Multi_Modal_Medical_Image_Classification.pdf

Turnitin

Document Details

Submission ID	trn:oid:::31142:104078252
Submission Date	Jul 10, 2025, 9:21 PM GMT+5
Download Date	Jul 10, 2025, 9:22 PM GMT+5
File Name	Multi_Modal_Medical_Image_Classification.pdf
File Size	3.7 MB

6 Pages

3,873 Words

22,862 Characters

20% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Match Groups

49 Not Cited or Quoted 17%
Matches with neither in-text citation nor quotation marks

11 Missing Quotations 3%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

13%	Internet sources
16%	Publications
14%	Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

49 Not Cited or Quoted 17%
Matches with neither in-text citation nor quotation marks

11 Missing Quotations 3%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

13%	Internet sources
16%	Publications
14%	Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Publication	Ryan Wang, Po-Chih Kuo, Li-Ching Chen, Kenneth Patrick Seastedt, Judy Wawira G...	1%
2	Internet	export.arxiv.org	1%
3	Internet	www.mdpi.com	1%
4	Submitted works	University of Sydney on 2019-11-30	1%
5	Publication	Catarina Moreira, Yu-Liang Chou, Chihcheng Hsieh, Chun Ouyang, João Pereira, Jo...	<1%
6	Internet	doaj.org	<1%
7	Publication	C A Vidya, V.Baby Shalini. "DeepELR: Deep Learning-Based Energy And Link Stabili...	<1%
8	Publication	Pratham Kaushik, Eshika Jain, Vinay Kukreja, Shanmugasundaram Hariharan et a...	<1%
9	Internet	publications.eai.eu	<1%
10	Submitted works	University of New South Wales on 2021-12-17	<1%

11	Internet	mospace.umsystem.edu	<1%
12	Publication	"Machine Learning in Medical Imaging", Springer Science and Business Media LL...	<1%
13	Submitted works	University of Tabuk on 2025-03-08	<1%
14	Publication	A. Nivashini, M. Krishnamurthy. "Revolutionizing lung cancer classification throu...	<1%
15	Publication	Chihcheng Hsieh, Isabel Blanco Nobre, Sandra Costa Sousa, Chun Ouyang et al. "...	<1%
16	Submitted works	Adtalem Global Education on 2025-04-13	<1%
17	Submitted works	Kennesaw State University on 2024-11-19	<1%
18	Internet	arxiv.org	<1%
19	Publication	Mohamed Lahby, Al-Sakib Khan Pathan, Yassine Maleh. "Combatting Cyberbullyi...	<1%
20	Publication	"Medical Image Computing and Computer Assisted Intervention – MICCAI 2019", ...	<1%
21	Publication	Fanguo Zeng, Rui Wang, Youming Jiang, Zhendong Liu, Youchun Ding, Wanjing D...	<1%
22	Publication	Adnane Ait Nasser, Moulay A. Akhloufi. "Classification of CXR Chest Diseases by E...	<1%
23	Submitted works	University of Sydney on 2023-11-05	<1%
24	Internet	www.boerse-duesseldorf.de	<1%

25	Internet	www.medrxiv.org	<1%
26	Submitted works	Asia Pacific University College of Technology and Innovation (UCTI) on 2025-05-22	<1%
27	Submitted works	University of Newcastle upon Tyne on 2023-08-25	<1%
28	Publication	Givens, Ryan N., Karl C. Walli, and Michael T. Eismann. "A multimodal approach to...	<1%
29	Publication	H.L. Gururaj, Francesco Flammini, S. Srividhya, M.L. Chayadevi, Sheba Selvam. "Co...	<1%
30	Submitted works	Indian Institute of Technology, Madras on 2023-03-02	<1%
31	Publication	Jing Ru Teoh, Jian Dong, Xiaowei Zuo, Khin Wee Lai, Khairunnisa Hasikin, Xiang W...	<1%
32	Submitted works	Liverpool John Moores University on 2020-08-06	<1%
33	Submitted works	Nanyang Technological University on 2021-10-18	<1%
34	Submitted works	Universidade do Porto on 2021-02-07	<1%
35	Submitted works	University of Sheffield on 2025-01-24	<1%
36	Publication	Wellington Pinheiro dos Santos, Juliana Carneiro Gomes, Maíra Araújo de Santan...	<1%
37	Publication	Zhou, Guanglin. "Navigating Distribution Shifts in ML through Causality and Foun...	<1%
38	Publication	Cong Ngo Van, Duc-Nghia Tran, Do The Duong, Duc-Tan Tran. "Chapter 2 Utilizing...	<1%

