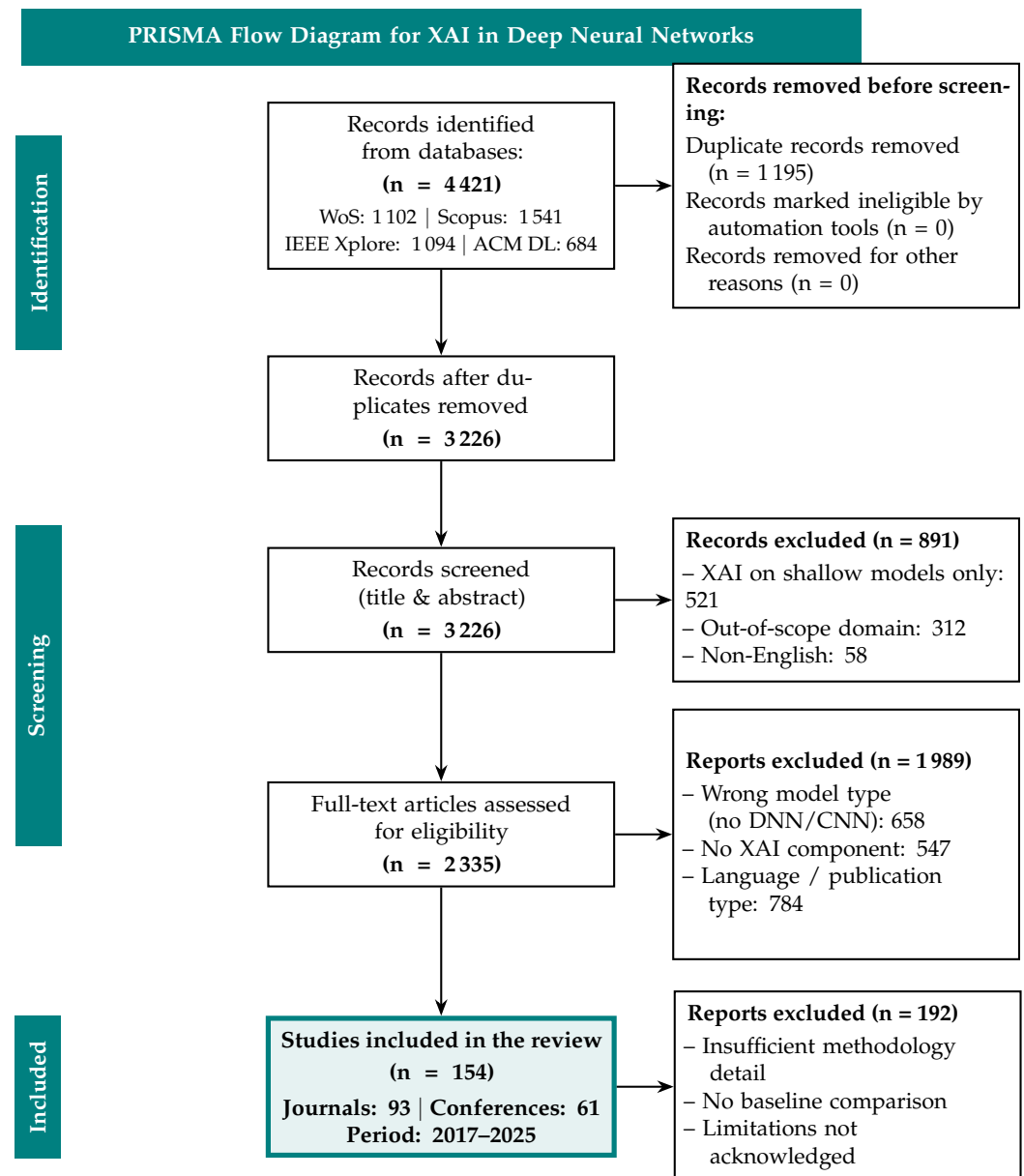


Figure 2. PRISMA flow diagram for the systematic selection of XAI studies in deep neural networks, spanning 2017–2025.

4. Interpretability/Explainability in AI Models: Concepts and the Trade-Off Between Explainability and Accuracy

Something is interpretable if it can be understood without any external aids. Consequently, we can say that ML models are interpretable if a human can understand them without external techniques or aids [40]. An example of an interpretable model in ML is the decision tree. It is a fine example to understand by traversing down all the nodes of a tree. However, as the model becomes more complex, it loses its interpretability. A model is explainable if it requires additional external techniques, functions, or other aids to be understood by humans. Therefore, an explainable model is essentially a process too complex for human comprehension. This model is considered a black-box model [41]. An extra strategy or approach becomes very necessary to examine the black box and comprehend the functioning of such a model. An example of these models is the random forest, which consists of many decision trees. To make the final prediction, all the individual trees' predictions should be taken into consideration. To give a clear picture, Alexnet, a CNN designed for image recognition, has an outstanding 62,378,344 parameters [42]. Therefore,

it is impossible to understand how a model like Alexnet operates with such a huge number of parameters. To comprehend this type of model, further exploration of other methods will be discussed later. Model accuracy and explainability are frequently traded off.

The trade-off between explainability and performance for various machine learning models is shown in Figure 3. Explainability is represented by the horizontal axis, which ranges from low to high, while performance is represented by the vertical axis, which likewise ranges from low to high.

- The top-left represents deep learning models, such as GANs, CNNs, RNNs, and ResNet, which exhibit excellent performance but low explainability.
- Although ensemble approaches (e.g., MLPs, random forests, and XGBoost) achieve a balance, their explainability is still limited.
- Statistical models that offer modest explainability and performance include SVMs, KNN, and naive Bayes.
- With a minor performance trade-off, graphical models (such as Bayesian belief networks, Markov logic networks, and Markov models) provide better explainability.
- Although rule-based and linear models are more explainable, they are usually less effective at complex tasks.
- An ideal model is offered as an aspirational objective—one that is distinguished by both high explainability and excellent performance.

In deep learning applications, this illustration emphasizes the necessity of striking a balance between interpretability and accuracy by highlighting the inherent trade-offs in model selection.

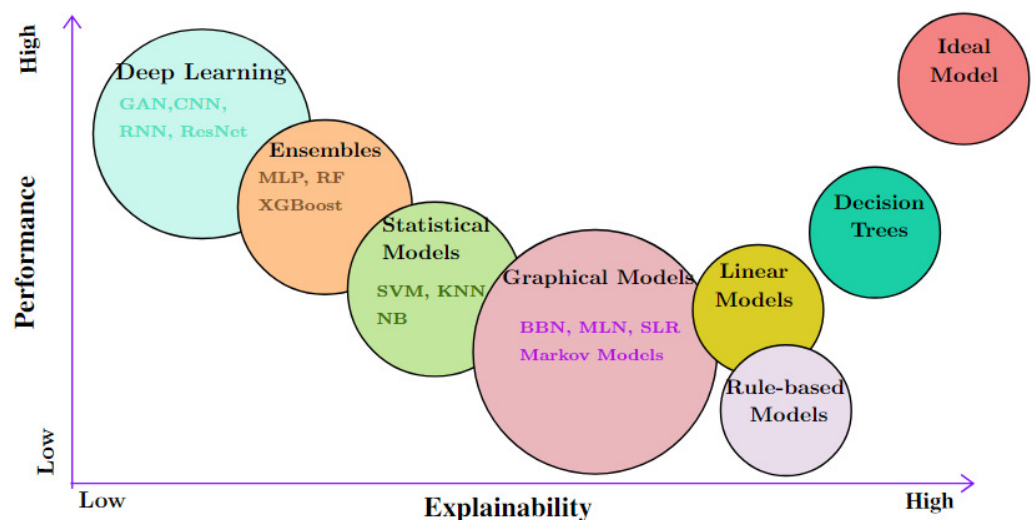


Figure 3. Trade-off between explainability and accuracy.

5. Objectives of Explainable Artificial Intelligence

XAI aims to enhance user understanding by providing detailed insights into its internal operations. The objectives of XAI [43], as illustrated in Figure 4, encompass various crucial aspects:

Interpretability: Interpretability makes it possible for users, including non-experts, to comprehend and interpret the ideas underlying the predictions and judgments made by AI models. User comprehension tests and qualitative comments on how well the model explains things may be used to evaluate interpretability [44–46].

Trust: Trust evaluates how consistently and effectively the model addresses problems. User surveys and performance indicators that assess users' confidence in the model's predictions are two ways of measuring trust [47–49].

Transparency: Transparency is the degree to which stakeholders can understand how a system makes decisions. It can be evaluated through user studies that examine how effectively users understand the model's decisions based on the provided explanations [50–53].

Reliability: Reliability evaluates how consistently a learning model produces outcomes under comparable circumstances. Statistical techniques that measure stability, including split-half reliability tests or the intraclass correlation coefficient (ICC), can be used to evaluate dependability [54–56].

Causality: Causality is the study of the connections between model variables and the effects of adjusting one on the others. To evaluate relationships, causal inference methods, namely those obtained from statistical analysis or causal diagrams, can be used.

Privacy: To secure user data, privacy evaluates whether other parties can access the original training data. Audits that evaluate data access restrictions and differential privacy measures are examples of privacy metrics.

Fairness: Fairness ensures that all user groups are treated equally by evaluating ethical issues or biases in the learning model. Statistical parity measures and disparity indices that compare results across demographic groupings can be used to measure fairness.

Usability: Usability evaluates how well and safely the system interacts with people, guaranteeing usability. Heuristic analyses and user experience surveys are examples of usability testing techniques that can reveal information about user satisfaction and interaction effectiveness.

Collaboration: Collaboration enables data scientists, domain experts, and end users to work together, promoting a shared understanding of AI model behavior. Metrics related to collaborative platform use and feedback systems can be used to assess collaboration efficacy.



Figure 4. Explainable AI objectives.

6. Foundations of Explainable AI (XAI)

6.1. Core Concepts and Principles

In the world of XAI, grasping foundational elements is paramount. Explaining power refers to the artificial intelligence's ability to elucidate its predictions or results, enabling

humans to understand them easily [57]. The pivotal components encompass clarity, interpretability, and accountability. These components guarantee that AI models do not seem like impenetrable black boxes but more like systems capable of clearly describing thought processes behind decision-making activity [58].

6.2. Necessity for Clear and Interpretable AI Decision-Making in AI Models

Due to widespread adoption rates across various fields in our everyday lives, the need for transparent AI decision-making has surged [59]. It is crucial not only to achieve accuracy but also to maintain transparency, as well-matched explanations are needed [60]. Therefore, considering the potential moral implications in social explanations is a humanistic strategy that stresses resolving critical issues to promote trust.

6.3. Taxonomy of Explainability in Artificial Intelligence

Diverse approaches aim to enhance interpretability and explainability in ML models [61], classifying them into two key categories: model-agnostic and model-specific. Model-agnostic methods [62] are flexible approaches that can interpret and clarify predictions across various ML models, irrespective of their architecture or complexity. LIME and SHAP tackle the problem of explainability of complex classification models from different angles. Perturbed samples of the input form a decision boundary and help to approximate the prediction region for LIME. In contrast, SHAP assigns a value to each feature based on game theory to explain the prediction made. Although both approaches regard the classification model as a black box, they differ significantly with regard to their stability, scope, and scalability, as elaborated in Section 9.2. However, model-specific techniques [63] are designed to evaluate and clarify predictions made by a given architecture or ML model. Methods such as layer-wise relevance propagation (LRP) are designed to provide insights into the inner workings of neural networks. However, their effectiveness may vary depending on the model and context [64]. These methods may require access to internal parameters or intermediate representations for generating explanations. A detailed comparison of the two methods is represented in Figure 5. The term “intrinsic interpretability” [65] describes ML models, like sparse linear models or short decision trees, considered interpretable because of their basic structure. Using interpretation techniques after model training is known as “post-hoc interpretability” [66]. Therefore, post hoc explanations concentrate on providing insights following the prediction-making of the model. They help users to comprehend model behavior and develop confidence in its judgments. Furthermore, there has been a significant interest in the application of intrinsic explainable models to improve scientific knowledge and open the way for new scientific discoveries [67,68].

As shown in Table 4, some earlier taxonomies for the XAI literature have differentiated in similar but distinct ways. For example, Arrieta et al. [13] classify methods focusing primarily on the level of transparency (intrinsic versus post hoc) and on the extent of the explanation (local versus global) without considering the architecture. Meanwhile, Schwalbe and Finzel [61], Graziani et al. [63], Yang et al. [69], and Bibal and Frenay [70] offer more extensive, model-agnostic taxonomies that encompass all of machine learning rather than just CNNs. Vilone and Longo [71] focus on taxonomies of evaluation approaches (application-, functionally, and human-grounded) rather than with the explanation generation methods. Saeed and Omlin [18] are more concerned with classifying the literature with respect to the open problems and the research agenda than with a taxonomy of methods. Unlike the others, Clement et al. [19] position XAI methods within the context of the software development lifecycle. Tjoa and Guan [72] propose a taxonomy that focuses on specifics of the domain, not the architecture. Unlike the former example, their taxonomy focuses on a specific application area (medical XAI) rather than model architecture. This demonstrates

that domain specificity and architecture specificity serve different organizational needs and cannot be applied interchangeably.

