Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

An Explainable Multimodal Framework for Chest X-Ray Alert Classification Using Radiology Reports and Images

Edy Winarno [1,*], Indah Manfaati Nur [2], Abdul Karim [3], Saeful Amri [4], Ismi Elya Wirdati [5], and Prajanto Wahyu Adi [6]

[1] Department of Information Technology, Faculty of Engineering and Computer Science, Universitas Muhammadiyah Semarang, Semarang 50273, Indonesia; e-mail : edywin@unimus.ac.id
[2] Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Muhammadiyah Semarang, Semarang 50273, Indonesia; e-mail : indahmnur@unimus.ac.id
[3] College of Information Science / AI-X, Hallym University, Chuncheon 24252, South Korea; e-mail : abdullkarim@korea.ac.kr
[4] Department of Data Science, Faculty of Science and Agricultural Technology, Universitas Muhammadiyah Semarang, Semarang 50273, Indonesia; e-mail : saefulamri@unimus.ac.id
[5] Faculty of Public Health, Universitas Muhammadiyah Semarang, Semarang 50273, Indonesia; e-mail : ismi.elya@unimus.ac.id
[6] Department of Computer Science / Informatics, Faculty of Science and Mathematics, Universitas Diponegoro, Semarang 50275, Indonesia; e-mail : prajanto@lecturer.undip.ac.id
* Corresponding Author : Edy Winarno

Abstract: Artificial intelligence has the potential to support radiology workflows by assisting in the identification of cases that may require additional clinical attention. However, alert-oriented medical AI systems should provide not only classification outputs but also interpretable evidence that can be reviewed and audited by clinicians. This study develops and evaluates an explainable multimodal framework for binary chest X-ray alert classification using paired radiology reports and chest X-ray images. The text branch employs TF-IDF n-gram features with a class-balanced Logistic Regression classifier, while the image branch fine-tunes a pretrained ResNet18 model. The two branches are integrated through probability-level late fusion using a validation-selected fusion weight. Explainability is implemented in a modality-specific manner: global coefficient analysis is used to identify influential textual cues, while Grad-CAM heatmaps are used to visualize salient image regions. Experiments were conducted on paired samples from the Open-i/IU X-Ray dataset using text-only, image-only, and fusion-based evaluation settings. Additional analyses include case-level complementarity analysis, bootstrap confidence intervals for ROC-AUC, shortcut-feature inspection, and qualitative Grad-CAM auditing. The results indicate that the text modality provides the dominant predictive signal under the current proxy-label setting. Late fusion produced a small descriptive improvement on the test set, increasing accuracy from 0.8533 to 0.8667, F1-score from 0.8817 to 0.8936, and ROC-AUC from 0.8936 to 0.9025 compared with the text-only baseline. However, the observed ROC-AUC improvement was not statistically conclusive based on bootstrap analysis. These findings suggest that the proposed framework is useful as a reproducible and auditable multimodal prototype, while also highlighting important limitations, including proxy-label ambiguity, potential label leakage from radiology reports, limited image-branch contribution, lack of external validation, and the need for stronger explanation and calibration assessment.

Keywords: Alert classification; Chest X-ray; Clinical NLP; Explainable artificial intelligence; Grad-CAM; Late fusion; Multimodal learning; Radiology reports.

Received: April, 22nd 2026
Revised: May, 16th 2026
Accepted: May, 17th 2026
Published: May, 23rd 2026

1. Introduction

Chest radiography is one of the most widely used imaging modalities in clinical practice and is commonly accompanied by free-text radiology reports. In high-volume radiology workflows, automated systems capable of prioritizing or flagging potentially abnormal cases

may assist clinical review, particularly when rapid decision support is required. However, alert-oriented systems in healthcare must be interpreted carefully because false alerts may increase unnecessary workload, whereas missed alerts may delay clinical attention. Therefore, clinically useful artificial intelligence (AI) systems should provide not only classification outputs or probability scores, but also interpretable evidence that enables clinicians to audit the basis of model predictions [1], [2].

Recent advances in medical AI have demonstrated promising performance in chest X-ray classification and clinical text analysis. Computer vision models are capable of learning visual representations from radiographic images, while natural language processing models can extract clinically relevant signals from radiology reports and other medical narratives [3]–[6]. Nevertheless, many AI systems remain difficult to inspect because their predictions are generated by complex architectures with limited transparency. This lack of interpretability is a critical concern in clinical decision support, where model outputs must be interpreted in relation to clinical context, data quality, and potential bias [7], [8]. Recent work [9] further emphasizes that healthcare AI evaluation should extend beyond predictive accuracy by considering generalizability, governance, interpretability, and workflow integration.

A multimodal approach is particularly relevant in radiology because image and text data provide different but related representations of the same clinical case. Chest X-ray images may contain anatomical or pathological visual evidence, whereas radiology reports summarize clinical interpretation in textual form. In principle, combining both modalities may improve robustness when one modality is incomplete, noisy, or ambiguous. However, such assumptions should be validated empirically rather than accepted implicitly. In addition, radiology reports may introduce methodological risks because they can contain direct or indirect descriptions of the target condition. When labels are derived from report-associated information, text-based models may learn report-writing patterns instead of independent clinical reasoning signals, thereby creating a potential label leakage problem that must be explicitly acknowledged and evaluated.

In this study, the term ALERT refers to a prototype binary classification category indicating that a chest X-ray case contains findings or report cues that may require additional clinical attention under the implemented labeling scheme. The term NORMAL refers to cases that do not meet the alert criteria. This task should not be interpreted as temporal forecasting, emergency triage prioritization, or a clinically validated early-warning system. Instead, the proposed framework performs binary alert classification using paired chest X-ray images and radiology reports. The term alert is therefore used to describe the intended decision-support output rather than to claim prospective clinical prediction.

The main hypothesis of this study is that radiology reports and chest X-ray images may provide complementary information for alert classification, although the degree of complementarity must be measured rather than assumed. Specifically, this work investigates whether: (1) report-based text classification provides a strong but potentially leakage-prone baseline; (2) image-based classification contributes additional visual evidence beyond textual cues; and (3) probability-level late fusion improves classification or ranking performance compared with unimodal approaches. The main contributions of this study are summarized as follows:

- A reproducible explainable multimodal framework for binary chest X-ray alert classification is developed using paired radiology reports and chest X-ray images processed through separate text and image branches.
- Text-only, image-only, and probability-level late-fusion models are systematically evaluated to analyze modality contribution and multimodal complementarity under a proxy-label setting.
- Modality-specific explainability mechanisms are integrated through Logistic Regression coefficient analysis for textual interpretation and Grad-CAM heatmaps for visual saliency inspection.
- Critical analysis of the proposed prototype is provided, including discussion of proxy-label ambiguity, potential label leakage from radiology reports, limited image-branch contribution, lack of external validation, and the need for calibration and quantitative explanation assessment.

The remainder of this paper is organized as follows. Section 2 reviews related work on chest X-ray classification, clinical NLP, multimodal medical AI, and explainable AI. Section 3 describes the proposed framework, dataset, label construction process, model architecture,

fusion strategy, explainability methods, and experimental protocol. Section 4 presents the experimental results and discussion, including predictive performance comparison, case-level multimodal complementarity analysis, bootstrap ROC-AUC uncertainty analysis, shortcut-feature inspection, Grad-CAM quality auditing, latency evaluation, synthesis of experimental insights, and study limitations. Finally, Section 5 concludes the paper and outlines directions for future research.

2. Related Work

2.1. Chest X-Ray Classification and Public Radiology Datasets

Chest X-ray classification has been extensively studied as a benchmark problem in medical computer vision. Public datasets such as Open-i/IU X-Ray, MIMIC-CXR, CheXpert, ChestX-ray8, and related repositories have enabled reproducible research in radiographic image classification and radiology report analysis [3], [10]–[12]. These datasets differ in terms of scale, annotation strategy, label quality, and the availability of paired image-report data. Large-scale chest X-ray datasets are valuable for training high-capacity deep learning models; however, many labels are generated through report-mining pipelines or weak annotation strategies, which may introduce uncertainty, ambiguity, and label noise.

Convolutional neural networks (CNNs) have been widely adopted for chest X-ray classification because they can learn hierarchical visual representations from radiographic images. Architectures such as ResNet, DenseNet, and other pretrained CNN backbones are frequently utilized due to their strong transfer-learning capability and competitive predictive performance [13], [14]. Nevertheless, image-only classification remains challenging when findings are subtle, image quality varies, supervision is weak, or labels are derived from report-level information rather than directly curated visual annotations. These challenges motivate careful interpretation of image-only performance and caution when assessing the clinical validity of weakly supervised chest radiograph classification systems. Related studies on chest X-ray pneumonia segmentation also demonstrate the relevance of transfer-learning approaches for radiographic analysis while emphasizing the importance of task-specific interpretation and evaluation [15].

