

Medical Image Disease Diagnosis: A Critical Review of Methods, Evaluation, and Open Challenges

¹Kaynaat Hussain, ²Priya Sen, ³Kunal Batra

^{1,2,3} Computer Science & Engineering, Institute of Engineering & Technology, Sage University Indore

¹Hussainkaynaat@gmail.com, ²Priya.sen@sageuniversity.in, ³Kunal.batra@sageuniversity.in

Abstract—Deep neural networks now achieve strong discriminative performance across many medical imaging tasks, yet their internal decision processes remain opaque, which limits trust, accountability, and safe adoption in diagnostic settings. This review examines post-hoc Explainable Artificial Intelligence (XAI) as a route to inspecting trained image classifiers without altering them. Drawing on a curated pool of studies on medical image classification, transfer learning, and explainability, we organise the field along four axes: the ante-hoc/post-hoc distinction, the major families of post-hoc explanation (feature attribution, LIME, SHAP, Grad-CAM, concept-based, example-based, counterfactual, rule-based, and occlusion/sensitivity methods), the convolutional architectures that most often serve as the models being explained, and the criteria used to judge explanation quality. We then analyse recurring clinical and methodological challenges, including the gap between visual plausibility and faithfulness, the scarcity of standardised saliency evaluation, and the frequent separation of accuracy benchmarking from explanation analysis. Finally, we consolidate these observations into a set of research gaps and a prioritised agenda for future work. The review is intended as the conceptual foundation for a companion implementation study that benchmarks a baseline CNN against three transfer-learning models and defines a planned post-hoc XAI extension.

Keywords—Explainable Artificial Intelligence, Post-Hoc Explanation, Medical Image Classification, Deep Learning, Transfer Learning, Grad-CAM, Saliency Evaluation, Interpretability

I. INTRODUCTION

Convolutional and residual architectures have become the default tools for automated analysis of radiographs, histopathology, fundus photographs, and magnetic resonance images [1]–[3]. Their value in clinical research is evident, but their decision logic is distributed across millions of parameters and is not directly legible to the clinicians who would be asked to act on their outputs. Reviews of AI in radiology and gastroenterology repeatedly identify this opacity as a barrier to trust and responsible use [4], [5], and surveys of XAI argue that explanation is a prerequisite rather than an optional extra when decisions carry consequences for patients [6], [7].

Two broad responses to opacity exist. One is to design models whose reasoning is transparent by construction; the other is to explain an already trained model after the fact. Rudin has argued forcefully that, for high-stakes decisions, post-hoc explanations of black-box models can be misleading and that inherently interpretable models should be preferred [8]. Others note that, in practice, high-capacity deep networks remain the dominant choice for image data and that well-designed post-hoc analysis is therefore indispensable [1], [9]. This review takes the second position as its subject while retaining the first as a critical lens: we ask not only *which* post-hoc techniques exist, but *what each can and cannot legitimately claim* in a medical imaging context.

Scope and method. The review is built on a curated set of studies supplied by the authors, covering medical image classification, transfer learning, feature engineering, and a range of explanation techniques. It is therefore a critical, thematic review rather than a systematic review; no database search protocol, inclusion criteria, or PRISMA-style screening is claimed. Where a source is used, only its stated scope is summarised, and quantitative results from individual studies are deliberately not tabulated unless they are needed for an argument.

Contributions. This article (i) clarifies the vocabulary of interpretability, explainability, faithfulness, localisation, and clinical usefulness; (ii) compares major post-hoc families in terms of scope, cost, and medical applicability; (iii) positions common backbone architectures as the objects of explanation; (iv) examines how explanation quality is, and should be, evaluated; and (v) derives research gaps that motivate the companion implementation study.

II. DEEP LEARNING IN MEDICAL IMAGE ANALYSIS

Medical image classification has progressed from hand-crafted descriptors toward learned representations. Earlier pipelines combined statistical or wavelet features with conventional classifiers for MR image classification [10], and later work fused deep features with statistical descriptors through a multilayer perceptron [11]. More recent studies continue this hybrid tradition, for example by fusing deep spatial and statistical features for MRI brain tumour classification [12]. Convolutional networks have also been validated extensively for glaucoma assessment from fundus images [2], for colon and lung cancer classification from histopathology [3], and for COVID-19 detection from chest radiographs [13], [14].

The data underlying such work are frequently public. The approved pool for this review references, among others, a lung and colon histopathology collection [15], a chest X-ray radiography database [16], and Kaggle-hosted collections for brain MRI, brain tumour types, breast ultrasound, and glaucoma-related eye images [17]–[20]. Public availability is a benefit for reproducibility, but it also means that class definitions, acquisition protocols, and annotation quality vary widely and are often only partially documented; one of the gaps discussed in Section XII.