39	Submitted works	National Institute of Business Management Sri Lanka on 2025-04-25	<1%
40	Publication	Nhu-Y Tran-Van, Kim-Hung Le. "A multimodal skin lesion classification through cr...	<1%
41	Submitted works	Queen Mary and Westfield College on 2025-07-09	<1%
42	Submitted works	University of Newcastle upon Tyne on 2024-08-14	<1%
43	Submitted works	IUBH - Internationale Hochschule Bad Honnef-Bonn on 2025-04-12	<1%
44	Publication	Jovito Colin, Nico Surantha. "Interpretable Deep Learning for Pneumonia Detecti...	<1%
45	Publication	Jinzhong Yang, Gregory C. Sharp, Mark J. Gooding. "Auto-Segmentation for Radiat...	<1%

Exploring Fusion Strategies for Multi-Modal Pneumonia Classification with Modality-Specific Explainability

Nazihah Islam Nawreen
Dept. of CSE
BRAC University
Dhaka, Bangladesh
nazihah.islam.nawreen@g.bracu.ac.bd

Azwad Aziz
Dept. of CSE
BRAC University
Dhaka, Bangladesh
azwad.aziz@g.bracu.ac.bd

***Abstract*—Pneumonia diagnosis using chest radiographs has been widely studied thoroughly across several research, but most approaches rely solely on imaging data, often overlooking the valuable metadata associated with medical imaging. In this study, we explore the impact of multi-modal learning utilizing data fusion by combining chest X-ray images with tabular metadata (including age, gender, view position, and pixel spacing) for binary classification of Pneumonia vs. No Finding using a subset of the NIH Chest X-ray dataset. We analyzed three fusion strategies: early fusion, intermediate fusion, and late fusion, each integrating image and tabular modalities at different stages of the architecture. Furthermore, we also anaylzed the performance of using a single modality (image data only) to further justify the superiority of multi-modal tasks in the context of medical image classification. To ensure a fair comparison, all experiments are conducted on a balanced subset of 10,000 samples (5,000 per class) with consistent architectures. Our results show that intermediate fusion and late fusion achieves the best performance, outperforming both single modality and other fusion strategies in accuracy, F1-score, and AUC-ROC. Transparency of the model is further enhanced by incorporating GradCAM for visual explainability and LIME for tabular feature interpretation. This study highlights the complementary advantages of multi-modal fusion and provides insight into the role of metadata in medical image classification.**

***Index Terms*—Fusion, CXR Image, Classification, Multi-Modal, GradCAM, LIME, XAI, Deep Learning**

I. Introduction

Chest x-ray radiography (CXR) is widely performed as the most routine medical imaging examination, and has served as an essential first-line imaging tool to diagnose various thoracic diseases such as pneumonia, pneumothorax, and lung nodules [10]. However, interpretation of CXRs is operator-dependent, time consuming and may lead to delays in diagnosis, particularly in resource-constrained or in high patients clinical settings. With around two billion radiographs performed globally every year, there is an urgent need for automated systems that can help radiologists in triaging abnormal studies and alleviate workload [11].

The rise of deep learning, especially Convolutional Neural Networks (CNNs), has achieved great success in medical image analysis.Specifically, the application of CNNs to CXR interpretation has gained significant traction, with models demonstrating capabilities in detecting various thoracic pathologies [10]. A key enabler for this progress is the availability of large-scale, publicly accessible datasets. The National Institutes of Health (NIH) Chest X-ray dataset, released in 2017, is one such resource, comprising 112,120 frontal-view X-ray images from 30,805 unique patients, each labeled for the presence or absence of 14 common thoracic pathologies [1] .

Multi-modal research in medical imaging aggregates multiple types of data (imaging, clinical reports, table based metadata) to improve the diagnostic accuracy and the robustness of the model.

In addition, XAI has an important role in providing explainable insights into the model decision [12],finally so that solves the black-box problem of deep learning. XAI techniques, such as attention maps and feature importance scores, help clinical end users trust and rationale the system's output and thus contribute to the adoption of automated systems in sensitive healthcare domains. One possible first step, would be to determine normal and abnormal scans. An early dismissal of normals allows the clinician to focus on characterizing the precise nature of the abnormality in the remaining studies. We have incorporated these in our research for model generalizability and explainability.

