International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Bridging the AI Explainability Gap in Cardiac Imaging: A Review of Hybrid Approaches Using Grad-CAM and SHAP

Umar Sani Dabai[1, 2], **Moussa Mahamat Boukar**[3], **Muhammad Aliyu Suleiman**[4]

Submitted:11/03/2024 **Revised**: 26/04/2024 **Accepted**: 03/05/2024

Abstract: The growing adoption of deep learning within the domain of medical imaging has yielded substantial enhancements in the capabilities of computer-assisted diagnosis and prognosis. However, the ability to interpret and explain the decisions made by these models continues to pose a significant obstacle, which impedes their successful integration into clinical practice. This review paper explores a hybrid approach that leverages the Gradient-weighted Class Activation Mapping technique and the Shapley Additive Explanations technique to bridge the artificial intelligence explainability gap in cardiac imaging. The paper discusses the strengths and limitations of these techniques, their application in cardiac imaging, and the potential for integrating them into a machine learning pipeline for robust and trustworthy artificial intelligence systems. Furthermore, it emphasizes the significance of developing artificial intelligence systems that are clinically translatable, addressing the explainability gap between clinical experts and non-experts. This ensures wider inclusion of diverse stakeholders involved in patient care, ultimately leading to improved patient outcomes and enhanced trust in AI-driven healthcare solutions.

Keywords: *Cardiac Imaging, Explainable AI, Grad-CAM; SHAP*

Introduction

Cardiovascular diseases have been a major death-leading cause worldwide, driving extensive research efforts to enhance diagnostic and prognostic capabilities [1], [2]. Technological breakthrough within the artificial intelligence (AI) domain have revolutionized the field of cardiac imaging, enabling more efficient and personalized patient care [3]. However, the opaque decision-making process of numerous AI models presents a substantial obstacle, hindering the absolute adoption of these techniques in clinical practice [1]. Explainable Artificial Intelligence (XAI), a recent paradigm, aims to address the issue of AI model opacity through making the decision processes more transparent and understandable [4].

Over the past decade, scholars have developed various XAI models to improve the interpretability of diverse machine learning (ML) algorithms, with a particularly focus on the domain of cardiac imaging. According to [5], Gradient-Weighted Class Activation Mapping (Grad-CAM), SHapley Additive exPlanations (SHAP), and Local Interpretable Model-Agnostic Explanations (LIME) are identified as three of the most effective XAI techniques employed in this domain. [5]. While these XAI techniques offer promising outcomes, it is apparent that they possess distinct limitations and drawbacks [6]. For example, saliency-based methods like Grad-CAM can have problems like localization errors and insensitivity to specific model architectures [5] as well as occlusion [7]. Similarly, Patricio et al. [8] posit that perturbation-based model explanations like SHAP may struggle to adequately capture the complex, non-linear relationships inherent to medical imaging data. A further limitation of these techniques is the challenge in quantifying the Grad-CAM output, despite its demonstrated effectiveness in interpreting and visualizing deep learning models [8]. Additionally, SHAP can be computationally demanding and less scalable when applied to high-dimensional data [9], [10], [11].

To address these limitations and leverage the strengths of these popular techniques, a number of researchers have attempted a hybrid model for enhanced interpretability [8], [12]. This study presents a non-exhaustive literature review to evaluate the effectiveness of a hybrid framework that integrates Grad-CAM and SHAP to augment the interpretability of deep learning models (DLM) applied to the domain of cardiac imaging. The goal is to bridge the divide that lies between AI-driven decisions and the understanding of clinical procedures among both expert and non-expert users.

This study offers a key contribution by addressing the fundamental need for intelligible and transparent explanations of AI-driven decisions in medical applications. Furthermore, the study identifies specific challenges associated with applying the SHAP technique to complex cardiac imaging data, such as computational demands and visual complexity. The proposed hybrid approach, which combines Grad-CAM and SHAP, is shown to mitigate these limitations, presenting a more practical and effective XAI

[1] *Department of Information Technology, Federal University Dutse, Jigawa, Nigeria*
[2,3,4] *Department of Computer Science, Faculty of Computing, Nile University of Nigeria, Abuja,, Nigeria*
[1,2] *ORCID ID : 0009-0003-4201-2333*
[3] *ORCID ID : 0000-0002-2494-2698*
[4] *ORCID ID : 0000-0003-0465-8689*
* *Corresponding Author Email: usdabai@yahoo.com*

solution. Finally, recognizing the significance of robust evaluation, this work emphasizes the importance of developing methodologies to quantify the impact of hybrid XAI models on clinical judgment and patient outcomes in the domain of cardiac imaging, paving the way for future research in this critical area.

Explainable AI (XAI)

XAI is a rapidly emerging field that focuses on developing AI models which are not only accurate but also transparent and interpretable [13], [14]. The ultimate goal of XAI is to create AI systems that can explain their decision-making process in a way that humans can understand and trust [15], [16], [17]. This is particularly crucial in sensitive domains like healthcare, where AI-based decisions can have significant impact on patient outcomes [18], [19].

Explainability AI Taxonomy

Researchers [19], [20] have broadly categorized the XAI models into two, intrinsic and post-hoc methods. While intrinsic methods like decision trees, linear regression, and logistic regression are inherently interpretable and are considered "ante-hoc" or "transparent" XAI models [20], [21] post-hoc methods are "black box" models that require additional explanations to understand the reasoning behind the predictions [20], [22]. The latter category is further subdivided into two, model-specific and model-agnostics.

XAI explanations can be further classified into two primary categories according to the scope of the explanation: global and local. An explanation focused on a specific prediction or output is considered a local perspective, whereas an explanation encompassing the entirety of the model is regarded as a global perspective [20]. This work references the comprehensive taxonomy of XAI models adopted in [20] work as summarized and illustrated in Figure 1.

Grad-CAM and SHAP as XAI Techniques

The two XAI techniques central to this review are Grad-CAM and SHAPley Additive Values.

Gradient-Weighted Class Activation Mapping

Gradient-Weighted Class Activation Mapping (Grad-CAM) is a popular visual explanation technique that generates saliency maps to highlight the regions in an input image that are most influential in the model's decision-making process [21], [23].

Grad-CAM technique essentially uses the gradients of the target class flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the input image [24]. Explanations in Grad-CAM are considered local as they provide insights into the model's logic for a specific input [25]. Although Grad-CAM is a powerful XAI tool, it lacks the ability to quantify its output, which can be challenging for both experts and non-experts to accurately interpret the results [26].

