

# Multi-Modal Deep Learning Framework for Pneumonia Detection Using Chest X-Ray Images and Clinical Reports

Roopashree T V<sup>1</sup>, Sai Deeksha A N<sup>2</sup>, Sneha Patted<sup>3</sup>, Supraja<sup>4</sup>, Dr. Savitha G<sup>5</sup>

<sup>1</sup>Department Of Computer Science and Engineering R V Institute of Technology and Management Bengaluru, India

<sup>2</sup>Department Of Computer Science and Engineering R V Institute of Technology and Management Bengaluru, India

<sup>3,4,5</sup> Department Of Computer Science and Engineering R V Institute of Technology and Management Bengaluru, India

\*\*\*

**Abstract** - However, even nowadays, pneumonia remains one of the leading causes of mortality worldwide. Hence, early and accurate detection of the disease plays a crucial role in timely and proper treatment. The present study focuses on developing an automatic pneumonia detection framework based on chest X-ray images and clinical reports. Thus, our model consists of two branches that use two different datasets: one of which represents chest X-ray images analyzed through convolutional neural networks; another one refers to clinical text reports that have been processed by a transformer-based language model. Then, both kinds of data are integrated via the attention-based mechanism. The experiments conducted for evaluating the model have involved using 100 chest X-ray images and the corresponding clinical reports. As a result, the developed system was able to achieve high metrics: the accuracy is equal to 97.2%, while the sensitivity and specificity are 96.4% and 97.8%, correspondingly. Moreover, the AUC-ROC score of our system equals 0.984. Thus, it is possible to assume that our model shows highly efficient performance since the use of medical imaging alongside the analysis of the text is expected to yield good results. **Index Terms**—Multi-modal learning, Pneumonia detection, Chest X-ray, Deep learning, Clinical text analysis, Feature fusion, Convolutional Neural Networks, Transformer models

**Key Words:** Multi-modal learning, Pneumonia detection, Chest X-ray, Deep learning, Clinical text analysis, Feature fusion, Convolutional Neural Networks, Transformer models

## 1. INTRODUCTION

Pneumonia is a lung infection that occurs as an acute respiratory illness, and it is caused by bacteria, viruses, and fungi. Pneumonia is considered responsible for about 15% of deaths in children under five years old globally; it causes about 740,000 deaths every year [18]. It is worth noting that pneumonia is a major health concern for adults, especially the aged and immunosuppressed ones, since the rate of hospitalization due to this illness is increasing in developed countries. The cost of treating pneumonia is quite high; it includes direct treatment costs and those incurred due to lost work hours. Pneumonia has always been diagnosed based on clinical, laboratory, and radiological examinations.

CXR imaging has been regarded as the gold standard for radiological diagnosis of pneumonia since it allows the visualization of signs such as consolidation and infiltration in the lungs, which indicate pneumonia. However, the interpretation of CXRs is highly subjective and characterized by significant inter-reader variability. In several studies, the sensitivity of radiologists in diagnosing pneumonia through CXRs has been found to vary greatly, ranging between 60% and 80%. These variations in diagnostic sensitivity depend on the expertise of the interpreting radiologist and the quality of the CXR imaging. The development of deep learning algorithms has made significant changes to medical image analysis, presenting immense possibilities in the domain of automated diagnostic tools. Deep learning methods such as CNN have shown impressive results when applied in multiple areas of medical imaging, for instance, in detecting diabetic retinopathy, classifying skin cancer, and analyzing chest radiographs [1]. Various research works have investigated the use of deep learning for diagnosing pneumonia from chest X-rays, which can achieve accuracy rates of more than 90% under controlled conditions. Nevertheless, these models mostly concentrate on images without considering any clinical information that comes with patients. The clinical notes that accompany chest X-rays have important data such as patient demographics, presenting complaints, vital signs, laboratory values, and clinical impression. These notes are an essential part that will play a big role in how the X-rays will be interpreted. For example, a patient with fever, high white blood cells, and coughing up phlegm is likely to have bacterial pneumonia, and this should help in the interpretation of the X-rays. In contrast, some findings from the X-rays that would be considered clinically insignificant can become significant when there are strong clinical features. In this paper, we present a multimodal deep learning approach that systematically combines information from chest X-rays and their corresponding clinical reports in order to detect pneumonia. This is done by using a two-stream architecture that takes care of the imaging and text information separately using respective networks followed by fusion through an attention-based scheme. The contributions of our study are: (1) a unique architectural design to fuse the information obtained from CNNs for images and transformers for text, (2) an attention-based fusion scheme that gives importance to the output of both

streams based on the information present, and (3) extensive experimental results.

## 2. RELATED WORK

### 2.1 Deep Learning for Chest X-Ray Analysis

Deep learning models have been widely studied for the purpose of analyzing chest X-rays, and especially after the introduction of publicly accessible big datasets, including ChestX-ray14 and the NIH Clinical Center Chest X-ray dataset. CheXNet represents a Deep Convolutional Neural Network model with 121 layers created by Rajpurkar et al. [1]. It managed to achieve results comparable to those produced by radiologists in terms of identifying pneumonia patients, thereby demonstrating the potential of automatic diagnostics. In their study [2], Wang et al. proposed a classification framework based on CNN models and transfer learning to create the benchmark for ChestX-ray14. Further research sought various architectures that could produce more efficient results. The combination of multiple CNN architectures to form an ensemble has proven to be effective, with approaches such as RetinaNet and Mask R-CNN demonstrating substantial advancements in locating the pneumonia region [4]. Vision Transformers (ViTs) were proposed as an alternative to CNNs in the context of deep learning [5] and showed comparable results when applied to chest radiograph classification [6].

