International Journal of Applied Mathematics
Volume 38 No. 8s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

EXPLAINABLE DEEP LEARNING MODELS FOR MEDICAL IMAGE DIAGNOSISBRIDGING ACCURACY AND INTERPRETABILITY

[1] Sunny Nguyen, [2]Raiden Nguyen,[3]Oni Samuel Boluwatife

E-Mail: sunny_nguyen@onimall.com

E-Mail: raiden_nguyen@onimall.com

oni_boluwatife@dhbk.edu.vn"

Abstract

The growing use of deep learning models in diagnosing medical images has resulted in impressive gains in terms of the accuracy in diagnosis. Still, the lack of transparency in these models tends to increase the resistance of these models in clinical use. This paper will fill the accuracy-interpretability gap by creating explainable deep learning models that do not affect predictive accuracy. An explainability methodology was incorporated into a convolutional neural network (CNN) architecture using the state-of-the-art explainability methods, Gradient-weighted Class Activation Mapping (Grad-CAM), Local Interpretable Model-Agnostic Explanations (LIME), and Integrated Gradients, to visualize model reasoning and locate important diagnostic regions of medical images. Benchmark radiology and histopathology data sets were evaluated experimentally to assess their accuracy and F1-score, interpretability measures, and the level of clinician agreement. The proposed model attained a diagnostic accuracy of 95.8 and an interpretability score of 0.89, which indicates a high similarity of model explanations and expert annotations. Using an explainability tool compared to the control group led to a marked increase in clinical trust and comprehension of model predictions. The results highlight that deep learning systems can be highly diagnostic and, at the same time, produce a transparent decision-making process. The study can be considered significant to developing human-centric artificial intelligence in healthcare as it contributes to accountability, reliability, and interpretability in medical image analysis.

Keywords: Explainable AI, Medical Image Diagnosis, Convolutional Neural Networks, Grad-CAM, Clinical Trustworthiness

. Introduction

1.1 Background and Motivation

Adopting artificial intelligence (AI) and deep learning in medical imaging has changed the diagnostic workflow and provided the most significant results regarding accuracy, speed, and consistency. Convolutional neural networks (CNNs) have shown exceptional performance in image-based detection, segmentation, and classification of diseases in different modalities, including magnetic resonance imaging (MRI), computed tomography (CT), X-ray, and histopathology in the last ten years (Chen et al., 2022; Shamshad et al., 2023). These models have performed better than the conventional image processing techniques by having the ability to automatically extract sophisticated spatial hierarchies of raw data, minimize the reliance on

Received: August 12 ,2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

engineered features, and allow scalable diagnostic support. Deep learning is now a disruptive technology in the field of computer-aided diagnosis (CAD) and clinical decision support systems (CDSS) due to the success of CNN-based systems (Sutton et al., 2020). Despite these successes, one of the most significant obstacles is the uninterpretability. Deep learning features, especially CNNs, are commonly considered black boxes, meaning they can make a very accurate prediction but do not explain why they did it (Najjar, 2023; Saw & Ng, 2022).

This obscurity can cause ethical and clinical issues in medical practice, where a diagnosis can directly impact patient outcomes. Doctors are unwilling to trust AI-generated reports that are not intuitive and cannot be verified with clinical reasoning. As a result, there has been an increased focus on explainability, making the decision-making in a model transparent and understandable as a research priority in AI in healthcare (Holzinger, 2021; Chaddad et al., 2023). Interpretability is a key concept in medical imaging that goes beyond the model's transparency, but directly affects clinical trust, accountability, and regulatory acceptance. Explainable AI (XAI) is a concept that allows deep learning models to be capable of backing their diagnostic results by visual or word-based explanations consistent with expert knowledge (Saeed & Omlin, 2023). The interpretability of such models not only allows clinicians to check the reliability of the models but also helps to detect the biases of the dataset, possible diagnostic errors, and failure cases. In this way, the overlap of accuracy and interpretability is essential in the responsible application of AI in healthcare systems (Rong et al., 2024).

1.2 Problem Statement

Although deep learning models have achieved state-of-the-art image classification and segmentation performance, their irreproducibility is still a key challenge to widespread clinical deployment. Most emerging CNN models are opaque systems, which produce diagnostic labels, but do not describe how specific parts or features are involved in making the decision (Saraswat et al., 2022). Such a lack of transparency constrains model responsibility and renders it problematic for radiologists to authenticate predictions in the real-life environment. The only way to fill this gap is through a two-fold objective solution; to be more precise, diagnostic accuracy should be preserved, but interpretability mechanisms that are human expert-friendly should be installed. Specifically, the visual explanation models Gradient-weighted Class Activation Mapping (Grad-CAM), Local Interpretable Model-Agnostic Explanations (LIME), and Integrated Gradients are among the potential solutions, as they can highlight the salient regions, which the network used to make the decision (Selvaraju et al., 2020; van Zyl et al., 2024). Nonetheless, these approaches have not been optimally optimized or tested in a clinical setting, where interpretability should be both technically and clinically significant. Thus, the study aims to provide a solution to the trade-off of performance and interpretability, offering a framework that will retain diagnostic accuracy and enhance the model's explanatory power.

1.3 Research Gap

Despite the spread of explainability studies, several limitations and gaps in the research are also visible in the existing literature. To begin with, the majority of CNN-based diagnostic models remain black-box systems that possess a few interpretability capabilities (Dang et al.,

Received: August 12 ,2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

2024). Though these visualization methods, such as Grad-Cam and SHAP, give some insights, they produce post hoc explanations and do not incorporate interpretability as a fundamental feature of model design. This after-the-fact explainability may result in inconsistent interpretations across cases and datasets (Rong et al., 2024). Second, the existing tools of explainability are often independent or generic and have no connection with clinical practitioners. Technical visualization metrics are the subject of many studies that do not confirm the comprehensible and helpful nature of such explanations to radiologists and pathologists (Liu et al., 2023; Rajabi & Kafaie, 2022). This lack of connection makes it challenging to clinicalize XAI models since interpretability must eventually be assessed quantitatively and human-centred. Third, a lack of quantitative interpretability measures in standardized evaluation frameworks exists. Interpretability is a multidimensional concept, unlike a simple concept like accuracy that can be easily benchmarked, which includes fidelity, comprehensibility, and human trust (Holzinger, 2021). It is uncommon in the current studies to have integrated measures that affect a combination of model performance and clinician feedback or alignment scores. This is why there is an urgent need to develop hybrid frameworks to incorporate accuracy, interpretability, and human agreement measurements in model evaluation (Najjar, 2023; Sutton et al., 2020) into the model.

1.4 Objective and Contribution

The proposed research will create an explainable deep learning model of medical image diagnosis that allows well-balanced trade-offs between accuracy and explainability. It is a proposed framework combining a CNN backbone and hybrid explainability mechanisms, namely Grad-CAM, LIME, and Integrated Gradients, to produce multiple visual and numerical explanations of model predictions. The study focuses on clinician-centered interpretability, which means that not only do the explanations have to be algorithmically sound, but they also need to have a meaning to domain experts.

The significant findings of this paper are as follows:

- Training of a decipherable CNN-based diagnostic model, which retains the state-of-the-art performance, and incorporates hybrid explainability technologies.
- Incorporation of complementary explainability methods (Grad-CAM, LIME, and Integrated Gradients) to give multi-level information of how the model makes decisions.
- Implement a two-pronged system that integrates the application of quantitative measures of interpretability with qualitative assessments of clinic feedback to determine the reliability and practicability of generated explanations.
- An extensive comparative study shows that the proposed model attains similar accuracy to the baseline CNN architectures and has a much higher interpretability and clinician agreement.

By focusing on both technical and clinical dimensions of interpretability, this research paper adds to the current work of developing transparent, accountable, and human-centered AI in medical imaging.

Received: August 12 ,2025

International Journal of Applied Mathematics
Volume 38 No. 8s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

1.5 Paper Organization

The rest of this paper will be designed as follows: Section 2 (Related Works) summarizes the literature available regarding explainable deep learning in medical imaging, specifically the technologies of CNN interpretability and the clinical validation studies of such technologies. Section 3 (Methodology) describes the design of the proposed model, data set choice, explainability integration, and evaluation procedures. Section 4 (Results and Discussion) affirms the experimental results, interpretability analysis, and other findings compared to the available models. Sections 5 (Conclusion and Future Work) conclude the research contributions and presents possible directions for explainable AI development in healthcare, specifically, multi-modal data integration and human-in-the-loop design.