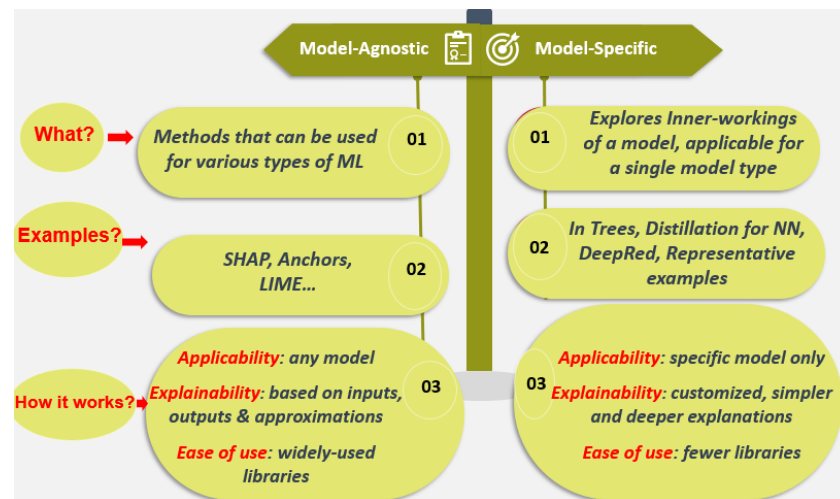


Figure 5. Model-agnostic vs. model-specific.

Table 4. Comparison of the proposed CNN-specific taxonomy with some prior XAI taxonomies and surveys.

Taxonomy (Reference)	Organizing Axis/Axes	Scope	CNN-Specific?	Deployment Lookup?
Arrieta et al. [13]	Transparency level (intrinsic vs. post hoc); scope (local vs. global)	General ML	No	No
Schwalbe and Finzel [61]	Comprehensive technical taxonomy of XAI concepts and methods	General ML, cross-domain	No	No
Graziani et al. [63]	Unified technical/social terminology for interpretable AI	General ML, technical + social	No	No
Vilone and Longo [71]	Evaluation approach: application-, functionally, and human-grounded	Evaluation-focused, general ML	No	No
Saeed and Omlin [18]	Current challenges and future research opportunities (meta-survey)	General ML, challenge-oriented	No	No
Yang et al. [69]	Approaches, limitations, and application aspects	General ML	No	No
Tjoa and Guan [72]	XAI methods cataloged for clinical interpretability needs	Medical-imaging domain, not architecture-specific	No	No
Clement et al. [19]	Software-development lifecycle stage	General XAI, software-engineering focus	No	No
Bibal and Frenay [70]	Interpretability of models and representations (heuristic- vs. user-based)	General ML	No	No
Present Review	Explanation timing; model dependency; output type; target CNN component	CNN-specific	Yes	Yes

All these studies individually approach axes in this review but do not bring them together in an individual framework that speaks to CNNs. These axes include the timing of the explanation, model dependency, the type of the output, and the targeted component of the CNN. Specifically, whether the method modifies the architecture, simplifies it, analyzes feature relevance, or generates a visual explanation. The added value of this taxonomy is that it discusses all four axes together with respect to CNNs, which is something that the

literature lacks. As a result, in constraining the application of this taxonomy to CNNs, a user can effectively address a specific application constraint (for example, ‘I cannot modify my deployed architecture’) and look through the taxonomies for available methods, which is something the previous general taxonomies did not account for.

7. Explainability Methods in AI

7.1. Data Explainability

Data clarity in AI relates to the transparency and openness regarding how data is leveraged for training, validating, and scrutinizing ML models that necessitate making all decisions by an AI system driven by this underlying data clear-cut and understandable for humans [73]. Unlike conventional explainability methods that primarily focus on elucidating the internal mechanisms of models, the data transparency allows us to understand the attributes or features of datasets used in evaluation and the quality biases that they may contain. These aspects are essential when considering the explainability of how data is used in AI [74]. Table 5 provides a detailed analysis of data explainability aspects. In short, data explainability is crucial for building trust in AI systems. It also guarantees that stakeholders can understand the foundation of the model’s predictions. It acknowledges that a wide range of data concerns, including data quality, bias, and quality, substantially impact the accuracy and reliability of AI outputs, going beyond the model.

Table 5. Key aspects of data explainability.

Aspect	Definition	Example
Data Characteristics	Providing insights into the nature and properties of the dataset.	Understanding the distribution of features, data types, and statistical properties.
Feature Importance	Identifying and communicating the importance of different characteristics in affecting the predictions of the model.	Highlighting central characteristics in the decision-making process.
Bias and Fairness	Assessing and addressing biases in the data that might lead to unfair or discriminatory outcomes.	Detecting and mitigating biases related to gender, race, or other sensitive attributes.
Data Quality and Preprocessing	Describing the quality of the data and any preprocessing steps applied.	Communicating how missing values, outliers, or other data issues were handled before training the model.
Data Governance	Providing information about the policies, standards, and practices governing data collection, storage, and usage.	Documenting data sources, ownership, and any regulations associated with the dataset.
Temporal Aspects	Taking into account the temporal aspects of the data and understanding how it changes over time.	Recognizing trends, seasonality, or changes in the dataset that may affect the model’s performance.
Explainable Data Processing Pipelines	Documenting and explaining the steps involved in processing and transforming the data before it is fed into the model.	Describing feature engineering, scaling, or any other preprocessing steps contributing to model training.

7.2. Model Explainability

Model explainability [75] pertains to the capability to comprehend and interpret decisions made by ML models, offering insight into how the model achieves particular predictions or distributions, resulting in a more transparent and understandable model performance to humans.

Diverse approaches and techniques have been used to achieve better interpretation. Some highlights of model explainability aspects are detailed in Table 6.

Table 6. Model explainability aspects.

Aspect	Definition	Example
Rule-Based Models	Models expressed as human-understandable rules, simplifying decision-making transparency.	Decision trees, outlining decision rules based on input features.
Feature Importance	Identifying and ranking the importance of features in influencing model predictions.	Permutation importance, SHAP values, or LIME emphasizing main features.
LIME	Using data perturbation to produce locally accurate explanations for individual forecasts.	Surrogate models approximating complex model behavior for specific instances.
Shapley Values	Determining the contribution of each feature to the model's output using cooperative game theory.	Fair distribution of feature importance considering all possible feature combinations.
Layer-wise Relevance Propagation (LRP)	Giving input features significance scores based on their impact on the model's output.	Frequently used in neural networks to grasp feature contributions at different layers.
Attention Mechanisms	Highlighting specific input parts focused on the model during decision-making.	Neural network attention mechanisms visualizing influential input regions.
Counterfactual Explanations	Generating alternative input scenarios resulting in different model outputs.	Instances where changing certain features leads to distinct predictions.
Global Surrogate Models	Training interpretable models approximating complex model decision boundaries.	Decision trees or linear models as surrogates for comprehending overall model behavior.

Model-explainability methods bridge the complexity of advanced ML models and the need for human-understandable annotations [69]. Thus, the choice of a particular technique is often based on the specifics of the model and the semantic requirements of a given project or application.

7.3. Post Hoc Explainability

Although post hoc methods are a part of the model-explainability techniques discussed in Section 7.2, we analyze them separately due to a specific characteristic that we observe with post hoc methods and not with other model-explainability methods that we analyze in Section 7.2 (such as intrinsic rule-based models): post hoc methods are applied externally after model training, and do not require access to or modification of model's internal structure. This characteristic is important in most real-world applications when we are explaining the use of a particular model, as it decides whether the technique can be used on a legacy or proprietary model, which is a restriction that does not apply to the other methods discussed previously.

Decisions or predictions can be defined after an ML model is semantically trained [76]. Specifically, the subsequent interpretive methods are applied externally to an already trained model to clarify how the model reaches conclusions, unlike approaches aimed at rendering the model inherently interpretable in the training phase. Table 7 summarizes key characteristics of post hoc explainability. Widespread post hoc explainability methods include LIME, SHAP, and Anchors [77]. Even when the model is not the main objective during training, these methods are crucial for creating complex models and guaranteeing that stakeholders can comprehend and trust the judgments made by the AI system.

Table 7. Post hoc explainability aspects.

Aspect	Definition	Example
External Interpretability	Post hoc methods applied externally to black-box models without altering internal structures.	Generating explanations for complex neural networks without modifying their architecture or training process.
Application to Any Model	Model-agnostic post hoc methods applicable to a wide range of machine learning models.	Applying methods like LIME to explain predictions from different types of models.
Flexibility in Explanation Types	Post hoc methods offer a variety of explanations for user flexibility.	Selecting rule-based explanations, feature importance, or instance-based explanations based on application needs.
Addressing Black Box Models	Useful for explaining complex black box models, such as deep neural networks.	Generating explanations for predictions from deep learning models without intricate details of hidden layers.
Local Interpretability	Focus on providing local interpretability for individual predictions.	LIME generates surrogate models for specific instances to approximate complex model behavior locally.
Useful for Legacy Systems	Applicable to existing and legacy models without retraining.	Explaining decisions made by machine learning models trained years ago without altering the original models.

7.4. Critical Assessment of Method Selection

The classification above provides a structure, but there are several practical issues that need to be addressed.

Data Explainability Methods: These are necessary, but in practice, they are few. Most of the literature centers around model-explainability methods, but data-centered methods respond to a somewhat different problem. Most of the literature asks the question “Why did the model make this prediction?” Data-centered methods ask “Is the data suitable for this model?” This difference is important because, in practice, explanations based on data that are biased or of low quality may be misleading, even if they are literally correct. For example, a CNN that was trained on medical images may produce correct saliency maps that indicate that the features correlated with the medical condition. However, if the training data is biased demographically, perhaps containing only images of fair-skinned cancer patients, this explanation will not indicate this bias.

Post hoc vs. Intrinsic Models: Post hoc methods are generally preferred in practice because they leave the model untouched, which is a critical advantage in regulated industries where model retraining triggers costly revalidation processes. However, this separation between model and explanation creates a fundamental tension: the explanation is produced by an external system that operates independently of the model and may therefore have its own limitations and failure modes. For example, in the medical domain, a post hoc explanation of a model prediction may be locally accurate relative to the model’s output while failing to surface clinically relevant features that the model itself did not encode. Intrinsic models, while more transparent by design, often require sacrificing predictive accuracy (see Section 4 for more detail on this trade-off). The choice between these approaches should therefore be driven by the application’s tolerance for the accuracy–explainability compromise.

Method selection implication: The choice among feature relevance, model simplification, and hybrid approaches depends not only on technical requirements but also on the following:

User expertise: Visual methods of explanation are suited to clinicians and radiologists. Rule-based methods of explanation are suited to auditors and regulators. Relevant feature methods are suited to data scientists and modelers.

Deployment Constraints: Applications that are time-sensitive cannot afford the computational cost of SHAP. Devices on the edge may not have the memory capacity to accommodate gradient-based methods.

Explanation fidelity vs. simplicity: Users may find complex explanations hard to understand, although they may be more faithful to the model.

8. Explainable AI in Deep Neural Networks

Deep neural models, especially multi-layer neural networks (NNs), have emerged as powerful tools, leading to breakthroughs in various fields [78]. The inherent complexity of these models, with complex architecture and multiple interconnected nodes, poses challenges in comprehending their decision-making processes.

8.1. Complexity of Multi-Layer Neural Networks (NNs)

DL researchers have studied the complexity of multi-layer neural networks in great detail [79]. Neural networks may tackle more difficult tasks as they get more proficient at modeling complex relationships and capturing complex patterns. However, this enhanced capacity is achieved at the expense of more computational complexity and overfitting risk. The network's capacity to generalize to previously unobserved data is greatly influenced by the interaction among the overall number of layers, the size of each layer, and the overall architecture. Scholars have investigated several strategies to address these difficulties, including regularization approaches and architectural improvements.

8.2. Need for Explainability in Deep Networks

Given that multilevel NNs attain state-of-the-art performance, the demand for explainability becomes increasingly significant. The inherent ambiguity of DL models raises concerns about their explainable, statistical, and ethical implications. Unlike classical models, DNNs not only find the mapping between the representation and the output but also figure out representations from unprocessed data (Figure 6). The explanation for DNNs mainly focuses on comprehending the representations gathered by the neurons at intermediate layers of DNNs since the deep representations learned are typically not interpretable by humans [80]. In what follows, we will introduce some explainable approaches for DNNs.

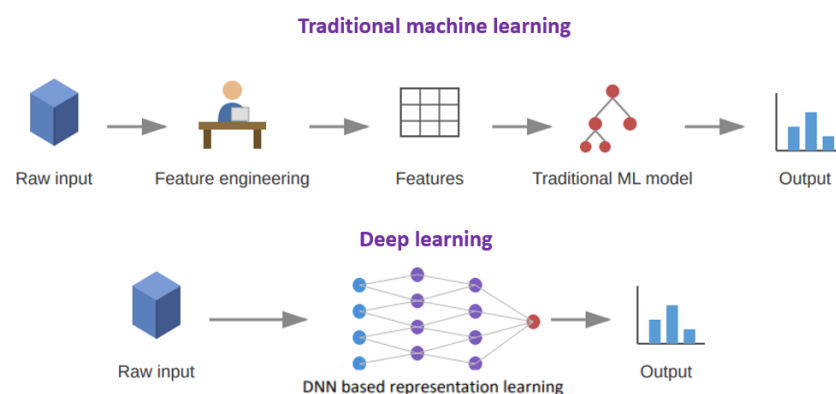


Figure 6. A deep learning pipeline vs. a typical machine learning pipeline inspired by the conceptual illustration in [81].