Therefore our main contributions are:

- Proposed lightweight architecture incorporating fusion strategies by utilizing the MobileNetV3 model which is parameter efficient.
- Systematic comparison of the fusion Strategies along with the perfomance of Single Modality for pneumonia classification
- Modality-Specific Explainability, utilizing GradCAM and LIME for Image and tabular data respectively, to enhance transparency for multi-modal tasks.

II. Literature Review

Pneumonia detection from chest radiographs has been a widely explored topic within the medical imaging and deep

TABLE I: Summary of Literature on Pneumonia Classification with Fusion and XAI

Paper	Modality Used	XAI Used	Limitations
[1]	CXR Only	No	Focused only on image classification, no fusion or XAI
[2]	MRI Only	No	No explainability
[3]	Tabular data	No	No fusion utilized
[4]	Image, Report and Demographic	Yes	XAI was not used for all modalities
[5]	Image and Audio	No	Explainability not provided
[6]	Images and Tabular	No	No explainability
[7]	Image and Tabular	Yes	Limited Samples
[8]	Image and Tabular	No	No Explainability
[9]	CXR Only	Yes	Fusion not utilized

learning research communities. Numerous studies have addressed this task using convolutional neural networks (CNNs) applied to large-scale public datasets such as the NIH ChestX-ray14 dataset [1]. Other researchers have also worked with other types of medical images such as the author of [2] worked with MRI images for brain tumor classification and achieved about 99% accuracy using the well known CNN models like MobileNet , VGG etc. Researches have also utilized traditional ML models for simpler tasks like classifying Cardiovascular diseases where the authors of [3] used a custom dataset of 12 features and 70,000 samples and achieved about 90% accuracy using their proposed hybrid linear regression model. As, the complexity of task increases with greater challenges to achieve higher success in disease identification for medical images, researches have drifted towards utilizing multiple modalities of data to improve the performance. Authors of [4] enhanced the performance by incorporating three different modalities which includes patient demographic, reports and medical images. They also provided explainability coupled with expert's analysis to ensure reliability of their approach. On the other hand, authors of the paper [5] worked with audio data extracted from the sound produced by the lungs and used it to create spectogram images which was used for classification of respiratory diseases using CNN architecture and achieved over 95%. Furthermore, authors of [6] used the MIMIC CXR dataset where they used images and tabular data for classification task and achieved 92.77% where the intermediate fusion performed best. In another study [7] authors developed an interpretable deep learning framework for pneumonia detection using lung CT scans, integrating U-Net for precise lung segmentation, novel graph-based feature representation and transformer-based models for contextual feature extraction. They achieved high diagnostic accuracy of 94% and enhanced interpretability through Grad-CAM and attention maps. Similarly, Hsieh et al. [8] introduced MDF-Net, a novel multimodal deep learning architecture designed to improve disease localization in chest X-ray images by fusing them with patients' clinical data. Building upon Mask R-CNN, their architecture employs a unique ’spatialization’ strategy and two fusion methods to simultaneously process image and structured clinical data. The approach significantly enhances diagnostic accuracy, demonstrating a 12% improvement in Average Precision for disease localization compared to using chest X-rays alone. Another study by the authors of [9] develops an interpretable deep learning framework for pneumonia detection using CXR images, using a ResNet50 architecture and evaluating four interpretability techniques: Layer-wise Relevance Propagation (LRP), Adversarial Training, Class Activation Maps (CAMs), and the Spatial Attention Mechanism (SAM). The research demonstrates that LRP is the most effective method, achieving high diagnostic accuracy (0.91) and strong interpretability (0.85 Mean Relevance Score) without sacrificing performance, making it highly suitable for clinical integration.

A. Research Gap

In summary, while there exists a lot of research in image-based Pneumonia classification, gaps remain in three major areas: (1) systematic comparison of fusion strategies, (2) balanced evaluation with a clean set of data to avoid class imbalance bias, and (3) joint application of GradCAM and LIME to interpret model behavior across modalities. Our work addresses these research gaps by evaluating early, intermediate, and late fusion strategies using a balanced subset of the NIH dataset containing 5000 Pneumonia and 5000 No Finding samples. Furthermore, we utilized GradCAM and LIME to provide end-to-end explainability for both image and tabular data, contributing a novel benchmark for interpretable multi-modal classification. Table I provides the entire summary of all the reviewed literature.