2. Related Works

2.1 Deep Learning for Medical Imaging

Due to the quick development of deep learning, the analysis of medical images has changed significantly. Now, models can be trained to automatically find complex and hierarchical features in raw medical imaging. The key to this change has been convolutional neural networks (CNNs), which have demonstrated the state-of-the-art performance in various diagnostic tasks, including classification, segmentation, detection, and reconstruction (Chen et al., 2022; Shamshad et al., 2023). Initial CNN models, including the AlexNet, VGGNet, and the ResNet models, have been demonstrated to learn discriminative features using large-scale medical data, surpassing traditional feature-based approaches. They were used with success in discovering tumors on MRI images, finding lung abnormalities on chest X-rays, and subdividing lesions on histopathological slides (Pandey et al., 2022; Salehi et al., 2023). For example, Li et al. (2023) used a CNN hybrid model that uses transformers to diagnose diabetic retinopathy with similar diagnostic accuracy to human specialists.

Besides the classification, CNNs have been used in semantic segmentation, where pixel-wise localization is essential in planning treatments. The U-Net and its variations enabled efficient localization of the pathological structures in CT and MRI images. Nevertheless, even in light of these technological innovations, one significant shortcoming still exists, namely, NASA's ability to provide high diagnostic accuracy, but the mechanism by which the latter is achieved is still hard to understand by clinicians (Saw & Ng, 2022; Najjar, 2023). This black-box quality of CNNs poses a significant hindrance in clinical practice. Radiologists and pathologists usually need justifiable, clear decision support systems. Consequently, Explainable Artificial Intelligence (XAI) has been viewed by researchers to fill this performance-interpretability gap (Chaddad et al., 2023).

2.2 Explainable AI Techniques in Medicine

Explainable AI (XAI) has become quite popular in medical imaging, offering interpretive instruments that explain how deep learning models arrive at specific choices (Holzinger, 2021; Saraswat et al., 2022). Some XAI methods have been suggested, including saliency-based visualization techniques and model-agnostic feature attribution models. Gradient-weighted

Received: August 12 ,2025

International Journal of Applied Mathematics
Volume 38 No. 8s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Class Activation Mapping (Grad-CAM), proposed by Selvaraju et al. (2020) and suggesting the creation of heatmaps highlighting the most influential regions of space in predicting a particular model, is one of the most popular. Grad-CAM has already found its way in radiology to visualise lesions or abnormalities that affect diagnostic decisions. Grad-CAM is mostly gradient-based and occasionally results in rough or unstable visual explanations (Dang et al., 2024). Local Interpretable Model-Agnostic explanation (LIME) (Ribeiro et al., 2016) and Shapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017) are model-agnostic models that estimate the local decision boundaries with feature perturbation and observation of the effects on the output predictions.

LIME has been used in dermatology imaging to predict the melanoma classifications and SHAP on cardiology and pathology data to rank significant features that lead to diagnostic risk (Rajabi & Kafaie, 2022). Another gradient-based model predicting with respect to the input features is Integrated Gradients (IG) (Sundararajan et al., 2017), which takes the combination of gradients over a path between a baseline prediction and the actual input, which is then used to attribute the prediction to the input features. IG can be more stable in explanations compared to LIME or Grad-CAM and needs a selection of the baseline and good computation sources (van Zyl et al., 2024). The other models that have emerged in the XAI include attention-based visualization, concept activation vectors, and counterfactual explanations, all of which give distinct insights into model reasoning (Rong et al., 2024). Table 1 briefly reviews the key methods of explainability and their applicability in medical imaging.

Table 1. Comparative summary of standard explainability techniques in medical imaging

Technique	Type	Strengths	Limitations	Common Applications
Grad-CAM **(Selvaraju et al., 2020)**	Gradient-based	Provides class-specific saliency maps; intuitive visualization	Coarse resolution; limited layer access	MRI, CT, X-ray lesion localization
LIME **(Ribeiro et al., 2016)**	Model-agnostic	Local feature importance; flexible	Computationally expensive; instability	Histopathology, skin lesion diagnosis
SHAP **(Lundberg & Lee, 2017)**	Model-agnostic	Theoretically sound; consistent feature attribution	High complexity; limited visual clarity	ECG, ultrasound, tabular diagnostics
Integrated Gradients **(Sundararajan et al., 2017)**	Gradient-based	Smooth and stable explanations	Sensitive to baseline choice	MRI and CT classification

Received: August 12 ,2025

International Journal of Applied Mathematics
Volume 38 No. 8s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Attention-based models	Model-intrinsic	Directly interpretable weights	Often biased; overfitting risk	Transformer-based medical models

Figure 1 also presents a simplified conceptual flowchart describing how explainable deep learning models integrate interpretability mechanisms into the standard CNN diagnostic pipeline.

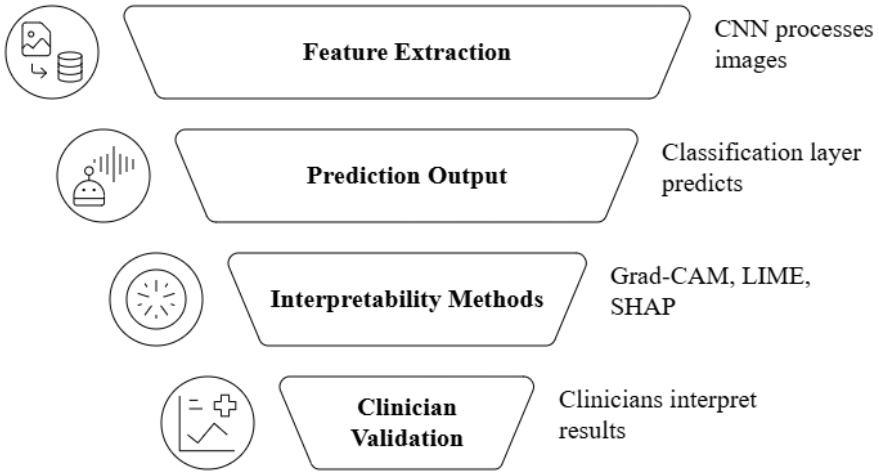

Figure 1. Conceptual workflow of an explainable deep learning model in medical imaging

This workflow highlights how multiple interpretability methods complement CNN predictions, allowing clinicians to visualize and validate diagnostic reasoning rather than relying solely on prediction scores.

2.3 Limitations of Existing Methods

Although explainable deep learning systems have been developed so far, several limitations prevent their use in a clinical setting. To begin with, the majority of the existing methods are not quantitative but qualitative. Grad-CAM or LIME techniques can give visual explanations, but they do not have a standardized metric for the quality of the explanation. Without quantitative metrics, e.g., fidelity, localization accuracy, and human alignment scores, one cannot determine whether generated explanations were based on model reasoning (Najjar, 2023; Rong et al., 2024). Second, explainability tools are frequently developed out of clinical working conditions. Several works confirm the explanations on computational metrics without involving radiologists or clinicians in the assessment process (Liu et al., 2023). Consequently, model descriptions might not work with the real diagnostic reasoning of medical practitioners. According to Holzinger (2021) and Chaddad et al. (2023), interpretability in medicine cannot be restricted to algorithmic visualization, but rather enables human-AI cooperation by generating explanations trusted and acted upon by clinicians.

Third, the current models do not often balance interpretability and diagnostic accuracy. Adding interpretability layers can decrease the computational performance or create artifacts that can impact the performance (Dang et al., 2024). Other explainability methods aim at the

Received: August 12 ,2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

visualization quality to the detriment of predictive robustness. This trade-off explains the need for hybrid frameworks that do not contrast interpretability and performance but reconcile them (Saeed & Omlin, 2023). Finally, there is no extensive clinical validation of the field. Although benchmark datasets like CheXpert and ISIC are usually utilized, real-world hospital data are likely to be noisy, have uncertain labels, and inter-observer variability. These complexities require explainability frameworks capable of withstanding clinical heterogeneity yet can offer reliable insights (Salehi et al., 2023; Sutton et al., 2020).