2.2. Clinical NLP and Radiology Report Classification

Natural language processing (NLP) has become an important approach for extracting clinically relevant information from medical narratives, including radiology reports. Traditional methods such as TF-IDF combined with linear classifiers remain useful baselines because they are computationally efficient, relatively interpretable, and straightforward to reproduce [16]. In particular, Logistic Regression trained on TF-IDF n-grams enables direct inspection of feature coefficients, making it suitable for identifying influential words or phrases associated with model predictions.

More advanced approaches, including contextual embeddings and transformer-based language models, may capture richer semantic representations from clinical text [5], [6], [17]. However, these approaches typically require larger datasets, greater computational resources, and more complex interpretability strategies. In radiology report classification, text-based models may achieve strong predictive performance because clinically relevant findings are often explicitly stated within the report narrative itself. At the same time, this creates a methodological concern because models may learn report-writing conventions or shortcut correlations rather than independent clinical reasoning signals. Consequently, strong performance from report-based classification should not automatically be interpreted as evidence of robust clinical understanding.

2.3. Multimodal Medical AI and Late Fusion

Multimodal medical AI aims to integrate complementary information from multiple data sources, including images, text, structured variables, and time-series measurements. In radiology, image-text integration is particularly relevant because a clinical examination commonly consists of both radiographic images and accompanying narrative reports. Multimodal learning may improve robustness when one modality is incomplete, ambiguous, or noisy [18], [19].

Several multimodal integration strategies have been explored, including early fusion, intermediate representation fusion, and late fusion. Early and intermediate fusion methods may

capture deeper cross-modal interactions, but they often require larger datasets, more complex architectures, and more difficult optimization and interpretation procedures. Studies on multimodal healthcare monitoring similarly demonstrate that multimodal systems must address challenges related to data heterogeneity, domain shift, computational constraints, and interpretability in deployment-oriented settings [20].

Late fusion combines the outputs of independently trained unimodal models, typically through averaging or weighted aggregation of predicted probabilities. Although comparatively simple, late fusion offers several practical advantages. Each modality-specific model can be developed, evaluated, and interpreted independently, while modality contribution can be analyzed through the fusion mechanism itself. This modular structure is particularly suitable for reproducible and deployment-oriented prototypes where interpretability, computational simplicity, and implementation flexibility are important considerations. However, late fusion should not be overstated as deep multimodal reasoning because it primarily represents decision-level aggregation of separate classifiers rather than joint cross-modal representation learning.

2.4. Explainable AI in Medical Imaging and Clinical Text

Explainable AI (XAI) plays an important role in medical decision support because clinicians must understand the basis of model predictions before integrating AI outputs into clinical workflows. In medical imaging, Grad-CAM and related saliency methods are commonly used to visualize image regions that contribute to CNN predictions [21]. These visualizations may help determine whether a model focuses on anatomically plausible regions or irrelevant artifacts. However, saliency maps should not be interpreted as direct causal clinical evidence. Prior studies have shown that visual explanations may become unstable or misleading without robustness analysis, sanity checks, and clinical validation [22], [23]. Broader studies on AI deployment in radiology, emergency systems, and remote healthcare also emphasize that operational advantages must be balanced against interpretability, compliance, and adoption challenges [24].

For clinical text models, explanation methods may include feature attribution, coefficient analysis, LIME, SHAP, or attention-based interpretation techniques, depending on the underlying model architecture [25], [26]. Linear classifiers such as Logistic Regression provide relatively transparent global inspection of influential terms and n-grams. Nevertheless, textual explanations should also be interpreted carefully because highly weighted features may reflect report-writing conventions, dataset artifacts, or shortcut correlations rather than clinically meaningful evidence.

In this study, explainability is implemented as modality-specific post-hoc audit support. The text branch is interpreted through Logistic Regression coefficient analysis, while the image branch is analyzed using Grad-CAM heatmaps. These explanations support inspection of each modality independently but do not constitute a unified explanation of the fused multimodal decision. This distinction is important because the proposed framework performs probability-level late fusion rather than deep cross-modal reasoning.

3. Proposed Method

To address the methodological considerations discussed in Section 2, this study adopts a reproducible and deployment-oriented multimodal framework designed to support modality-specific inspection and auditability. Rather than implementing deep cross-modal representation learning, the proposed system combines independently trained text and image classifiers through probability-level late fusion. This design prioritizes interpretability, modularity, and experimental transparency while enabling analysis of modality contribution, potential shortcut behavior, and the limitations of multimodal alert classification under a proxy-label setting.

3.1. Research Design

This study adopts a development-oriented experimental design to build and evaluate an explainable multimodal framework for binary chest X-ray alert classification. The proposed framework processes two paired modalities from the same radiology case: a chest X-ray image and its corresponding radiology report. The task is formulated as a binary classification

problem in which each sample is assigned to either ALERT or NORMAL according to the prototype labeling scheme described in Section 3.2.

The framework is designed with three primary objectives. First, predictive performance is evaluated across report-based text classification, image-based classification, and probability-level late fusion. Second, the framework investigates whether the image modality contributes complementary information beyond the radiology report. Third, modality-specific explanation artifacts are generated to support model auditing, including textual feature coefficient inspection and Grad-CAM visual heatmaps.

As illustrated in Figure 1, the workflow consists of five stages: (1) paired image-report data preparation; (2) ALERT/NORMAL label construction; (3) independent development of text and image classifiers; (4) probability-level late fusion; and (5) modality-specific explainability and evaluation. The proposed framework should therefore be interpreted as a reproducible proof-of-concept prototype rather than a clinically validated alerting system.

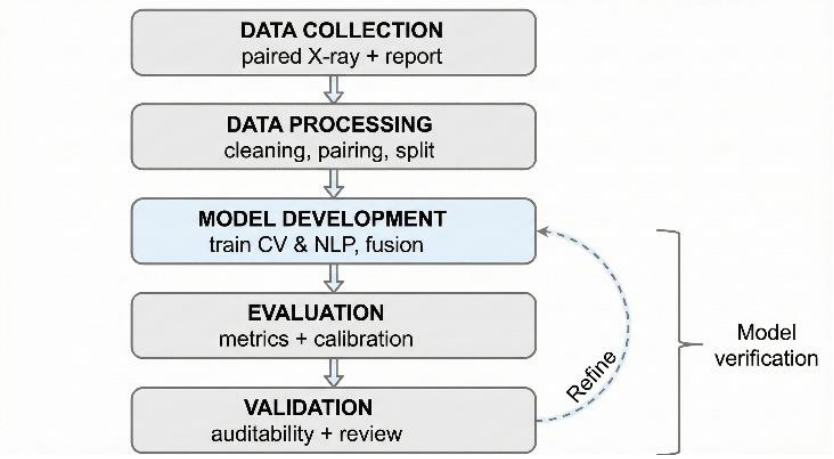


Figure 1. Research workflow of the proposed explainable multimodal chest X-ray alert-classification framework.

3.2. Dataset and Label Definition

The dataset used in this study is derived from the Open-i/IU X-Ray collection, which provides paired chest X-ray images and radiology reports [3]. This dataset was selected because it supports direct multimodal experimentation in which each clinical case contains both a visual modality and a textual modality. Such pairing enables consistent comparison among text-only, image-only, and fusion-based models using the same sample-level split. Each record is treated as a multimodal sample consisting of three components: radiology report text, the corresponding chest X-ray image, and a binary class label. The positive class is defined as ALERT, whereas the negative class is defined as NORMAL. A summary of the dataset characteristics and labeling configuration is presented in Table 1.

Table 1. Dataset and label summary.

Item	Description
Dataset source	Open-i/IU X-Ray collection
Data type	Paired chest X-ray image and radiology report
Task	Binary alert classification
Positive class	ALERT
Negative class	NORMAL
Label source	Proxy labels derived from available dataset and report-related information
Main limitation	Potential label ambiguity and label leakage from report text
Clinical status	Prototype classification label, not a final clinical diagnosis

In this study, ALERT refers to a prototype classification category indicating that a case contains report cues or associated findings that may require additional clinical attention under the implemented labeling scheme. NORMAL refers to cases that do not meet the alert criteria.

The ALERT label should not be interpreted as a definitive clinical diagnosis, emergency triage decision, or temporal early-warning outcome.

The current labeling process relies on proxy labels derived from available dataset information and report-associated indicators. Consequently, the labels may contain ambiguity, particularly when reports include non-acute findings, device- or tube-related descriptions, or abnormal but clinically stable observations. This limitation is especially important because the report text is also used as an input feature. As a result, the text branch may learn report-writing patterns that are strongly correlated with the target label, thereby introducing potential label leakage. This issue is explicitly considered during experimental interpretation and further discussed in the limitations section. An illustrative example of a paired multimodal sample consisting of a chest X-ray image and the corresponding radiology report is shown in Figure 2.