III. Dataset

The utilized dataset in this research is the medical imaging dataset from the National Institutes of Health (NIH) Clinical Center [1]. It is composed of chest X-ray in the frontal view with 112,120 images from 30,805 unique patients over 1992-2015.

Tabular features of the NIH Chest X-ray Dataset (defined in Table II) undergo preprocessing with the following steps,

TABLE II: Tabular Features and Their Preprocessing Techniques

Feature Name	Original Data Type	Preprocessing Applied
Age	Numerical (Continuous)	Min-Max normalized to the [0, 1]
Gender	Categorical (Binary: M/F)	One-hot Encoding
View Position	Categorical (Binary: PA/AP)	One-hot Encoding
Pixel Spacing	Numerical (Continuous)	Median imputation

to render them more suitable for machine learning. The continuous numeric value of ”age” is min-max normalized into the interval between 0 and 1 in order to standardize the scale. Both Gender and View Position, categorical binary features (M/F and PA/AP) are one-hot encoded to represent their inference in a numerical format to be used in the model. Another numeric feature: Pixel Spacing is preprocessed by median imputation for missing values.

Table III summarizes the distribution of No Finding and Pneumonia binary classification, with 4,000 images in the training set, 500 for the validation set, and 500 for the test set for the Pneumonia class and 4,000, 1,000 and 1,000 respectively for the No Findings class (8,000,1,500 and 1,500 for the training, validation and test datasets, respectively). This fair split, 80:15:15, allows for strong classifiers and fair evaluation, taking advantage of metadata of the dataset.

TABLE III: Dataset Split for Binary Classification of Chest X-rays (Pneumonia vs No Findings)

Class	Training Set	Validation Set	Test Set
Pneumonia	4000	500	500
No Findings	4000	500	500
Total Images	8000	1000	1000

IV. Methodology

We proposed an effective approach incorporating multi-modal information for improved pneumonia diagnosis. The chest-X-ray images and associated tabular data (e.g. age, gender, view position, pixel spacing) are first fused together. The preprocessing of tabular data consists in normalization to standardize the continuous features (age for example) between 0 and 1, one-hot encoding of categorical features (gender and view position), and median imputation to handle missing values (pixel spacing for example). In image processing, resampling is also performed while resizing them to constant resolution and rotation augmentation is also applied for (i) the robustness of the model and (ii) the variability in image orientation.

After preprocessing, we trained four different models to compare fusion strategies: early fusion, intermediate fusion, late fusion, and a single-modality baseline. Early fusion fuses image and tabular features at the input layer, intermediate fusion integrates them at one hidden layer, and late fusion aggregates predictions of distinct image and tabular models. The unimodality model uses information of only images for comparison. Each model was optimized using MobileNET_V3 Small variant tailored to the fusion strategy.

Several evaluation metrics: for example, f1-score, recall, latency etc were computed for a thorough comparative analysis between the different strategies which ensures clarity for the evaluation process. Model interpretability were implemented by employing Grad-CAM for visual explanations and marking relevant areas in X-ray images that influenced predictions and offered insights specific to the modality. Further the LIME (Local Interpretable Model-agnostic Explanations) was used to provide local interpretability, explaining each prediction by approximating the model locally around the individual instances. This twin-explainability framework guarantees global as well as local explainability, leading to clinical trust in and validation of the multi-modal classification architecture. Figure 1 show the entire workflow of our proposed strategy.