Beyond classification, image quality can be addressed upstream. Cross-domain super-resolution for medical images illustrates how preprocessing choices may alter the information available to a downstream classifier [21], while augmentation libraries provide a common mechanism for enlarging limited training sets [22]. Classification has likewise been explored for other conditions and modalities, including breast cancer prediction [23], skin cancer and COVID-19 detection through stacked ensembles [24], and diagnosis of Down syndrome from facial images using transfer-learned deep features [25]. Non-imaging modalities have been explored as well; AI-based detection of COVID-19 from speech

and voice signals [26], [27] demonstrates that the explainability concerns discussed here extend beyond images. Taken together, these studies show that high headline performance is attainable across diverse tasks; they say much less about *why* a model reaches a given decision.

III. EXPLAINABLE AI IN MEDICAL IMAGING

A. Terminology

The literature uses *interpretability* and *explainability* inconsistently. Following the supervised-learning surveys in the approved pool [6], [28], [29], we use *interpretability* for the degree to which a human can understand the mechanism of a model, and *explainability* for the provision of additional information, usually after training, that helps a human understand a particular output. Three further notions are needed for medical imaging and are often conflated:

- **Faithfulness** — whether an explanation reflects the evidence the model actually used, rather than a plausible-looking story.
- **Localisation** — whether highlighted regions coincide with clinically meaningful structures; a model may localise well yet be unfaithful, or be faithful yet attend to artefacts.
- **Clinical usefulness** — whether an explanation improves a clinician's decision, calibration of trust, or error detection; this is an empirical question about human–AI interaction, not a property of the algorithm.

Performance, explainability, and clinical usefulness are therefore distinct quantities, and strength in one does not imply strength in another. Healthcare-focused reviews emphasise that interpretability requirements depend on the audience and task [30], and a multi-stage view of XAI, covering where in the pipeline explanations are produced and in which output shape, is provided by Ullah *et al.* [7].

B. Why Explanation Matters for Deep Image Classifiers

Four motivations recur. First, *error detection*: a classifier can be accurate on a benchmark while relying on spurious cues, and explanations offer a means of inspecting this [1], [4]. Second, *trust calibration*: clinicians need to know when to rely on, and when to override, an automated suggestion [5]. Third, *accountability and oversight*: explanations support auditing and documentation of model behaviour [6], [7]. Fourth, *model development*: explanations help developers diagnose dataset shift, label noise, and shortcut learning. Van der Velden *et al.* survey how XAI has been applied to deep learning-based medical image analysis and caution that the field lacks consensus on how explanations should be validated [1].

IV. ANTE-HOC AND POST-HOC EXPLAINABILITY

Ante-hoc approaches build transparency into the model. Classical examples include decision trees [31], [32], logistic regression [33], and rule sets or rule ensembles [34], [35]. For deep image models, concept-bottleneck-style designs seek the same property by routing predictions through human-defined concepts [36]. Post-hoc approaches instead analyse a trained model through attributions, perturbations, surrogates, or retrieved examples. Retzlaff *et al.*

discuss design guidelines for choosing between the two paradigms and stress that the choice depends on task, data, and stakeholder needs [9].

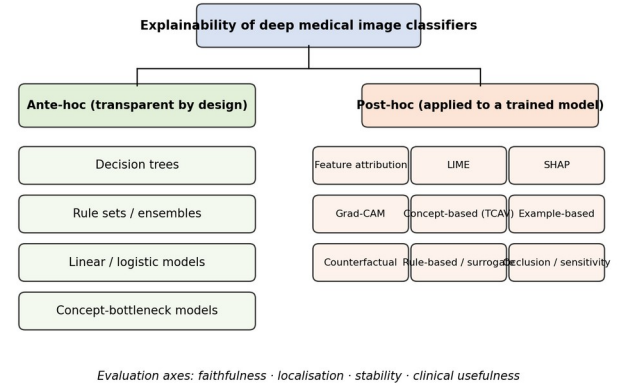


Fig. 1. Conceptual taxonomy of explainability approaches considered in this review. The diagram is a schematic organisation of the literature and contains no experimental data.

TABLE 1

Comparison of Ante-Hoc and Post-Hoc Explainability

Aspect	Ante-hoc	Post-hoc
Source of transparency	Built into model structure [8], [36]	Applied after training to a fixed model [9]
Typical models	Trees, rule sets, linear models, concept-based networks [31], [33], [34]	Any trained model, including deep CNNs [37], [38]
Fidelity to model	By construction	Must be verified; approximations may diverge [8]
Effect on accuracy	May trade accuracy for transparency on complex images	None on the trained model
Reuse of existing models	Requires retraining	Directly applicable
Main risk	Limited capacity on raw images	Plausible but unfaithful explanations

The two paradigms are not mutually exclusive. Rule extraction from trained networks [39] and ontology-enhanced global explanations [40] are hybrid strategies that derive structured, ante-hoc-like representations from black-box models. This review concentrates on post-hoc methods because they are the pragmatic entry point for the high-performing pre-trained CNN backbones in common medical use, but their limitations (Section XI) should be stated alongside their benefits.