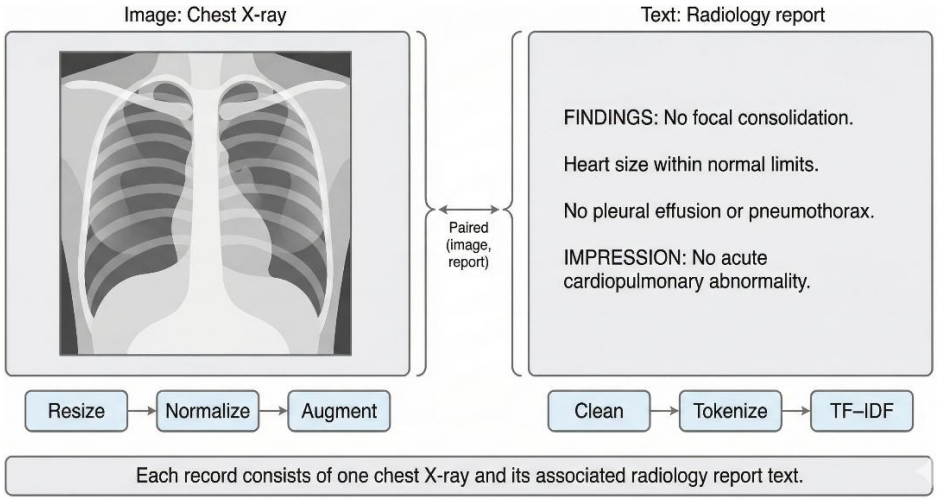


Figure 2. Schematic illustration of a paired chest X-ray image and radiology report from the Open-i/IU X-Ray dataset and the corresponding modality-specific preprocessing workflow.

3.3. Data Preprocessing

For the text modality, radiology reports are preprocessed prior to feature extraction. The preprocessing steps include lowercasing, removal of non-informative characters, whitespace normalization, and token preparation for n-gram extraction. The cleaned reports are subsequently transformed into numerical representations using Term Frequency-Inverse Document Frequency (TF-IDF). The use of TF-IDF is motivated by three considerations. First, it provides a strong and reproducible baseline for clinical text classification. Second, it is computationally efficient and suitable for deployment-oriented prototype systems. Third, when combined with Logistic Regression, TF-IDF supports transparent global interpretation through feature coefficient inspection [16].

For the image modality, chest X-ray images are preprocessed before being passed to the convolutional neural network. The preprocessing pipeline includes resizing to a fixed input resolution, pixel-intensity normalization, and standard image transformations or augmentations during training. These operations are intended to reduce input variability and improve model robustness. Some low-level training configurations exported from the original prototype were incomplete; this limitation is reported transparently in Table 2 and discussed as part of the reproducibility considerations.

3.4. Unimodal Classification Branches

The proposed framework consists of two independently trained unimodal branches: a text classification branch and an image classification branch. The overall multimodal architecture and probability-level late-fusion workflow are illustrated in Figure 3.

3.4.1. Text Classification Branch

The text branch uses TF-IDF feature representations combined with a Logistic Regression classifier. Logistic Regression was selected because it is lightweight, reproducible, and relatively interpretable. Compared with more complex deep language models, a linear

classifier enables direct inspection of the contribution of individual n-gram features through learned coefficients. The classifier estimates the probability of the ALERT class from the radiology report text. Class balancing is applied during training to reduce the effect of class imbalance in the dataset. The trained text model is subsequently evaluated on the validation and test sets using the same sample-level split as the image branch.

3.4.2. Image Classification Branch

The image branch uses a pretrained ResNet18 model fine-tuned for binary chest X-ray classification [13]. ResNet18 was selected because it provides a practical balance between model capacity, training efficiency, and interpretability support through Grad-CAM visualization. Compared with deeper CNN architectures, ResNet18 is relatively lightweight and suitable for proof-of-concept systems in which inference latency, reproducibility, and deployment simplicity are important considerations.

The image branch estimates the ALERT probability directly from the chest X-ray image. Fine-tuning is performed to adapt pretrained visual representations to the chest radiograph classification task. However, because the labels are proxy labels and may be derived from report-related information, the image branch may receive weaker or noisier supervision than the text branch. This consideration is important when interpreting the relatively lower performance of the image-only model.

3.5. Probability-Level Late Fusion

The overall multimodal architecture, including the independently trained text and image branches, probability-level late fusion, and modality-specific explainability workflow, is illustrated in Figure 3.

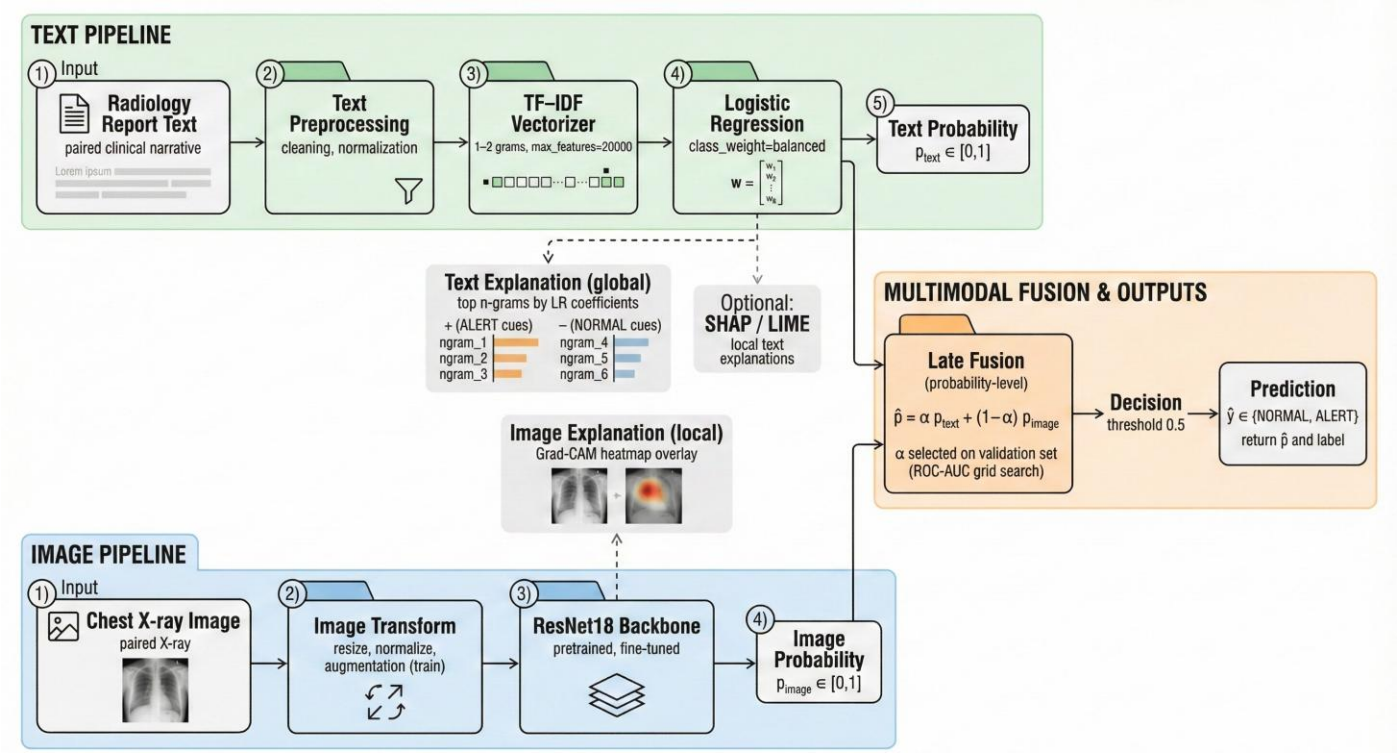


Figure 3. Overview of the proposed multimodal alert-classification framework consisting of independently trained text and image branches, probability-level late fusion, and modality-specific explainability outputs.

The multimodal framework combines the output probabilities of the text and image branches using probability-level late fusion. The fused ALERT probability is computed as:

$$p_{fusion} = \alpha p_{text} + (1 - \alpha) p_{image} \quad (1)$$

where p_{fusion} represents the ALERT probability produced by the text branch, p_{image} represents the ALERT probability produced by the image branch, and α controls the relative contribution of the text modality.

The fusion weight α is selected on the validation set using ROC-AUC as the optimization criterion. Candidate values are evaluated across a predefined range, and the value that maximizes validation ROC-AUC is selected for the final prototype configuration. In the implemented system, the selected value is α=0.90, indicating that the fused model is strongly text-dominant. This result suggests that the radiology report contributes most of the predictive signal, while the image branch provides limited complementary information under the current experimental setting. It is important to clarify that the proposed fusion mechanism does not perform deep cross-modal reasoning or joint representation learning. Instead, the system consists of two independently trained classifiers whose probabilities are combined through weighted averaging.

3.6. Explainability Strategy

The explainability component is designed to support post-hoc auditing of modality-specific model behavior rather than to provide a unified explanation of the fused multimodal decision. As illustrated in Figure 3, explanation outputs are generated independently for the text and image branches to facilitate inspection of influential textual patterns and visually salient image regions.