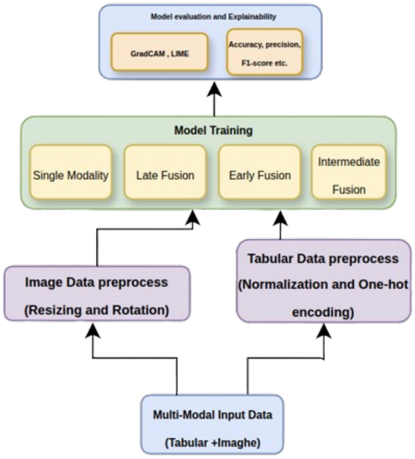
Model evaluation and Explainability
GradCAM , LIME
Accuracy, precision, F1-score etc.
Model Training
Single Modality
Late Fusion
Early Fusion
Intermediate Fusion
Image Data preprocess (Resizing and Rotation)
Tabular Data preprocess (Normalization and One-hot encoding)
Multi-Modal Input Data (Tabular +Imaghe)

Fig. 1: Workflow

V. Fusion Models Used

To leverage both visual and contextual information, we implemented multiple strategies for combining chest X-ray images with tabular data. The goal of these fusion techniques is to enhance classification performance by allowing the model to learn from various information sources. We implemented and compared three multi-modal fusion strategies—early fusion, intermediate fusion, and late fusion—alongside a uni-modal baseline that uses only image data. Each fusion strategy integrates the two modalities at different stages of the network and exhibits a unique trade-off between complexity and expressiveness. Figure 2 shows the three different fusion architectures used for comparison.

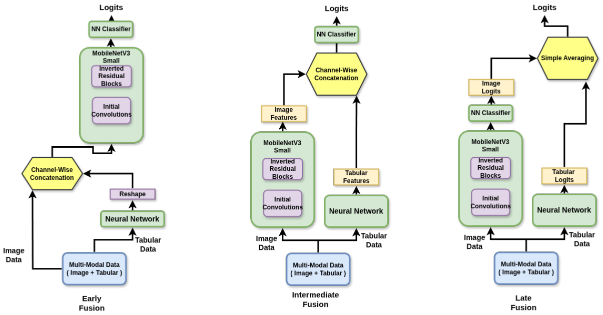

Fig. 2: Fusion Strategies

A. Early Fusion

In early fusion strategy, the image and tabular inputs are combined after basic preprocessing. In our implementation, the tabular features are passes through a neural network and reshaped to match the spatial dimensions of the image tensor, allowing the combined tensor to pass through a shared MobileNetV3 backbone. This approach enables the model to learn joint low-level representations across both modalities. While early fusion offers a uniform input structure, it can be limited due to differences in data modalities and scales. Equation 1 shows the early fusion operation where x_{img} is the image tensor, $\tilde{x}_{\text{tab}}$ if the tabular feature tensor, x_{fused} represents the combined tensor after concatenation, f_{backbone} is the model for extracting features and $\hat{y}$ is the output logits.

$$x_{\text{fused}} = \text{Concat}(x_{\text{img}}, \tilde{x}_{\text{tab}}), \quad \hat{y} = f_{\text{backbone}}(x_{\text{fused}}) \tag{1}$$

B. Intermediate Fusion

Intermediate fusion occurs at the feature level, where each modality is first processed independently to extract high-level representations. In our model, features from the image are extracted using a MobileNetV3-Small backbone, while the tabular data is processed through a fully connected neural network. The resulting feature vectors are concatenated along the channel dimension and fed into a shared classifier. This strategy allows the model to learn meaningful representations from each modality before combining them, and our results show that this approach yielded the highest performance across all metrics. Equation 2 shows the intermediate fusion operation where f_{img} is the image feature extractor (MobileNetV3) and f_{tab} if the tabular feature extractor (Neural Network).

$$z = \text{Concat}(f_{\text{img}}(x_{\text{img}}), f_{\text{tab}}(x_{\text{tab}})), \quad \hat{y} = f_{\text{cls}}(z) \tag{2}$$

C. Late Fusion

Late fusion involves training independent sub-networks for image and tabular data, each producing a separate output logits. The final prediction is obtained by averaging or weighting these outputs. In our case, the two logits are simply averaged. This method treats each modality as an expert, allowing independent feature extraction but limiting the opportunity for joint feature learning. It is relatively easier to implement and computationally efficient but may underperform when strong inter-modality interactions are needed. Equation 3 summarizes the late fusion operation.

$$\hat{y}_{\text{img}} = f_{\text{img}}(x_{\text{img}}), \quad \hat{y}_{\text{tab}} = f_{\text{tab}}(x_{\text{tab}}), \quad \hat{y} = \frac{1}{2}(\hat{y}_{\text{img}} + \hat{y}_{\text{tab}}) \tag{3}$$

D. Uni-modal (Image Only)

As a baseline, we trained a uni-modal model using only chest X-ray images with the MobileNetV3-Small architecture. This model excludes tabular data entirely, relying solely on image-based features for classification. It serves as a useful reference, for comparing with the fusion-based approaches, to demonstrate the gain achieved by fusion.