SHapley Additive exPlanations

SHAPley Additive Values is a model-agnostic technique that provides both local and global explanations [27]. SHAP values quantify the contribution of each feature to the model's output for a specific instance, providing a detailed understanding of how the model arrived at a particular prediction [27], [28]. SHAP technique emerged from the game theory concept of Shapley values, which provides a principled way to allocate the output of a model among its input features [29]. SHAP values have been widely adopted in the field of XAI due to their theoretical grounding and ability to handle complex, non-linear relationships [30]. Comparative studies of SHAP, LIME and other permutation-based models have demonstrated the superior performance of SHAP in terms of accuracy, consistency, and computational efficiency [31], [32].

Explainability in AI-Driven Cardiac Imaging Diagnosis

In researching the intersection of AI and healthcare, particularly in cardiovascular diagnosis, it becomes crucial to explore how explainability can enhance trust and understanding among diverse stakeholders [29]. Gunning et al. [29] further state that the establishment of trust is pivotal for the smooth integration of AI-driven technologies into clinical practice. This trust promotes collaborative partnerships between healthcare professionals and patients, ensuring that critical medical decisions are informed by transparent and interpretable data [19], [30].

In recent years, Grad-CAM has gained significant attention as a visualization technique that facilitates the interpretation of convolutional neural network decisions, particularly within the medical imaging domain [33], [34]. Jahmunah et al. and Sakai et al. [35], [36] add that the technique has proven to be invaluable in providing insights into the areas of an image that most substantially contribute to a model's predictions, thereby enabling clinicians to better comprehend and validate AI-driven assessments. Although Grad-CAM visualizations have revealed considerable promise, challenges persist in providing coherent explanations that can be comprehended by all stakeholders engaged in patient care. A very serious concern of these visual interpretable models is their lack of quantitative evaluation metrics, which complicates the process of assessing their reliability and effectiveness in clinical settings [34]. Other challenges associated with heatmap-based interpretable methods, like Grad-CAM, include the potential for misinterpretation of the results, which can lead to erroneous conclusions that may adversely impact patient outcomes [37].

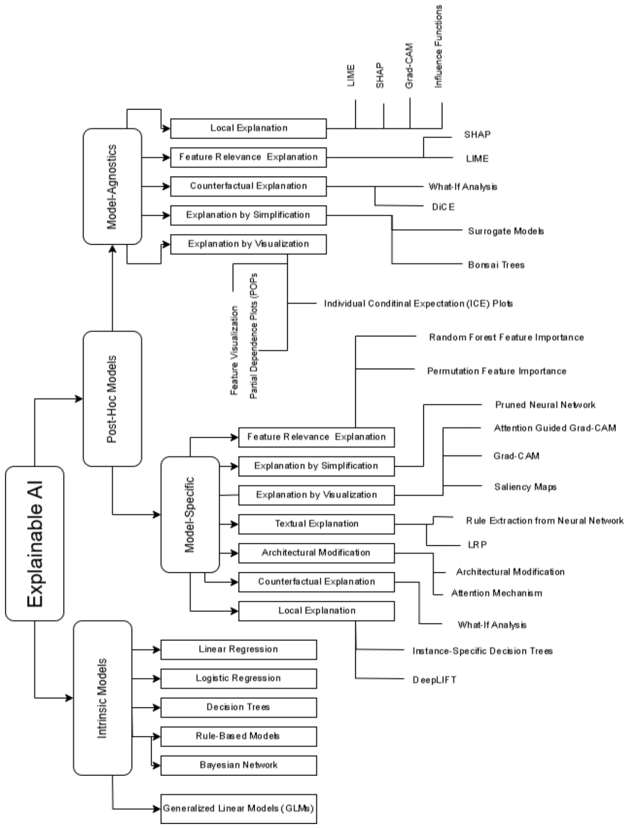
Explainable AI
Model-Agnostics
Local Explanation
Feature Relevance Explanation
Counterfactual Explanation
Explanation by Simplification
Explanation by Visualization
LIME
SHAP
Grad-CAM
Influence Functions
SHAP
LIME
What-If Analysis
DiCE
Surrogate Models
Bonsai Trees
Feature Visualization
Partial Dependence Plots (PDPs
Individual Conditinal Expectation (ICE) Plots
Post-Hoc Models
Random Forest Feature Importance
Permutation Feature Importance
Model-Specific
Feature Relevance Explanation
Explanation by Simplification
Explanation by Visualization
Textual Explanation
Architectural Modification
Counterfactual Explanation
Local Explanation
Pruned Neural Network
Attention Guided Grad-CAM
Grad-CAM
Saliency Maps
Rule Extraction from Neural Network
LRP
Architectural Modification
Attention Mechanism
What-If Analysis
Instance-Specific Decision Trees
DeepLIFT
Intrinsic Models
Linear Regression
Logistic Regression
Decision Trees
Rule-Based Models
Bayesian Network
Generalized Linear Models (GLMs)

Fig. 1. An illustration of Comprehensive XAI Taxonomy

SHapley Additives values exPlanations (SHAP), on the other hand, offers a more robust approach to understanding model predictions by providing consistent and theoretically grounded explanations [38], [39]. Dewi et al. and Sheu et al. [24], [27] describe SHAP as a XAI technique that quantifies the individual contribution of each input feature to the model's predictions, allowing clinicians to understand more the underlying factors critical to model's decision. SHAP explanations are more intuitive and easier to communicate to stakeholders as every feature's impact is presented in a clear and comprehensible manner, facilitating informed decision-making in clinical settings [22].

A Hybrid XAI Approach for Cardiac Imaging Explainability

Salau et al. [40] characterize medical data as extensive and intricate, which poses obstacles to efficient analysis and interpretation. Hou et al. and Li et al.'s [11], [41] study describe the heterogeneous nature of cardiac data and the intricacy of DLM as the instrumental factors that necessitated the development of various XAI techniques to address the black-box problem within this domain. Several studies have explored the individual application of Grad-CAM [25], [42], [43], [44] and SHAP [9], [45], [46], [47], [48] in the domain of cardiac imaging, outlining their distinct drawbacks and caveats [9], [10]. Recognizing the complementary nature of these techniques, researchers have proposed hybrid approaches that leverage the strengths of both methods [49], [50] or either of them with other methods to enhance the interpretability of AI-driven cardiac diagnostics [51].