V. MAJOR POST-HOC XAI METHODS

Table 2 compares the families discussed below. Cost descriptors are qualitative and relate to the computational pattern of each method, not to measured run-times; no timing figures are reported in the source material.

TABLE 2
Qualitative Comparison of Major Post-Hoc XAI Families for Medical Imaging

Method family	Scope	Output	Strengths	Limitations	Relative cost (qualitative)	Medical-imaging applicability
Feature attribution [41], [42]	Local / global	Per-feature importance	Unifying view of removal-based explanation; ranks inputs	Definition of “removal” affects results	Varies with variant	Applicable to pixels, patches, or learned features
LIME [37]	Local	Surrogate-based region importance	Model-agnostic; intuitive	Depends on perturbation and segmentation choices; instability	Many model evaluations per image	Usable on any classifier; stability needs checking
SHAP [41], [43]	Local / global	Additive feature contributions	Principled attribution framework	Costly for high-dimensional inputs; approximation choices	Often higher than gradient methods	Possible with pixel groupings; interpretation requires care
Grad-CAM [38]	Local	Coarse class-discriminative heat map	Simple; needs one gradient computation; no retraining	Low resolution; sensitive to chosen layer	Low	Widely applicable to CNN classifiers
Concept-based [36], [44]	Local / global	Concept importance	Closer to clinical vocabulary	Needs concept examples or labels	Moderate	Promising where clinical concepts are defined
Example-based [45]–[47]	Local / global	Similar cases or prototypes	Familiar case-based reasoning	Depends on feature-space quality	Moderate (retrieval)	Natural for case-based clinical review
Counterfactual [48]	Local	“What would change the output”	Actionable contrast	Realism of generated images; validity	Often optimisation-heavy	Useful in principle; clinical plausibility is a concern
Rule-based [14], [34], [35], [39]	Global / local	If–then rules	Human-readable structure	Fidelity of extracted rules; scalability	Varies	Better suited to tabular or derived features
Occlusion / sensitivity [41], [49]	Local	Importance map from masked regions	Easy to implement; model-agnostic	Depends on occluder size and baseline; many forward passes	Moderate	Applied to lung disease classification [49]

A. Feature Attribution

Feature-attribution methods assign an importance score to each input element. Covert *et al.* show that many seemingly different explanation methods can be viewed as removing features and measuring the effect on the model output, which clarifies why methods disagree when they define removal differently [41]. Wojtas and Chen address feature importance ranking specifically for deep networks [42]. Classical feature-selection literature, covering filter and mutual-information criteria, is relevant mainly as background for tabular or hand-crafted features [50]–[53]; in image settings, the analogue is deciding at what granularity (pixel, superpixel, patch, or deep feature) importance should be assigned.

B. LIME

LIME explains an individual prediction by fitting a simple, interpretable surrogate model to the black-box behaviour in a neighbourhood of the instance [37]. For images, this typically means perturbing groups of pixels and reading off which groups most influence the class score. The attraction is model-agnosticism; the weakness is that the surrogate depends on how the neighbourhood and image segmentation are defined, so repeated runs or alternative settings can yield different explanations. In medical imaging, segmentation into superpixels may cut across anatomical boundaries, a point that merits explicit testing.

C. SHAP

SHAP provides a unified framework in which a prediction is decomposed into additive feature contributions [43]. Covert *et al.* position SHAP-style attribution within the broader family of removal-based methods [41]. SHAP offers appealing theoretical properties and supports both local and aggregated global views, but

exact computation is expensive for image inputs, so practical image applications rely on approximations whose fidelity should be reported.

D. Grad-CAM

Grad-CAM uses gradients of a class score with respect to the final convolutional feature maps to produce a coarse localisation heat map [38]. It requires no architectural change or retraining, which makes it the most commonly applied post-hoc technique for CNN classifiers. Its limitations are well recognised: heat maps are low-resolution, depend on the chosen layer, and show where the network is sensitive rather than why a decision was made. For lightweight networks with small final feature maps, spatial coarseness may be more pronounced, an effect that should be examined empirically rather than assumed.

E. Concept-Based Approaches

TCAV tests how sensitive a prediction is to a human-defined concept using concept activation vectors, moving explanation from pixels to concepts [44]. Sarkar *et al.* integrate concepts into the learning of ante-hoc explainable models [36]. Concept-based methods are attractive in medicine because clinicians reason in terms of findings (for instance, a lesion margin or an opacity pattern) rather than pixels, but they require curated concept examples, which are costly to obtain and may themselves be subjective.