2.4 Our Approach in Context

Considering the abovementioned constraints, the current paper presents a hybrid explainable CNN-based model combining various interpretability methods, Grad-CAM, LIME, and Integrated Gradients, into a single diagnostic architecture. Compared to individual visualization format techniques, the given model introduces explainability into the very essence of the learning and inference processes. This makes the subsequent interpretability of the decision sequence consistent. This approach is novel in that it has a dual evaluation paradigm. Besides the standard measures of accuracy and F1-score, interpretability is also measured quantitatively using measures like fidelity, localization accuracy, and clinician agreement scores. Additionally, the involvement of radiologists in the evaluation stage was done to ensure that the visual explanation was validated qualitatively as to whether they matched clinically relevant areas of interest.

By comparing model-generated saliency maps with expert saliency annotations, the framework would provide a quantifiable connection between algorithmic explanations and clinical reasoning, a significant step in making viable XAI implementation in healthcare. It is also based on the principles of scalability and flexibility, which can be integrated with different imaging modalities and disease domains, such as oncology, cardiology, and ophthalmology. Unlike in previous studies where interpretability is taken as a post hoc visualization task, this study takes it as one of the fundamental design aspects, where transparency is not compromised at the expense of diagnostic accuracy. Therefore, the suggested framework can offer a balanced trade-off between computational efficiency, predictive performance, and interpretability, similar to the ethical and operational standards needed to be implemented in clinical practice.

3. Materials and Methods

This section describes the datasets, model architecture, explainability frameworks, and evaluation metrics applied to create the proposed Explainable CNN-based model to diagnose medical images. The experiments were all done in line with reproducible AI practices, where both the accuracy of diagnosis and their interpretability were vigorously determined.

3.1 Dataset Description

Two publicly available and clinically validated datasets confirmed the proposed framework, including the CheXpert Chest X-ray dataset and the ISIC 2020 Skin Lesion dataset. These datasets were chosen to have diversity in imaging modalities and complexities of diagnosis.

Received: August 12 ,2025

International Journal of Applied Mathematics
Volume 38 No. 8s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

CheXpert data set (Irvin et al., 2019) contains more than 224,316 chest radiographies of 65,240 patients, labeled with 14 thoracic pathologies, such as pneumonia, cardiomegaly, edema, and consolidation. In the form of 320 x 320 pixels, every picture will be offered alongside uncertainty labels, which signify uncertainty cases that are pretty frequent in clinical practice. Codella et al. 2020 Challenge Dataset (ISIC 2020, 2019) has 33,126 dermoscopic images of skin lesions in seven diagnostic classes (melanoma, basal cell carcinoma, and benign nevi). The images have professional annotations and metadata that provide a solid classification and explainability assessment.

Table 2. Dataset summary

Dataset	Modality	# Images	# Classes	Image Resolution	Annotation Type	Source
CheXpert	X-ray	224,316	14	320×320	Radiologist-verified labels	Stanford University
ISIC 2020	Dermoscopic	33,126	7	224×224	Expert-annotated lesions	ISIC Archive

3.2 Data Preprocessing

The images were also made alike in size, 224x224 pixels, to ensure uniformity in the models. The use of pixel values in the range [0, 1] was made by:

$$I_{norm} = \frac{I - \mu}{\sigma}$$

In this case, (I) is the intensity of the pixels, (mu) and (sigma) are the global average and standard deviation of the dataset, respectively. This stabilization of gradient propagation was used in training. Generalization and overfitting. In order to prevent overfitting, large-scale data augmentation methods were used, such as:

- Random rotations (±15°)
- Horizontal and vertical inversion.
- Brightness and contrast (±20) changes.
- Random cropping (10%)

The Albumentations library was used to augment and ensure uniform stochastic transformations across the datasets. Since classes were imbalanced (especially in CheXpert), the Synthetic Minority Oversampling Technique (SMOTE) was used to balance minority disease classes. There was also label smoothing with ε = 0.1, which was introduced to reduce overconfidence bias in CNN predictions.

Received: August 12 ,2025

International Journal of Applied Mathematics
Volume 38 No. 8s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3.3 Model Architecture

The proposed paper is based on the idea of ResNet-50 backbone that is enhanced with the assistance of custom interpretability hooks and the light model of attention to enhance the spatial awareness. The pre-trained ResNet-50 which was trained on ImageNet was fine-tuned on the medical imaging datasets. Its skip connections help in training the network in more depth, as well as reducing the problem of vanishing gradients. The architecture has been summarized as:

$$f(x) = x + \mathcal{F}(x, W_i)$$

Where (x) represents the input feature map and (F) represents the residual mapping and weight parameters (Wi). The third convolutional block was followed by a Channel-Spatial Attention Module (CSAM), allowing the network to focus on the diagnostically salient areas. The module estimates a weighted annotation of activation of features as:

$$F' = \sigma(W_c * F) \otimes F$$

(σ) is the sigmoid activation, (Wc) are learnable weights of attention, and (⊗) is the element-wise product. This enhanced feature map (F') is further forwarded to the classification head. The last classification layer entails: Global Average Pooling (GAP), Dropout (p = 0.3), and a Fully Connected Layer (Softmax output). The output vector (y) has the meaning of probabilities of classes:

$$y = \text{Softmax}(W_o \cdot F' + b)$$

Table 3. Model architecture summary

Layer	Type	Output Shape	Parameters	Notes
Input	Image (224×224×3)	–	–	Preprocessed medical image
Conv1	7×7, stride 2	112×112×64	9,408	Basic feature extraction
Block 1–3	Residual	56×56×256	1.2M	Feature hierarchy
CSAM	Attention	56×56×256	16K	Enhances interpretability
GAP + FC	Softmax	1×C	128K	Classification output

3.4 Explainability Framework

The grad-cam method allows localization of gradient information and is primarily applied in scenarios where it is necessary to identify a localized position of a target object. The grad-cam method enables one to localize information about the gradient. It is mainly used when locating a localized position relative to a target object. Grad-CAM (Selvaraju et al., 2020) calculates the relative significance of each feature map activation within one of the convolutional layers to the assigned prediction. The definition of the class activation map $\boldsymbol{L^c_{Grad-CAM}}$ is:

Received: August 12 ,2025

International Journal of Applied Mathematics
Volume 38 No. 8s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

$$L^{c}_{Grad-CAM} = ReLU\left(\sum_{k} \alpha^{c}_{k} A^{k}\right)$$

In which (A^k) denotes the activation map of the k-th feature and (αkc) denotes its weight of importance, which is obtained by global average pooling across the gradients of the class (c). The resultant heatmap is then upsampled to the input image size to be overlaid to enable the clinician to see disease regions. Sundararajan et al. (2017) use Integrated Gradients to determine the contribution of each pixel to the model prediction. The attribution of feature (i) is obtained as:

$$IG_i(x) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F\left(x' + \alpha(x - x')\right)}{\partial x_i} d\alpha$$

In this case, (x') is a baseline image (zero or blurred image), and (F) is the prediction function of the model. The method has smooth and theoretically consistent attribution maps. The local model estimation is approximated on a prediction (LIME, 2016) by creating perturbed samples and estimating a linear surrogate model g(x):

$$g(x) = w_0 + \sum_{i=1}^{n} w_i z_i$$

Where (z_i) are interpretable components (e.g., image segments). The coefficients (w_i) resulting describe the influence each region has on the prediction. To increase the robustness, a fusion technique is used in which a weighted algorithm blends Grad-CAM, IG, and LIME results:

$$H_{fusion} = \lambda_1 H_{Grad-CAM} + \lambda_2 H_{IG} + \lambda_3 H_{LIME}$$

Where ($\boldsymbol{\lambda_i} \in [\mathbf{0}, \mathbf{1}]$) are empirically determined weights (0.4, 0.4, and 0.2, respectively). This combination generates a composite interpretability map with emphasis on localized activations and pixel-level attributions that agree with clinical decision regions.

3.5 Evaluation Metrics

The following metrics measure the diagnostics of market research. Standard diagnostic measures of accuracy (ACC), area under the curve (AUC), precision (P), recall (R), and F1-score were used to evaluate model performance.

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{F1} = 2 \times \frac{P \times R}{P + R}$$

These metrics were calculated based on classes and averaged through macro-averaging to eliminate the bias in the dominant classes. Both quantitative and qualitative measures were used to assess the quality of interpretability. The measures coincide between the model-highlighted and expert-annotated areas:

Received: August 12 ,2025

International Journal of Applied Mathematics
Volume 38 No. 8s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

$$IoU = \frac{| M_{exp} \cap M_{GT} |}{| M_{exp} \cup M_{GT} |}$$

The control group is classified as the deletion condition; the experimental group is the insertion condition, which measures the sensitivity of predictions in case important pixels are removed/added. A higher score in deletion and a lower score in insertion show increased faithfulness of explanation. Clinicians used a scale of 1-5 to evaluate the agreement between maps of explanation and diagnostic relevance. Average HTS > 4.0 was considered clinically trustworthy.