8.3. Brief Overview of Explainable Deep Learning Models

In this section, we provide an overview of XAI techniques created for DNNs. In several applications, NNs attain state-of-the-art performance as highly expressive computational models. Unfortunately, once we move beyond simple models such as single-layer perceptrons, they do not satisfy any transparency level, since their architecture and learning

procedure align with a complex computational pipeline. Most of these techniques can be categorized as feature-relevant or model simplification. Rule extraction is one of the most used methods for simplifying models. Compositional rule extraction strategies work at the level of individual neurons instead of the entire model. The authors in [82] present a technique that uses NNs to generate if-else rules while simultaneously training the model. An alternative method, called CRED [83], uses decision trees to describe the extracted rules. A related technique that generates if-else rules layer by layer is called KT [84]. One of the most often used methods for enhancing CRED is DeepRED [85]. The suggested algorithm includes intermediate rules and extra decision trees for each hidden layer. It can be compared to a divide-and-conquer strategy where the goal is to use the results of the preceding layer to describe each layer and then aggregate them to explain the entire network. Nonetheless, analogous methods are referred to as educational when the internal structure of an NN is ignored. In other words, methods that do not examine the network at the neuronal level to provide an explanation instead handle the whole network as a black-box function. Such a technique is TREPAN [86], which combines a query and sample strategy with decision trees. Based on a network inversion technique, an algorithm named HYPINV [87] can generate rules in the form of hyperplane conjunctions and disjunctions. Also, the RxREN algorithm uses reverse engineering approaches to analyze the output and identify the individual components that lead to the final outcome [88]. The preceding methods can be combined to create eclectic rule extraction methods. One such technique is RX [89], which clusters a neural network's hidden units and then extracts the logical rules that link the input to the resulting clusters. Similar eclectic methods are used in [90] to use algorithms from the so-called artificial immune system (AIS) [91] to derive rules from an NN.

The rule extraction methods examined above show a similar fidelity-versus-scalability trade-off. DeepRED and similar compositional methods work at the level of individual neurons. Because of this, they tend to generate rules that are more faithful to the actual internal computations of the network. However, analyzing the network at the level of individual neurons does not scale well with respect to the depth of the network. Non-compositional methods, in contrast, work at a black-box level with respect to the network. Among pedagogical methods, TREPAN directly queries the network and fits rules to its I/O behavior, while, among eclectic methods, RX, combines rule extraction with internal clustering. Both methods are more scalable than compositional methods and trade scalability for a less direct link between the rules extracted and the internal computations of the network.

Other methods have been proposed alongside rule extraction techniques to help clarify NN decisions. To approximate the original NN with a simpler interpretable model, interpretable mimic learning is devised in [92], which expands on concepts from model distillation. Various works, such as [93–95], explored the concept of knowledge transfer from a complicated model (the instructor) to a simpler one (the learner). One basic observation regarding NNs is that creating model simplification techniques becomes more challenging as the number of layers increases, which is why feature relevance approaches have become increasingly common in the last few years. Methods for estimating neuron-wise signals in NNs were developed [96]. They offer a method for superimposing neuron-wise explanations to generate more comprehensive explanations using these estimators. A method for breaking down an NN's prediction is provided in [97]. It is accomplished by dissecting a neuron's activity and then backpropagating its score to the input layer, producing a vector that represents the significance of each feature. When employing NNs, another method for assigning significance scores is DeepLIFT [98]. This approach computes the relevance of a feature by initially comparing the activation of a neuron with a reference one and then

calculating the difference between the two. Another popular method is presented by the authors in [99], where they discuss integrated gradients. Analyzing the model's behavior while following a line joining the instance to be explained and a base instance, which serves as a "neutral" instance, is the primary goal of this effort. Moreover, this approach has several attractive theoretical attributes that guarantee the provided explanations, such as completeness and symmetry preservation.

8.4. Challenges in Model Simplification and Feature Relevance with Increasing NN Complexity

As neural networks become increasingly complex, the simple modeling approach poses a daunting challenge [100]. As NNs evolve to incorporate more layers and intricate architectures, the development of model simplification algorithms becomes progressively more demanding. The intricate nature of DL models, particularly when internal structures are not considered, complicates the task of creating interpretable and transparent rules. The challenges in simplifying these complex models highlight the need for alternative solutions that can yield significant conclusions without sacrificing accuracy. Additionally, the importance of feature relevance techniques in interpreting neural network decisions cannot be overstated. Methods such as estimation of neuron-wise signals, decomposition of NN predictions, Deep LIFT, and integrated gradients are crucial instruments for deciphering the intricate NN decision-making processes [101]. These techniques provide a detailed comprehension of the contribution of individual features and neurons, enhancing the interpretability of NNs. The ability to assign importance scores and decompose predictions contributes to a comprehensive understanding of the factors influencing model outcomes. Therefore, the field of XAI has evolved significantly, and the combination of previous approaches showcases a multifaceted approach toward achieving comprehensive explainability.

9. Explainable AI Models in Convolutional Neural Networks

Recent studies suggest that CNNs are the most attractive deep neural networks [102]. These models excel in numerous domains, including image classification and object recognition. These models are loosely inspired by the hierarchical processing of the human visual system, enabling them to see phenomena by moving images through their layers. However, the lack of explanation leads to a search for clarifications to justify their predictions. Research on XAI has advanced in providing knowledge and insight into neurons [72]. In this section, we classify some methods that have attempted to improve the understanding of CNNs. We present and discuss a classification of XAI models that modify CNN structures, simplify CNN representations, analyze relevant features, and visualize annotations.

CNNs exhibit vigorous performance across tasks in the computer vision domain, such as image classification, object detection, face recognition, and semantic augmentation. Their structure allows them to mimic human behavior. As they move data through hundreds of convolutional and pooling layers, they gradually catch the significant features of an object. In the end, fully connected layers link these features to identify the output class scores. Despite their outstanding achievements, the complicated architecture of CNNs restricts their capacity to explore their internal structure and understand their decision-making process. Consequently, there is a growing need for CNNs' explainability in advanced machine vision fields. In this section, we examine the most recent studies about explainability models in CNNs [103]. We classify the literature works in this context into four subclasses.

9.1. Explainable CNN Models Based on Architecture Modification

To improve the explainability of CNNs, some explainable models alter the architecture of the entire network [104]. They can alter CNNs by replacing some parts in the network,

like loss functions and some layers. In addition, they can also add new components, namely deconvolutional layers or autoencoders. Many mechanisms in this category were noted in the literature over the past few years. We focus on the most used of these models.

Dynamic-K Activation [105]: This method used a customized stochastic gradient descent (SGD) to elucidate CNN models' predictions and decisions. This method adopted another optimization function called adaptive activation thresholding. The ResNet architecture was modified and trained using Dynamic-K activation. Dynamic-K outperformed traditional ResNet in interpretability and saliency map effectiveness. However, this model was just evaluated on the ResNet network. Thus, it remains unclear if this model can be adapted to other CNN types. Its contribution to explainability focuses on its adaptive activation thresholding mechanism. This was shown to result in sparser and more spatially localized saliency maps than a standard ResNet built to achieve the same objective. We emphasize that the architectural change was motivated by explainability and not solely by a desire to improve performance. However, its evaluation against established saliency baselines (e.g., Grad-CAM) was limited to one single architecture.

Residual Attention Network [106]: This method integrated attention modules within networks, such as ResNet, to generate features that are sensitive to attention. Each attention module, acting as a residual unit, comprised a mask branch and a trunk branch. The former enhanced the latter using top-down and bottom-up feedforward mechanisms to highlight output features. The latter, however, focused on feature processing. This model, including more stacks of attention modules, improved accuracy. However, despite this improvement in accuracy, there was a lack of analysis that deals with the scenario complexity, measuring the alternate-off involved in adding extra stacks of residual attention to the CNN. The contribution to explainability is the attention mask, which can be represented as a soft, learned saliency map. However, unlike Grad-CAM, this map is associated with the trained attention branch, so it is not possible to obtain this map for models that were not constructed with this architecture.

ALL-CNN [107]: The authors switched out the usual max-pooling layers with large-stride convolutional layers, where the stride size changed to 2×2 to optimize the network's complexity. The authors asserted that even though max-pooling was often used to reduce overfitting and improve CNN regularization, it did not yield the expected benefits on smaller datasets. Moreover, they demonstrated that using max-pooling layers was unimportant for training extensive CNNs. The model included deconvolutional layers and guided backpropagation to create saliency maps. Nevertheless, choosing whether to keep or drop off max-pooling layers is hard, as it depends on factors such as domain area, dataset characteristics, and the overall network structure. Its contribution to explainability is not the elimination of max-pooling, which is a choice of performance/regularization, but the saliency generation deconvolutional and guided-backpropagation pipeline that this design facilitates. Differing from Grad-CAM, this pipeline is only applicable to models that have been re-trained with max-pooling removed, and not to those that have it applied post hoc.

NIN [108]: This approach chose to substitute convolutional layers and linear filters using a smaller neural network. The authors asserted that convolutional layers have a limited capability to seize spatial invariance. The micro-convolutional layers, known as mlpconv layers, are made up of several fully connected layers with non-linear activation functions. NIN followed a comparable sliding window method to convolutional layers, allowing every mlpconv layer to generate its function map. Subsequently, the averaged function map was transferred to processing through an average pooling layer, and the resulting vector was sent to a SoftMax function. The authors proved that NIN had less accuracy than state-of-the-art networks, but its saliency maps were more explainable. Its explainability contribution is the mlpconv layers, which, compared to the standard

convolutional layers, have function maps that are explainable at the expense of reduced classification performance. As seen with ALL-CNN, this type of explainability is inherent to the architecture after training, and cannot be derived post hoc, like LIME or SHAP.

Interpretable CNN [109]: This method aimed to enhance interpretability by improving the CNN loss function. In this approach, the authors added the loss of the feature map to all filters in the final convolutional layer. The idea was to force each filter to encode particular item elements, eliminating the need for any additional annotations on object parts. Therefore, the explainable CNN performed better than state-of-the-art networks in object localization and location instability. However, the classification accuracy was lower than state-of-the-art networks. Therefore, this method entails a compromise between exactness and explainability. In terms of model explainability, the contribution is a modified loss function that causes each filter to learn to encode a part of a semantic object, with no external human annotation required. This offers a greater level of interpretability in comparison to post hoc methods such as Grad-CAM, which provides saliency maps of features that the network has, in some cases, learned in a highly entangled way. Classification accuracy is, however, lower than state-of-the-art networks, which demonstrates the trade-off between exactness and explainability.

9.2. Explainable CNN Models Based on Architecture Simplification

In this category, explainable models rely on the extraction technique to generate policies comprehensible to various stakeholders. These models consist of applying network distillation and compression. Researchers in this category reduce redundant functions to simplify the CNN model. Previous studies have approached CNN interpretation by adding hybrid models that integrate linear models into their architecture.

Decision Trees [110]: this method was attached to high-degree functions to decompose them into semantic object elements. These decision trees quantified the effect of each filter on the CNN output score and related each filter with a semantic object part label. However, this model required manual labeling of item parts for each filter, which may be challenging, specifically in medical imaging packages, in which objects are tissues and cells.

DeepKNN [111]: This model incorporated linear classifiers with every intermediate layer in CNNs. This hybrid version used training data to assess the non-conformity of predictions on tested inputs, ensuring that the intermediate layers at some stage in training aligned with CNN predictions. A D-KNN classifier became linked to every layer to identify training data points like the test input. Subsequently, the learned training inputs were compared to the CNN output throughout check time to improve explainability. The authors confirmed that the Deep KNN model furnished deeper insights and exhibited more robustness than standard CNNs. However, it is crucial to acknowledge that including a KNN classifier in each layer can impact the community's computational performance. Therefore, engaging in a complexity evaluation is essential to affirm the feasibility of training CNNs with connected KNN classifiers.

LIME [112], SHAP [113]: LIME and SHAP are the two dominant model-agnostic post hoc methods across our corpus, appearing in 61 and 74 studies, respectively. Despite their similar goals, they differ in three practically important ways. **Stability:** LIME explanations are known to vary across runs on identical inputs because they depend on random perturbation sampling, a reproducibility problem documented in 12 of the reviewed studies. SHAP, grounded in cooperative game theory, produces consistent attributions and is therefore preferred in high-stakes domains such as healthcare and finance, where explanation reproducibility is a regulatory requirement. **Scalability:** LIME is computationally lightweight and suitable for real-time applications, while SHAP's exact computation grows exponentially with feature count, making KernelSHAP approximations necessary for deep

models. **Scope:** LIME provides local explanations only; SHAP can aggregate local Shapley values to produce global feature importance rankings, giving it a decisive advantage in model debugging and audit contexts.

In short, LIME is the practical choice when speed matters. SHAP is the rigorous choice when consistency and global interpretability are required.

Subnet Extraction [114]: This method was introduced in the context of neural network pruning. The authors aimed to detect semantic convolutional layers. There exist two levels of subnet extraction: sample level and class level. The sample level extraction guaranteed the uniformity of distinct predictions in CNNs. However, the class-level extraction measured the network prediction within a singular class.