For the text branch, explainability is implemented through inspection of the learned Logistic Regression coefficients. Positive coefficients indicate n-grams associated with the ALERT class, whereas negative coefficients indicate n-grams associated with the NORMAL class. This approach was selected because the combination of TF-IDF and Logistic Regression provides relatively transparent global feature inspection while maintaining computational simplicity and reproducibility. In the proposed framework, coefficient analysis is used to identify whether the model captures clinically meaningful report cues or potentially problematic shortcut patterns.

Although coefficient-based inspection is straightforward and interpretable, the resulting explanations must still be interpreted cautiously. Highly weighted terms may represent clinically relevant language, but they may also reflect report-writing conventions, dataset artifacts, shortcut correlations, or potential label leakage due to the proxy-label setting.

For the image branch, Grad-CAM is used to generate local visual explanations from the fine-tuned ResNet18 model [21]. Grad-CAM highlights image regions that contribute strongly to the CNN prediction by projecting activation importance onto the original chest X-ray image. In the proposed framework, Grad-CAM is used to support qualitative inspection of whether the CNN attends to anatomically plausible regions or potentially irrelevant visual artifacts.

The generated heatmaps are overlaid on the original chest X-ray images to facilitate visual interpretation. However, Grad-CAM is used as an attention-oriented visualization tool rather than causal clinical evidence. Consequently, highlighted regions should not be interpreted as definitive indicators of pathology, and non-highlighted regions should not be interpreted as proof of normality.

The proposed explainability strategy does not introduce a new XAI algorithm. Instead, established explanation methods are integrated into a multimodal classification workflow to improve transparency, auditability, and interpretability of the prototype system. As shown in Figure 3, explainability remains modality-specific because the proposed framework performs probability-level late fusion rather than deep cross-modal reasoning. Therefore, a complete multimodal explanation of the fused decision, including cross-modal attribution, explanation faithfulness analysis, and quantitative robustness evaluation, remains outside the scope of the current prototype and is identified as future work.

3.7. Experimental Protocol

To support reproducibility, the experimental configuration available from the implemented prototype is summarized in Table 2. The reported configuration distinguishes between parameters explicitly available from the exported manuscript artifacts and parameters that still require verification from the original Colab execution logs. This reporting strategy

avoids introducing unverified implementation details while transparently documenting the reproducibility limitations identified during revision.

The experimental setup evaluates three configurations: text-only classification, image-only classification, and multimodal probability-level late fusion. All experiments use the same paired sample split to ensure consistent comparison across modalities and fusion settings. The text branch is implemented using TF-IDF unigram and bigram features combined with class-balanced Logistic Regression, while the image branch uses a pretrained ResNet18 backbone fine-tuned for binary classification. Fusion weights are selected on the validation set using ROC-AUC optimization. The overall implementation environment and major experimental components are summarized in Table 2.

Table 2. Experimental configuration of the implemented prototype.

Component	Configuration
Dataset	Open-i/IU X-Ray paired image-report samples
Task	Binary classification: ALERT vs. NORMAL
Validation/test size	75 validation samples and 75 test samples based on confusion-matrix totals
Text preprocessing	Lowercasing, cleaning, token normalization
Text features	TF-IDF unigram and bigram features
Text classifier	Class-balanced Logistic Regression
Image backbone	Pretrained ResNet18 fine-tuned for binary classification
Image preprocessing	Fixed-resolution resizing, intensity normalization, and standard image transformations/augmentations
Fusion method	Probability-level late fusion
Fusion weight selection	Validation ROC-AUC
Selected fusion weight	$\alpha = 0.90$
Text explanation	Logistic Regression coefficient analysis
Image explanation	Grad-CAM heatmaps
Implementation environment	Google Colab, Python, scikit-learn, and PyTorch
Reproducibility note	Low-level CNN training settings such as optimizer, learning rate, batch size, and epochs were not fully reported in the exported manuscript artifacts and should be documented in future releases

3.8. Evaluation Metrics

Model performance is evaluated using accuracy, ROC-AUC, F1-score, precision, recall, and confusion matrices. These metrics are used to compare the text-only, image-only, and fusion-based systems under the same evaluation setting. Accuracy measures the proportion of correctly classified samples, precision measures the proportion of predicted ALERT cases that are truly ALERT, recall measures the proportion of true ALERT cases successfully detected, and F1-score summarizes the balance between precision and recall. ROC-AUC is used to evaluate the ability of the model to rank ALERT cases above NORMAL cases across varying classification thresholds.

Beyond standard classification metrics, additional analyses are conducted to evaluate multimodal behavior and model interpretability more critically. Case-level complementarity analysis is performed to examine whether the fusion model corrects errors produced by unimodal systems. Bootstrap confidence interval analysis is also conducted to assess whether the observed ROC-AUC differences between fusion and text-only models are statistically meaningful rather than descriptive fluctuations.

Because alert-oriented systems rely on probability thresholds, calibration is also an important consideration. However, the current prototype does not yet include a complete calibration analysis using reliability diagrams, Expected Calibration Error (ECE), or Brier score. Consequently, the predicted probabilities should not be interpreted as clinically calibrated risk estimates. Additional explainability-oriented audits are also conducted. A shortcut-feature audit is performed for selected textual features to assess whether influential n-grams represent clinically meaningful cues or report-writing artifacts. In parallel, qualitative Grad-CAM

auditing is conducted for representative image predictions to evaluate whether the highlighted regions correspond to anatomically plausible visual patterns or potential spurious activations.

3.9. Implementation Environment and Artifact Generation

The proposed prototype is implemented in Google Colab using Python. The text-processing pipeline is implemented using scikit-learn, while the image classification and Grad-CAM generation components are implemented using PyTorch. The workflow produces multiple research artifacts to support traceability, reproducibility, and experimental analysis. These artifacts include model performance tables, confusion matrices, ROC-AUC summaries, case-level prediction outputs, bootstrap ROC-AUC summaries, shortcut-feature inspection tables, top n-gram coefficient summaries, Grad-CAM heatmaps, Grad-CAM quality audit records, and inference latency measurements.

These generated artifacts are used to support both quantitative evaluation and qualitative inspection of multimodal model behavior. Nevertheless, the current implementation remains a retrospective experimental prototype and has not yet been evaluated within a prospective clinical workflow or real-world deployment environment.

4. Results and Discussion

4.1. Experimental Overview

This section presents the experimental evaluation of the proposed explainable multimodal framework for binary chest X-ray alert classification. The evaluation compares three model configurations: (1) a text-only model based on TF-IDF features and Logistic Regression, (2) an image-only model using a fine-tuned ResNet18 backbone, and (3) a probability-level late-fusion model combining the outputs of both modalities. In addition to standard classification metrics, several complementary analyses are conducted to evaluate multimodal behavior, uncertainty, and explainability characteristics. These analyses include case-level multimodal complementarity evaluation, bootstrap confidence interval analysis for ROC-AUC improvement, shortcut-feature inspection in the text branch, and qualitative Grad-CAM explanation auditing for the image branch.

The results are interpreted cautiously because the ALERT/NORMAL labels used in this prototype are proxy labels derived from available dataset and report-associated information. Consequently, strong text-based performance may reflect clinically meaningful report cues, but it may also reflect report-writing conventions or potential label leakage. Therefore, the purpose of this evaluation is not to claim clinical deployment readiness, but rather to analyze the behavior, limitations, interpretability, and auditability of a reproducible multimodal prototype system.

4.2. Predictive Performance and Fusion Behavior

Table 3 summarizes the recomputed validation and test performance of the text-only, image-only, and probability-level late-fusion models using the available experimental artifacts. Overall, the text-only model remains the strongest unimodal branch across most evaluation metrics, whereas the image-only model produces substantially lower performance on both validation and test sets.

Table 3. Performance comparison of text-only, image-only, and probability-level late-fusion models.

Model	Val Acc	Val AUC	Val F1	Test Acc	Test AUC	Test F1
Text-only	0.8800	0.9582	0.9011	0.8533	0.8936	0.8817
Image-only	0.7200	0.7409	0.7742	0.6800	0.7279	0.7600
Fusion (α=0.90)	0.8667	0.9514	0.8889	0.8667	0.9025	0.8936

The relatively strong performance of the text branch suggests that the proxy labels are more closely aligned with radiology report semantics than with image-only visual evidence. In contrast, the image branch achieves lower accuracy, ROC-AUC, and F1-score, indicating that visual learning under the current dataset and labeling configuration remains comparatively more challenging.

The probability-level late-fusion model achieves the highest test-set performance among the evaluated configurations. Compared with the text-only baseline, the fusion model improves test accuracy from 0.8533 to 0.8667, test ROC-AUC from 0.8936 to 0.9025, and test F1-score from 0.8817 to 0.8936. However, the observed improvements remain relatively small and should be interpreted cautiously, particularly because the validation metrics do not consistently outperform the text-only configuration. The performance comparisons across validation and test sets are visualized in Figures 4–6. Across all evaluation settings, the image-only model consistently underperforms compared with the text-only branch, while the fusion model provides only modest additional gains over the report-based baseline.