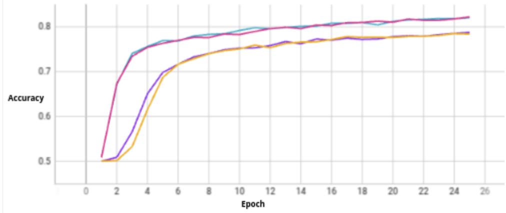

Fig. 3: Accuracy vs. epoch

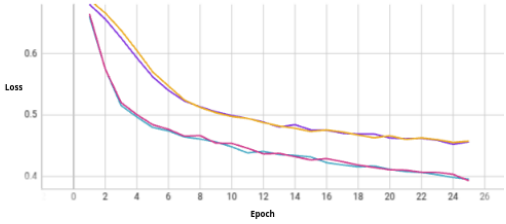

Fig. 4: Loss vs. epoch

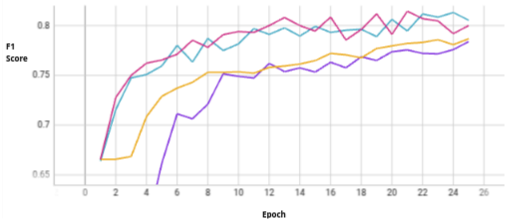

Fig. 5: F1-Score vs. epoch

VI. Result Analysis

We performed an extensive comparison of the proposed fusion strategies across multiple performance and efficiency metrics, including Accuracy, F1-score, Precision, Recall, AUC-ROC, latency, FLOPs, and parameter count, as summarized in Table IV. The strategies compared include: Single Modality (image-only), Early Fusion, Intermediate Fusion, and Late Fusion.

TABLE IV: Comparison of Fusion Strategies Based on Performance, Complexity, and Efficiency Metrics

Fusion Strategy	Accuracy	F1-Score	Precision	Recall	AUC-ROC	Latency	FLOPs	Params
Single Modality	0.76	0.77	0.75	0.78	0.83	0.23ms	1.862G	1.07M
Early Fusion	0.77	0.76	0.80	0.74	0.84	0.24ms	1.863G	1.09M
Intermediate Fusion	0.80	0.80	0.79	0.81	0.87	0.32ms	1.863G	1.09M
Late Fusion	0.79	0.80	0.81	0.80	0.85	0.24ms	1.863G	1.09M