The existing literature features studies that have paired Grad-CAM or SHAP with other techniques to augment the interpretability of AI-driven explanations in cardiac imaging applications. For instance, [52] utilized SHAP and LIME in their study to understand the rationale behind the outcome of their stroke prediction model. [53] employed Grad-CAM and LIME to elucidate how a DLM's uncovers the Aortic elongation of chest radiographic images. Additionally, [12] deployed Tree SHAP technique and extreme gradient boosting technique to evaluate their model for predicting the risk of fatal or non-fatal cardiovascular events among individuals diagnosed with Type 2 Diabetes Mellitus.

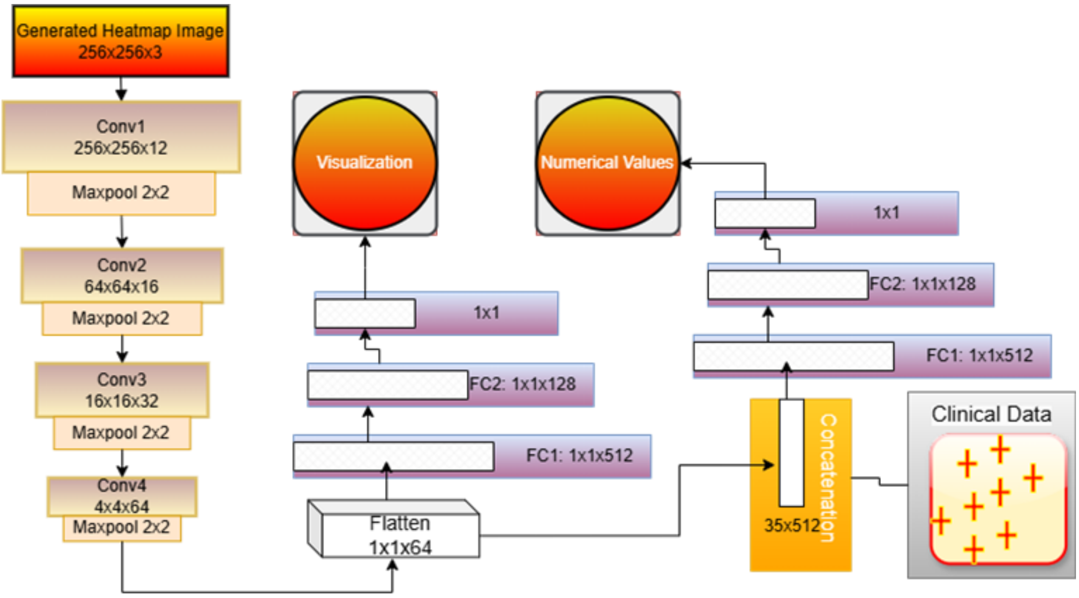

Fig. 2. A description of the hybrid DL model with two explainability display windows

The study by Teuho et al. [54] provides a representative example of the hybrid setup discussed above. They developed a deep learning-based classifier that utilizes polar map images to identify flow-limiting coronary artery disease for detection of ischemia. Figure 2 mimics their hybrid approach where the model's rationale is presented through two distinct output windows. One window provides visual explanations, while the other offers numerical values to elucidate the model's decision-making process.

According to [54], the image classification model, illustrated in figure 2, accepts a JPEG polar map only as input and employs Grad-CAM technique to generate visual images. The polar map data undergoes feature extraction through convolutional layers prior to the initial flattening layer. These extracted features are then fed into a dual-input model, which also incorporates raw tabulated clinical data as a secondary input. The image features and tabulated data are concatenated, and an NN conducts the final classification task [54].

The remaining parts of this paper are structured as follows: Section III outlines the method and materials used to select articles that specifically utilized both Grad-CAM and SHAP techniques. Section IV presents the discussion and research challenges. Finally, the paper concludes with Section V.

Materials and Method

This section describes the methodology employed in this non-exhaustive literature review on the hybrid use of Grad-CAM and SHAP in cardiac imaging. The eligibility inclusion criteria, article identification process, and data extraction are explained.

Inclusion Criteria

To gather relevant data from various studies in a comprehensive and unbiased manner, aligned with the aims of this paper, the following criteria were established:

i. Only research papers written in English are eligible for selection.

ii. The study must focus on cardiovascular disorders and utilize a hybrid interpretable model that employs the SHAP XAI technique and Grad-CAM. Additionally, studies that employ more than these two XAI tools are also considered eligible.

iii. The publications must be released between 2018 and 2024.

iv. The research papers should provide details on the dataset used, including its source and size.

Identifying Potential Research Articles

In October 2024, the researchers conducted a literature search to identify relevant studies on the application of explainable deep learning techniques in the cardiovascular domain. An initial exploratory search on the PubMed research database [55]using the keywords "Explainable Deep Learning AND Cardiovascular" was performed on October 19, 2024. This preliminary search yielded a limited number of articles, which were then subject to further refinement through the application of inclusion and exclusion criteria. As the initial search results were potentially insufficient, the researchers employed a more targeted search query, "Grad-CAM AND Cardiovascular", as outlined in the table 1, to obtain a more satisfactory set of relevant studies. On the same day, an additional targeted search was conducted in the ScienceDirect research database [56] using the query "Grad-CAM AND Cardiovascular AND Explainable AND SHAP" to further identify relevant studies. Subsequently, the Semantic Scholar research database [57] was also queried on October 20, 2024, using the same search criteria to supplement the pool of potentially relevant literature.

Included Studies

The literature search yielded a total of 87 studies, with 17 from PubMed, 46 from ScienceDirect, and 24 from Semantic Scholar.

A systematic screening process was followed to identify relevant studies:

1. Duplicate Removal: Two duplicate records were excluded, reducing the total to 85 studies.
2. Relevance Screening: Twenty-six articles were removed due to lack of relevance to cardiovascular diseases, leaving 59 studies.
3. Availability Screening: Two inaccessible studies were further excluded, resulting in 57 remaining studies.
4. Exclusion of Review Articles: An additional 20 review papers were removed, leaving 37 studies for further evaluation.
5. Final Selection: After a comprehensive abstract review, only six studies met the inclusion criteria by employing both Grad-CAM and SHAP as hybrid XAI techniques for enhanced interpretability. Among these, two studies also incorporated the LIME method alongside Grad-CAM and SHAP.