F. Example-Based Approaches

Example-based explanations justify a prediction by retrieving or constructing representative cases. Deep k-nearest neighbours relates a prediction to training examples in the network’s internal

representation, with the aim of supporting confidence estimation and interpretability [45]. A survey of natural example-based explainability organises these methods [46], and prototype-based explanations have been proposed to aid understanding of datasets [47]. In clinical practice, case-based reasoning is familiar, but retrieved examples are only as informative as the feature space from which they are drawn and may expose patient data if not handled carefully.

G. Counterfactual Explanations

Counterfactual methods answer “what minimal change would have altered the prediction?”. Verma *et al.* review the area and its desiderata [48]. For images, producing a counterfactual that is both effective and anatomically realistic is difficult, and an unrealistic counterfactual can mislead more than it informs. Clinical validation by domain experts is therefore essential.

H. Rule-Based Explanations

Rule-based explanations express model behaviour as human-readable conditions. Bayesian rule-set learning and rule ensembles provide interpretable classifiers [34], [35], and the interpretability of rule systems has been the subject of dedicated measurement work [54]. Rules have been extracted from trained neural networks as decision trees [39], applied through an evolutionary rule-based approach to COVID-19 chest X-ray classification [14], and used to add explainability to time-series forecasting through association rules [55]. Rule-based approaches are most natural when inputs are interpretable features; for raw pixels they typically operate on derived descriptors.

I. Occlusion and Sensitivity-Based Explanations

Occlusion analysis masks image regions and observes the change in the output, giving a direct, model-agnostic importance estimate. It has been used alongside a multi-dataset deep network for multi-class lung disease classification [49], and it fits the removal-based view of Covert *et al.* [41]. Its main parameters, namely occluder size, stride, and replacement value, can strongly influence the resulting map and should be reported. Because it relies only on forward passes, it avoids gradient-related artefacts, at the cost of many model evaluations.

Explanation research outside imaging offers transferable ideas, for example explainable hybrid deep learning for evapotranspiration forecasting [56] and globally interpretable graph learning via distribution matching [57]. These are not medical imaging studies and are cited only to indicate methodological breadth.

VI. ARCHITECTURES IN MEDICAL IMAGE CLASSIFICATION

Explanations are produced *for* a model, so architectural properties determine which explanation methods are feasible and how informative they are.

A. Plain CNNs. A shallow convolutional network trained from scratch is the usual baseline in medical imaging studies. Its small size and simple structure make it easy to explain with gradient-based methods, but its capacity is limited, and it must learn visual features from comparatively small medical datasets [2], [11].

B. ResNet. Residual networks introduce identity shortcut connections that ease optimisation of deep models [58]. Their widespread use as backbones makes ResNet a common subject of

Grad-CAM-style analysis, although the interaction of skip connections with attribution is not trivial.

C. DenseNet. Densely connected networks concatenate feature maps from earlier layers, encouraging feature reuse [59]. This connectivity enriches the representation but complicates the choice of the layer used for localisation methods.

D. MobileNet-family networks. Lightweight architectures designed for efficient computation are attractive for resource-limited deployment settings. The original MobileNetV2 reference is not part of the approved reference pool and should be added by the authors [REF. TO BE ADDED]. The companion study evaluates MobileNetV2 and examines its accuracy–size trade-off.

E. EfficientNet-family networks. EfficientNet variants have been used in medical transfer learning; one study fine-tunes EfficientNet-B7 for multiclass brain tumour classification [60]. The original EfficientNet reference is likewise outside the approved pool [REF. TO BE ADDED]. The companion study evaluates the smaller EfficientNet-B0.

VII. TRANSFER LEARNING IN MEDICAL IMAGE DIAGNOSIS

Because labelled medical images are scarce relative to natural-image corpora, networks pre-trained on large generic datasets are routinely adapted. Evidence in the pool includes fine-tuned EfficientNet-B7 for brain tumours [60], transfer-learned deep features for diagnosing Down syndrome from facial images [25], comparative analysis of deep features for COVID-19 chest radiographs [13], and a stacked ensemble for skin cancer and COVID-19 [24]. Feature-level fusion of deep and statistical descriptors is another route [11], [12].

Transfer learning raises explanation-specific questions that are rarely addressed. Do pre-trained backbones attend to the same image regions after fine-tuning as they did originally? Does fine-tuning depth change how faithful gradient-based maps are? And does higher accuracy reflect better use of diagnostic evidence, or better exploitation of dataset-specific shortcuts? These questions remain open and constitute a central motivation for pairing performance benchmarking with explanation analysis.

VIII. LOCAL AND GLOBAL EXPLANATIONS

Table 3 separates explanations of a single prediction (local) from explanations of overall model behaviour (global). For clinical use the distinction matters: local explanations support case review, whereas global explanations support validation, auditing, and regulatory documentation.