Self-Explaining Neural Networks (SENNs) [115]: This method used a bottom-up mechanism to explain predictions in CNNs. It encompassed three components: a concept encoder, an input-dependent parameterizer, and a combination function. The image was transformed into a set of representative features. Afterward, the relevant scores were computed to make the right prediction. Although this mechanism was considered a robust and reliable model, there is no evidence of classification accuracy.

9.3. Explainable CNN Models Based on Feature Relevance

The primary idea behind these models lies in emphasizing the importance of features against CNN prediction. These techniques look for important characteristics and cluster them.

EBANO model [116]: This mechanism grouped hyper-column features selected from high-level layers using the K-means method. Each cluster of pixels presented an interpretable feature. Then, explainable features perturbed the trained input, and finally, the mechanism provided explainable details of the classified image.

Feature Similarity [117]: This model used the K-nearest algorithm to measure feature similarity. It distinguished features from the first pooling layer by training a CNN and grouping them in a database. Afterward, this database of features was compared to the extracted image using the same CNN. Nonetheless, it extracted features from the low-level pooling layer and ignored high-level features.

DGN-AM model [118]: This scheme used a CNN to recognize features learned by neurons. Then, the authors applied a backpropagation mechanism on the generated image to seek the neuron with maximum activations. This model improved the CNN's ability to learn features on the neuron level. However, DGN-AM could depict features if the images were standards.

DeepLIFT model [119]: This model decomposed CNN prediction with respect to the initial image using a backpropagation mechanism through each feature. This decomposition is referred to as an input reference relying on domain knowledge.

Feature Extraction and Reconstruction CNN (FER-CNN) [120]: This different approach attached a feedback CNN to the input image to reconstruct features hierarchically. Two subnets made up this mechanism: an encoder for feature extraction and a decoder for feature reconstruction. Regarding classification accuracy, the authors verified that FER-CNN performed better than alternative techniques.

9.4. Explainable CNN Models Based on Visual Explanation

These models aim to elucidate CNN predictions by causing class activation maps or heatmaps such as saliency maps.

Saliency maps [121]: This explainable method generated saliency maps to explain CNN predictions. The authors considered the pixel influence on the gradient score to classify the input image. The saliency maps were visually assessed for different categories.

Class Activation Maps (CAMs) [122]: CAMs and one of their gradient-based successors, Grad-CAM, are the most surveyed visual explanation methods from our set ($n = 47$ studies). By swapping the last fully connected layer with a global average pooling (GAP) layer, CAMs offer a form of saliency localization. However, this is a deliberate structural edit that decreases the classification accuracy by approximately 1 to 3 % and, as noted in the surveyed studies, imposes the limitation of its application to GAPs' layer-less architectures. Dissolving this constraint, Grad-CAM derives class-discriminative maps directly from the gradients, thus making them architecture-agnostic and substantially more useful in practical applications.

Making Visualizations [123]: These methods present a masking approach. An example of these models is the Score-CAM model, which confirms that using gradients can have some drawbacks, such as gradient saturation and false certainty. These drawbacks can produce noisy saliency maps.

Intrinsic Visualization [124]: The previous visualization models were post hoc that added an extra component to interpret CNN predictions. However, intrinsic visualization models modified the CNN architecture to search for an explanation of network predictions. The teacher–student model used autoencoders to explain the class prediction. It used a network to encode the input image and another to reconstruct it.

CAM variants have known limitations that are likely to affect many studies in our corpus that utilize these methods, despite their popularity. First, the localization quality of CAM variants is highly dependent on the neural network architecture. CAM methods that rely on global average pooling (GAP) and those that depend on gradients computed at a specific convolutional layer can produce highly variable results depending on the depth of the network and the specific target layer. Second, localization errors and saturation of gradients affect CAM variants as noted for Score-CAM in this section, and can create explanations that are focused on visually prominent areas that are not causally related to the decision of the neural network. Third, there is no guarantee of faithfulness, as a heatmap may look good, but may not represent the features that the model actually used to make its decision. This concern is framed in the discussion of the infidelity and sensitivity metrics in Section 10.1. Finally, like other post hoc explanation methods, CAM variants can be affected by adversarial attacks/perturbations. In these attacks, small perturbations that are imperceptible to humans may lead to drastic changes in the CAM saliency maps, while related predictions of the neural network remain unchanged. This is explored more in the challenge map in Section 11.

Table 8 illustrates a thorough comparison of the explainability methodologies under four dimensions: complexity of computation, model access, scope of explanation, and architecture applicability. From this comparison, some trends can be observed. First, most methods require white-box access. LIME and SHAP are the only true black-box methods and are therefore the only implementable methods in the absence of model gradients and/or in the absence of intermediate representations. Second, the majority of the methods presented provide local explanations. In the context of the table presented, only SHAP is able to consolidate local explanations to provide a global feature importance summary, which makes SHAP particularly useful in the context of model evaluation and model regulations. Third, the methods presented in the table also show a clear separation of specificity to a particular model architecture. The methods (Dynamic-K activation, residual attention networks, ALL-CNN, NIN, interpretable CNNs, and CAMs) that require modifications of the model architecture are specific to certain models and cannot be applied to a model that has already been deployed. However, the other methods (LIME, SHAP, Grad-CAM, DeepLIFT, saliency maps, EBANO, DGN-AM) are applicable regardless of the model architecture. Lastly, the complexity of computation for these methods varies, for

example, from methods that are lightweight and suitable for real-time inference (LIME, saliency maps, CAMs, and Grad-CAM) to methods that are too computationally expensive for practical use and require approximation (SHAP and DeepKNN).

Table 8. Comprehensive summary of some explainability techniques for CNN models.

Technique	Approach	Strengths (+)/ Challenges (–)	Ref.	Complexity	Model Access	Scope	Architecture Applicability
Dynamic-K Activation	Modifies ResNet using adaptive activation thresholds.	+ Improves interpretability and saliency maps. – Limited to ResNet only.	[105]	Medium	White-box	Local	Specific (ResNet)
Residual Attention Network	Integrates attention modules within CNNs.	+ Improves accuracy and feature focus. – No complexity trade-off analysis.	[106]	Medium	White-box	Local	Specific (ResNet)
ALL-CNN	Replaces max-pooling with strided convolutions; guided backpropagation for saliency.	+ Cleaner saliency maps; reduces overfitting. – Performance varies with dataset.	[107]	Medium	White-box	Local	Specific (no max-pooling)
NIN	Replaces convolutional layers with mlpconv micro-networks.	+ More explainable saliency maps. – Lower classification accuracy.	[108]	Medium	White-box	Local	Specific (mlpconv)
Interpretable CNN	Modifies loss to force each filter to encode one semantic part.	+ Localization without manual annotation. – Accuracy–explainability trade-off.	[109]	Medium	White-box	Local	Specific (final conv)
Decision Trees	Attaches semantic labels to CNN filters via decision trees.	+ Global, human-readable explanations. – Manual per-filter labeling required.	[110]	Low	White-box	Global	Agnostic
DeepKNN	Incorporates k-nearest neighbors into every intermediate layer.	+ Robustness and deeper interpretability. – High computational overhead.	[111]	High	White-box	Local	Agnostic
LIME	Fits a local surrogate model on perturbed input samples.	+ Lightweight; model-agnostic. – Unstable across runs; local only.	[112]	Low	Black-box	Local	Agnostic
SHAP	Assigns feature contributions via Shapley values.	+ Consistent; supports global explanations. – Exponential cost with feature count.	[113]	High	Black-box	Local + Global	Agnostic
Saliency Maps	Gradient-based pixel-influence visualization.	+ Widely applicable; simple to compute. – Sensitive to gradient noise.	[121]	Low	White-box	Local	Agnostic
CAM	GAP layer generates class activation maps.	+ Highlights discriminative regions. – Requires GAP; reduces accuracy 1–3%.	[122]	Low	White-box	Local	Specific (GAP)
Grad-CAM	Gradient-based maps; no architectural modification.	+ Architecture-agnostic; widely applicable. – Gradient saturation; adversarial vulnerability.	[122]	Low	White-box	Local	Agnostic
SENN	Concept encoding and input-dependent parameterization.	+ Produces interpretable features. – Classification accuracy not reported.	[115]	Medium	White-box	Local	Agnostic
DeepLIFT	Backpropagation relative to a reference activation.	+ Contrastive, fine-grained attribution. – Sensitive to reference input.	[119]	Medium	White-box	Local	Agnostic
EBANO	K-means clustering of hyper-column features.	+ Interpretable pixel clusters. – Limited generalization.	[116]	Medium	White-box	Local	Agnostic
DGN-AM	Backpropagation on generated images for neuron-level features.	+ Neuron-level understanding. – Requires standardized images.	[118]	Medium	White-box	Local	Agnostic

10. Review of XAI Evaluation Metrics

Even though XAI research is booming and explainable models have shown promise across many application fields, there is still a lack of standardization in evaluation methodologies. A plethora of strategies have been proposed and tested to address this issue. The emergence of numerous explainability techniques has caused academics to concentrate on their evaluation. Different evaluation metrics have been proposed and discovered in the literature, and numerous evaluation types have been carried out. These metrics consider not only explainability but also accessibility and content inclusion, among other elements of explanations [125,126].

Our corpus reveals a fundamental misalignment between the evaluation and practical application of XAI methods. Objective metrics, particularly fidelity, stability, and sensitivity, dominate the existing literature, appearing in 98 of the reviewed studies; however, these metrics assess proxy properties of an explanation rather than its actual utility to the end user. By contrast, only 31 studies incorporate human-centered metrics such as task performance, calibration, and subjective trust, and fewer than 10 report both metric families in conjunction.

This disparity is particularly significant given that the two families frequently yield conflicting assessments: a method that scores highly on fidelity may simultaneously receive low human trust ratings, indicating that users find the resulting explanation confusing, misleading, or inadequate for practical decision-making. As long as studies continue to report these dimensions in isolation, the field will lack the empirical foundation required to converge on best practices and enable principled method selection.

We recommend the dual-reporting convention, where studies report at least one objective metric (e.g., fidelity, stability, sensitivity) and one human-centered metric (e.g., task performance, subjective trust, user calibration). There are three cases presented below and summarized in Table 9. Objective (or quantitative) and human-centered (or qualitative) metrics lead to conflicting conclusions when measured using the same method. Reporting only one set of metrics can lead to over-claiming the study's effectiveness or missing key study limitations, which can be revealed using the other set of metrics.

Conflict 1—Fidelity vs. Decision Support Value: Many reviewed studies in the CAM-family report high fidelity when estimated using objective metrics of infidelity/sensitivity. However, in the human-centered study, it was found that participants who were shown visual explanations did not outperform participants who were shown no visual explanation in the prediction task. Thus, the explanation can be locally faithful to the model, but can provide no decision support to the end user.

Conflict 2—Objective Quality vs. User Trust: In one study, rule-based explanations were evaluated and performed poorly on the criteria of objective correctness and redundancy. However, they were found to be more favorable in subjective trust when compared to the feature attribution maps. This is because if-then rule-based reasoning better aligns with lay users' mental models of decision rules than heat maps do.

Conflict 3—Stability vs. Practical Utility: The instability of LIME across instances has been noted, and it is identified as an objective weakness in relation to stability in the reviewed literature. However, this does not always signify a decrease in utility. Particularly in the context of exploratory model-debugging, different explanations can illustrate and help capture different aspects of model behavior. This thus shows that the practical meaning of an objective metric depends primarily on the use case in which it is applied.

Table 9. Documented conflicts between objective and human-centered evaluation metrics in the reviewed corpus.

References	Objective Metric Result	Human-Centered Result	Type of Conflict
[127,128]	High fidelity scores reported for CAM-family methods under infidelity/sensitivity metrics.	Participants exposed to visual explanations did not outperform participants given no explanation on a downstream prediction task.	Fidelity $\neq$ Decision-support value
[129–131]	Poor objective correctness and redundancy scores for rule-based explanations.	Higher subjective trust ratings than feature-attribution maps; if-then structure matches users' mental models more closely.	Quality $\neq$ User trust
[112] + corpus	Low stability scores for LIME across runs (instability documented in reviewed studies).	Varying explanations prove useful in exploratory model-debugging contexts by revealing complementary facets of model behavior.	Stability $\neq$ Practical utility

- **Objective evaluations:** This category includes studies that use automated techniques and objective criteria to assess explainability techniques.
- **Human-centered evaluations:** These studies use a human-in-the-loop methodology to assess explainability techniques by incorporating end users and taking advantage of their comments or knowledgeable decision-making.

In [70], a similar dual-categorization scheme was proposed; however, the two categories were denoted as heuristic-based and user-based metrics. Model sizes and other numerical measurements made up of mathematical objects are included in the former category [132,133]. This clear explainability indicator is based on the idea that as the size of a model increases, people will find it more difficult to understand. This assumption was proven to be false, though. Because larger models provide more information pertinent to categorization and smaller models are unlikely to confuse users, users perceived some larger models as being more intelligible than some smaller ones. This finding emerged from the human-centered evaluation study carried out in [134]. A different classification scheme was introduced in [71], comprising three assessment metrics categories: application-grounded, functionally grounded, and human-grounded. The functionally grounded and human-grounded metrics are equivalent to the user and heuristic-based metrics suggested in [135]. By comparing the productivity enhancement of an insignificant number of users who adhere to these explanations over those provided by human engineers, as in [136], application-grounded metrics evaluate the quality of machine-made explanations of data-driven models. They can be combined with human-centered ones since they include people.