Table 1. Search Equations

#	Database URL Date Accessed	Search Words (Query)	Number of Papers
1	PubMed https://pubmed.ncbi.nlm.nih.gov 19th October 2024	“Explainable Deep Learning AND cardiovascular”	3
		"Grad-CAM AND Cardiovascular"	14
2	Science Direct https://www.sciencedirect.com 19th October 2024	“Grad-CAM AND Cardiovascular AND Explainable AND SHAP"	46
3	Semantic Scholar https://www.semanticscholar.org 20th October 2024	“Grad-CAM AND Cardiovascular AND Explainable AND SHAP”.	24

As reported in Table 2, a total of 31 studies were excluded based on their methodological approaches. Specifically, the excluded studies employed either Grad-CAM alone, SHAP alone, Grad-CAM in combination with other XAI techniques while excluding

Table 2. Inclusion and exclusion criteria

Database	Search word	Search result	Not retrieved	Non-Cardiovascular	Duplicates	Review Articles	Non-Hybridized Grad-CAM and	Grad-CAM and SHAP	Purely Grad-CAM	Grad-CAM and others	Purely SHAP	SHAP and Others	Others	Total Non-Hybridized Grad-
PubMed	Explainable Deep Learning AND cardiovascular	3	0	2	0	0	15	0	13	2	0	0	0	**15**
	Grad-CAM AND Cardiovascular	14	0	0										
Science Direct	Grad-CAM AND Cardiovascular AND Explainable AND SHAP	46	0	16	0	19	7	4	1	0	3	1	2	**7**
Semantic Scholar	Grad-CAM AND Cardiovascular AND Explainable AND SHAP	24	2	8	2	1	9	2	2	2	4	1	0	**9**
Total		**87**	**2**	**26**	**2**	**20**	**31**	**6**	**16**	**4**	**7**	**2**	**2**	**31**
Remaining		**87**	**85**	**59**	**57**	**37**	**6**							

SHAP, SHAP in conjunction with other interpretable models while excluding Grad-CAM, or neither Grad-CAM nor SHAP.

Following the final stage of the screening process, the number of eligible studies remained unchanged at six.

Data Extraction

The following parameters were considered important for assessing the literature: author name, year of publication, dataset type and size, open-source dataset use, region/country, algorithms, classifiers, disease type predicted, XAI techniques used, and algorithm and XAI tool performance.

The selected study articles all focused on the prediction and diagnosis of cardiovascular diseases using DL approaches, and evaluated the associated metrics related to these tasks. A summary of these parameters is illustrated in Table 3.

Outcomes of Grad-CAM and SHAP Hybridization for Cardiac Imaging Explainability

The examined study by Le et al. [58] leveraged both Grad-CAM and SHAP techniques to improve the interpretability of the machine learning models employed. Specifically, [58] applied ML approaches on CT angiography images to predict the symptoms of carotid artery disease (CAD). First ML approach they embraced was radiomics, a technique that used medical images to extracts quantitative features. SHAP was used to analyze the radiomic features behind the prediction. GRAD-CAM was utilized for the second approach to visualize the regions of interest where the DL model focuses. Also, [54] developed a DL-based classifier, which uses polar maps images, to identify flow-limiting CAD for ischemia detection. The authors [54] utilized Grad-CAM and SHAP to visualize the regions and contributing variables that the model deemed important for its predictions, respectively.

According to [58] findings, the SHAP interpretable model consistently identified high-value radiomic features that were integral to the classifier's decision-making mechanism. They further described the utilization of Grad-CAM as a means to enhance the robustness of the analysis for the black box algorithms employed in the DL approach, in which the features extracted are not predetermined by the user, providing an additional layer of analytic rigor. This approach helped the algorithms concentrated on the pertinent regions of interest. It was observed that the visualizations of the GRAD-CAM technique for both the DL models and VGG-16 models together with the simpler models were consistently centered on the carotid arteries [58]. Similarly, the SHAP values in Teuho et al's study [54] categorically revealed how individual features contributed to the model's prediction of ischemia detection for each input data point. Furthermore, the Grad-CAM outputs highlighted perfusion abnormalities on the polar map images, which were segmented according to the respective coronary artery territories [54].

Singh et al. study attempted to ensure accurate classification of arrhythmia from ECG datasets, the authors [32] utilized the ECANet attention module and proposed a new interpretable model K-Grad-CAM, after careful utilization, evaluation and comparison of SHAP, LIME and the Grad-CAM models. The proposed interpretable model, K-Grad-CAM which is an extension of the Grad-CAM, is developed to accommodate the strengths of both perturbation and gradient-backpropagation approaches. Further, K-Grad-CAM addresses some of Grad-CAM's flaws, as [32] described. The implementation of this approach was conducted through segmentation of ECG signals. Authors in [32] divided the ECG segment window into 10 distinct sections. The researchers [32] then evaluated the post-hoc explanation techniques and calculated the average saliency value for each of the designated ECG signal segments, thereby obtaining one saliency value per segment across the ECG window. In a relatively similar classification task goal, [10] employed image-based ECG recordings to detect different types of arrhythmias. To enhance the interpretability and trustworthiness of the diagnostic process, the authors [10] utilized cascading deep neural networks (CDNNs) in combination with both SHAP and Grad-CAM. The authors justified this approach as a means of providing a more comprehensive and transparent explanation, thereby facilitating the generation of clinically meaningful interpretations for therapeutic applications [10].

Also, Singh and Sharma's study [32] initially highlighted the strength of both perturbation, such as SHAP and LIME, and gradient-backpropagation, such as the Grad-CAM, approaches which they claimed as the most extensively XAI tools used in ECG time series classification. They also mentioned a number of drawbacks associated with each of the approaches describing SHAP's explanations as, sometimes, unreliable because of its random perturbation nature of operation. High demand of large computational resources due to combinatorial complexities is another drawback that limits its applicability in a setting where the input data is large.

Grad-CAM method, on the other hand, falls short in accurately pinpointing the importance of a feature in the data when multiple instances of the same feature exist. Additionally, it can only partially localize certain features due to the unweighted averaging of partial derivatives [32]. Similarly, the algorithm's interpretability is limited in scenarios involving occlusion or overlap of specific features. K-GradCam, however, was proposed to ensemble the benefits of the former models while addressing their specific drawbacks [32]. The technique was quantitatively compared with the post-hoc XAI methods, namely SHAP, Grad-CAM and LIME, using dice loss. Results of the measurement demonstrated that the new K-GradCam outperformed all the three of these post-hoc XAI techniques. The authors claimed that the model combines Grad-CAM and SHAP techniques in a highly informative and effective manner. Similarity between K- GradCam and Grad-CAM was measured at 82% which outweighs other figures in comparison of the new method with the other two methods other than Grad-CAM [32].