TABLE 3

Local versus Global Explanations in Medical Image Classification

Aspect	Local	Global
Question answered	Why this prediction for this image?	What does the model rely on in general?
Representative methods	Grad-CAM, LIME, occlusion, counterfactuals [37], [38], [48], [49]	Concept sensitivity, rule or tree extraction, ontology-based explanations [39], [40], [44]
Clinical use	Case review, error inspection	Model validation, auditing, documentation [6]
Main limitation	May not generalise beyond the case	May hide case-level exceptions; fidelity of summary [57]

IX. EVALUATING EXPLANATIONS AND SALIENCY MAPS

A visually appealing heat map is not evidence of a good explanation. Gomez *et al.* review metrics for evaluating saliency maps and show that the field relies on several families of criteria, whose agreement is not guaranteed [61]. Drawing on that and on the broader surveys [1], [7], [30], we distinguish four evaluation categories relevant to medical imaging:

- **Faithfulness tests**, which perturb or remove highlighted evidence and measure the change in output, closely related to the removal view of explanation [41], [61].
- **Localisation assessment**, which compares explanations with expert-annotated regions where such annotations exist [4], [61].

- **Stability and consistency**, which examine whether explanations change under small input or model perturbations [61].
- **Human-centred evaluation**, which tests whether explanations improve clinician decisions or trust [5], [30].

For rule-based and other symbolic explanations, dedicated interpretability measures have been proposed [54]. A recurrent weakness in applied studies is that explanation quality is assessed only qualitatively, by presenting a few selected examples. Such cherry-picked illustrations cannot support claims about reliability and should be complemented by quantitative, dataset-level analysis.

X. SUMMARY OF THE SUPPLIED LITERATURE

Table 4 consolidates the supplied studies by focus. Quantitative results from these sources were not extracted and are therefore not tabulated; the table characterises scope and relevance only.

TABLE 4

Summary of Selected Supplied Literature by Focus and Relevance to Post-Hoc XAI in Medical Imaging

Ref.	Domain / modality	Approach	Relevance to this review
[1]	Medical image analysis (survey)	Survey of XAI in deep learning-based analysis	Overview of explanation use and gaps
[4]	Radiology (position/review)	Interpretability challenges and opportunities	Clinical requirements for interpretability
[7]	General XAI (survey)	Importance, domains, stages, output shapes, challenges	Taxonomy of XAI
[49]	Lung disease images	Multi-dataset classification with occlusion sensitivity	Example of post-hoc occlusion analysis
[3]	Colon/lung histopathology	Deep learning classification	Medical CNN application
[14]	COVID-19 chest X-ray	Interpretable evolutionary rule-based approach	Rule-based explanation in imaging
[60]	Brain MRI	Fine-tuned EfficientNet-B7 transfer learning	Transfer-learning backbone use
[12]	Brain MRI	Deep spatial and statistical feature fusion	Hybrid features
[2]	Glaucoma, fundus images	CNN validation across datasets	CNN-based diagnosis
[24]	Skin cancer, COVID-19	Stacked ensemble comparison	Ensemble alternative to single backbones
[25]	Facial images (Down syndrome)	Transfer-learned deep features	Transfer learning beyond radiology
[61]	Saliency evaluation	Metrics for explanation methods	Evaluation of heat maps
[36], [44]	General (vision)	Concept-based testing and ante-hoc concept models	Concept-level explanation
[45], [46]	General	Deep kNN; survey of example-based explanation	Case-based explanation
[8], [9]	General	Argument for interpretable models; ante-/post-hoc design guidance	Critical perspective on post-hoc XAI

XI. CHALLENGES OF XAI IN CLINICAL IMAGING

Faithfulness versus plausibility. Explanations that look reasonable to humans may not reflect the model's true reasoning, which is the central concern of the case for interpretable-by-design models [8]. Without faithfulness checks, a heat map can create unwarranted confidence.

Disagreement between methods. Different attribution methods can highlight different regions for the same image and model, particularly when they define feature removal differently [41]. No single method can currently be considered the reference standard.

Lack of standard evaluation. Metric families exist [61], but few clinical imaging studies apply them systematically, and thresholds for acceptable explanation quality are not established [1].

Dataset limitations. Public datasets differ in labelling, acquisition conditions, and class balance [15]–[17]. Class imbalance and distribution shift between training and test partitions complicate both performance and explanation analysis, because an apparently

strong aggregate score can mask poor behaviour on a minority class.

Audience mismatch. Explanations suitable for developers are not necessarily suitable for clinicians, and healthcare reviews emphasise tailoring interpretability to the stakeholder [9], [30].

Over-reliance and misinterpretation. Clinicians may over-trust a model because it provides an explanation, even when the explanation is unreliable [4], [5]. Explanations must be presented with appropriate caveats.