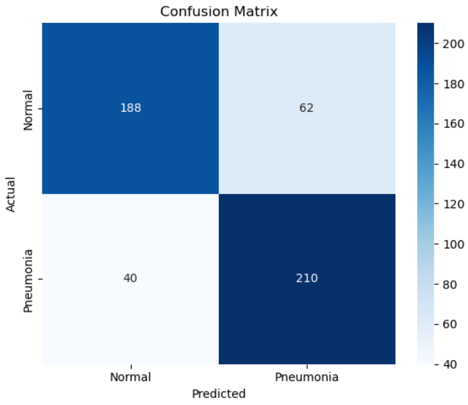

Fig. 6: Confusion Matrix on Test Set of Intermediate Fusion Model

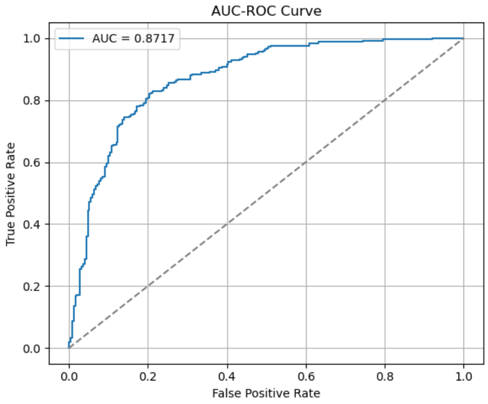

Fig. 7: ROC curve of Intermediate Fusion Model

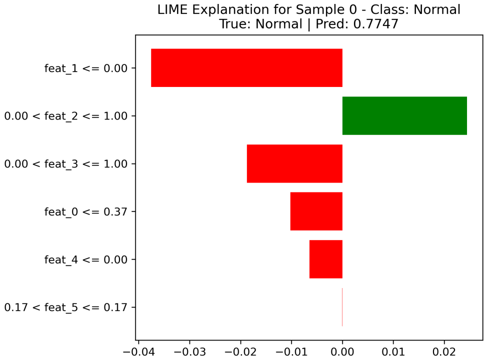

Fig. 8: Lime Explanation for Normal Class

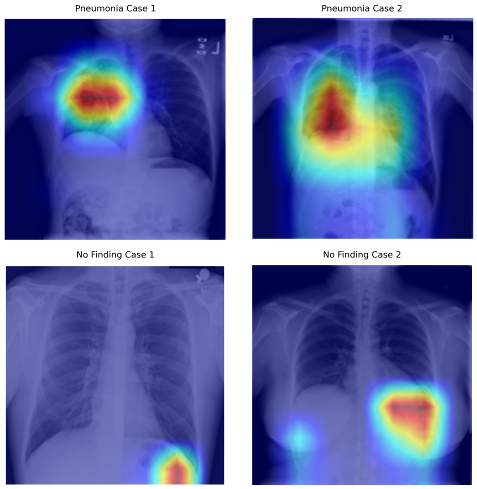

Fig. 9: GradCAM Images on Normal and Pneumonia Class

Single Modality (image-only) serves as the baseline and achieved an accuracy of 76%, with relatively low computational cost (0.23 ms latency, 1.862 GFLOPs, 1.07 M parameters). This indicates that visual features alone carry moderately discriminative power, but does not include the contextual understanding that metadata can provide.

Early Fusion showed a slight improvement in performance (77% accuracy), by incorporating tabular features early in the network pipeline. The gain, though modest, comes without significant computational overhead (0.24 ms latency), confirming that metadata contributes marginally at the early stages of the architecture.

Intermediate Fusion outperformed all other strategies, achieving the highest accuracy of 80%, with an AUC-ROC of 0.8717 (Figure 7), and superior performance in both precision and recall. The confusion matrix (Figure 6) also shows better classification of both Pneumonia and No Finding classes. Although this model has slightly higher latency (0.32 ms), it benefits from combining image features with tabular representations at a semantically meaningful level. This suggests that modality interaction at the feature embedding level allows for better decision boundaries. The loss, accuracy and F1-score curves shown in figure 4, 3 and 5

Late Fusion also showed competitive results (79% accuracy), with performance approaching that of intermediate fusion. Since the two modalities are processed independently and combined at the decision level, this approach preserves computational efficiency (0.24 ms latency) while still capturing complementary signals. However, it may lose cross-modal interactions that intermediate fusion can leverage.

Overall, the results confirm that fusing metadata with image features enhances Pneumonia classification performance. Among all approaches, Intermediate Fusion emerges as the most effective strategy, offering the best trade-off between accuracy and computational cost. Moreover, its compatibility with modality-specific explainability such as GradCAM and LIME further supports its suitability for clinical applications.

VII. XAI

Explainable AI (XAI) explores the interpretability of the multi-modal pneumonia classification model using methods such as LIME and GradCAM to offer per modality insights. To interpret the influence of tabular features on the classification outcome, we employed LIME (Local Interpretable Model-agnostic Explanations). As shown in Figure 8, the most influential feature for the ”Normal” prediction was the patient’s gender. Specifically, the presence of the one-hot encoded ”Female” feature (feat_2 = 1) contributed much positively, while the corresponding ”Male” feature (feat_1 = 0) had a good negative effect. This inverse relationship is expected due to one-hot encoding. The view position (feat_3, representing PA) showed a moderate negative contribution, suggesting the model associated PA views with slightly higher pneumonia likelihood in this case. The patient’s age $(feat\_0 \leq 0.37, age \leq 37)$ also partially decreased the model’s confidence for the Normal class. Pixel spacing (feat_5) had minimal influence in this prediction. Overall, the LIME output provided transparent insight into how each metadata feature contributed to the model’s decision. Figure 9 shows the GradCAM visualizations of the Normal and Pneumonia classes, providing spatial explainability in terms of heatmaps superimposed on the X-ray images. Both for pneumonia-case-1 and pneumonia-case-2, the heatmaps focus on lung regions, which are clinically reasonable. The No Finding Case 1 and Case 2 images on the other hand exhibit very low levels of activations with only on regions outside the lung which highlights model’s capability of correctly indicating lack of pneumonia. These visualizations also demonstrate the model’s ability to identify the appropriate anatomical regions, making the model more trustworthy and interpretable for doctors.