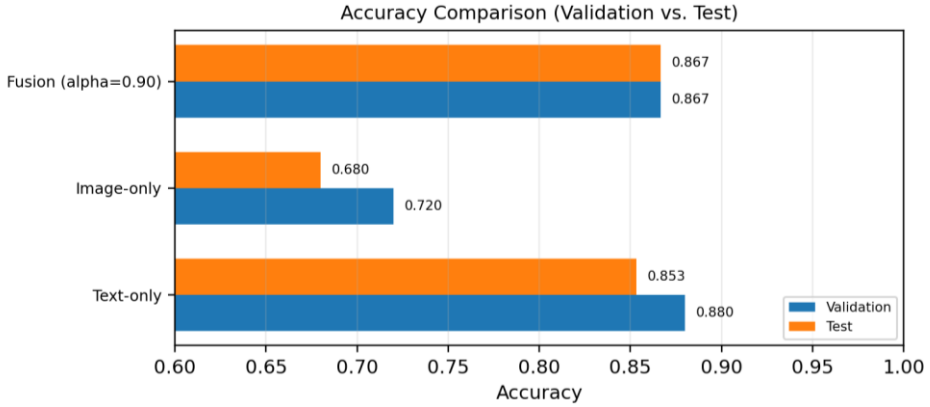


Figure 4. Accuracy comparison across text-only, image-only, and probability-level late-fusion models on validation and test sets.

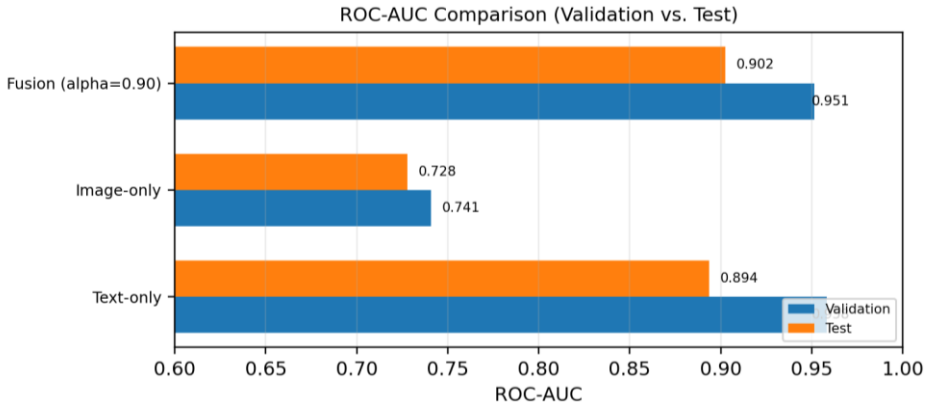


Figure 5. ROC-AUC comparison across text-only, image-only, and probability-level late-fusion models on validation and test sets.

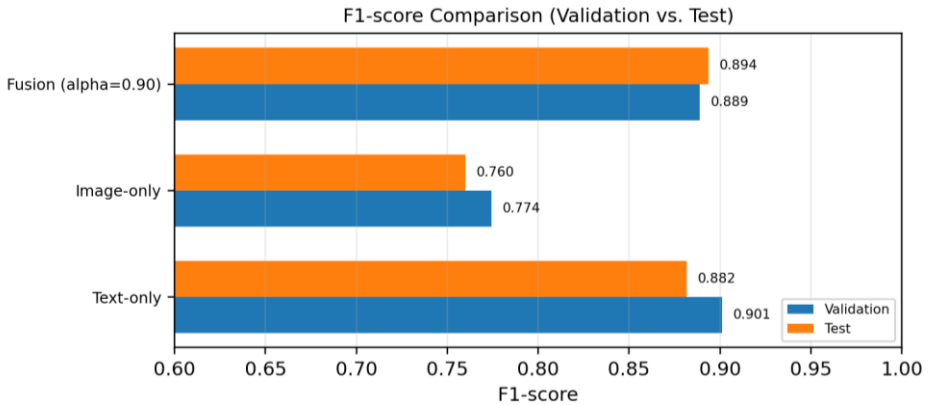


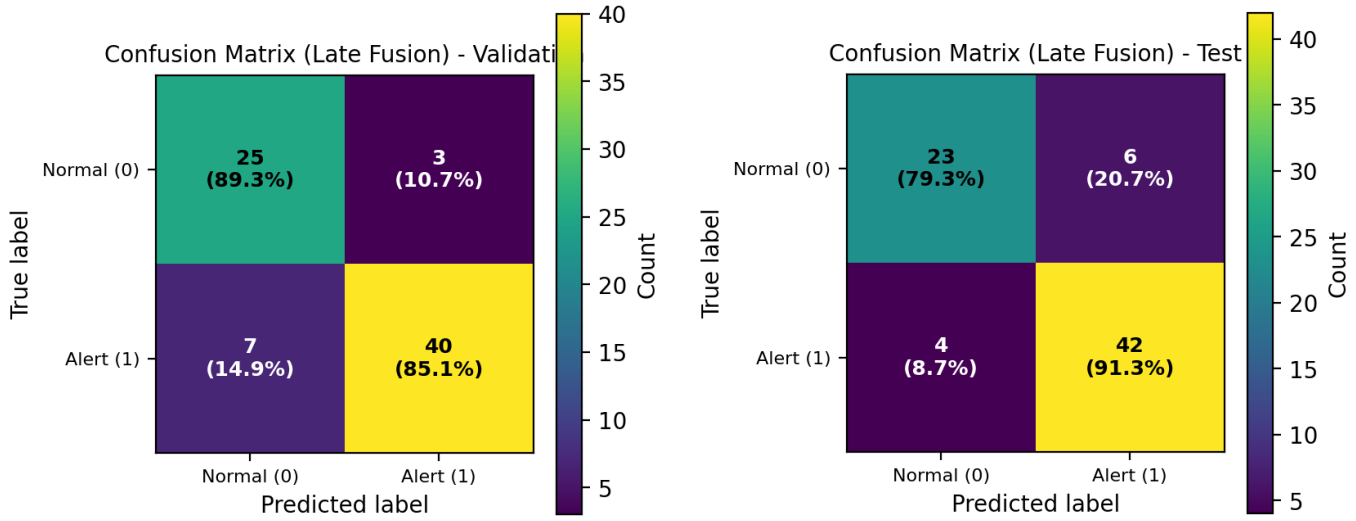
Figure 6. F1-score comparison across text-only, image-only, and probability-level late-fusion models on validation and test sets.

To evaluate whether the observed ROC-AUC improvement from fusion was statistically meaningful, bootstrap confidence interval analysis was conducted for both validation and test performance. The test ROC-AUC of the fusion model was 0.9025, compared with 0.8936 for the text-only baseline, resulting in a delta AUC of +0.0090. However, the 95% bootstrap confidence interval for the delta AUC was [−0.0044,0.0270], with a two-sided bootstrap p-value of 0.2184, for more details see Tabel 4. Since the confidence interval includes zero, the observed improvement should be interpreted as a descriptive performance gain rather than statistically conclusive evidence of multimodal superiority. Similar behavior is also observed on the validation set, where the delta AUC between fusion and text-only performance is slightly negative.

Table 4. Bootstrap ROC-AUC confidence intervals and delta-AUC analysis.

Split	Comparison/Model	AUC or Delta AUC	95% CI	Bootstrap p-value
Validation	Text-only AUC	0.9582	[0.9060, 0.9937]	–
Validation	Fusion AUC	0.9514	[0.8973, 0.9907]	–
Validation	Fusion − Text	−0.0068	[−0.0235, 0.0073]	0.3236
Test	Text-only AUC	0.8936	[0.8081, 0.9616]	–
Test	Fusion AUC	0.9025	[0.8211, 0.9688]	–
Test	Fusion − Text	0.0090	[−0.0044, 0.0270]	0.2184

Additional error-profile analysis was conducted using confusion matrices for the late-fusion model. Figures 7 summarize the validation and test confusion matrices at the selected decision threshold.



(a) (b)

Figure 7. Confusion matrix of the probability-level late-fusion model (**a**) on the validation set; (**b**) on the test set.

On the validation set, the fusion model achieves an accuracy of 0.8667, precision of 0.9302, recall of 0.8511, and F1-score of 0.8889. On the test set, the model achieves an accuracy of 0.8667, precision of 0.8750, recall of 0.9130, and F1-score of 0.8936. The test confusion matrix shows that the fusion model correctly classified 23 NORMAL cases and 42 ALERT cases while producing 6 false positives and 4 false negatives.