10.1. Objective Evaluations

Academics proposed several metrics to formally and objectively assess the explainability techniques. There is a general agreement in the scientific literature that simpler learning methods, such as decision trees and linear regression, can produce conclusions compared to more intricate methods, like neural networks, since they are inherently self-interpreting. However, these less sophisticated methods typically do not generate models as accurate as those produced by more sophisticated ones. Numerous variables, namely learning architecture, learning algorithm, and hyper-parameter setting, affect models' interpretability.

- Two approaches of explainability, Explain and Ime [137,138], were used to measure the increase in sales department productivity using a few formal metrics or sales

performance indicators. In a real-world business, the system was tested extensively, and the sales performance metrics were tracked. The indicators corroborated the idea that, unlike the mental models of sales experts based on their prior experience, data-driven explanations facilitate unbiased decision-making. The forecasts based on the Explain and Ime explanations outperformed initial sales forecasts.

- The explainability of methods creating visual explanations of a neural network trained to identify images was evaluated using two quantitative criteria presented in [139]. The first evaluation metric, i.e., filter interpretability, takes into account six different categories of semantics for CNN filters that need human annotation at the pixel level: components, sceneries, materials, textures, colors, and objects. The metric calculates the areas where these annotations intersect with the distribution of a filter's activation values over a heatmap. It can be said that the filter represents these semantic notions if there are areas of overlap since the distances between an item's parts must be nearly invariant. The second metric, i.e., location instability, determines whether a CNN locates the pertinent portions of the same object, shown in several images, at a virtually constant relative distance.
- The authors in [127] suggested two substitute measurements: the first, i.e., infidelity, is the anticipated variation between the dot product of a notable input disturbance and the subsequent output perturbation; the second is sensitivity, measuring how much a minor perturbation in the input examples affects a visual explanation and the ensuing output disturbance.
- The effectiveness of rule-based explanations in [129] was evaluated by determining the percentage of invalid rules that are said to be false. Instead, the rules that were shown to be correct were divided into two categories: rules that were minimum and correct and those that were redundant, i.e., they could be eliminated without affecting the rule set's ability to forecast the future.
- In [140], a general evaluation metric was suggested for explainability in post hoc methods. It is based on the possibility of presenting circumstances that are artifacts, or "counterfactual examples," that are not justified for the classifier and independent of prior knowledge. Such a situation may arise when a model has to predict something beyond its training set. These examples are generated via an algorithm that uses the lowest perturbation that alters the anticipated observation classes. Therefore, these classes remain linked to the supplied data and prevent the creation of examples that logically or practically reflect impossible scenarios.
- Automatically evaluating textual explanations [141] generated by explainability algorithms via three automated quantitative criteria aimed at evaluating the quality of text documents: using pre-trained word embeddings, the automatic NT translation metric (METEOR) estimates the semantic similarity between words in sentences, and consensus-based image description evaluation (CIDE) compares sentences generated by neural networks to human-written explanations by counting "Term Frequency-Inverse Document Frequency"-weighted n-grams. The last one is the bilingual evaluation understudy (BLEU) metric, which measures sentence similarity based on the average percentage of n-gram matches.

10.2. Human-Centered Evaluation

When explanations assist users in creating a thorough and accurate mental model of a model's inferential process, they are considered efficient. Numerous scientific papers examined the extent to which one or more approaches could be explained [142]. Figure 7 provides a diagrammatic representation of a human-in-the-loop approach. Two types of human subjects were used in these investigations. First, participants were encouraged to en-

gage with one or more explanatory tools and provide feedback by filling out questionnaires. Participants were randomly recruited from the general public without previous technical or domain knowledge. Second, domain experts were asked to provide knowledgeable evaluations of the explanations generated by these methods and evaluate the uniformity of the explanations with the domain knowledge. Depending on the type of questions asked of participants, the scientific papers can be divided into two groups (see Table 10). Quantitative studies rely on closed-ended questions that are simple to analyze statistically. However, qualitative research uses open-ended questions to gain deeper insights.

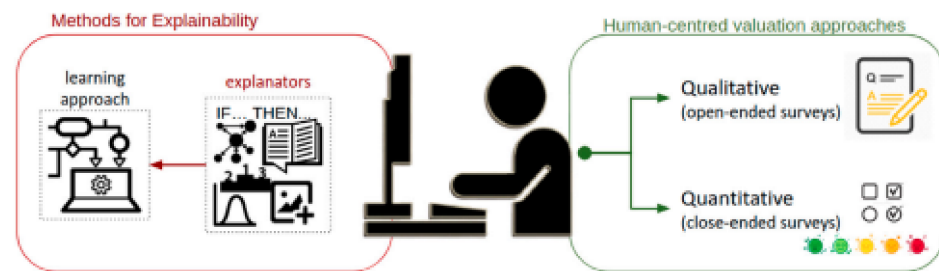


Figure 7. Overview of explainability methods and human-centered evaluation approaches in the XAI assessment framework: qualitative (open-ended surveys) and quantitative (closed-ended surveys) strategies used to assess machine-generated explanations across the reviewed studies [71].

Table 10. Categorization of some scientific papers aiming to evaluate explainable AI models based on human-centered metrics.

Construction Approach	Measure Type	Output Format	Reference
Autonomous Agents	Qualitative	Textual	[142]
Data Clustering	Qualitative	Mixed	[143]
Learning Systems	Qualitative	Textual	[144]
Neural Networks	Qualitative	Visual	[145]
Case-based Reasoning	Quantitative	Mixed	[146]
Neural Networks	Quantitative	Visual	[128]
ML (Model-agnostic)	Quantitative	Rules	[131]
Rule-based-classifier	Quantitative	Mixed	[130,147]

10.2.1. Qualitative Methods

- To evaluate end users' situation awareness when engaging with autonomous intelligent agents, the Situation Awareness-based Global Assessment Technique (SAGAT) was proposed in [143]. The objective was to evaluate whether the explanations offered have sufficient details to enable people to carry out their responsibilities. Users were asked to assess the situation after simulations of representative tasks, which require the assistance of an autonomous agent, were frozen at randomly chosen moments in the SAGAT test.
- In [144], a qualitative evaluation of the Mind the Gap Model (MGM) approach's interpretability for explainability was conducted. Using the most pertinent attributes, the MGM grouped the input dataset's data. Participants had access to both raw data and data clustered using the MGM and K-means. They were allotted five minutes to draft an executive summary of each data representation, consisting of two to three sentences. They could not summarize the raw data within the allotted time frame; however, they could do it for the data that contained the clustering MGM and K-means.
- The need for XAI in intelligent tutoring systems (ITSs) was investigated in [148]. In an interactive simulation, study participants were instructed to use an ITS that offered tools for exploring and elucidating an algorithm for constraint fulfillment problems.

The exploration tool provided instructions to the participants through a textual hint message. They were allowed to choose whether they wanted an explanation provided for each hint or not.

- Three further methods, activation maximization, sampling units, and linear combination, were contrasted in [146] to produce heatmaps that highlight the key input image elements. Activation maximization involves selecting input properties that increase the activation of a single hidden neuron. A sampling unit calculates the likelihood that each sample will be assigned to a class and sets the value of a neuron to one. Finally, linear combination entails choosing the neuronal connections with the highest weights between two adjacent layers. Rather than employing any objective criteria, the authors subjectively contrasted these methods through a visual examination of the heatmaps and an analysis of the connections between them.

10.2.2. Quantitative Methods

- To test the Bayesian case model (BCM) [149], participants classified sixteen recipes that were only composed of a set of ingredients into the appropriate category (for example, a recipe for cookies had to be categorized under the category “cookie”). The number of recipes that were successfully classified was then counted. The BCM was contrasted with latent Dirichlet allocation (LDA), a clustering technique focused on identifying comparable features in the data.
- A test of GAN dissection was performed in [150] using a similar setup. Images reporting highlighted patches displaying the most strongly activated areas for every unit at every intermediate convolutional layer of a DNN were shown to the participants. Each layer received labels for various parts, objects, scenes, textures, materials, and colors. Then, it was matched with a semantic. For instance, the semantic might comprise this list, and images with flowers would be labeled “flower” if a DNN was trained to identify a list of objects in input images, such as vehicles and flowers. Regarding whether the highlighted patches were related to the label, participants were asked to answer yes or no.
- In [145], the ability of Anchors and LIME to assist end users in projecting the predictions of an image classifier was evaluated. Ten random test instances were presented with two rounds of explanations produced by either LIME or Anchors. The participants were asked to predict which output class the classifier would assign to each of the ten instances.
- In the human-centered experiment conducted in [128], participants were divided into four groups and given no explanations apart from textual, verbal, visual (created with LIME), and vocal stimuli. A 7-point Likert scale and open-ended questions were used to obtain participants’ feedback regarding the explanations.

In conclusion, these metrics provide a comprehensive evaluation of the explainable convolutional neural model, considering the architecture of the entire network and human understanding. This combination is considered as a challenge to ensure efficiency and reliability in complex neural networks.

10.3. Benchmarking Gap

Among the 154 reviewed studies, none focused on a community agreed-upon XAI benchmark suite that is analogous to the standard accuracy benchmarks regarding computer vision. Most objective evaluations are restricted to the studies’ particular datasets, and other protocol choices are made arbitrarily. Examples include the selection of perturbation strategy to apply when calculating the infidelity/sensitivity metrics, or the selection of annotation granularity to apply for the filter interpretability metrics. The diversity in

human-centered evaluations is even more pronounced in terms of participant recruitment, task design, and sample size. The lack of common benchmarks and standardized metrics results in most of the cross-study comparisons being estimates. For instance, the two studies may evaluate “fidelity” independently, and thereby use different datasets and perturbation strategies, and as a result, their conclusions would be literally incomparable. Establishing community agreed-upon XAI benchmark suites for saliency maps, feature attribution, rule-based and counterfactual explanations, and other XAI methods is one of the most pressing infrastructure demands of the domain. In Section 11, we revisit the gap and propose concrete strategies.

Table 11 puts the dual-reporting recommendation into practice. For the four explanation-output types detailed in the previous taxonomy, it describes the most suitable objective and human-centered metrics. An evaluation protocol designer can pick the most suitable metric pair from this table that aligns with both the explanation type being presented and the intended use case of the evaluation. This approach avoids the problem of using the most common metrics for the sake of using metrics.

Table 11. Recommended evaluation metrics by explanation type.

Explanation Type	Recommended Objective Metric	Recommended Human-Centered Metric
Saliency/visual maps	Infidelity, sensitivity [127]; location instability [139]	Qualitative visual-comparison studies [146]
Rule-based explanations	Rule correctness and redundancy [129]	Task-performance studies [145]
Feature-attribution methods	Shapley-value consistency checks [113]	Subjective trust/Likert-scale studies [128]
Counterfactual explanations	Validity-of-counterfactual checks [140]	User-comprehension studies [142]

11. Challenges and Future Directions

This review offered an overview of various progressions and facets within the field of XAI. However, because XAI is a relatively new and expanding field, many unresolved research and operational issues, as shown in Figure 8, will need to be taken into account as the field develops.

The first thing that springs to mind is how various explanation models mesh with one another. A closer examination of the proposed techniques reveals that although there is some overlap among the different explanation types, they seem to be widely divided, each focusing on a distinct subject. Furthermore, it is not evident how to combine them to create a more comprehensive explanation, making it more difficult to agree on what constitutes a comprehensive explanation or build pipelines that attempt to automate explanations. Practically speaking, there are not many XAI techniques with effective implementations, which may be explained by the fact that the area is still young and developing, yet it still increases the difficulty of implementing XAI in large-scale applications. Creating more robust model-specific strategies is another area that will attract further focus. The benefit of this approach is that the ensuing methods could use the unique characteristics of a model to generate explanations, likely increasing fidelity and enabling a greater understanding of the model’s internal dynamics rather than just describing its output. In addition, since the new algorithms do not rely on expensive approximations, it would likely be easier to find efficient algorithmic implementations. Building trust in the explanations themselves is a broader issue raised by this last point and needs to be addressed. As previously indicated, current studies have demonstrated how several frequently used XAI techniques are susceptible to adverse attacks. Hence, questions arise regarding the credibility of an XAI technique result. Furthermore, there are other relevant questions in the literature regarding the suitability of some of the suggested methods to handle broad explainability [131]. Hybrid models combining the expressiveness of opaque models with the interpretable semantics of transparent models [113], where neural networks and linear regression are

combined, represent another research area that has recently gained momentum. This approach may facilitate the development of explainable models that close the divide between opaque and transparent models. Finally, as XAI develops, new methods should incorporate ideas from causal analysis [151,152], which is already a significant factor in basic issues in other fields, such as bias and fairness in machine learning [153,154]. Thus, it is expected to play a crucial role in the development of the XAI literature.