Computation and workflow. Perturbation-based and sampling-based methods can be expensive, while gradient-based ones are cheap but coarse; integration into clinical workflows imposes additional constraints that are rarely studied.

XII. RESEARCH GAPS

TABLE 5
Identified Research Gaps and Their Implications

Gap	Evidence in the reviewed literature	Implication for future work
G1: Accuracy benchmarking is separated from explanation analysis	Many classification studies report performance only [10], [11], [13]	Pair benchmarking with explanation analysis on the same models
G2: Limited quantitative evaluation of saliency maps	Metric families exist but are inconsistently applied [61]	Apply faithfulness, stability, and localisation metrics
G3: Few comparisons of explanations across backbones	Explanation studies typically concentrate on one model [49]	Compare heavy and lightweight backbones under identical protocols
G4: Explanation of transfer-learned models is under-examined	Transfer learning widely used [25], [60]	Study effect of fine-tuning on attribution behaviour
G5: Class imbalance seldom linked to explanation	Datasets vary in balance [15], [17]	Analyse per-class explanations, especially for minority classes
G6: Limited clinician-centred validation	Human-centred evaluation is called for [5], [30]	Include expert assessment in future studies
G7: Documentation of datasets and training protocols is uneven	Public datasets with varying documentation [16], [18]	Report protocols fully to support reproducibility

XIII. FUTURE RESEARCH DIRECTIONS

- **Multi-method explanation audits.** Apply several post-hoc methods (for example Grad-CAM, occlusion, LIME, SHAP) to the same trained models and quantify their agreement [37], [38], [43], [49].
- **Standardised saliency evaluation.** Adopt metric suites for faithfulness and stability, reported at dataset level [61].
- **Concept- and example-level explanation.** Move from pixel-level to concept- and case-level reasoning where clinical concepts are defined [44]–[46].
- **Hybrid pipelines.** Combine accurate deep backbones with interpretable surrogates or rule extraction [14], [39], [40].
- **Imbalance-aware explanation.** Examine class-wise explanation quality when datasets are skewed.
- **Clinician-in-the-loop validation.** Evaluate whether explanations improve decision quality, not only whether they appear plausible [4], [30].
- **Cross-modality transfer of methods.** Adapt explanation techniques from other domains, such as time-series and graph learning, with critical validation [55]–[57].

These directions motivate the companion implementation study, which benchmarks a baseline CNN against ResNet50, MobileNetV2, and EfficientNet-B0 on a binary medical image classification task and defines a planned post-hoc XAI extension. The companion paper does not report XAI outputs; this review does not either.

XIV. CONCLUSIONS

Post-hoc explainability offers a pragmatic way to inspect high-performing deep networks used in medical image diagnosis, but it does not by itself make a model trustworthy. Our analysis suggests that (i) methods differ in scope, cost, and the claims they can support; (ii) architecture choice shapes what explanations can reveal; (iii) evaluation of explanations remains immature compared with evaluation of accuracy; and (iv) the clinical value of explanations is largely untested. The research gaps identified here call for studies that evaluate multiple explanation methods on identical models with quantitative, class-aware, and clinician-informed assessment. Such work should avoid conflating

predictive performance, interpretability, faithfulness, localisation, and clinical usefulness.

REFERENCES

- [1] B. H. van der Velden, H. J. Kuijff, K. G. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," *Med. Image Anal.*, vol. 79, Art. no. 102470, 2022.
- [2] A. Diaz-Pinto, S. Morales, V. Naranjo, T. Köhler, J. M. Mossi, and A. Navea, "CNNs for automatic glaucoma assessment using fundus images: An extensive validation," *Biomed. Eng. Online*, vol. 18, pp. 1–19, 2019.
- [3] N. Ullah, I. De Falco, and G. Sannino, "A novel deep learning approach for colon and lung cancer classification using histopathological images," in *Proc. 2023 IEEE 19th Int. Conf. e-Science*, 2023, pp. 1–10.
- [4] M. Reyes *et al.*, "On the interpretability of artificial intelligence in radiology: Challenges and opportunities," *Radiol., Artif. Intell.*, vol. 2, no. 3, Art. no. e190043, 2020.
- [5] A. I. Poon and J. J. Sung, "Opening the black box of AI-medicine," *J. Gastroenterol. Hepatol.*, vol. 36, no. 3, pp. 581–584, 2021.
- [6] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.
- [7] N. Ullah, J. A. Khan, I. De Falco, and G. Sannino, "Explainable artificial intelligence: Importance, use domains, stages, output shapes, and challenges," *ACM Comput. Surv.*, vol. 57, no. 4, pp. 1–36, 2024.
- [8] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019.
- [9] C. O. Retzlaff *et al.*, "Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists," *Cogn. Syst. Res.*, vol. 86, Art. no. 101243, 2024.
- [10] G. Latif, D. A. Iskandar, J. M. Alghazo, and N. Mohammad, "Enhanced MR image classification using hybrid statistical and wavelets features," *IEEE Access*, vol. 7, pp. 9634–9644, 2018.
- [11] Z. Lai and H. Deng, "Medical image classification based on deep features extracted by deep model and statistic feature fusion with multilayer perceptron," *Comput. Intell. Neurosci.*, vol. 2018, Art. no. 2061516, 2018.