Zeng et al.'s research demonstrated consistent findings with the results discussed above. SHAP values offer a comprehensive understanding of the significance and impact of individual features on the final classification of arrhythmia, rendering the decision process interpretable [10]. Conversely, Grad-CAM visualizes specific regions of interest and offers important insights into the internal mechanisms of the cascading deep neural networks. As Zeng and friends describe [10], the technique further illuminates the crucial regions in the transformed signal matrices that are instrumental to the model's classification choices. Commenting on the SHAP model's performance, [10] firmly believe that this integrated methodology has improved the clinical applicability and credibility of the arrhythmia classification system. The authors [10] add that the hybrid approach allows healthcare professionals to assess both the final prognosis and the ranked importance of the contributing features. The SHAP-based approach provides valuable insights by harmonizing clinical logic with the model outputs, thereby and promoting a better comprehension of the variables driving each classification decision. Turning to Grad-CAM visualization window, Zeng et al. [10] further state that specific regions of relevance in the relative positioning matrices transformed signals are visibly shown, and these localized features have a substantial impact on the CDNN model's predictions for different types of arrhythmias. In addition, the authors state that analyzing these visual explanations for each of the seven arrhythmia categories shows substantial variations in the model's areas of focus, thereby enabling a more holistic comprehension of each decision.

Ribeiro et al. [49] approach was to qualitatively assess Grad-CAM, SHAP, and LIME interpretable models using DL models for detecting cardiomegaly, a condition marked by abnormal enlargement of the cardiac muscle. Instead of integrating the models to improve interpretability, the objective of this study was to assess the performance of each model alone and its appropriateness for deployment in diverse settings, including those necessitating prompt results [49].

Valsaraj et al. [50] leveraged expert-curated echocardiographic measurements and video data from a patient cohort comprising individuals with and without heart failure (HF) to develop DL and gradient boosting-based models to forecast 1-year, 3-year, and 5-year mortality prognoses. The authors used Grad-CAM and SHAP to understand the inner workings of their model. [50] claim that no prior research has examined HF patients, simulated long-term mortality beyond one-year timeframe, and been externally validated in separate cohorts. Valsaraj et al. [50] accomplished amazing work by analyzing the behavior of the heart from echo films and forecasting the severity of the condition in terms of life expectancy, in contrast to other research where a specific disease is observed.

Table 3. Extracted Data from the selected articles

Reference	Disease	Dataset Size	Origin of Dataset	Range of Dataset	Model	Dataset Type	Main Purpose of Research (Disease	Performance (Qualitative/ Quantitative)	Grad-CAM and SHAP Interpretable Model
Le et al. 2024 [58]	Carotid Artery Disease (CAD)	132 individuals, 1848 images	Addenbrooke' s Hospital, Cambridge, UK	2011-2016	CNNs	CT and CTA images	Disease Identification	Asymptomatic vs symptomatic AUC of 0.96(± 0.02) Culprit vs Non-Culprit arteries, AUC of 0.75(± 0.09)	Yes
Teuho et al., 2024 [54]		138 individuals	Turku University Hospital, Finland	2007–2011	DL pipeline	CTA & PET images	Disease Identification	Accuracy = 0.8478, AUROC = 0.8481, F1-Score = 0.8293, SPE = 0.8846, SEN = 0.8500, Precision = 0.8500	Yes

Singh & Sharma, 2022 [32]	Arrhythmia	48 patients	Boston' s Beth Israel Hospital (MIT-BIH)	1975-1979	ResNet	ECG Recordings	Proposes a better Explainability models (K-Grad-CAM)	K-GradCam has demonstrated advantages over gradient-based and perturbation-based approaches in terms of interpreting the models' decisions.	Yes + LIME
Zeng et al., 2024 [10]		1000++	Shaoxing–Chapman ECG database	1980-2018	CDNNs	Image-based ECG	Disease detection	Exceptional classification performance and further boosts model transparency and trustworthiness	Yes
Ribeiro et al., 2023 [49]	Cardiomegaly	18000 + 1233	VinDr-Chest X-Ray (CXR) , Vietnam + Picture Archiving and Communication System (PACS) CXR from Sau Paulo Hospital in	2018-2020	ImageNet, RestNet 50 V2	Chest X-Ray (CXR) images	Assess Explainable tool	Accuracy, Precision, Sensitivity, Specificity, F1S and AUC = 91.8 ± 0.7%, 74.0 ± 2.7%, 87.0 ± 5.5%, 92.9 ± 1.2%, 79.8 ± 1.9%, and 90.0 ± 0.7%. Interpretable models identified the expected location for cardiomegaly, except for the SHAP approach.	Yes + LIME
Valsaraj et al., 2023 [50]	All Cardiovascular Diseases (Heart Failure)	7080 echos, 4221 patients	Mackay dataset, Taiwan + Alberta Dataset, Canada	Undetermined	ResNet and CatBoost	Echo videos and echo measurements	Predict patients' all-cause mortality	The ResNet and CatBoost models achieved AUCs of 85% and 92% respectively during internal validation. When tested on external data, the AUROCs for the ResNet (82%, 82%, and 78%) CatBoost (78%, 73%, and 75%), for predicting 1-year, 3-year & 5-year mortality, respectively	Yes

Results from [49] demonstrate the excellent performance of Grad-CAM and LIME by emphasizing the heart area as the most significant characteristic pertinent to the prediction. This is accurate from a physiological perspective because cardiomegaly is correlated with the size of the heart. Conversely, SHAP did not identify cardiac regions as the most important attribute for the prediction. Nonetheless, they ascribed this disparity to the segmentation technique used in the LIME and SHAP approaches when the images were perturbed, adding that results can be segmentation method-dependent. Their work also revealed that Grad-CAM responded quickly, with output for the cardiomegaly and non-cardiomegaly detection tasks appearing in 15 and 16 seconds, respectively, compared to LIME (6.74 minutes, 6.42 minutes) and SHAP (3.36 minutes, 3.40 minutes) [49]. The employment of these perturbation techniques might not be feasible in a real-world situation where results must be produced as quickly as possible.