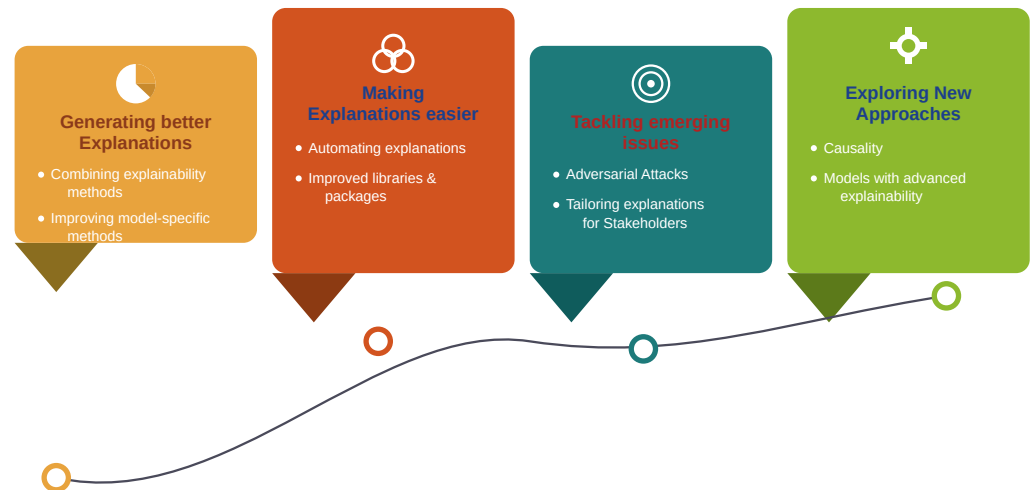


Figure 8. Unresolved research and future operational issues in XAI.

12. Summary and Outlook

A consolidated overview of XAI methods for CNNs is provided, including their assessments and how they are integrated and used across different fields. An overview of intrinsic and post hoc methods is provided, which includes modification of architectures, simplification of architectures, feature relevance, and visual explanation. Both objective measurements and human-centered evaluation metrics are discussed. The application domain for practical explanations, organized by methodological approach and assessment framework, is detailed, with special attention to works published between 2020 and 2025 and a small number of prior seminal works.

Patterns show across multiple domains. In healthcare, post hoc visual explanation methods such as Grad-CAM and its variants are utilized almost exclusively. This is largely due to the heatmaps being directly interpretable by clinicians, as well as their use not requiring infrastructure changes to existing diagnostic models. In computer vision, gradient-based methods are the most prevalent in tasks that involve object detection and image classification. In these cases, SHAP and DeepLIFT, which are methods of feature relevance, are preferred in the context of model audits. In finance, due to the requirements of regulations for explanation to be consistent and reproducible, the model-agnostic methods are preferred, and due to the link to cooperative game theory, SHAP outlines the most frequent examples. In cybersecurity and in autonomous systems, explanation methods remain largely unexplored, and the studies considered the most use post hoc methods in a general framework. In all areas, especially when it comes to the simultaneous review of objective and human-centered metrics, research is still in its infancy, and there are no standard frameworks for the comparison of explanation methods across these disciplines.

This review offers practical advice based on a systematic synthesis of XAI methodologies and their relevance to CNNs. Four analytical dimensions of the reviewed methodologies describe their explanation timing, model dependency, output type, and target component. In some scenarios, lightweight, post hoc methods like LIME are adequate for local interpretability, while, in other scenarios, model-specific methods like LRP or

Grad-CAM are necessary to provide higher fidelity or architectural transparency. The persistent gap between objective evaluation metrics and human-centered assessments is made explicit, together with a dual-reporting recommendation to support more rigorous and comparable evaluations across studies.

We present a perspective that attempts to address identified gaps. First, we claim that, within the context of CNNs, the formulation of standardized XAI benchmarks is an urgent matter because it will allow for the rigorous and scalable comparison of explanation methods, datasets, and applications. We claim that the scalability constraints of post hoc methods for large models need to be resolved within the context of Shapley values and/or gradient-based attribution computing, via approximation and/or parallelization. We state that the main design goal for explanation methods should be the robustness of adversarial attacks, especially given the vulnerabilities of LIME and SHAP, which are among the most popular. We also claim that the design of XAI will be enhanced with early human alignment and coworking participatory evaluations of end users in the design of XAI, throughout the process, rather than at the end. We also claim that the priority of the design of XAI, along with deep hybrid architectures, should be the exploration of the transparency of these architectures. We conclude that the design of CNNs within the scope of federated learning, generative modeling, and multi-modal synthesis remains an important and largely unexplored research area.

For practitioners, we outline a path from application scope to XAI method selection, coupled with specific evaluation and assessment needs, while for researchers, we establish a research and development agenda to help lower the barriers to the design of embedded, transposable, and human-centered explainability across deep learning models in order to enhance the transparency, trust, and accountability of AI systems, especially in the fields of healthcare, finance, and computer vision.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/make8080220/s1>, Table S1: PRISMA 2020 Checklist.

Author Contributions: Conceptualization, Z.T. and M.Z.; data curation, formal analysis, investigation, methodology, Z.T. and M.Z.; project administration, Z.T. and M.Z.; resources, software, Z.T. and M.Z.; validation, visualization, Z.T. and M.Z.; writing—original draft, Z.T. and M.Z.; writing—review and editing, Z.T. and M.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
CAM	Class Activation Map
CNN	Convolutional Neural Network
DARPA	Defense Advanced Research Projects Agency
DL	Deep Learning
DNN	Deep Neural Network
GAN	Generative Adversarial Network
Grad-CAM	Gradient-weighted Class Activation Mapping
IoT	Internet of Things

LIME	Local Interpretable Model-agnostic Explanations
LRP	Layer-wise Relevance Propagation
ML	Machine Learning
NN	Neural Network
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RNN	Recurrent Neural Network
SHAP	SHapley Additive exPlanations
SLR	Systematic Literature Review
XAI	eXplainable Artificial Intelligence

References

- Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjel, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [\[CrossRef\]](#) [\[PubMed\]](#)
- Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent trends in deep learning based natural language processing. *IEEE Comput. Intell. Mag.* **2018**, *13*, 55–75. [\[CrossRef\]](#)
- Andaur Navarro, C.L.; Damen, J.A.A.; Takada, T.; Nijman, S.W.; Dhiman, P.; Ma, J.; Collins, G.S.; Bajpai, R.; Riley, R.D.; Moons, K.G.; et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: Systematic review. *BMJ* **2021**, *375*, n2281. [\[CrossRef\]](#) [\[PubMed\]](#)
- Prabhu, S.P. Ethical challenges of machine learning and deep learning algorithms. *Lancet Oncol.* **2019**, *20*, 621–622. [\[CrossRef\]](#) [\[PubMed\]](#)
- Chen, R.; Yang, L.; Goodison, S.; Sun, Y. Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data. *Bioinformatics* **2020**, *36*, 1476–1483. [\[PubMed\]](#)
- Sayres, R.; Taly, A.; Rahimy, E.; Blumer, K.; Coz, D.; Hammel, N.; Krause, J.; Narayanaswamy, A.; Rastegar, Z.; Wu, D.; et al. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *Ophthalmology* **2019**, *126*, 552–564. [\[CrossRef\]](#) [\[PubMed\]](#)
- Boles, A.; Rad, P. Voice biometrics: Deep learning-based voiceprint authentication system. In *2017 12th System of Systems Engineering Conference (SoSE)*; IEEE: Piscataway, NJ, USA, 2017; pp. 1–6.
- Panwar, S.; Das, A.; Roopaei, M.; Rad, P. A deep learning approach for mapping music genres. In *2017 12th System of Systems Engineering Conference (SoSE)*; IEEE: Piscataway, NJ, USA, 2017; pp. 1–5.
- Sahba, A.; Das, A.; Rad, P.; Jamshidi, M. Image graph production by dense captioning. In *2018 World Automation Congress*; IEEE: Piscataway, NJ, USA, 2018; pp. 1–5.
- Bendre, N.; Ebadi, N.; Prevost, J.J.; Najafirad, P. Human action performance using deep neuro-fuzzy recurrent attention model. *IEEE Access* **2020**, *8*, 57749–57761. [\[CrossRef\]](#)
- Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [\[CrossRef\]](#) [\[PubMed\]](#)
- Preece, A.; Harborne, D.; Braines, D.; Tomsett, R.; Chakraborty, S. Stakeholders in explainable AI. *arXiv* **2018**, arXiv:1810.00184.
- Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* **2018**, *51*, 93. [\[CrossRef\]](#)
- Wanner, J.; Herm, L.-V.; Heinrich, K.; Janiesch, C.; Zschech, P. White, grey, black: Effects of XAI augmentation on the confidence in AI-based decision support systems. In *ICIS 2020 Proceedings*; Association for Information Systems: Atlanta, GA, USA, 2020.
- Rai, A. Explainable AI: From black box to glass box. *J. Acad. Mark. Sci.* **2020**, *48*, 137–141.
- Ferreira, J.J.; Monteiro, M.S. What are people doing about XAI user experience? A survey on AI explainability research and practice. In *Design, User Experience, and Usability. Design for Contemporary Interactive Environments*; Marcus, A., Rosenzweig, E., Eds.; Springer International Publishing: Cham, Switzerland, 2020; Volume 12201, pp. 56–73.
- Kaur, H.; Nori, H.; Jenkins, S.; Caruana, R.; Wallach, H.; Vaughan, J.W. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*, Honolulu, HI, USA, 25–30 April 2020; pp. 1–14.
- Saeed, W.; Omlin, C. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowl.-Based Syst.* **2023**, *263*, 110273. [\[CrossRef\]](#)
- Clement, T.; Kemmerzell, N.; Abdelaal, M.; Amberg, M. XAIR: A Systematic Metareview of Explainable AI (XAI) Aligned to the Software Development Process. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 78–108. [\[CrossRef\]](#)
- Naeem, H.; Alshammari, B.M.; Ullah, F. Explainable artificial intelligence-based IoT device malware detection mechanism using image visualization and fine-tuned CNN-based transfer learning model. *Comput. Intell. Neurosci.* **2022**, *2022*, 7671967. [\[CrossRef\]](#) [\[PubMed\]](#)

21. Langer, M.; Oster, D.; Speith, T.; Hermanns, H.; Kastner, L.; Schmidt, E.; Sasing, A.; Baum, K. What do we want from explainable artificial intelligence (XAI)?—A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artif. Intell.* **2021**, *296*, 103473. [CrossRef]
22. Alicioglu, G.; Sun, B. A survey of visual analytics for explainable artificial intelligence methods. *Comput. Graph.* **2022**, *102*, 502–520. [CrossRef]
23. De, T.; Giri, P.; Mevawala, A.; Nemani, R.; Deo, A. Explainable AI: A hybrid approach to generate human-interpretable explanation for deep learning prediction. *Procedia Comput. Sci.* **2020**, *168*, 40–48. [CrossRef]
24. Martins, T.A.; Almeida, A.; Cardoso, E.; Nunes, L. Explainable artificial intelligence (XAI): A systematic literature review on taxonomies and applications in finance. *IEEE Access* **2023**, *12*, 618–629. [CrossRef]
25. Wei, K.; Chen, B.; Zhang, J.; Fan, S.; Wu, K.; Liu, G.; Chen, D. Explainable deep learning study for leaf disease classification. *Agronomy* **2022**, *12*, 1035. [CrossRef]
26. Garrido, M.C.; Cadenas, J.M.; Bueno-Crespo, A.; Martínez-España, R.; Giménez, J.G.; Cecilia, J.M. Evaporation forecasting through interpretable data analysis techniques. *Electronics* **2022**, *11*, 536. [CrossRef]
27. Zhang, Z.; Huang, J.; Duan, S.; Huang, Y.; Cai, J.; Bian, J. Use of interpretable machine learning to identify the factors influencing the nonlinear linkage between land use and river water quality in the Chesapeake Bay watershed. *Ecol. Indic.* **2022**, *140*, 108977. [CrossRef]
28. Kakogeorgiou, I.; Karantzalos, K. Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing. *J. Appl. Earth Obs. Geoinf.* **2021**, *103*, 102520. [CrossRef]
29. Meena, J.; Hasija, Y. Application of explainable artificial intelligence in the identification of squamous cell carcinoma biomarkers. *Comput. Biol. Med.* **2022**, *146*, 105505. [CrossRef] [PubMed]
30. Miró-Nicolau, M.; Moyà-Alcover, G.; Jaume-i-Capó, A. Evaluating Explainable Artificial Intelligence for X-Ray Image Analysis. *Appl. Sci.* **2022**, *12*, 4459. [CrossRef]
31. Lombardi, A.; Diacono, D.; Amoroso, N.; Biecek, P.; Monaco, A.; Bellantuono, L.; Pantaleo, E.; Logroscino, G.; De Blasi, R.; Tangaro, S.; et al. A robust framework to investigate the reliability and stability of explainable artificial intelligence markers of mild cognitive impairment and Alzheimer’s disease. *Brain Inf.* **2022**, *9*, 17. [CrossRef]
32. Chen, H.-Y.; Lee, C.-H. Vibration Signals Analysis by Explainable Artificial Intelligence (XAI) Approach: Application on Bearing Faults Diagnosis. *IEEE Access* **2020**, *8*, 134246–134256. [CrossRef]
33. Kim, D.; Lee, J. Predictive evaluation of spectrogram-based vehicle sound quality via data augmentation and explainable artificial intelligence: Image color adjustment with brightness and contrast. *Mech. Syst. Signal Process.* **2022**, *179*, 109363. [CrossRef]
34. Rozanec, J.M.; Fortuna, B.; Mladenec, D. Knowledge graph-based rich and confidentiality preserving explainable artificial intelligence (XAI). *Inf. Fusion* **2022**, *81*, 91–102. [CrossRef]
35. Mehta, H.; Passi, K. Social media hate speech detection using explainable artificial intelligence (XAI). *Algorithms* **2022**, *15*, 291. [CrossRef]
36. Lim, S.-Y.; Chae, D.-K.; Lee, S.-C. Detecting deepfake voice using explainable deep learning techniques. *Appl. Sci.* **2022**, *12*, 3926. [CrossRef]
37. Necula, S.-C.; Dumitriu, F.; Greavu-Şerban, V. A Systematic Literature Review on Using Natural Language Processing in Software Requirements Engineering. *Electronics* **2024**, *13*, 2055. [CrossRef]
38. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Moher, D. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef] [PubMed]
39. Gunning, D. *Explainable Artificial Intelligence (XAI)*; DARPA: Arlington, VA, USA, 2017. Available online: <https://onlinelibrary.wiley.com/doi/full/10.1002/aii2.61> (accessed on 27 June 2026).
40. Mosqueira-Rey, E.; Hernández-Pereira, E.; Alonso-Ríos, D.; Bobes-Bascarán, J.; Fernández-Leal, Á. Human-in-the-loop machine learning: A state of the art. *Artif. Intell. Rev.* **2023**, *56*, 3005–3054. [CrossRef]
41. Malandri, L.; Mercorio, F.; Mezzanzanica, M.; Nobani, N. ConvXAI: A System for Multimodal Interaction with Any Black-Box Explainer. *Cogn. Comput.* **2023**, *15*, 613–644. [CrossRef]
42. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105.
43. Chamola, V.; Hassija, V.; Sulthana, A.R.; Ghosh, D.; Dhingra, D.; Sikdar, B. A Review of Trustworthy and Explainable Artificial Intelligence (XAI). *IEEE Access* **2023**, *11*, 78994–79015. [CrossRef]
44. Chazette, L.; Schneider, K. Explainability as a non-functional requirement: Challenges and recommendations. *Requir. Eng.* **2020**, *25*, 493–514. [CrossRef]
45. Cirqueira, D.; Nedbal, D.; Helfert, M.; Bezbradica, M. Scenario-based requirements elicitation for user-centric explainable AI. In *Machine Learning and Knowledge Extraction*; Springer: Cham, Switzerland, 2020; pp. 321–341. [CrossRef]
46. Retzlaff, C.O.; Angerschmid, A.; Saranti, A.; Schneeberger, D.; Roettger, R.; Mueller, H.; Holzinger, A. Post-hoc vs. Ante-hoc Explanations: XAI Design Guidelines for Data Scientists. *Cogn. Syst. Res.* **2024**, *86*, 101243. [CrossRef]