The resulting error profile suggests a relatively balanced but still clinically incomplete trade-off between false positives and false negatives. False negatives remain important because potentially relevant cases may be missed, whereas false positives may increase workload and contribute to alert fatigue. Because the current prototype does not include cost-sensitive threshold optimization or clinician-defined operating points, the threshold-based metrics should still be interpreted as preliminary.

4.3. Multimodal Complementarity Analysis

To further investigate whether the image modality provides complementary information beyond the radiology report, a case-level multimodal complementarity analysis was conducted on both validation and test predictions. Rather than relying solely on aggregate metrics, this analysis examines how the fusion model behaves at the individual-case level, particularly in situations where unimodal models disagree.

The analysis focuses on three aspects: (1) whether the fusion model corrects errors produced by the text-only branch, (2) whether the image branch is correct when the text branch is incorrect, and (3) whether fusion degrades predictions that were previously correct in the text-only configuration.

On the test set, the text-only model correctly classified 64 of 75 cases, whereas the fusion model correctly classified 65 of 75 cases. Fusion successfully corrected two cases that were previously misclassified by the text branch. However, fusion also converted one previously correct text prediction into an incorrect prediction. The image branch correctly identified five cases in which the text branch failed, but only two of these cases were successfully recovered by fusion because the image modality contributed only a small portion of the final fused probability.

A similar pattern was observed on the validation set. The image branch correctly identified five cases that were misclassified by the text branch, but these corrections did not propagate effectively into the final fusion output. This behavior indicates that potentially useful visual information exists in selected cases, but its influence remains limited under the current late-fusion weighting strategy.

Overall, the complementarity analysis suggests that the image modality can provide supportive information in specific situations; however, the contribution remains relatively weak and inconsistent compared with the dominant text modality. These findings are consistent with the selected fusion weight (α=0.90) and reinforce the observation that the current multimodal behavior is strongly report-driven under the proxy-label setting. The results of the case-level complementarity analysis are summarized in Table 5.

Table 5. Case-level multimodal complementarity analysis.

Split	n	Text correct	Image correct	Fusion correct	Fusion corrects text error	Image correct when text wrong	Fusion hurts text-correct
Validation	75	66 (88.0%)	54 (72.0%)	65 (86.7%)	0 (0.0%)	5 (6.7%)	1 (1.3%)
Test	75	64 (85.3%)	51 (68.0%)	65 (86.7%)	2 (2.7%)	5 (6.7%)	1 (1.3%)

4.4. Explainability, Shortcut Learning, and Error Analysis

To improve model auditability, explainability analysis was conducted for both the text and image branches. The explainability strategy in this study is designed primarily as a post-hoc audit mechanism rather than as a clinically validated reasoning framework. Consequently, the generated explanations should be interpreted as supportive inspection artifacts that help analyze model behavior, potential shortcut learning, and possible sources of prediction error.

For the text branch, explainability is obtained through Logistic Regression coefficient analysis over TF-IDF n-gram features. Positive coefficients indicate terms associated with the ALERT class, whereas negative coefficients indicate terms associated with the NORMAL class. This approach enables direct inspection of globally influential textual features contributing to the classification boundary. A targeted shortcut-feature audit was conducted for several generic or artifact-prone features highlighted during the review process. Table 6 summarizes the coefficient values and prevalence statistics for selected features across training, validation, and test sets.

The prevalence analysis shows that several generic terms, such as right and left, occur substantially more frequently in ALERT cases than in NORMAL cases. For example, in the test set, right appears in 47.8% of ALERT reports but only 3.4% of NORMAL reports, whereas left appears in 45.7% of ALERT reports and 3.4% of NORMAL reports. Although these laterality-related terms may carry contextual clinical meaning, they are not sufficiently specific to function as standalone clinical indicators. The anonymization token xxxx is

particularly problematic because it reflects dataset formatting rather than medical semantics. Several false-positive predictions also receive positive logit contributions from these features, suggesting that the text branch may partially exploit report-writing conventions, formatting artifacts, or shortcut correlations instead of fully learning clinically grounded reasoning patterns. To further inspect the text classifier behavior, the most influential positive and negative n-grams contributing to the decision boundary are summarized in Table 7.

Table 6. Targeted shortcut-feature audit for selected text features.

Feature	Coefficient	Train ALERT/ NORMAL (%)	Validation ALERT/ NORMAL (%)	Test ALERT/ NORMAL (%)	Interpretation
right	1.1138	30.6 / 2.3	40.4 / 3.6	47.8 / 3.4	Laterality cue; not clinically specific by itself
left	0.9462	26.4 / 2.3	42.6 / 0.0	45.7 / 3.4	Laterality cue; possible shortcut when used alone
changes	0.9007	23.1 / 1.5	21.3 / 3.6	26.1 / 10.3	May reflect abnormality or non-acute report wording
xxxx	0.8133	50.9 / 22.6	55.3 / 14.3	43.5 / 44.8	Anonymization artifact; not a clinical concept

Table 7. Top n-grams contributing to the text classifier decision boundary.

ALERT-associated n-gram	Weight	NORMAL-associated n-gram	Weight
right	1.3182	lungs clear	−1.3985
changes	1.0986	clear	−1.3737
left	1.0953	normal	−1.3475
calcified	0.9840	normal limits	−1.0479
mild	0.9565	limits	−1.0204
xxxx	0.9288	normal lungs	−1.0188
degenerative changes	0.8467	pneumothorax	−0.9880
spine	0.7905	lungs	−0.9870
degenerative	0.7145	silhouette normal	−0.9607
aorta	0.6991	cardiomediastinal	−0.8994

These findings do not imply that the identified features are clinically irrelevant. Some terms, including laterality cues and change-related expressions, may indeed reflect meaningful radiological context when interpreted within complete report narratives. However, their strong coefficients and prevalence within false-positive predictions indicate that coefficient-based explanations should primarily be interpreted as audit-support tools rather than direct clinical justification mechanisms.

Explainability analysis was also conducted for the image branch using Grad-CAM heatmaps [21]. Grad-CAM was applied to visualize image regions contributing most strongly to the CNN prediction. To support qualitative interpretation, three representative examples were selected for detailed inspection: (1) a high-confidence ALERT case, (2) a high-confidence NORMAL case, and (3) a borderline or error-prone ALERT case. Representative Grad-CAM visualizations are shown in Figure 8, while the qualitative explanation audit is summarized in Table 8.

The Grad-CAM audit reveals mixed explanation quality across representative cases. The high-confidence ALERT example demonstrates partially plausible activation over thoracic regions; however, the highlighted regions remain broad and insufficiently localized to support disease-level interpretation. The borderline ALERT example does not clearly localize the expected pathological pattern, whereas the NORMAL example shows activation concentrated near a lateral image boundary, suggesting potential sensitivity to peripheral image artifacts.