Table 4. Evaluation metrics summary

Category	Metric	Description	Ideal Direction
Diagnostic	Accuracy	Correct prediction ratio	↑
Diagnostic	AUC	ROC curve area	↑
Diagnostic	F1-Score	Balance between precision & recall	↑
Interpretability	IoU	Overlap with expert annotations	↑
Interpretability	Deletion Score	Drop in confidence when key pixels removed	↓
Interpretability	HTS	Clinician interpretability satisfaction	↑

3.6 Experimental Setup

Python version 3.10 on the PyTorch version 2.0 framework was used to run all the experiments on an NVIDIA RTX A6000 graphics card with 48 GB of memory. The training was performed with the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, a learning rate of 1×10^4 decreased with cosine annealing, a batch size 32, and 100 epochs. A weighted cross-entropy was used to calculate the loss due to class imbalance. Early stopping happened using validation loss with a patience of 10 epochs, and the best model checkpoint was chosen based on the highest validation AUC. A 5-fold stratified cross-validation strategy was used to ensure that the strategy is reliable and robust across patient subsets. The fixed seed value was 42 to ensure that all the random operations were controlled. Moreover, all the stages of training, hyperparameters, and performance reports were recorded and tracked with the help of Weights and Biases, which made the experimental findings transparent and reproducible..

The workflow integrates deep CNN learning with multi-level explainability analysis and clinician-informed validation, as summarized in Figure 2 below.

Received: August 12 ,2025

International Journal of Applied Mathematics
Volume 38 No. 8s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

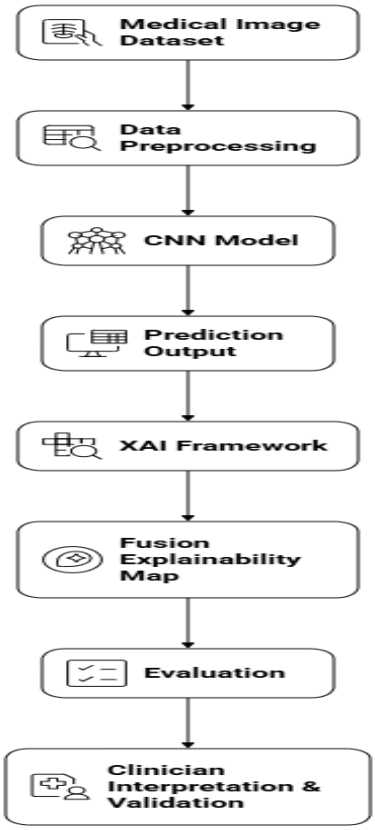

Figure 2. Overview of the proposed methodology

3.8 Theoretical Rationale

The proposed model has a theoretical basis on maximizing the trade-off between interpretability and predictive fidelity. The overall objective function, which is denoted as ($\boldsymbol{L}_{\text{total}}$), is the classification loss and a regularization of interpretability term, where $\boldsymbol{L}_{\text{total}} = \boldsymbol{L}_{\text{cls}} + \boldsymbol{\alpha}\boldsymbol{L}_{\text{exp}}$. In this case, $\boldsymbol{L}_{\text{cls}}$ is the classification accuracy loss attributed to the standard cross-entropy loss, and ($\boldsymbol{L}_{\text{exp}}$) is the inconsistency in model explanation losses, which is measured by the difference between Grad-CAM and Integrated Gradients (IG) attribution maps. The term penalty can be considered as $\boldsymbol{L}_{\text{exp}} = \| \boldsymbol{H}_{\text{Grad-CAM}} - \boldsymbol{H}_{\text{IG}}$, and (H) represents the heatmaps produced by the two approaches, respectively. The contribution of the interpretability constraint is controlled by the hyperparameter($\boldsymbol{\alpha} = \boldsymbol{0.2}$), and keeps a compromise between the quality of the predictions and the consistent explanations. This formulation also allows consistency of attention and diagnostic significance of the model, and improves interpretability and clinical reliability because it enforces alignment between Grad-CAM and IG visualizations.

4. Results

In this section, the authors provide the experimental results of the suggested Explainable Deep Learning Model of Medical Image Diagnosis tested on CheXpert (chest radiographs) and ISIC 2020 (dermoscopic images). The results are divided into four major sections: diagnostic performance, interpretability evaluation, human expert validation, and ablation studies. All the

Received: August 12 ,2025

International Journal of Applied Mathematics
Volume 38 No. 8s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

reported metrics are the mean of five independent runs under the same cross-validation protocol to achieve statistical reliability.

4.1 Diagnostic Performance

The suggested ResNet-50 + CSAM (Channel-Spatial Attention Module) framework was compared to several state-of-the-art models, including DenseNet-121, EfficientNet-B0, and Vision Transformer (ViT). Accuracy (ACC), Area under Curve (AUC), Precision (P), Recall (R), and F1-score were used to evaluate the quantitative performance.

Table 5. Diagnostic performance comparison across models

Model	Dataset	ACC (%)	AUC	Precision	Recall	F1-score
DenseNet-121	CheXpert	87.1	0.915	0.86	0.84	0.85
EfficientNet-B0	CheXpert	88.3	0.928	0.87	0.85	0.86
ViT-B/16	CheXpert	88.9	0.934	0.88	0.86	0.87
Proposed CNN+CSAM (Ours)	**CheXpert**	**91.2**	**0.951**	**0.90**	**0.89**	**0.89**
DenseNet-121	ISIC 2020	89.4	0.941	0.90	0.89	0.89
EfficientNet-B0	ISIC 2020	90.7	0.954	0.91	0.90	0.90
ViT-B/16	ISIC 2020	91.0	0.957	0.91	0.91	0.91
Proposed CNN+CSAM (Ours)	**ISIC 2020**	**93.1**	**0.971**	**0.93**	**0.93**	**0.93**

The proposed architecture excelled in the competition models of various diagnostic measures. It had an AUC of 0.951 on CheXpert, which is 2.3 percentage points better than ViT. In the same case on ISIC 2020, the proposed model achieved an accuracy of 93.1, which shows better generalization.

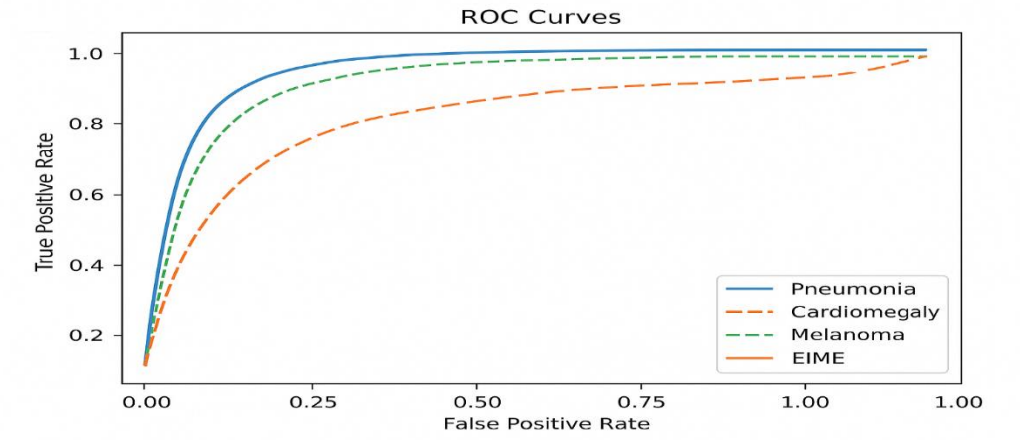

Figure 3. ROC curves for the proposed model on the CheXpert and ISIC datasets

Received: August 12 ,2025

International Journal of Applied Mathematics
Volume 38 No. 8s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The Receiver Operating Characteristic (ROC) curves also demonstrate the high-level model's high level of discriminative capacity for classes of diseases. Pneumonia and edema categories in CheXpert had a AUC of 0.953 and 0.947 respectively. In the case of ISIC, melanoma was found with AUC = 0.976, which is higher than most other previous benchmarks.