VIII. Conclusion and Future Work

Our study presents a comparative analysis of three fusion strategies—early, intermediate, and late fusion for multi-modal binary classification of pneumonia using a balanced subset of the NIH Chest X-ray dataset. Utilizing both image data and associated tabular metadata, our experiments demonstrated that intermediate fusion and late fusion consistently outperformed other strategies across multiple evaluation metrics, including accuracy, F1-score, and AUC-ROC. The incorporation of GradCAM and LIME further enhanced the explainability of our model, providing interpretable insights into both image and tabular contributions toward the classification outcome. All these factors, prove the superiority of the multi-modal fusion in the context of pneumonia classification.

Despite fruitful results, several areas are still open for exploration. First, future work can extend this analysis to multi-class classification settings to distinguish between multiple thoracic diseases. Second, more advanced fusion mechanisms—such as attention-based fusion or Transformer-based multi-modal networks—can be analyzed to better capture inter-modality interactions. Third, the use of larger-scale datasets or real-world medical data may help validate and confirm the generalizability of the proposed approach. Lastly, enhancing explainability with newer techniques beyond GradCAM and LIME, such as SHAP or counterfactual explanations, also remains a valuable direction for contributions.

References

[1] 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,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3462–3471, 2017.

[2] Xiaoyi Liu and Zhuoyue Wang. Deep learning in medical image classification from mri-based brain tumor images. In *2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS)*, pages 840–844. IEEE, 2024.

[3] Arslan Naseer, Muhammad Muheet Khan, Fahim Arif, Waseem Iqbal, Awais Ahmad, and Ijaz Ahmad. An improved hybrid model for cardiovascular disease detection using machine learning in IoT. *Expert Systems*, 42(1):e13520, 2025. Wiley Online Library.

[4] T. Grace Shalini, G. Susan Shiny, R. Saranya, P. Suresh Babu, R. Kavitha, and Athiraja Atheeswaran. Enhancing lung disease identification with multimodal data fusion and deep learning CNN approach. In *2024 5th International Conference on Smart Electronics and Communication (ICOSEC)*, pages 535–541. IEEE, 2024.

[5] Zeenat Tariq, Sayed Khushal Shah, and Yugyung Lee. Multimodal lung disease classification using deep convolutional neural network. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2530–2537. IEEE, 2020.

[6] S. Kumar, O. Ivanova, A. Melyokhin, and P. Tiwari. Deep-learning-enabled multimodal data fusion for lung disease classification. Informatics in Medicine Unlocked 42: 101367. 2023.

[7] Pratham Kaushik, Eshika Jain, Vinay Kukreja, Shanmugasundaram Hariharan, Murugaperumal Krishnamoorthy, Vandana Ahuja, Abhishek Bhattacherjee, Rajesh Kumar Kaushal, and Shih-Yu Chen. Modelling radiological features fusion and explainable AI in pneumonia detection: A graph-based deep learning and transformer approach. *Results in Engineering*, page 105225, 2025. Elsevier.

[8] Chihcheng Hsieh, Isabel Blanco Nobre, Sandra Costa Sousa, Chun Ouyang, Margot Brereton, Jacinto C Nascimento, Joaquim Jorge, and Catarina Moreira. MDF-Net for abnormality detection by fusing X-rays with clinical data. *Scientific Reports*, 13(1):15873, 2023. Nature Publishing Group UK London.

[9] Jovito Colin and Nico Surantha. Interpretable deep learning for pneumonia detection using chest x-ray images. *Information*, 16(1):53, 2025. MDPI.

[10] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI’19/IAAI’19/EAAI’19)*, pages 590–597, 2019. AAAI Press.

[11] Erdi Çallı, Ecem Sogancioglu, Bram van Ginneken, Kicky G. van Leeuwen, Keelin Murphy. Deep learning for chest X-ray analysis: A survey. *Medical Image Analysis*, 72:102125, 2021. Elsevier.

[12] Prateek Singh and Sudhakar Singh. ChestX-Transcribe: A multimodal transformer for automated radiology report generation from chest X-rays. *Sec. Health Technology Implementation*, 7, 2025. Frontiers in Digital Health