The Grad-CAM technique, however, was unable to uncover distinct and clinically meaningful patterns in Valsaraj et al.'s study [50]. Despite this, the model was able to detect key cardiac anatomical features that are essential for the interpretation task, such as the left atrium or the mitral and aortic valves [50]. Second, using the SHAP approach, the authors additionally reported interpretability for CatBoost models, which identified a small number of echocardiographic measurement parameters as important factors contributing to mortality risk. Overall, the interpretable models in [50] study failed to capture distinct features relevant to the model decision.

Discussion and Research Challenges

The combined use of Grad-CAM and SHAP as hybrid XAI techniques for cardiovascular imaging has showcased substantial progress in addressing the interpretability gap inherent to artificial intelligence models. A key insight drawn from the reviewed literature is the complementary strengths of these methodologies. Grad-CAM excels at providing region-based visual representations that clinicians can intuitively comprehend, while SHAP contributes by delivering feature-level explanations that align with the theoretical foundations of predictive modeling. Collectively, these techniques constitute a robust framework tailored to the diverse needs of stakeholders, ranging from healthcare professionals to patients.

The review highlights several challenges that warrant further consideration, despite the advancements in the field. Grad-CAM's reliance on visual saliency can occasionally lead to oversights, particularly in scenarios involving overlapping or occluded features. While SHAP is powerful in quantifying feature importance, its computational intensity may limit its applicability in resource-constrained environments. Additionally, there is a pressing need to establish standardized evaluation metrics that can quantifiably assess the efficacy of hybrid models in enhancing diagnostic precision and patient-oriented outcomes.

The integration of Grad-CAM and SHAP techniques in a hybrid framework has been shown to augment the interpretability of DLMs, thereby promoting trust in AI-driven diagnostic tools. This trust is crucial, as it ensures healthcare professionals maintain confidence in leveraging AI tools for critical decision-making. The variability in model performance observed across diverse study contexts underscores the importance of conducting rigorous, context-specific evaluations of these explainable AI techniques. Factors such as the dataset type, size, and imaging modality significantly influence the reliability of Grad-CAM and SHAP outputs, indicating that a one-size-fits-all solution may not exist for explainable artificial intelligence in cardiac imaging.

The findings also suggest that hybrid models could play a pivotal role in democratizing AI in healthcare. By providing both high-level visual explanations and granular feature contributions, these models can potentially bridge the expertise gap, enabling non-specialist stakeholders to comprehend AI outputs. This democratization may prove crucial in promoting broader adoption of AI technologies within clinical practice.

In light of these observations, future research should focus on refining hybrid methods to address their current limitations. Efforts to enhance the computational efficiency of SHAP, improve the localization accuracy of Grad-CAM, and establish robust evaluation metrics will be critical. Moreover, the potential of combining these techniques with other XAI approaches, such as LIME or perturbation-based models, warrants further exploration. Such endeavors could pave the way for more holistic and versatile explainability frameworks in cardiac imaging.

A key limitation of this review is the small number of studies examining hybrid Grad-CAM and SHAP models. Additionally, the diversity in imaging modalities, datasets, and evaluation metrics across studies complicates direct comparisons. While the reviewed studies offer valuable qualitative insights, there is a scarcity of quantitative evaluations measuring the impact of these hybrid models on clinical outcomes. Finally, the computational demands of SHAP and the contextual dependencies of Grad-CAM pose challenges for their broader adoption. Despite these limitations, this review provides a comprehensive overview of the current state of Grad-CAM and SHAP in bridging the AI explainability gap in cardiac imaging.

Conclusion

The hybridization of Grad-CAM and SHAP represents a promising step forward in making AI-driven diagnostics more interpretable and accessible. By leveraging the unique strengths of both techniques, researchers have demonstrated improved model transparency, which is crucial for cultivating trust and broader utilization within clinical environments. Nevertheless, challenges such as computational complexity, localization inaccuracies, and variability in performance highlight the need for continued innovation in this space. As XAI continues to evolve, hybrid models like those reviewed in this paper have the potential to revolutionize not only cardiac imaging but also the broader field of AI in healthcare.

Author contributions

Umar Dabai Sani: Conceptualization, Methodology, Writing-Original draft, Explainable AI **Moussa Mahamat Boukar:** Writing-Reviewing and Editing, ML Algorithms **Muhammed Aliyu Suleiman:** Writing-Reviewing and Editing

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] P. Covas et al., “Artificial Intelligence Advancements in the Cardiovascular Imaging of Coronary Atherosclerosis,” Front Cardiovasc Med, vol. 9, Mar. 2022, doi: 10.3389/fcvm.2022.839400.

[2] X. Wang and H. Zhu, "Artificial Intelligence in Image-based Cardiovascular Disease Analysis: A Comprehensive Survey and Future Outlook," Feb. 2024.

[3] Lin, M. Kolossváry, I. Išgum, P. Maurovich-Horvat, P. J. Slomka, and D. Dey, "Artificial intelligence: improving the efficiency of cardiovascular imaging," Expert Rev Med Devices, vol. 17, no. 6, pp. 565–577, Jun. 2020, doi: 10.1080/17434440.2020.1777855.

[4] J. D. Fuhrman, N. Gorre, Q. Hu, H. Li, I. El Naqa, and M. L. Giger, "A review of explainable and interpretable AI with applications in COVID-19 imaging," Med Phys, vol. 49, no. 1, pp. 1–14, Jan. 2022, doi: 10.1002/mp.15359.

[5] Amin, K. Hasan, S. Zein-Sabatto, D. Chimba, I. Ahmed, and T. Islam, "An Explainable AI Framework for Artificial Intelligence of Medical Things," in 2023 IEEE Globecom Workshops (GC Wkshps), IEEE, Dec. 2023, pp. 2097–2102. doi: 10.1109/GCWkshps58843.2023.10464798.

[6] S. B. Mallampati and H. Seetha, "A Review on Recent Approaches of Machine Learning, Deep Learning, and Explainable Artificial Intelligence in Intrusion Detection Systems," Majlesi Journal of Electrical Engineering; Isfahan, vol. 17, no. 1, pp. 29–54, Mar. 2023.

[7] Saporta et al., "Benchmarking saliency methods for chest X-ray interpretation," Nat Mach Intell, vol. 4, no. 10, pp. 867–878, Oct. 2022, doi: 10.1038/s42256-022-00536-x.

[8] Patrício, J. C. Neves, and L. F. Teixeira, "Explainable Deep Learning Methods in Medical Image Classification: A Survey," ACM Comput Surv, vol. 56, no. 4, pp. 1–41, Apr. 2024, doi: 10.1145/3625287.