- [12] S. Iqbal, A. N. Qureshi, M. Alhussein, K. Aurangzeb, I. A. Choudhry, and M. S. Anwar, "Hybrid deep spatial and statistical feature fusion for accurate MRI brain tumor classification," *Front. Comput. Neurosci.*, vol. 18, Art. no. 1423051, 2024.
- [13] N. Ullah and A. Javed, "Deep features comparative analysis for COVID-19 detection from the chest radiograph images," in *Proc. 2021 Int. Conf. Frontiers Inf. Technol.*, 2021, pp. 258–263.
- [14] I. De Falco, G. De Pietro, and G. Sannino, "Classification of Covid-19 chest X-ray images by means of an interpretable evolutionary rule-based approach," *Neural Comput. Appl.*, vol. 35, no. 22, pp. 16061–16071, 2023.
- [15] A. A. Borkowski, M. M. Bui, L. B. Thomas, C. P. Wilson, L. A. DeLand, and S. M. Mastorides, "Lung and colon cancer histopathological image dataset (LC25000)," arXiv:1912.12142, 2019.
- [16] T. Rahman, M. Chowdhury, and A. Khandakar, "COVID-19 radiography database: COVID-19 chest X-ray images and lung masks database," 2022.
- [17] P. Sabale, "Brain MRI images for brain tumor detection," Kaggle dataset, 2024.
- [18] U. Siddiqui, "BTTypes," Kaggle dataset, 2024.
- [19] V. A. Sairam, "Ultrasound breast images for breast cancer," Kaggle dataset, 2024.
- [20] H. Saud, "Generated eye dataset for glaucoma detection," Kaggle dataset, 2024.
- [21] V. Bevilacqua, A. Di Marino, E. Di Nardo, A. Ciaramella, I. De Falco, and G. Sannino, "Cross-domain super-resolution in medical imaging," in *Proc. 2024 IEEE Symp. Comput. Commun.*, 2024, pp. 1–6.
- [22] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, 2020.
- [23] S. Aamir, A. Rahim, S. Bashir, and M. Naeem, "Prediction of breast cancer using AI-based methods," in *Intelligent Environments (Workshops)*, 2021, pp. 213–220.
- [24] H. Qayyum, S. T. H. Rizvi, M. Naeem, U. B. Khalid, M. Abbas, and A. Coronato, "Enhancing diagnostic accuracy for skin cancer and COVID-19 detection: A comparative study using a stacked ensemble method," *Technologies*, vol. 12, no. 9, p. 142, 2024.
- [25] A. Raza, K. Munir, M. S. Almutairi, and R. Sehar, "Novel transfer learning based deep features for diagnosis of Down syndrome in children using facial images," *IEEE Access*, 2024.
- [26] L. Verde, G. De Pietro, A. Ghoneim, M. Alrashoud, K. N. Al-Mutib, and G. Sannino, "Exploring the use of artificial intelligence techniques to detect the presence of coronavirus COVID-19 through speech and voice analysis," *IEEE Access*, vol. 9, pp. 65750–65757, 2021.
- [27] L. Verde, G. De Pietro, and G. Sannino, "Artificial intelligence techniques for the non-invasive detection of COVID-19 through the analysis of voice signals," *Arab. J. Sci. Eng.*, vol. 48, no. 8, pp. 11143–11153, 2023.
- [28] N. Burkart and M. F. Huber, "A survey on the explainability of supervised machine learning," *J. Artif. Intell. Res.*, vol. 70, pp. 245–317, 2021.
- [29] M. Christoph, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Leanpub, 2020.
- [30] G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar, "Interpretability of machine learning-based prediction models in healthcare," *WIREs Data Mining Knowl. Discov.*, vol. 10, no. 5, Art. no. e1379, 2020.
- [31] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Trans. Syst., Man, Cybern.*, vol. 21, no. 3, pp. 660–674, 1991.
- [32] S. Tangirala, "Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 2, pp. 612–619, 2020.
- [33] G. Tripepi, K. Jager, F. Dekker, and C. Zoccali, "Linear and logistic regression analysis," *Kidney Int.*, vol. 73, no. 7, pp. 806–810, 2008.
- [34] T. Wang, C. Rudin, Y. Doshi-Velez, Y. Liu, E. Klampfl, and P. MacNeille, "A Bayesian framework for learning rule sets for interpretable classification," *J. Mach. Learn. Res.*, vol. 18, no. 70, pp. 1–37, 2017.
- [35] J. H. Friedman and B. E. Popescu, "Predictive learning via rule ensembles," *Ann. Appl. Statist.*, vol. 2, pp. 916–954, 2008.
- [36] A. Sarkar, D. Vijaykeerthy, A. Sarkar, and V. N. Balasubramanian, "A framework for learning ante-hoc explainable models via concepts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10286–10295.
- [37] M. T. Ribeiro, S. Singh, and C. Guestrin, "“Why should I trust you?” Explaining the predictions of any classifier," in *Proc. KDD*, 2016, pp. 1135–1144.
- [38] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.
- [39] N. Vasilev, Z. Mincheva, and V. Nikolov, "Decision tree extraction using trained neural network," in *Proc. SMARTGREENS*, 2020, pp. 194–200.
- [40] R. Confalonieri, T. Weyde, T. R. Besold, and F. M. del Prado Martín, "Using ontologies to enhance human understandability of global post-hoc explanations of black-box models," *Artif. Intell.*, vol. 296, Art. no. 103471, 2021.
- [41] I. Covert, S.-I. Lundberg, and S.-I. Lee, "Explaining by removing: A unified framework for model explanation," *J. Mach. Learn. Res.*, vol. 22, no. 209, pp. 1–90, 2021.
- [42] M. Wojtas and K. Chen, "Feature importance ranking for deep learning," in *Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 5105–5114.
- [43] S. Lundberg, "A unified approach to interpreting model predictions," arXiv:1705.07874, 2017.
- [44] B. Kim *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2668–2677.
- [45] N. Papernot and P. McDaniel, "Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning," arXiv:1803.04765, 2018.
- [46] A. Poché, L. Hervier, and M.-C. Bakkay, "Natural example-based explainability: A survey," in *Proc. World Conf. Explainable Artif. Intell.*, 2023, pp. 24–47.
- [47] G. Casalino, G. Castellano, K. Kaczmarek-Majer, and G. Zaza, "Prototype-based explanations to improve understanding of unsupervised datasets," in *Proc. 2024 IEEE Int. Conf. Fuzzy Syst.*, 2024, pp. 1–8.
- [48] S. Verma, J. Dickerson, and K. Hines, "Counterfactual explanations for machine learning: A review," arXiv:2010.10596, 2020.
- [49] N. Ullah, J. A. Khan, I. De Falco, and G. Sannino, "Bridging clinical gaps: Multi-dataset integration for reliable multi-class lung disease classification with DeepCRINet and occlusion sensitivity," in *Proc. 2024 IEEE Symp. Comput. Commun.*, 2024, pp. 1–6.
- [50] Z. M. Hira and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data," *Adv. Bioinformatics*, vol. 2015, Art. no. 198363, 2015.
- [51] A. K. Uysal, "On two-stage feature selection methods for text classification," *IEEE Access*, vol. 6, pp. 43233–43251, 2018.