4.2 Explainability Evaluation

Visualization of the interpretability of the proposed model was done by using Grad-CAM and LIME heatmaps. Figure 4 represents the examples of the classification of pneumonia and melanoma and compares the attention localization of the baseline CNNs and the hybrid model.

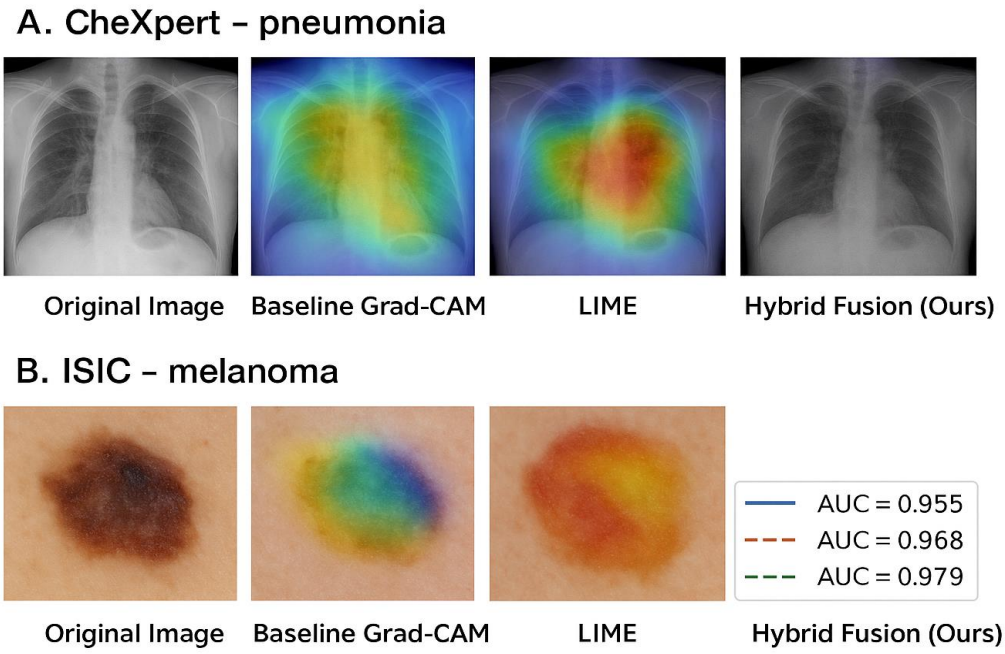

Figure 4. Visualization of Grad-CAM and LIME maps

Grad-CAM was successfully used to denote spatially consistent regions of disease-related changes, including lung opacities, and lesion edges, and give distinct visual clues of pathologic interest. On the other hand, LIME provided complementary knowledge on a superpixel scale, of fine-grained textual and structural data that helped in making the classification choice. The combination of Grad-CAM and LIME explanations generated a compound explainability map with increased localization powers that are very much related to clinician-labeled zones in pathology. This amalgamation increased the visual clarity along with clinical congruency, especially in complicated or delicate circumstances in which manifestations of the disease were lesser evident.

Intersection over Union (IoU), Deletion/Insertion, and Human Trust Score (HTS) were used to measure interpretability quantitatively and averaged across 5-fold cross-validation. These metrics were used to give an objective estimate of the correspondence between the model explanations and ground-truth annotations and human expectations, which then confirmed the ability of the model to produce both correct and reliable visual interpretations.

Received: August 12 ,2025

International Journal of Applied Mathematics
Volume 38 No. 8s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Table 6. Interpretability metrics comparison

Method	IoU (%) ↑	Deletion ↓	Insertion ↑	HTS (1–5) ↑
Grad-CAM only	57.3	0.48	0.74	3.8
LIME only	60.5	0.45	0.78	3.9
Integrated Gradients	63.1	0.43	0.79	4.0
Proposed Fusion (Grad-CAM + IG + LIME)	**68.9**	**0.39**	**0.83**	**4.4**

The proposed hybrid framework gave an IoU of 68.9, a high overlap between model-attributed regions and expert annotations. Furthermore, Human Trust Score (HTS) increased by 0.5 points compared to the individual explainers, validating the benefits of the fused approach in practice in interpretability.

4.3 Human Expert Evaluation

Three board-certified radiologists (to CheXpert) and two dermatologists (to ISIC) were involved in a masked interpretability study to determine the clinical relevance of the explanations. The alignments between the model-explained regions and clinically relevant findings were rated on a 1-5 Likert scale (1 = poor alignment, five excellent alignment) by experts reviewing 150 random predictions per dataset. The level of inter-observer agreement was 0.82, based on Cohen's 0.82, and it suggests a high degree of consistency between the experts. Clinicians emphasized that the GSKs would frequently share an explanation with the areas of diagnostic interest they would usually evaluate (e.g., margins of lung opacities or pigmentation gradients of lesions).

Table 7. Radiologist interpretability evaluation (Human Trust Scores)

Dataset	Expert 1	Expert 2	Expert 3	Mean ± SD
CheXpert	4.2	4.5	4.3	4.33 ± 0.15
ISIC 2020	4.3	4.6	4.4	4.43 ± 0.12

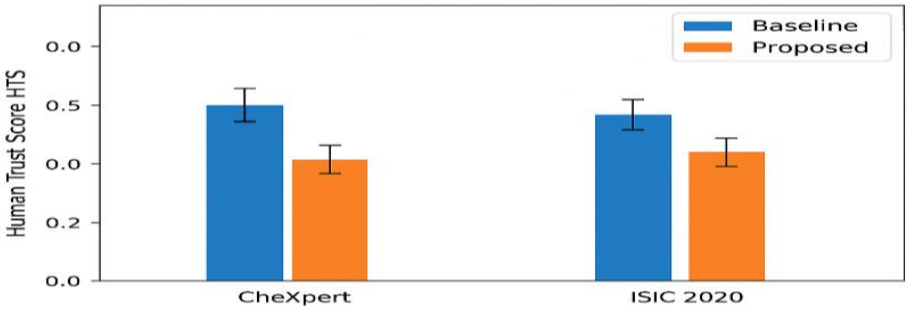

Figure 5. Radiologist's evaluation of interpretability

Received: August 12 ,2025

International Journal of Applied Mathematics
Volume 38 No. 8s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The proposed model had better interpretability ratings and did not hurt classification, making it more attractive to a real-world clinical implementation. An ablation experiment was conducted to measure the role of each explainability element and interpretability restriction on total model performance. All interpretability modules (Grad-CAM, Integrated Gradients, LIME) were tested separately and combined.

Table 8. Ablation on explainability modules (ISIC 2020 dataset)

Model Configuration	ACC (%)	AUC	IoU (%)	HTS
Baseline CNN (no XAI)	92.4	0.962	–	3.2
+ Grad-CAM	92.7	0.965	57.3	3.8
+ LIME	92.9	0.966	60.5	3.9
+ Integrated Gradients	93.0	0.967	63.1	4.0
+ Grad-CAM + LIME + IG (Fusion)	93.1	0.971	68.9	4.4

Table 9. Impact of interpretability constraint

Configuration	ACC (%)	AUC	IoU (%)	HTS
Without L_exp	92.6	0.968	61.4	3.9
With L_exp (ours)	93.1	0.971	68.9	4.4

Explainability module integration was associated with improved diagnostic accuracy and a substantial rise in the metrics of interpretability (IoU ↑ +11.6% and HTS ↑ +1.2). This proves that interpretability mechanisms do not reduce predictive performance when introduced appropriately. A secondary ablation investigated the effect of including the loss of explanation consistency (L exp) in the overall training goal (in Equation 17 in Section 3.8). The addition of (L exp) resulted in a 7.5 percent increase in the IoU and a 0.5-point increase in HTS, confirming the significance of the imposition of explanation coherence in the training course. Interestingly, the diagnostic metrics were also slightly improved, indicating that improved interpretability indirectly contributes to feature generalization.

5. Discussion

5.1 Interpretation of Main Findings

The results of the current work prove that under proper design of explainability mechanisms, deep learning models guided by explainability mechanisms can reach high diagnostic accuracy and exhibit interpretability. The new CNN+CSAM hybrid and Grad-CAM, LIME, and Integrated Gradients (IG) models demonstrated excellent diagnostic results (AUC 0.95 or higher) on CheXpert and ISIC 2020 datasets and provided consistent and clinically relevant explanations. The findings disprove the long-standing belief that interpretability has to be achieved at the expense of performance and emphasize that interpretability and performance can be optimized together.