47. Brennan, A. What do people really want when they say they want “explainable AI?” We asked 60 stakeholders. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 1–7. [\[CrossRef\]](#)
48. Ehsan, U.; Tambwekar, P.; Chan, L.; Harrison, B.; Riedl, M.O. Automated rationale generation: A technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*; Association for Computing Machinery: New York, NY, USA, 2019; pp. 263–274. [\[CrossRef\]](#)
49. Kale, A.; Nguyen, T.; Harris, F.C.; Li, C.; Zhang, J.; Ma, X. Provenance documentation to enable explainable and trustworthy AI: A literature review. *Data Intell.* **2023**, *5*, 139–162. [\[CrossRef\]](#)
50. Ehsan, U.; Riedl, M.O. Explainability pitfalls: Beyond dark patterns in explainable AI. *Patterns* **2024**, *5*, 100971. [\[CrossRef\]](#) [\[PubMed\]](#)
51. Ahmed, I.; Jeon, G.; Piccialli, F. From Artificial Intelligence to Explainable Artificial Intelligence in Industry 4.0: A Survey on What, How, and Where. *IEEE Trans. Ind. Inform.* **2022**, *18*, 5031–5042. [\[CrossRef\]](#)
52. Rosenberger, J.; Kuhlemann, S.; Tiefenbeck, V.; Kraus, M.; Zschech, P. The Impact of Transparency in AI Systems on Users’ Data-Sharing Intentions: A Scenario-Based Experiment. *arXiv* **2025**, arXiv:2502.20243.
53. Schrills, T.; Franke, T. Color for characters-effects of visual explanations of AI on trust and observability. In *Artificial Intelligence in HCI*; Springer: Cham, Switzerland, 2020; pp. 121–135. [\[CrossRef\]](#)
54. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Benetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [\[CrossRef\]](#)
55. Ngo, T.; Kunkel, J.; Ziegler, J. Exploring mental models for transparent and controllable recommender systems: A qualitative study. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation, and Personalization*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 183–191. [\[CrossRef\]](#)
56. Oh, C.; Song, J.; Choi, J.; Kim, S.; Lee, S.; Suh, B. I lead, you help, but only with enough details: Understanding user experience of co-creation with artificial intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*; Association for Computing Machinery: New York, NY, USA, 2018; pp. 1–13. [\[CrossRef\]](#)
57. Ali, S.; Abuhmed, T.; El-Sappagh, S.; Muhammad, K.; Alonso-Moral, J.M.; Confalonieri, R.; Guidotti, R.; Del Ser, J.; Díaz-Rodríguez, N.; Herrera, F. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Inf. Fusion* **2023**, *99*, 101805. [\[CrossRef\]](#)
58. Hosain, M.T.; Anik, M.H.; Rafi, S.; Tabassum, R.; Insia, K.; Siddiky, M.M. Path to Gain Functional Transparency in Artificial Intelligence with Meaningful Explainability. *J. Metaverse* **2023**, *3*, 166–180. [\[CrossRef\]](#)
59. Habbal, A.; Ali, M.K.; Abuzaraida, M.A. Artificial Intelligence Trust, Risk and Security Management (AI TRiSM): Frameworks, applications, challenges and future research directions. *Expert Syst. Appl.* **2024**, *240*, 122442. [\[CrossRef\]](#)
60. Ueda, D.; Kakinuma, T.; Fujita, S.; Kamagata, K.; Fushimi, Y.; Ito, R.; Matsui, Y.; Nozaki, T.; Nakaura, T.; Fujima, N.; et al. Fairness of artificial intelligence in healthcare: Review and recommendations. *Jpn. J. Radiol.* **2024**, *42*, 3–15. [\[CrossRef\]](#) [\[PubMed\]](#)
61. Schwalbe, G.; Finzel, B. A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. *Data Min. Knowl. Discov.* **2024**, *38*, 3043–3101. [\[CrossRef\]](#)
62. Notovich, A.; Chaltut-Ben Gal, H.; Ben-Gal, I. Explainable Artificial Intelligence (XAI): Motivation, Terminology, and Taxonomy. In *Machine Learning for Data Science Handbook*; Rokach, L., Maimon, O., Shmueli, E., Eds.; Springer: Cham, Switzerland, 2023. [\[CrossRef\]](#)
63. Graziani, M.; Dutkiewicz, L.; Calvaresi, D.; Amorim, J.P.; Yordanova, K.; Vered, M.; Nair, R.; Abreu, P.H.; Blanke, T.; Pulignano, V.; et al. A global taxonomy of interpretable AI: Unifying the terminology for the technical and social sciences. *Artif. Intell. Rev.* **2023**, *56*, 3473–3504. [\[CrossRef\]](#) [\[PubMed\]](#)
64. Khot, A.; Neubauer, M.S.; Roy, A. A detailed study of interpretability of deep neural network based top taggers. *Mach. Learn. Sci. Technol.* **2023**, *4*, 035003. [\[CrossRef\]](#)
65. Nagy, M.; Molontay, R. Interpretable Dropout Prediction: Towards XAI-Based Personalized Intervention. *Int. J. Artif. Intell. Educ.* **2024**, *34*, 274–300. [\[CrossRef\]](#)
66. Jalali, A.; Haslhofer, B.; Kriglstein, S.; Rauber, A. Predictability and Comprehensibility in Post-Hoc XAI Methods: A User-Centered Analysis. In *Intelligent Computing. SAI 2023*; Arai, K., Ed.; Lecture Notes in Networks and Systems; Springer: Cham, Switzerland, 2023; Volume 711. [\[CrossRef\]](#)
67. Kitano, H. Nobel Turing Challenge: Creating the engine for scientific discovery. *npj Syst. Biol. Appl.* **2021**, *7*, 29. [\[CrossRef\]](#) [\[PubMed\]](#)
68. Krenn, M.; Pollice, R.; Guo, S.; Aldeghi, M.; Cervera-Lierta, A.; Friederich, P.; dos Passos Gomes, G.; Häse, F.; Jinich, A.; Nigam, A.; et al. On scientific understanding with artificial intelligence. *Nat. Rev. Phys.* **2022**, *4*, 761–769. [\[CrossRef\]](#) [\[PubMed\]](#)
69. Yang, W.; Wei, Y.; Wei, H.; Chen, Y.; Huang, G.; Li, X.; Li, R.; Yao, N.; Wang, X.; Gu, X.; et al. Survey on Explainable AI: From Approaches, Limitations and Applications Aspects. *Hum.-Cent. Intell. Syst.* **2023**, *3*, 161–188. [\[CrossRef\]](#)

70. Bibal, A.; Frenay, B. Interpretability of machine learning models and representations: An introduction. In Proceedings of the European Symposium on Artificial Neural Networks, ESANN, Bruges, Belgium, 27–29 April 2016; pp. 77–82.
71. Vilone, G.; Longo, L. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inf. Fusion* **2021**, *76*, 89–106. [\[CrossRef\]](#)
72. Tjoa, E.; Guan, C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 4793–4813. [\[CrossRef\]](#) [\[PubMed\]](#)
73. Bharati, S.; Mondal, M.R.H.; Podder, P.; Kose, U. Explainable Artificial Intelligence (XAI) with IoHT for Smart Healthcare: A Review, dans Interpretable Cognitive Internet of Things for Healthcare, édité par U. In *Interpretable Cognitive Internet of Things for Healthcare*; Kose, U., Gupta, D., Khanna, A., Rodrigues, J.J.P.C., Eds.; Internet of Things; Springer: Cham, Switzerland, 2023. [\[CrossRef\]](#)
74. Dwivedi, R.; Dave, D.; Naik, H.; Singhal, S.; Omer, R.; Patel, P.; Qian, B.; Wen, Z.; Shah, T.; Morgan, G.; et al. Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Comput. Surv.* **2023**, *55*, 194. [\[CrossRef\]](#)
75. Pradhan, B.; Dikshit, A.; Lee, S.; Kim, H. An explainable AI (XAI) model for landslide susceptibility modeling. *Appl. Soft Comput.* Vol. **2023**, *142*, 110324. [\[CrossRef\]](#)
76. Bourgaïs, M.; Giustozzi, F.; Vercouter, L.; Zanni-Merk, C. Towards the use of post-hoc explainable methods to define and detect semantic situations of importance in medical data. *Procedia Comput. Sci.* **2023**, *225*, 2312–2321. [\[CrossRef\]](#)
77. Mavrepis, P.; Makridis, G.; Fatouros, G.; Koukos, V.; Separdani, M.M.; Kyriazis, D. XAI for All: Can Large Language Models Simplify Explainable AI? *arXiv* **2024**, arXiv:2401.13110.
78. Love, P.E.D.; Fang, W.; Matthews, J.; Porter, S.; Luo, H.; Ding, L. Explainable artificial intelligence (XAI): Precepts, models, and opportunities for research in construction. *Adv. Eng. Inform.* **2023**, *57*, 102024. [\[CrossRef\]](#)
79. Goodfellow, I.J.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016. Available online: <http://www.deeplearningbook.org> (accessed on 27 June 2026).
80. Liu, N.; Du, M.; Hu, X. Representation interpretation with spatial encoding and multimodal analytics. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*; Association for Computing Machinery: New York, NY, USA, 2019.
81. Du, M.; Liu, N.; Hu, X. Techniques for interpretable machine learning. *Commun. ACM* **2019**, *63*, 68–77. [\[CrossRef\]](#)
82. Ozbakir, L.; Baykasoglu, A.; Kulluk, S. A soft computing-based approach for integrated training and rule extraction from artificial neural networks: Difaconn-miner. *Appl. Soft Comput.* **2010**, *10*, 304–317. [\[CrossRef\]](#)
83. Sato, M.; Tsukimoto, H. Rule extraction from neural networks via decision tree induction. In *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222)*; IEEE: Piscataway, NJ, USA, 2001; Volume 3, pp. 1870–1875.
84. Fu, L. Rule generation from neural networks. *IEEE Trans. Syst. Man Cybern.* **1994**, *24*, 1114–1124. [\[CrossRef\]](#)
85. Zilke, J.-R.; Mencia, E.L.; Janssen, F. Deepred—Rule extraction from deep neural networks. In *Discovery Science*; Springer International Publishing: Cham, Switzerland, 2016; pp. 457–473.
86. López-Martínez, D.; Fernández, M.; González Briones, A.; Chamoso, P.; Corchado, E. A Brief Review of Explainable Artificial Intelligence (XAI) Techniques. In *Trends in Sustainable Smart Cities and Territories*; Lecture Notes in Computer Science (LNCS); Springer: Cham, Switzerland, 2023; pp. 442–452. [\[CrossRef\]](#)
87. Jacobsson, H. Rule Extraction from Recurrent Neural Networks: A Taxonomy and Review. *Neural Comput.* **2005**, *17*, 1223–1263. [\[CrossRef\]](#)
88. Augasta, M.G.; Kathirvalavakumar, T. Reverse Engineering the Neural Networks for Rule Extraction in Classification Problems. *Neural Process. Lett.* **2012**, *35*, 131–150. [\[CrossRef\]](#)
89. Hruschka, E.R.; Ebecken, N.F. Extracting rules from multilayer perceptrons in classification problems: A clustering-based approach. *Neurocomputing* **2006**, *70*, 384–397. [\[CrossRef\]](#)
90. Kahramanli, H.; Allahverdi, N. Rule extraction from trained adaptive neural networks using artificial immune systems. *Expert Syst. Appl.* **2009**, *36*, 1513–1522. [\[CrossRef\]](#)
91. Aickelin, U.; Dasgupta, D.; Gu, F. Artificial Immune Systems. In *Search Methodologies*; Springer: Boston, MA, USA, 2014; pp. 187–212. [\[CrossRef\]](#)
92. Che, Z.; Purushotham, S.; Khemani, R.; Liu, Y. Interpretable deep models for ICU outcome prediction. *AMIA Annu. Symp. Proc.* **2017**, *2016*, 371–380 [\[PubMed\]](#)
93. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
94. Bucila, C.; Caruana, R.; Niculescu-Mizil, A. Model Compress. In *KDD '06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, USA, 2006; Volume 2006, pp. 535–541. [\[CrossRef\]](#)
95. Micaelli, P.; Storkey, A. Zero-shot Knowledge Transfer via Adversarial Belief Matching. *arXiv* **2019**, arXiv:1905.09768.
96. Kindermans, P.J.; Schütt, K.T.; Alber, M.; Müller, K.-R.; Erhan, D.; Kim, B.; Dähne, S. Learning how to explain neural networks: PatternNet and PatternAttribution. *arXiv* **2017**, arXiv:1705.05598.