[9] Z. Wang, K. Qian, H. Liu, B. Hu, B. W. Schuller, and Y. Yamamoto, "Exploring interpretable representations for heart sound abnormality detection," Biomed Signal Process Control, vol. 82, Apr. 2023, doi: 10.1016/j.bspc.2023.104569.

[10] W. Zeng, L. Shan, C. Yuan, and S. Du, "Advancing cardiac diagnostics: Exceptional accuracy in abnormal ECG signal classification with cascading deep learning and explainability analysis," Appl Soft Comput, vol. 165, p. 112056, Nov. 2024, doi: 10.1016/j.asoc.2024.112056.

[11] L. Li, W. Ding, L. Huang, X. Zhuang, and V. Grau, "Multi-modality cardiac image computing: A survey," Med Image Anal, vol. 88, p. 102869, Aug. 2023, doi: 10.1016/j.media.2023.102869.

[12] M. Athanasiou, K. Sfrintzeri, K. Zarkogianni, A. C. Thanopoulou, and K. S. Nikita, "An explainable XGBoost–based approach towards assessing the risk of cardiovascular disease in patients with Type 2 Diabetes Mellitus," in 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE), IEEE, Oct. 2020, pp. 859–864. doi: 10.1109/BIBE50027.2020.00146.

[13] J. , Amann et al., "To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems.," PLOS Digital Health, vol. 1, no. 2, p. 80, Nov. 2022.

[14] S. Sh. Taher, S. Y. Ameen, and J. A. Ahmed, "Advanced Fraud Detection in Blockchain Transactions: An Ensemble Learning and Explainable AI Approach," Engineering, Technology & Applied Science Research, vol. 14, no. 1, pp. 12822–12830, Feb. 2024, doi: 10.48084/etasr.6641.

[15] F. Markus, J. A. Kors, and P. R. Rijnbeek, "The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies," Jan. 01, 2021, Academic Press Inc. doi: 10.1016/j.jbi.2020.103655.

[16] R. Tiwari, "Explainable AI (XAI) and its Applications in Building Trust and Understanding in AI Decision Making," INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT, vol. 07, no. 01, Jan. 2023, doi: 10.55041/ijsrem17592.

[17] T. Vaiyapuri, "Utilizing Explainable AI and Biosensors for Clinical Diagnosis of Infectious Vector-Borne Diseases," Engineering, Technology & Applied Science Research, vol. 14, no. 6, pp. 18640–18648, Dec. 2024, doi: 10.48084/etasr.9026.

[18] J. Wang, L. Zhang, Y. Huang, and J. Zhao, "Safety of Autonomous Vehicles," 2020, Hindawi Limited. doi: 10.1155/2020/8867757.

[19] D. Gunning and D. W. Aha, "DARPA's Explainable Artificial Intelligence Program Deep Learning and Security," 2019.

[20] S. Bai, S. Nasir, R. A. Khan, S. Arif, A. Meyer, and H. Konik, "Breast Cancer Diagnosis: A Comprehensive Exploration of Explainable Artificial Intelligence (XAI) Techniques," Jun. 2024, [Online]. Available: http://arxiv.org/abs/2406.00532

A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," Oct. 2019, [Online]. Available: http://arxiv.org/abs/1910.10045

[21] Salih et al., “A Review of Evaluation Approaches for Explainable AI With Applications in Cardiology,” Nov. 22, 2023. doi: 10.36227/techrxiv.24573304.v1.

[22] Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, “Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks,” Oct. 2017, doi: 10.1109/WACV.2018.00097.

[23] R. K. Sheu and M. S. Pardeshi, “A Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach, Human Interaction and Scoring System,” Oct. 01, 2022, MDPI. doi: 10.3390/s22208068.

[24] Z. Wang, L. Wu, and X. Ji, “An Interpretable Deep Learning System for Automatic Intracranial Hemorrhage Diagnosis with CT Image,” in Proceedings of the 2021 International Conference on Bioinformatics and Intelligent Computing, BIC 2021, Association for Computing Machinery, Inc, Jan. 2021, pp. 338–357. doi: 10.1145/3448748.3448803.

[25] H. Mankodiya, D. Jadav, R. Gupta, S. Tanwar, W. C. Hong, and R. Sharma, “OD-XAI: Explainable AI-Based Semantic Object Detection for Autonomous Vehicles,” Applied Sciences (Switzerland), vol. 12, no. 11, Jun. 2022, doi: 10.3390/app12115310.

[26] Dewi, R. C. Chen, H. Yu, and X. Jiang, “XAI for Image Captioning using SHAP,” Journal of Information Science and Engineering, vol. 39, no. 4, pp. 711–724, Jul. 2023, doi: 10.6688/JISE.202307_39(4).0001.

[27] S. M. Hussain et al., “Shape-Based Breast Lesion Classification Using Digital Tomosynthesis Images: The Role of Explainable Artificial Intelligence,” Applied Sciences (Switzerland), vol. 12, no. 12, Jun. 2022, doi: 10.3390/app12126230.

[28] Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G. Z. Yang, “XAI-Explainable artificial intelligence,” Sci Robot, vol. 4, no. 37, Dec. 2019, doi: 10.1126/scirobotics.aay7120.

[29] Salih et al., “Explainable Artificial Intelligence and Cardiac Imaging: Toward More Interpretable Models,” Circ Cardiovasc Imaging, vol. 16, no. 4, p. E014519, Apr. 2023, doi: 10.1161/CIRCIMAGING.122.014519.

[30] Neves et al., “Interpretable heartbeat classification using local model-agnostic explanations on ECGs,” Comput Biol Med, vol. 133, Jun. 2021, doi: 10.1016/j.compbiomed.2021.104393.

[31] P. Singh and A. Sharma, “Interpretation and Classification of Arrhythmia Using Deep Convolutional Network,” IEEE Trans Instrum Meas, vol. 71, 2022, doi: 10.1109/TIM.2022.3204316.

[32] N. I. Papandrianos, A. Feleki, S. Moustakidis, E. I. Papageorgiou, I. D. Apostolopoulos, and D. J. Apostolopoulos, “An Explainable Classification Method of SPECT Myocardial Perfusion Images in Nuclear Cardiology Using Deep Learning and Grad-CAM,” Applied Sciences (Switzerland), vol. 12, no. 15, Aug. 2022, doi: 10.3390/app12157592.