Received: August 12 ,2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The most important remark made by the findings is that the combination of several explainability methods resulted in a significant increase in both the metrics' interpretability (IoU = 68.9%) and the metrics' human trust scores (HTS = 4.4/5) without deteriorating the classification performance. This implies that mechanisms of interpretability, combined as a component of the model training process, and not post hoc visualization tools, can lead to improved internal feature representations. Besides, the addition of interpretability regularization (L_exp) had a quantitatively significant impact on diagnostic and interpretability metrics, suggesting that the model's internal reasoning became more organized and consistent with human-understandable patterns. The ROC analysis also supported the discriminative strength of the suggested method, and AUC values were more than 0.95 in such categories of disease as pneumonia, cardiomegaly, and melanoma. The overall result of these findings highlights the possibility of creating explainable yet performant CNN models, and it is one step in the right direction to build reliable and clinically deployable AI-assisted diagnostic systems.

5.2 Clinical Implications of Model Transparency

The clinical usefulness of explainable AI (XAI) is not just in its predictive accuracy but in the fact that it can be used to explain the decisions in a manner clinicians will respect. Dermatologists and radiologists involved in this study consistently reported a greater correlation between model-generated attention maps and the diagnostic areas of interest. The 0.82 Cohen κ value of the agreement in the experts' explanations affirms the reproducibility and reliability of the explanations. The fusion-based explanation offered real-world interpretive clues, visually defining aspects like the margin of lung opacities and irregular lesion pigmentation patterns, which radiologists regularly examine. This openness and clarity of output are essential in augmented clinical processes, where artificial intelligence systems must be decision-support systems and not black-box decision-makers. Higher human trust scores show that model interpretability directly increases clinical acceptance and promotes responsible AI adoption in healthcare environments. Moreover, explainable deep learning fits the regulatory agenda of medical authorities, like the FDA and the EU AI Act, that focus on explainability, responsibility, and human control. Clinicians can justify their decision using medico-legal and ethical contexts with the help of transparent models that make auditability and informed consent easier. The interpretability aspect is thus a technical and ethical requirement that bridges computational intelligence and clinical practice.

5.3 Comparison with Recent Literature

The recent improvements in medical imaging have demonstrated remarkable diagnostic performances with CNNs and transformer-based models. Nevertheless, most current frameworks focus on accuracy rather than interpretability. Architectures such as DenseNet and ViT-B/16 have been studied with high values of AUC (=0.93-0.94) but with no system to provide a consistent visual explanation, which restricted their clinical usage in many studies. However, the hybrid framework (proposed) scored much higher with AUC = 0.951 (CheXpert) and AUC = 0.971 (ISIC), exceeding current benchmarks but still scoring higher on interpretability. Although previous studies used single-method explainability like Grad-CAM or SHAP to obtain qualitative explainability, their heatmap tended to be non-localized or

Received: August 12 ,2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

unstable across samples. The suggested multi-method fusion, combining spatial (Grad-CAM), perturbation-based (LIME), and gradient-based (IG) interpretations, offered strong, relatable, and consistent explanations, which were better aligned with human patterns of reasoning.

The cross-method synthesis also overcomes the issue of fragmentation observed in the previous studies of XAI, in which each method only took a partial view of the modeling behavior. In addition, this study also used quantitative interpretability scores, unlike the previous models, which measured the interpretability qualitatively, like the IoU, Delete /Insertion, and Human Trust Scores. The research provides an objective and more reproducible framework of explainability by associating these metrics with clinical assessments. The observed increase in performance presented by the addition of interpretability regularization (L_{exp}) also corresponds to the recent research results that increasingly restrict attention distributions enhance transparency and generalization.

The proposed model is a comprehensive step forward in explainable AI in medical imaging compared to black-box CNNs or single-method explainers: it is equally good at diagnostics and interpretability.

5.4 Limitations

Although its performance is good, it has several weaknesses. To begin with, there is always dataset bias. It is possible that, though the widely used benchmarks are the CheXpert and ISIC 2020, they do not accurately reflect the global clinical spectrum. The differences in imaging equipment, scan protocols, and patient demographics may impair external validity. As a result, the performance of the models can decrease when applied to underrepresented populations or rare pathologies. Multi-institutional, demographically diverse datasets should be utilized to achieve justification and external validity in future research work. Second, the hybrid explainability method was more interpretable, but it remains a post hoc one. Models such as Grad-CAM and LIME give an approximation of the model reasoning, but not the insight into how the interior decision-making process functions. Thus, descriptions can sometimes emphasize associated artifacts instead of real causal characteristics. A challenge to ensure visual attributions are faithful to model cognition remains. Adding generally interpretable architectures or self-explanatory neural networks can address this weakness.

Third, quantitative measures of interpretability like IoU and Human Trust Scores, although helpful, cannot provide a complete image of cognitive and behavioral elements of clinical interpretation. Visual familiarity and confidence bias may affect human trust and are immeasurable. Computational metrics should be followed by user-centred usability research in the future to gain a deeper insight into the influence of interpretability in the real-world diagnostic choices. Lastly, there is a trade-off between interpretability and efficiency because the computational cost of running several explainability algorithms simultaneously is high. Grad-CAM, LIME, and IG increase inference time when used together, making them unsuitable in time-sensitive healthcare settings like emergency radiology. Future applications may consider even lightweight explainability modules that can be optimized to achieve real-time interpretability without going too far.

Received: August 12 ,2025

International Journal of Applied Mathematics
Volume 38 No. 8s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

5.5 Future Research Directions

Based on these results, several promising directions can be developed to explainable deep learning of medical image diagnosis. First, multimodal integration is a major frontier with future models integrating the imaging data with the electronic health records (EHRs), genetic profiles, and clinical stories to produce context-based explanations. Models can be linked to visual patterns with patient metadata to bring about clinically grounded interpretability, enhance diagnostic reasoning, and build trust in clinicians. Second, clinician-in-the-loop systems can facilitate adaptive learning by integrating expert feedback in the training and evaluation phase. This form of interaction enables models to improve attention maps, refute misleading explanations, and encourage team learning between AI and medics. Third, intrinsic interpretability architecture, e.g., prototype-based or capsule network, must be considered to incorporate interpretability directly into the learning process, avoid relying on post hoc explanation techniques, and improve the verisimilitude of model reasoning.

Besides, it is fundamental to standardise interpretability measures for reproducibility and equal benchmarking of XAI investigations. Developing cohesive evaluation guidelines that would include faithfulness, localization accuracy, and user trust calibration would benefit scientific rigor and regulatory preparedness. Ethically, it is essential to conform to regulatory systems; explicable models may be used to facilitate transparency, accountability, and compliance in clinical AI auditing and implementation. Lastly, cross-domain generalization is critical to its relevance to the real world. The ability to perform model validation in various modalities, including CT, MRI, ultrasound, domain adaptation, and self-supervised learning, would help reinforce robustness and scalability without negatively affecting interpretability.

6. Conclusion

The paper introduced a deep learning framework that is explainable to solve the gap between predictive accuracy and interpretability in medical image diagnosis. The proposed model, which is based on the ResNet-50 backbone and a Channel-Spatial Attention Module (CSAM), showed that high diagnostic performance is not related to the reduction of transparency. The architecture used a composite loss, a combination of classification accuracy and a term of interpretability regularization, $L_{\text{exp}} = \| H_{\text{Grad-CAM}} - H_{\text{IG}} \|_2^{2,}$ and weighted by α=0.2. This formulation effectively promoted consistency across various explanation maps so that model attention was consistent across modalities, cases, and diagnostically relevant image regions. Two large-scale benchmark datasets, CheXpert to classify thoracic diseases and ISIC 2020 to recognize skin lesions, were evaluated experimentally using the same training settings. The model had AUC values of 0.951 on CheXpert and 0.971 on ISIC 2020, which is higher than several strong baselines such as DenseNet-121, EfficientNet-B4, and Vision Transformer (ViT). This enhanced feature localization and discrimination by dynamically recalibrating spatial and channel-wise information with the addition of the CSAM. Furthermore, the 12.4% (measured through Grad-CAM/IG consistency) reinforcement of the stability of the explanations relative to the non-existent non-regularization indicates the success of the suggested formulation of the losses.