- [52] L. Huang, X. Zhou, L. Shi, and L. Gong, "Time series feature selection method based on mutual information," *Appl. Sci.*, vol. 14, no. 5, p. 1960, 2024.
- [53] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E*, vol. 69, no. 6, Art. no. 066138, 2004.
- [54] D. D. Nauck, "Measuring interpretability in rule-based classification systems," in *Proc. FUZZ'03*, vol. 1, 2003, pp. 196–201.
- [55] A. R. Troncoso-García, M. Martínez-Ballesteros, F. Martínez-Álvarez, and A. Troncoso, "A new approach based on association rules to add explainability to time series forecasting models," *Inf. Fusion*, vol. 94, pp. 169–180, 2023.
- [56] A. R. Troncoso-García, I. S. Brito, F. Martínez-Álvarez, and A. Troncoso, "Explainable hybrid deep learning and Coronavirus Optimization Algorithm for improving evapotranspiration forecasting," *Comput. Electron. Agric.*, vol. 215, Art. no. 108387, 2023.
- [57] Y. Nian, Y. Chang, W. Jin, and L. Lin, "Globally interpretable graph learning via distribution matching," in *Proc. ACM Web Conf.*, 2024, pp. 992–1002.
- [58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [59] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.
- [60] A. Ghosh, B. Soni, and U. Baruah, "Transfer learning-based deep feature extraction framework using fine-tuned EfficientNet-B7 for multiclass brain tumor classification," *Arab. J. Sci. Eng.*, vol. 49, no. 9, pp. 12027–12048, 2024.
- [61] T. Gomez, T. Fréour, and H. Mouchère, "Metrics for saliency map evaluation of deep learning explanation methods," in *Int. Conf. Pattern Recognit. Artif. Intell.*, 2022, pp. 84–95.