[33] Tang, J. Chen, L. Ren, X. Wang, D. Li, and H. Zhang, “Reviewing CAM-Based Deep Explainable Methods in Healthcare,” May 01, 2024, Multidisciplinary Digital Publishing Institute (MDPI). doi: 10.3390/app14104124.

[34] V. Jahmunah, E. Y. K. Ng, R. S. Tan, S. L. Oh, and U. R. Acharya, “Explainable detection of myocardial infarction using deep learning models with Grad-CAM technique on ECG signals,” Comput Biol Med, vol. 146, Jul. 2022, doi: 10.1016/j.compbiomed.2022.105550.

[35] Sakai et al., “Medical Professional Enhancement Using Explainable Artificial Intelligence in Fetal Cardiac Ultrasound Screening,” Biomedicines, vol. 10, no. 3, Mar. 2022, doi: 10.3390/biomedicines10030551.

[36] Petch, S. Di, and W. Nelson, “Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology,” Feb. 01, 2022, Elsevier Inc. doi: 10.1016/j.cjca.2021.09.004.

[37] K. Kırboğa and E. U. Küçüksille, “Identifying Cardiovascular Disease Risk Factors in Adults with Explainable Artificial Intelligence,” Anatol J Cardiol, vol. 27, no. 11, pp. 657–663, Nov. 2023, doi: 10.14744/AnatolJCardiol.2023.3214.

[38] Y. M. Ayano, F. Schwenker, B. D. Dufera, and T. G. Debelee, “Interpretable Machine Learning Techniques in ECG-Based Heart Disease Classification: A Systematic Review,” Jan. 01, 2023, Multidisciplinary Digital Publishing Institute (MDPI). doi: 10.3390/diagnostics13010111.

[39] Salau, N. Agwu Nwojo, M. Mahamat Boukar, and O. Usen, “Advancing Preauthorization Task in Healthcare: An Application of Deep Active Incremental Learning for Medical Text Classification,” Engineering, Technology & Applied Science Research, vol. 13, no. 6, pp. 12205–12210, Dec. 2023, doi: 10.48084/etasr.6332.

[40] J. Hou et al., “Self-eXplainable AI for Medical Image Analysis: A Survey and New Outlooks,” Oct. 2024.

[41] X. Li, Y. Huang, Y. Ning, M. Wang, and W. Cai, “Multi-branch myocardial infarction detection and localization framework based on multi-instance

learning and domain knowledge," Physiol Meas, vol. 45, no. 4, Apr. 2024, doi: 10.1088/1361-6579/ad3d25.

[42] H. Shin, G. Noh, and B. M. Choi, "Photoplethysmogram based vascular aging assessment using the deep convolutional neural network," Sci Rep, vol. 12, no. 1, Dec. 2022, doi: 10.1038/s41598-022-15240-4.

[43] Jafari et al., "Automatic Diagnosis of Myocarditis Disease in Cardiac MRI Modality using Deep Transformers and Explainable Artificial Intelligence," Oct. 2022, [Online]. Available: http://arxiv.org/abs/2210.14611

[44] C. M. Frade et al., "Toward characterizing cardiovascular fitness using machine learning based on unobtrusive data," PLoS One, vol. 18, no. 3, p. e0282398, Mar. 2023, doi: 10.1371/journal.pone.0282398.

[45] P. Elias et al., "Deep Learning Electrocardiographic Analysis for Detection of Left-Sided Valvular Heart Disease," J Am Coll Cardiol, vol. 80, no. 6, pp. 613–626, Aug. 2022, doi: 10.1016/j.jacc.2022.05.029.

[46] Alamatsaz, L. Tabatabaei, M. Yazdchi, H. Payan, N. Alamatsaz, and F. Nasimi, "A lightweight hybrid CNN-LSTM explainable model for ECG-based arrhythmia detection," Biomed Signal Process Control, vol. 90, Apr. 2024, doi: 10.1016/j.bspc.2023.105884.

[47] Zamora et al., "Prognostic Stratification of Familial Hypercholesterolemia Patients Using AI Algorithms: [1] A Gender-Specific Approach," Oct. 14, 2024. doi: 10.1101/2024.10.11.24315359.

[48] Ribeiro, D. A. C. Cardenas, J. E. Krieger, and M. A. Gutierrez, "Interpretable Deep Learning Model For Cardiomegaly Detection with Chest X-ray Images," in Anais do XXIII Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS 2023), Sociedade Brasileira de Computação - SBC, Jun. 2023, pp. 340–347. doi: 10.5753/sbcas.2023.229943.

[49] Valsaraj et al., "Development and validation of echocardiography-based machine-learning models to predict mortality," EBioMedicine, vol. 90, p. 104479, Apr. 2023, doi: 10.1016/J.EBIOM.2023.104479.

[50] D. Kusumoto et al., "A deep learning-based automated diagnosis system for SPECT myocardial perfusion imaging," Sci Rep, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-64445-2.

[51] S. Bouazizi and H. Ltifi, "Enhancing accuracy and interpretability in EEG-based medical decision making using an explainable ensemble learning framework application for stroke prediction," Decis Support Syst, vol. 178, Mar. 2024, doi: 10.1016/j.dss.2023.114126.

[52] Ribeiro, D. A. C. Cardenas, F. M. Dias, J. E. Krieger, and M. A. Gutierrez, "Explainable AI in Deep Learning-based Detection of Aortic Elongation on Chest X-ray Images," Aug. 31, 2023. doi: 10.1101/2023.08.28.23294735.

[53] J. Teuho et al., "Explainable deep-learning-based ischemia detection using hybrid O-15 H2O perfusion positron emission tomography and computed tomography imaging with clinical data," Journal of Nuclear Cardiology, vol. 38, Aug. 2024, doi: 10.1016/j.nuclcard.2024.101889.

[54] Elsevier, "ScienceDirect Research Database." Accessed: Oct. 19, 2024. [Online]. Available: https://www.sciencedirect.com

[55] [Allen Institute for AI (AI2), "Semantic Scholar Research Database." Accessed: Oct. 20, 2025. [Online]. Available: https://www.semanticscholar.org

[56] I. of H. (NIH) U.S. National Library of Medicine (NLM), "PubMed Research Database." Accessed: Oct. 19, 2025. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov

[57] E. P. V. Le et al., "Using machine learning to predict carotid artery symptoms from CT angiography: A radiomics and deep learning approach," Eur J Radiol Open, vol. 13, p. 100594, Dec. 2024, doi: 10.1016/j.ejro.2024.100594.