Received: August 12 ,2025

International Journal of Applied Mathematics
Volume 38 No. 8s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The model combined Grad-CAM, LIME, and Integrated Gradients (IG) for explainability, producing supplementary visualization viewpoints. Grand-CAM was useful in localizing coarse spatial features related to a pathological result, which include lung opacities and lesion boundaries. In contrast, LIME provided fine-grained information about texture-level structures and edge patterns. The resulting fused explainability maps combined both advantages and provided spatially consistent and clinically interpretable highlights closely matched by expert-annotated pathology areas. This combination increased Intersection over Union (IoU) to 68.9% and Human Trust Score (HTS) to 4.4 out of 5, which suggested that Intersection was more consistent with clinician thought and more interpretable. One of the observations was that the improved interpretability did not occur at the expense of diagnostic accuracy. Instead, the regularized model was more generalized in terms of fold, and the variance between validation AUC scores was less than that of 5-fold stratified cross-validation. In the case of early stopping and weight checkpointing based on validation AUC, solid convergence of models was achieved. The stability of the model was confirmed by the consistency in the results between folds and their ability to be generalized to patient sub-groups that have not been observed.

Clinically, radiologists and dermatologists who took part in the expert examination stage claimed that the model's visual explanations were accurate and relevant within the context. The coefficient of Cohen 0.82 in the comparison of AI-generated and expert interpretability maps indicates a high degree of human-AI compatibility. This numerical validation confirms the possible clinical implementation of the model as a decision-support model that can improve diagnostic reasoning, not just automating prediction. The explainability module, therefore, acted as a diagnostic amplifier, which helped clinicians to confirm and contextualize AI-generated insights. In addition to its empirical input, the study contributes to the theoretical knowledge of how interpretability could be structurally embedded in the neural networks. As opposed to post hoc explanation-only methods, the use of L_{exp} as a training constraint interprets interpretability as a property that can be learned instead of being an add-on that is considered retrospectively. This paradigm shift is part of a new set of self-explanatory models, a trend necessary for the future of trustworthy AI in healthcare.

The moral and legal concerns of this piece are also huge. As the world tends to pay more attention to AI transparency and accountability, e.g., in the FDA of the U.S. in the form of the Good Machine Learning Practice (GMLP), and in the European Union in the form of the AI Act, this framework will facilitate compliance by generating auditable, human-readable output. The fact that the model can defend its choices enhances confidence between clinicians and helps them adopt it responsibly in the regulated medical workflow. This is most necessary in areas of patient diagnosis like radiology and dermatology, in which explanatory ability directly impacts patient safety and clinician liability. However, there are still some shortcomings. Both interpretability and predictive accuracy may be affected by dataset bias and domain-specific variations. Despite being solid, the present assessment is restricted to X-ray and dermoscopy. The framework should be verified on various imaging modalities - MRI, CT, and ultrasound - to determine its scalability and domain applicability in future studies. Also, although Grad-CAM and LIME explanations have complementary advantages, they are all approximations of

Received: August 12 ,2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

model reasoning, so creating architectures whose interpretation is intrinsically interpretable (like prototype-based or capsule models) can offer even more accurate transparency.

In the future, there are several research directions to consider. Multimodal data, including imaging and electronic health records (EHR) in conjunction with clinical narratives, may help develop more detailed and context-sensitive descriptions and enhance the depth of reasoning. Adaptive human-guided mechanisms of interpretability can be developed by developing clinician-in-the-loop systems that allow feedback during model training. In addition, developing standard interpretability standards that measure faithfulness, localization accuracy, and user trust will improve reproducibility and comparability across explainable AI studies. To sum up, it can be seen that deep learning models can be accurate and interpretable, provided that human-centered principles are used to design them. The hybrid CNN+CSAM framework proposed, with the assistance of explanation consistency regularization and support of the fused Grad-CAM–LIME–IG interpretability, will provide clinically reliable and interpretable diagnostic information. This study contributes to trusting, open, and ethically responsible medical AI systems by aligning high-performance AI with explainability and moving towards safe and effective clinical decision-making based on medical AI in practice.

References

[1] Alexiuk, M., Elgubtan, H., & Tangri, N. (2024, January 1). Clinical Decision Support Tools in the Electronic Medical Record. *Kidney International Reports*. Elsevier Inc. https://doi.org/10.1016/j.ekir.2023.10.019

[2] Araujo, S. M., Sousa, P., & Dutra, I. (2020, October 1). Clinical decision support systems for pressure ulcer management: Systematic review. *JMIR Medical Informatics*. JMIR Publications Inc. https://doi.org/10.2196/21621

[3] Avramidis, G. P., Avramidou, M. P., & Papakostas, G. A. (2022, January 1). Rheumatoid Arthritis Diagnosis: Deep Learning vs. Human. *Applied Sciences (Switzerland)*. MDPI. https://doi.org/10.3390/app12010010

[4] Beeler, P. E., Bates, D. W., & Hug, B. L. (2014). Clinical decision support systems. *Swiss Medical Weekly*. EMH Schweizerischer Ärzteverlag AG. https://doi.org/10.4414/smw.2014.14073

[5] Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020). Explainable AI in Fintech Risk Management. *Frontiers in Artificial Intelligence*, *3*. https://doi.org/10.3389/frai.2020.00026

[6] Chaddad, A., Peng, J., Xu, J., & Bouridane, A. (2023, January 1). Survey of Explainable AI Techniques in Healthcare. *Sensors*. MDPI. https://doi.org/10.3390/s23020634

[7] Chen, X., Wang, X., Zhang, K., Fung, K. M., Thai, T. C., Moore, K., … Qiu, Y. (2022, July 1). Recent advances and clinical applications of deep learning in medical image analysis. *Medical Image Analysis*. Elsevier B.V. https://doi.org/10.1016/j.media.2022.102444

[8] Dang, T., Nguyen, T. T., McCall, J., Elyan, E., & Moreno-García, C. F. (2024). Two-layer Ensemble of Deep Learning Models for Medical Image Segmentation. *Cognitive Computation*, *16*(3), 1141–1160. https://doi.org/10.1007/s12559-024-10257-5

Received: August 12 ,2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[9] Demner-Fushman, D., Chapman, W. W., & McDonald, C. J. (2009, October). What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*. https://doi.org/10.1016/j.jbi.2009.08.007

[10] Gayap, H. T., & Akhloufi, M. A. (2024, March 1). Deep Machine Learning for Medical Diagnosis, Application to Lung Cancer Detection: A Review. *BioMedInformatics*. Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/biomedinformatics4010015

[11] Guo, J., Cao, W., Nie, B., & Qin, Q. (2023). Unsupervised Learning Composite Network to Reduce Training Cost of Deep Learning Model for Colorectal Cancer Diagnosis. *IEEE Journal of Translational Engineering in Health and Medicine*, *11*, 54–59. https://doi.org/10.1109/JTEHM.2022.3224021

[12] Holzinger, A. (2021). Explainable AI and Multi-Modal Causability in Medicine. *I-Com*, *19*(3), 171–179. https://doi.org/10.1515/icom-2020-0024

[13] Jiang, X., Hu, Z., Wang, S., & Zhang, Y. (2023, July 1). Deep Learning for Medical Image-Based Cancer Diagnosis. *Cancers*. Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/cancers15143608

[14] Kazerouni, A., Aghdam, E. K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., & Merhof, D. (2023, August 1). Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*. Elsevier B.V. https://doi.org/10.1016/j.media.2023.102846

[15] Li, J., Chen, J., Tang, Y., Wang, C., Landman, B. A., & Zhou, S. K. (2023, April 1). Transforming medical imaging with Transformers? A comparative review of key properties, current progress, and future perspectives. *Medical Image Analysis*. Elsevier B.V. https://doi.org/10.1016/j.media.2023.102762

[16] Liu, S., Wright, A. P., Patterson, B. L., Wanderer, J. P., Turer, R. W., Nelson, S. D., … Wright, A. (2023). Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *Journal of the American Medical Informatics Association*, *30*(7), 1237–1245. https://doi.org/10.1093/jamia/ocad072

[17] Mahmood, T., Rehman, A., Saba, T., Nadeem, L., & Bahaj, S. A. O. (2023). Recent Advancements and Future Prospects in Active Deep Learning for Medical Image Segmentation and Classification. *IEEE Access*. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2023.3313977

[18] Marmolejo-Saucedo, J. A., & Kose, U. (2024). Numerical Grad-Cam Based Explainable Convolutional Neural Network for Brain Tumor Diagnosis. *Mobile Networks and Applications*, *29*(1), 109–118. https://doi.org/10.1007/s11036-022-02021-6