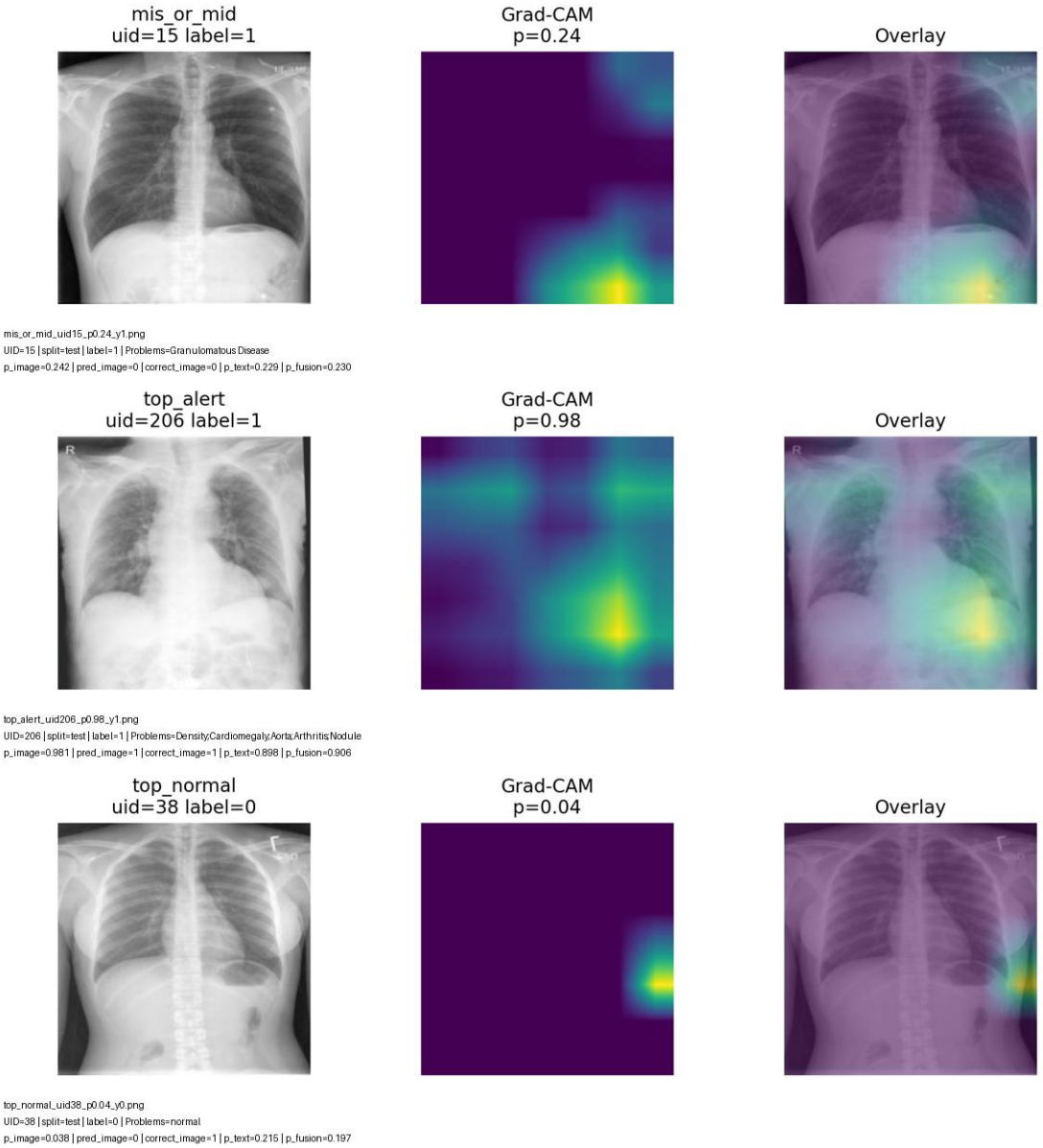


Figure 8. Representative Grad-CAM visualization examples for high-confidence and error-prone image predictions. From top to bottom: (1) borderline/error-prone ALERT case, (2) high-confidence ALERT case, and (3) high-confidence NORMAL case. From left to right: original chest X-ray image, Grad-CAM activation map, and Grad-CAM overlay visualization. The reported p_{image} values indicate the ALERT probability produced by the image branch.

These observations support the use of Grad-CAM as a qualitative inspection and audit tool rather than as validated evidence of clinically reliable visual reasoning. The generated heatmaps provide useful insight into model attention behavior, but they do not establish causal clinical interpretation. The combined explainability and error analyses also reveal evidence of potential label leakage and shortcut learning. Several false-positive predictions occur when reports contain ambiguous or artifact-prone expressions such as stable, changes, left, or xxxx. These features may resemble abnormality-related language or reporting conventions without necessarily indicating acute clinical concern. Conversely, false negatives may occur when the ALERT label is assigned to cases containing largely normal pulmonary descriptions but additional device-related findings, chronic abnormalities, or subtle non-pulmonary observations.

This behavior highlights two important methodological limitations. First, the ALERT/NORMAL categorization does not necessarily represent a single clinically coherent concept. ALERT cases may include heterogeneous findings ranging from pulmonary

abnormalities to chronic findings, device-related observations, or report-level cues. Second, because radiology reports are used both as model input and as part of the proxy-label generation process, the text branch may partially learn report-writing patterns that are strongly correlated with the target label.

Table 8. Representative Grad-CAM explanation examples and qualitative interpretation.

UID	Case type	p_{image}	Grad-CAM focus	Plausibility	Audit interpretation
15	Borderline/error-prone ALERT	0.2418	Lower thoracic / diaphragmatic-peripheral region	Limited	Does not clearly localize a clinically specific granulomatous pattern; consistent with an error-prone case
206	High-confidence ALERT	0.9807	Lower lung / cardiomediastinal and bilateral lung-field regions	Partially plausible	Overlaps with plausible thoracic regions but remains broad and not disease-specific
38	High-confidence NORMAL	0.0384	Right lateral image edge / peripheral region	Weak	Strongest activation near image boundary suggests possible edge or peripheral artifact focus

Taken together, the explainability analyses demonstrate that explanation methods are valuable not only for increasing transparency, but also for identifying potential weaknesses in dataset construction, shortcut learning behavior, and modality imbalance. In the current prototype, explainability functions primarily as an audit-support mechanism that helps characterize model behavior and limitations rather than as definitive evidence of clinically grounded reasoning.

4.5. Latency and Deployment Considerations

Table 9 summarizes the average per-sample inference latency measured in the implemented prototype environment. The results indicate that text inference contributes only negligible computational overhead compared with image inference. The image branch dominates the total computational cost because it requires a convolutional neural network forward pass, whereas the TF-IDF and Logistic Regression pipeline remains computationally lightweight. The probability-level fusion stage also introduces minimal additional overhead, supporting the practicality of the proposed framework for prototype-scale multimodal inference.

Table 9. Average inference latency measured in the prototype environment.

Component	Latency (ms/sample)
Text-only inference	0.0590
Image-only inference	11.5430
Fusion computation	9.7190

Although the measured latency supports near real-time inference under the current experimental configuration, these results should be interpreted cautiously. All experiments were conducted in a Google Colab environment and therefore do not directly represent deployment conditions in clinical infrastructure, edge devices, or production hospital systems. Actual inference latency may vary substantially depending on hardware acceleration, batching strategy, preprocessing overhead, model optimization, memory constraints, and system-level integration.

From a deployment perspective, the relatively lightweight computational profile of the text branch and the modularity of probability-level late fusion may provide practical advantages for reproducible prototype implementation. However, deployment feasibility should not be interpreted as evidence of clinical readiness. Reliable real-world deployment would

additionally require prospective validation, calibration analysis, workflow integration assessment, and clinician-centered usability evaluation.

4.6. Synthesis of Experimental Insights

The experimental analyses provide several important insights regarding multimodal alert classification under a proxy-label setting. First, the results demonstrate that proxy-label construction can naturally produce text-dominant multimodal behavior because radiology reports are more directly aligned with the label-generation process than image pixels. This observation explains why the selected fusion weight strongly favors the text modality and why the image branch contributes only limited complementary information under the current configuration.

Second, the explainability analyses demonstrate that explanation methods are useful not only for increasing model transparency, but also for identifying shortcut learning behavior and potential dataset artifacts. The coefficient audit and shortcut-feature analysis reveal that generic or formatting-related features such as right, left, changes, and xxxx can influence ALERT predictions despite lacking sufficient standalone clinical specificity. Similarly, the Grad-CAM audit shows that visual explanations may sometimes focus on broad thoracic regions or peripheral image areas rather than clearly localized pathological evidence.

Third, the findings show that probability-level late fusion does not automatically produce strong multimodal complementarity. Although fusion corrected several text-only errors, the image branch failed to consistently improve final predictions because its contribution remained relatively small within the selected weighting configuration. In several cases, image predictions were correct when text predictions failed, yet the final fusion output remained dominated by the report-based branch.

These observations reinforce the importance of evaluating multimodal systems beyond aggregate performance metrics alone. In clinical AI settings, robustness, traceability, explanation quality, and label validity are often as important as predictive accuracy itself [27]. The present findings also support broader observations in multimodal healthcare AI that additional modalities are beneficial only when they contribute sufficiently independent and reliable information for the target prediction task [28]. Under the current prototype setting, the proposed framework functions primarily as an auditable report-driven multimodal system with limited but observable image-based complementary behavior.

4.7. Limitations and Future Work

Several limitations should be considered when interpreting the findings of this study. First, the ALERT/NORMAL labels are proxy labels rather than expert-curated clinical outcome labels. Consequently, the semantic meaning of ALERT may remain heterogeneous, potentially combining pulmonary abnormalities, chronic findings, device-related observations, and report-level cues within a single binary category. Second, the use of radiology reports as both model input and partial label source introduces a potential label-leakage pathway. The strong performance of the text branch may therefore reflect not only clinically meaningful report interpretation but also report-writing conventions and label-associated textual patterns. The shortcut-feature audit supports this concern by showing that generic report terms and anonymization artifacts can influence ALERT predictions.

Third, the experimental evaluation relies on a single public dataset without external validation or cross-dataset testing. Therefore, the generalizability of the proposed framework across institutions, reporting styles, imaging protocols, and patient populations remains uncertain. In addition, the relatively small paired subset used for multimodal experiments may limit the robustness of image representation learning. Fourth, although the multimodal evaluation was strengthened through case-level complementarity analysis and bootstrap uncertainty analysis, the observed fusion improvement remains relatively modest and statistically inconclusive under the current experimental setting. Consequently, the proposed framework should not yet be interpreted as strong evidence of multimodal superiority.

Fifth, the explainability analysis remains qualitative and modality-specific. Logistic Regression coefficients and Grad-CAM heatmaps are useful for auditing model behavior, but they do not provide validated clinical explanations or unified cross-modal reasoning interpretation. Calibration analysis is also incomplete, and the current prototype has not been evaluated prospectively or reviewed directly by clinicians in operational workflows. Therefore, the

proposed framework should be interpreted as a reproducible and auditable research prototype rather than a clinically deployable decision-support system.