97. Montavon, G.; Lapuschkin, S.; Binder, A.; Samek, W.; Müller, K.-R. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognit.* **2017**, *65*, 211–222. [[CrossRef](#)]
98. Shrikumar, A.; Greenside, P.; Kundaje, A. Learning Important Features Through Propagating Activation Differences. *arXiv* **2019**, arXiv:1704.02685.
99. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. *arXiv* **2017**, arXiv:1703.01365.
100. Adamson, G. Ethics and the Explainable Artificial Intelligence (XAI) Movement. *TechRxiv* **2022**. [[CrossRef](#)]
101. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **2021**, *23*, 18. [[CrossRef](#)] [[PubMed](#)]
102. Khan, A.; Sohail, A.; Zahoor, U.; Qureshi, A.S. A survey of the recent architectures of deep convolutional neural networks. *Artif. Intell. Rev.* **2020**, *53*, 5455–5516. [[CrossRef](#)]
103. Bhakte, A.; Pakkiriswamy, V.; Srinivasan, R. An explainable artificial intelligence based approach for interpretation of fault classification results from deep neural networks. *Chem. Eng. Sci.* **2022**, *250*, 117373. [[CrossRef](#)]
104. Wang, Z.J.; Turko, R.; Shaikh, O.; Park, H.; Das, N.; Hohman, F.; Kahng, M.; Chau, D.H. CNN Explainer: Learning Convolutional Neural Networks with Interactive Visualization. *IEEE Trans. Vis. Comput. Graph.* **2021**, *27*, 1396–1406. [[CrossRef](#)] [[PubMed](#)]
105. Sun, Y.; Ravi, S.; Singh, V. Adaptive activation thresholding: Dynamic routing type behavior for interpretability in convolutional neural networks. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2019; pp. 4937–4946.
106. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; Tang, X. Residual attention network for image classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2017; pp. 6450–6458.
107. Springenberg, J.T.; Dosovitskiy, A.; Brox, T.; Riedmiller, M. Striving for Simplicity: The All Convolutional Net. 2015. Available online: <http://lmb.informatik.uni-freiburg.de/Publications/2015/DB15a> (accessed on 27 June 2026).
108. Lin, M.; Chen, Q.; Yan, S. Network in Network. *arXiv* **2014**, arXiv:1312.4400.
109. Zhang, Q.; Wang, X.; Wu, Y.N.; Zhou, H.; Zhu, S.-C. Interpretable CNNs for object classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3416–3431. [[CrossRef](#)]
110. Zhang, Q.; Yang, Y.; Ma, H.; Wu, Y.N. Interpreting CNNs via decision trees. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2019; pp. 6254–6263.
111. Papernot, N.; McDaniel, P. Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning. *arXiv* **2018**, arXiv:1803.04765.
112. Ribeiro, M.T.; Singh, S.; Guestrin, C. “Why should I trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 1135–1144.
113. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*; Curran Associates, Inc.: Red Hook, NY, USA, 2017.
114. Wang, Y.; Su, H.; Zhang, B.; Hu, X. Interpret neural networks by extracting critical subnetworks. *IEEE Trans. Image Process.* **2020**, *29*, 6707–6720. [[CrossRef](#)]
115. Alvarez-Melis, A.; Jaakkola, T. Towards Robust Interpretability with Self-Explaining Neural Networks. *arXiv* **2017**, arXiv:1703.04730.
116. Ventura, F.; Cerquitelli, T. What’s in the box? Explaining the black-box model through an evaluation of its interpretable features. *arXiv* **2019**, arXiv:1908.04348.
117. Tamajka, M.; Benesova, W.; Kompanek, M. Transforming convolutional neural network to an interpretable classifier. In *Proceedings of the 2019 International Conference on Systems, Signals and Image Processing (IWSSIP)*; IEEE: Piscataway, NJ, USA, 2019; pp. 255–259.
118. Nguyen, A.; Dosovitskiy, A.; Yosinski, J.; Brox, T.; Clune, J. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *arXiv* **2016**, arXiv:1605.09304.
119. Schlegel, U.; Arnout, H.; El-Assady, M.; Oelke, D.; Keim, D.A. Towards a Rigorous Evaluation of XAI Methods on Time Series. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, Seoul, Republic of Korea, 27–28 October 2019; pp. 4197–4201. [[CrossRef](#)]
120. Islam, N.U.; Lee, S. Interpretation of deep CNN based on learning feature reconstruction with feedback weights. *IEEE Access* **2019**, *7*, 25195–25208. [[CrossRef](#)]
121. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv* **2014**, arXiv:1312.6034.
122. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2016; pp. 2921–2929.
123. Wang, H.; Wang, Z.; Du, M.; Yang, F.; Zhang, Z.; Ding, S.; Mardziel, P.; Hu, X. Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*; IEEE: Piscataway, NJ, USA, 2020; pp. 111–119. [[CrossRef](#)]

124. Brahimi, M.; Mahmoudi, S.; Boukhalfa, K.; Moussaoui, A. Deep interpretable architecture for plant diseases classification. In *2019 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*; IEEE: Piscataway, NJ, USA, 2019; pp. 111–116.
125. Díaz-Rodríguez, N.; Pisoni, G. Accessible Cultural Heritage through Explainable Artificial Intelligence. In *Proceedings of the Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20 Adjunct)*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 317–324. [\[CrossRef\]](#)
126. Pisoni, G.; Diaz-Rodriguez, N.; Gijlers, H.; Tonolli, L. Human-centred artificial intelligence for designing accessible cultural heritage. *Appl. Sci.* **2021**, *11*, 870. [\[CrossRef\]](#)
127. Ignatiev, A. Towards trustable explainable AI. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*; Association for the Advancement of Artificial Intelligence (AAAI): Marina del Rey, CA, USA, 2020; pp. 5154–5158.
128. Stock, P.; Cisse, M. Convnets and Imagenet Beyond Accuracy: Understanding Mistakes and Uncovering Biases. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 498–512. [\[CrossRef\]](#)
129. Laugel, T.; Lesot, M.-J.; Marsala, C.; Renard, X.; Detyniecki, M. The dangers of post-hoc interpretability: Unjustified counterfactual explanations. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI*, Macao, China, 10–16 August 2019; pp. 2801–2807.
130. Huysmans, J.; Dejaeger, K.; Mues, C.; Vanthienen, J.; Baesens, B. An Empirical Evaluation of the Comprehensibility of Decision Table, Tree and Rule Based Predictive Models. *Decis. Support Syst.* **2011**, *51*, 141–154. [\[CrossRef\]](#)
131. Slack, D.; Hilgard, S.; Jia, E.; Singh, S.; Lakkaraju, H. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*; Association for Computing Machinery: New York, NY, USA, 2020.
132. Dam, H.K.; Tran, T.; Ghose, A. Explainable software analytics. In *Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results*; Association for Computing Machinery: New York, NY, USA, 2018; pp. 53–56.
133. Freitas, A.A. Comprehensible classification models: A position paper. *SIGKDD Explor. Newsl.* **2014**, *15*, 1–10. [\[CrossRef\]](#)
134. Preece, A. Asking “why” in AI: Explainability of intelligent systems—perspectives and challenges. *Intell. Syst. Account. Financ. Manag.* **2018**, *25*, 63–72. [\[CrossRef\]](#)
135. Houzé, E. A generic and adaptive approach to explainable AI in autonomic systems: The case of the smart home. In *Ubiquitous Computing*; Institut Polytechnique de Paris: Palaiseau, France, 2022.
136. Robnik-Sikonja, M.; Kononenko, I. Explaining classifications for individual instances. *IEEE Trans. Knowl. Data Eng.* **2008**, *20*, 589–600. [\[CrossRef\]](#)
137. Robnik-Sikonja, M. Explanation of prediction models with explain prediction. *Informatica* **2018**, *42*, 13–22.
138. Zhang, Q.S.; Zhu, S.-C. Visual interpretability for deep learning: A survey. *Front. Inf. Technol. Electron. Eng.* **2018**, *19*, 27–39. [\[CrossRef\]](#)
139. Yeh, C.-K.; Hsieh, C.-Y.; Suggala, A.; Inouye, D.I.; Ravikumar, P.K. On the (in)fidelity and sensitivity of explanations. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2019; pp. 10965–10976.
140. Barratt, S. InterpNET: Neural Introspection for Interpretable Deep Learning. *arXiv* **2017**, arXiv:1710.09511.
141. Rong, Y.; Leemann, T.; Nguyen, T.T.; Fiedler, L.; Qian, P.; Unhelkar, V.; Seidel, T.; Kasneci, G.; Kasneci, E. Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 2104–2122. [\[CrossRef\]](#) [\[PubMed\]](#)
142. Sanneman, L.; Shah, J.A. A situation awareness-based framework for design and evaluation of explainable AI. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*; Springer: Cham, Switzerland, 2020; pp. 94–110.
143. Kim, B.; Shah, J.A.; Doshi-Velez, F. Mind the gap: A generative approach to interpretable feature selection and extraction. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation, Inc.: Montréal, QC, Canada, 2015; pp. 2260–2268.
144. Putnam, V.; Conati, C. Exploring the need for explainable artificial intelligence (XAI) in intelligent tutoring systems (ITS). In *Proceedings of the ACM IUI 2019 Workshops Co-Located with the 24th ACM Conference on Intelligent User Interfaces*, Los Angeles, CA, USA, 16–20 March 2019; Volume 2327.
145. Weitz, K.; Schiller, D.; Schlagowski, R.; Huber, T.; Andre, E. “Let me explain!”: Exploring the potential of virtual agents in explainable AI interaction design. *J. Multimodal User Interfaces* **2021**, *15*, 87–98.
146. Kim, B.; Rudin, C.; Shah, J.A. The Bayesian case model: A generative approach for case-based reasoning and prototype classification. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation, Inc.: Montréal, QC, Canada, 2014; pp. 1952–1960.
147. Allahyari, H.; Lavesson, N. User-Oriented Assessment of Classification Model Understandability. In *11th Scandinavian Conference on Artificial Intelligence*; IOS Press: Trondheim, Norway, 2011; pp. 11–19. [\[CrossRef\]](#)
148. Erhan, D.; Bengio, Y.; Courville, A.; Vincent, P. Visualizing higher-layer features of a deep network. *Univ. Montr.* **2009**, *1341*, 1. [\[CrossRef\]](#)

149. Bau, D.; Zhou, B.; Khosla, A.; Oliva, A.; Torralba, A. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2017; pp. 6541–6549.
150. Ribeiro, M.T.; Singh, S.; Guestrin, C. Anchors: High-Precision Model-Agnostic Explanations. *Proc. Aaai Conf. Artif. Intell.* **2018**, *32*, 1527–1535. [[CrossRef](#)]
151. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **2019**, *267*, 1–38. [[CrossRef](#)]
152. Pearl, J. Theoretical Impediments to Machine Learning with Seven Sparks from the Causal Revolution. *arXiv* **2018**, arXiv:1801.04016.
153. Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; Zemel, R. Fairness through awareness. In *ITCS '12: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*; Association for Computing Machinery: New York, NY, USA, 2012; pp. 214–226. [[CrossRef](#)]
154. Kusner, J.R.; Loftus, A.; Luntama, C.R.H.G.; Weinberger, K.D. Counterfactual Fairness. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*; Curran Associates, Inc.: Red Hook, NY, USA, 2017. Available online: <https://arxiv.org/abs/1703.06856> (accessed on 27 June 2026).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.