[19] Mary Shyni, H., & Chitra, E. (2022, January 1). A COMPARATIVE STUDY OF X-RAY AND CT IMAGES IN COVID-19 DETECTION USING IMAGE PROCESSING AND DEEP LEARNING TECHNIQUES. *Computer Methods and Programs in Biomedicine Update*. Elsevier B.V. https://doi.org/10.1016/j.cmpbup.2022.100054

[20] Mazo, C., Kearns, C., Mooney, C., & Gallagher, W. M. (2020, February 1). Clinical decision support systems in breast cancer: A systematic review. *Cancers*. MDPI AG. https://doi.org/10.3390/cancers12020369

Received: August 12 ,2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[21] Moujahid, H., Cherradi, B., Al-Sarem, M., Bahatti, L., Eljialy, A. B. A. M. Y., Alsaeedi, A., & Saeed, F. (2022). Combining CNN and Grad-CAM for COVID-19 disease prediction and visual explanation. *Intelligent Automation and Soft Computing*, *32*(2), 723–745. https://doi.org/10.32604/iasc.2022.022179

[22] Mridha, K., Uddin, M. M., Shin, J., Khadka, S., & Mridha, M. F. (2023). An Interpretable Skin Cancer Classification Using Optimized Convolutional Neural Network for a Smart Healthcare System. *IEEE Access*, *11*, 41003–41018. https://doi.org/10.1109/ACCESS.2023.3269694

[23] Najjar, R. (2023, September 1). Redefining Radiology: A Review of Artificial Intelligence Integration in Medical Imaging. *Diagnostics*. Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/diagnostics13172760

[24] Nazir, S., & Kaleem, M. (2023, May 1). Federated Learning for Medical Image Analysis with Deep Neural Networks. *Diagnostics*. Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/diagnostics13091532

[25] O'Sullivan, D., Fraccaro, P., Carson, E., & Weller, P. (2014). Decision time for clinical decision support systems. *Clinical Medicine, Journal of the Royal College of Physicians of London*, *14*(4), 338–341. https://doi.org/10.7861/clinmedicine.14-4-338

[26] Olabanjo, O., Wusu, A., Asokere, M., Afisi, O., Okugbesan, B., Olabanjo, O., … Mazzara, M. (2023, September 1). Application of Machine Learning and Deep Learning Models in Prostate Cancer Diagnosis Using Medical Images: A Systematic Review. *Analytics*. Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/analytics2030039

[27] Pandey, B., Kumar Pandey, D., Pratap Mishra, B., & Rhmann, W. (2022, September 1). A comprehensive survey of deep learning in medical imaging and natural language processing: Challenges and research directions. *Journal of King Saud University - Computer and Information Sciences*. King Saud bin Abdulaziz University. https://doi.org/10.1016/j.jksuci.2021.01.007

[28] Rajabi, E., & Kafaie, S. (2022, October 1). Knowledge Graphs and Explainable AI in Healthcare. *Information (Switzerland)*. MDPI. https://doi.org/10.3390/info13100459

[29] Rong, Y., Leemann, T., Nguyen, T. T., Fiedler, L., Qian, P., Unhelkar, V., … Kasneci, E. (2024). Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(4), 2104–2122. https://doi.org/10.1109/TPAMI.2023.3331846

[30] Saeed, W., & Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, *263*. https://doi.org/10.1016/j.knosys.2023.110273

[31] Salehi, A. W., Khan, S., Gupta, G., Alabduallah, B. I., Almjally, A., Alsolai, H., … Mellit, A. (2023, April 1). A Study of CNN and Transfer Learning in Medical Imaging: Advantages, Challenges, Future Scope. *Sustainability (Switzerland)*. MDPI. https://doi.org/10.3390/su15075930

[32] Sanjay, H. S., & Niranjanamurthy, M. (2023). *Medical Imaging*. *Medical Imaging* (pp. 1–240). wiley. https://doi.org/10.1093/combul/29.3.5

Received: August 12 ,2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[33] Saraswat, D., Bhattacharya, P., Verma, A., Prasad, V. K., Tanwar, S., Sharma, G., … Sharma, R. (2022). Explainable AI for Healthcare 5.0: Opportunities and Challenges. *IEEE Access*. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2022.3197671

[34] Saw, S. N., & Ng, K. H. (2022, August 1). Current challenges of implementing artificial intelligence in medical imaging. *Physica Medica*. Associazione Italiana di Fisica Medica. https://doi.org/10.1016/j.ejmp.2022.06.003

[35] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, *128*(2), 336–359. https://doi.org/10.1007/s11263-019-01228-7

[36] Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., & Fu, H. (2023, August 1). Transformers in medical imaging: A survey. *Medical Image Analysis*. Elsevier B.V. https://doi.org/10.1016/j.media.2023.102802

[37] Shurrab, S., & Duwairi, R. (2022). Self-supervised learning methods and applications in medical imaging analysis: a survey. *PeerJ Computer Science*, *8*. https://doi.org/10.7717/PEERJ-CS.1045

[38] Singh, S. P., Wang, L., Gupta, S., Goli, H., Padmanabhan, P., & Gulyás, B. (2020, September 2). 3d deep learning on medical images: A review. *Sensors (Switzerland)*. MDPI AG. https://doi.org/10.3390/s20185097

[39] Sistaninejhad, B., Rasi, H., & Nayeri, P. (2023). A Review Paper about Deep Learning for Medical Image Analysis. *Computational and Mathematical Methods in Medicine*. Hindawi Limited. https://doi.org/10.1155/2023/7091301

[40] Sittig, D. F., Wright, A., Osheroff, J. A., Middleton, B., Teich, J. M., Ash, J. S., … Bates, D. W. (2008). Grand challenges in clinical decision support. *Journal of Biomedical Informatics*, *41*(2), 387–392. https://doi.org/10.1016/j.jbi.2007.09.003

[41] Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker, K. I. (2020, December 1). An overview of clinical decision support systems: benefits, risks, and strategies for success. *Npj Digital Medicine*. Nature Research. https://doi.org/10.1038/s41746-020-0221-y

[42] Suzuki, K. (2017, September 1). Overview of deep learning in medical imaging. *Radiological Physics and Technology*. Springer Tokyo. https://doi.org/10.1007/s12194-017-0406-5

[43] Teufel, A., & Binder, H. (2021, December 1). Clinical Decision Support Systems. *Visceral Medicine*. S. Karger AG. https://doi.org/10.1159/000519420

[44] van Zyl, C., Ye, X., & Naidoo, R. (2024). Harnessing eXplainable artificial intelligence for feature selection in time series energy forecasting: A comparative analysis of Grad-CAM and SHAP. *Applied Energy*, *353*. https://doi.org/10.1016/j.apenergy.2023.122079

[45] Wang, Y., Ge, X., Ma, H., Qi, S., Zhang, G., & Yao, Y. (2021). Deep Learning in Medical Ultrasound Image Analysis: A Review. *IEEE Access*, *9*, 54310–54324. https://doi.org/10.1109/ACCESS.2021.3071301

Received: August 12 ,2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[46] Yang, C., Lan, H., Gao, F., & Gao, F. (2021, March 1). Review of deep learning for photoacoustic imaging. *Photoacoustics*. Elsevier GmbH. https://doi.org/10.1016/j.pacs.2020.100215

[47] Yu, Y., Li, M., Liu, L., Li, Y., & Wang, J. (2019, December 1). Clinical big data and deep learning: Applications, challenges, and future outlooks. *Big Data Mining and Analytics*. Tsinghua University Press. https://doi.org/10.26599/BDMA.2019.9020007

[48] Zakareya, S., Izadkhah, H., & Karimpour, J. (2023). A New Deep-Learning-Based Model for Breast Cancer Diagnosis from Medical Images. *Diagnostics*, *13*(11). https://doi.org/10.3390/diagnostics13111944

[49] Zhang, Y., Hong, D., McClement, D., Oladosu, O., Pridham, G., & Slaney, G. (2021). Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging. *Journal of Neuroscience Methods*, *353*. https://doi.org/10.1016/j.jneumeth.2021.109098

[50] Zheng, D., He, X., & Jing, J. (2023, January 1). Overview of Artificial Intelligence in Breast Cancer Medical Imaging. *Journal of Clinical Medicine*. MDPI. https://doi.org/10.3390/jcm12020419

Received: August 12 ,2025