Future work should prioritize clinically curated labels independent from report-derived shortcuts, separation of heterogeneous alert categories into clinically meaningful subtypes, larger-scale multimodal training, external validation across institutions, calibration-aware evaluation, quantitative explainability assessment, and robustness analysis under missing-modality conditions. Additional studies are also needed to determine when image information provides genuinely complementary value beyond report-based prediction alone..

5. Conclusions

This study developed and evaluated an explainable multimodal framework for binary chest X-ray alert classification using paired radiology reports and chest X-ray images. The proposed framework combines a TF-IDF and Logistic Regression text branch, a fine-tuned ResNet18 image branch, and a probability-level late-fusion strategy. To support model auditability, modality-specific explainability mechanisms were incorporated through coefficient-based text inspection and Grad-CAM visualizations.

The experimental results indicate that the radiology report provides the dominant predictive signal under the current proxy-label setting. Although the late-fusion model achieved modest improvements over the text-only baseline, the bootstrap confidence interval analysis suggests that the observed ROC-AUC gain is not statistically conclusive. The selected fusion weight further confirms that the current multimodal behavior remains strongly report-driven, with the image modality contributing only limited complementary information.

The additional analyses also demonstrate that explainability can serve as an effective audit-support mechanism for identifying shortcut learning, modality imbalance, and potential dataset artifacts. The text audit revealed the influence of generic report-writing patterns and anonymization artifacts, while the Grad-CAM inspection showed mixed localization quality across representative image cases. These findings highlight that explanation outputs should be interpreted as qualitative inspection tools rather than validated clinical reasoning evidence.

Several limitations remain important. The ALERT/NORMAL labels are proxy labels rather than expert-curated clinical outcomes, and the use of radiology reports as both model input and partial label source introduces a potential label-leakage pathway. In addition, the study relies on a single public dataset without external validation, comprehensive calibration analysis, quantitative explanation evaluation, or clinician-centered assessment. Consequently, the proposed framework should be interpreted as a reproducible and auditable research prototype rather than a clinically deployable decision-support system.

Future work should focus on clinically curated labels, stronger separation of heterogeneous alert categories, external validation across institutions, calibration-aware evaluation, quantitative explainability assessment, and more rigorous multimodal robustness analysis. These directions may help clarify when multimodal integration provides clinically meaningful complementary value beyond report-based prediction alone..

Author Contributions: Conceptualization: E.W.; Methodology: E.W., I.M.N., and A.K.; Software: E.W. and S.A.; Validation: E.W., I.M.N., A.K., S.A., I.E.W., and P.W.A.; Formal analysis: E.W. and I.M.N.; Investigation: E.W. and S.A.; Resources: E.W.; Data curation: E.W. and S.A.; Writing-original draft preparation: E.W.; Writing-review and editing: E.W., I.M.N., A.K., S.A., I.E.W., and P.W.A.; Visualization: E.W. and S.A.; Supervision: E.W.; Project administration: E.W. All authors have read and approved the revised version of the manuscript.

Funding: This research was funded by the International Collaborative Research Grant (Hibah Penelitian Kerjasama Luar Negeri) provided by Universitas Muhammadiyah Semarang, Indonesia, during the 2025-2026 funding period (059/UNIMUS.L/PG/PKLN/PJ.INT/2025).

Data Availability Statement: The data used in this study were derived from publicly available chest X-ray resources, in-cluding the Open-i/IU X-Ray collection. Processed experimental artifacts supporting the reported findings are available from the corresponding author upon reasonable request, subject to data-use and ethical considerations

Acknowledgments: The authors thank the institutions and collaborators who supported this research. Parts of the implementation and manuscript preparation involved standard computational tools for data processing and visualization. Some illustrative diagrams and language editing were assisted by AI-based tools; all scientific content, experimental analyses, interpretations, and final manuscript decisions were reviewed and approved by the authors.

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

References

[1] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Med.*, vol. 17, no. 1, p. 195, Dec. 2019, doi: 10.1186/s12916-019-1426-2.

[2] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.

[3] D. Demner-Fushman *et al.*, "Preparing a collection of radiology examinations for distribution and retrieval," *J. Am. Med. Informatics Assoc.*, vol. 23, no. 2, pp. 304–310, Mar. 2016, doi: 10.1093/jamia/ocv080.

[4] N. Dewaswala *et al.*, "Natural language processing for identification of hypertrophic cardiomyopathy patients from cardiac magnetic resonance reports," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, p. 272, Oct. 2022, doi: 10.1186/s12911-022-02017-y.

[5] B. Zhou, G. Yang, Z. Shi, and S. Ma, "Natural Language Processing for Smart Healthcare," *IEEE Rev. Biomed. Eng.*, vol. 17, pp. 4–18, 2024, doi: 10.1109/RBME.2022.3210270.

[6] S. Sheikhalishahi, R. Miotto, J. T. Dudley, A. Lavelli, F. Rinaldi, and V. Osmani, "Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review," *JMIR Med. Informatics*, vol. 7, no. 2, p. e12239, Apr. 2019, doi: 10.2196/12239.

[7] D. Jin, E. Sergeeva, W. Weng, G. Chauhan, and P. Szolovits, "Explainable deep learning in healthcare: A methodological survey from an attribution view," *WIREs Mech. Dis.*, vol. 14, no. 3, May 2022, doi: 10.1002/wsbm.1548.

[8] G. Huang, Y. Li, S. Jameel, Y. Long, and G. Papanastasiou, "From explainable to interpretable deep learning for natural language processing in healthcare: How far from reality?," *Comput. Struct. Biotechnol. J.*, vol. 24, pp. 362–373, Dec. 2024, doi: 10.1016/j.csbj.2024.05.004.

[9] A. S. Egbunu and A. M. Okedoye, "Harnessing Artificial Intelligence for Early Disease Detection: Opportunities and Challenges in Modern Healthcare," *J. Comput. Theor. Appl.*, vol. 3, no. 3, pp. 384–401, Feb. 2026, doi: 10.62411/jcta.15367.

[10] A. E. W. Johnson *et al.*, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Sci. Data*, vol. 6, no. 1, p. 317, Dec. 2019, doi: 10.1038/s41597-019-0322-0.

[11] J. Irvin *et al.*, "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, pp. 590–597, Jul. 2019, doi: 10.1609/aaai.v33i01.3301590.

[12] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 3462–3471. doi: 10.1109/CVPR.2017.369.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, vol. 2016-Decem, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[14] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 2261–2269. doi: 10.1109/CVPR.2017.243.

[15] K. Pyar, "Segmentation Performance Analysis of Transfer Learning Models on X-Ray Pneumonia Images," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 1, pp. 64–74, Jun. 2024, doi: 10.62411/faith.2024-10.

[16] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manag.*, vol. 24, no. 5, pp. 513–523, Jan. 1988, doi: 10.1016/0306-4573(88)90021-0.

[17] N. Fadul, M. F. Alaskar, K. B. Jillahi, and D. B. El-Khaled, "Generative AI in Healthcare: An Analytical Review of Models, Clinical Applications, and Decision-Support Implications," *J. Futur. Artif. Intell. Technol.*, vol. 2, no. 4, pp. 587–615, Dec. 2025, doi: 10.62411/faith.3048-3719-298.

[18] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998, doi: 10.1109/34.667881.

[19] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 2011, pp. 689–696.

[20] E. O. Ibam and J. B. Oluwagbemi, "Multimodal Deep Learning for Pneumonia Detection Using Wearable Sensors: Toward an Edge-Cloud Framework," *J. Comput. Theor. Appl.*, vol. 3, no. 3, pp. 314–333, Jan. 2026, doi: 10.62411/jcta.14944.

[21] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, vol. 128, no. 2, pp. 618–626. doi: 10.1109/ICCV.2017.74.

[22] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 9525–9536. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/file/294a8ed24b1ad22ec2e7efea049b8737-Paper.pdf

[23] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, "A benchmark for interpretability methods in deep neural networks," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/fe4b8556000d0f0cae99daa5c5c5a410-Paper.pdf

[24] D. S. Stamoulis and C. Papachristopoulou, "Artificial Intelligence in Radiology, Emergency, and Remote Healthcare: A Snapshot of Present and Future Applications," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 3, pp. 228–234, Oct. 2024, doi: 10.62411/faith.3048-3719-38.

[25] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?' : Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, vol. 13-17-Augu, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[26] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, Nov. 2017, pp. 4768–4777. [Online]. Available: http://arxiv.org/abs/1705.07874

[27] K. Lekadir *et al.*, "FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare," *BMJ*, vol. 388, p. e081554, Feb. 2025, doi: 10.1136/bmj-2024-081554.

[28] L. R. Soenksen *et al.*, "Integrated multimodal artificial intelligence framework for healthcare applications," *npj Digit. Med.*, vol. 5, no. 1, p. 149, Sep. 2022, doi: 10.1038/s41746-022-00689-4.