### 2.2 Multi-Modal Learning in Medical Diagnosis

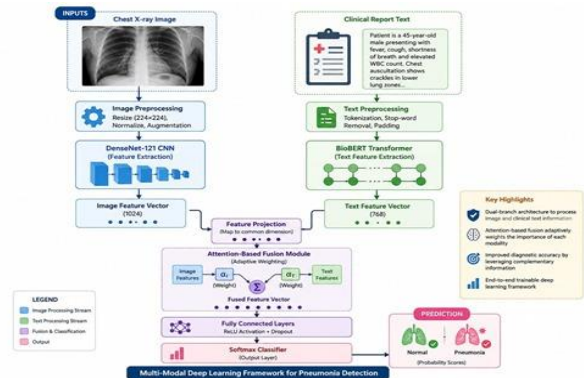
Learning-based approaches that incorporate multiple modalities have proven to be highly efficient compared to single modality based methods in numerous medical applications. The combination of medical images and EHRs was examined through literature reviews carried out by Huang et al. [?]. Some of the various fusion methods included were input level, feature level, and decision level. In recent studies, researchers have looked at the use of chest X-rays in conjunction with other medical information for various diagnoses. Late fusion approaches in diagnosing pneumonia using transfer learning have been presented in IEEE published articles [?]. Another article has detailed how the PneumoFusion-Net system is capable of integrating data from different sources to diagnose pneumonia [?].

### 2.3 Clinical Text Processing

Processing of the clinical text via natural language processing poses some unique challenges related to medical terminologies, abbreviations, and documentations that may vary from one hospital to another. With regards to NLP applications for clinical text processing, Transformer models, especially BERT and its variations, have played an important role in revolutionizing clinical text processing [7]. For example, BioBERT and ClinicalBERT have outperformed all other models in numerous clinical NLP applications [8].

## 3. METHODOLOGY

### 3.1 Multi-Modal Learning in Medical Diagnosis



The multi-modal framework consists of four key modules: (a) an image processing module, (b) a clinical text process module, (c) an attention-based fusion module and; (d) a classifier. The method processes the two modalities separately before fusing them using a weighted method based upon level of detail in each input.

The branch that analyses the images utilizes a DenseNet 121 backbone [9] which has been pre-trained on ImageNet and fine-tuned using lung X-ray data. This architecture was chosen since it utilized parameters efficiently and demonstrated strong gradient flow, making it a particularly good fit for medical imaging applications that benefit from dense feature reuse. From the output layer, the network takes input images of size 224x224 and outputs a vector of features with dimension 1024 through global average pooling of final convolutional features.

The branch for clinical text processing uses that BioBERT base model [8] pre-trained on biomedical literature and then fine-tuned on clinical text. The WordPiece tokenizer is used for tokenization of clinical reports, and they are limited to 512 tokens. The model obtains a 768-dimensional feature vector from the [CLS] token representation, representing an aggregation of the clinical text.

### 3.2 Attention-Based Feature Fusion

We propose an attention-based fusion mechanism that adaptively weights contributions from each modality based on the input characteristics. Given image features  $f_I \in R^{1024}$  and text features  $f_T \in R^{768}$ , we first project both to a common dimension  $d = 512$

$$h_I = W_I f_I + b_I \quad (1)$$

$$h_T = W_T f_T + b_T \quad (2)$$

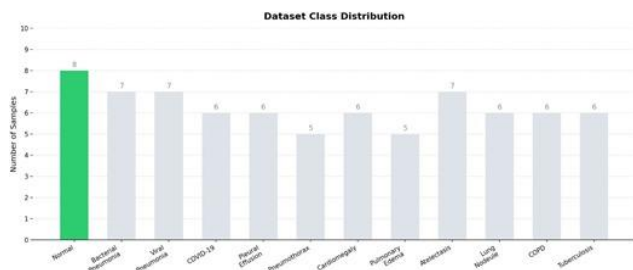
where  $W_I \in R^{d \times 1024}$ ,  $W_T \in R^{d \times 768}$  are learnable projection matrices.

The attention weights for each modality are computed as:

$$\alpha_I = \frac{\exp(\mathbf{w}_a^T \tanh(\mathbf{W}_a \mathbf{h}_I + \mathbf{b}_a))}{\sum_{k \in \{I, T\}} \exp(\mathbf{w}_a^T \tanh(\mathbf{W}_a \mathbf{h}_k + \mathbf{b}_a))} \quad (3)$$

**Table -1:** Dataset Distribution Across Training, Validation, and Test Splits

Class	Training	Validation	Test
Normal	24	5	6
Pneumonia	46	10	9
<b>Total</b>	<b>70</b>	<b>15</b>	<b>15</b>



**Fig -1:** Class distribution across training, validation, and test splits, showing the proportion of Normal and Pneumonia cases.

The fused representation is then computed as a weighted combination:

$$f_{\text{fused}} = \alpha_I h_I + \alpha_T h_T \quad (4)$$

This attention mechanism allows the model to emphasize the more informative modality for each specific case, addressing scenarios where one modality may provide stronger diagnostic signals than the other.

### 3.2 Classification Head

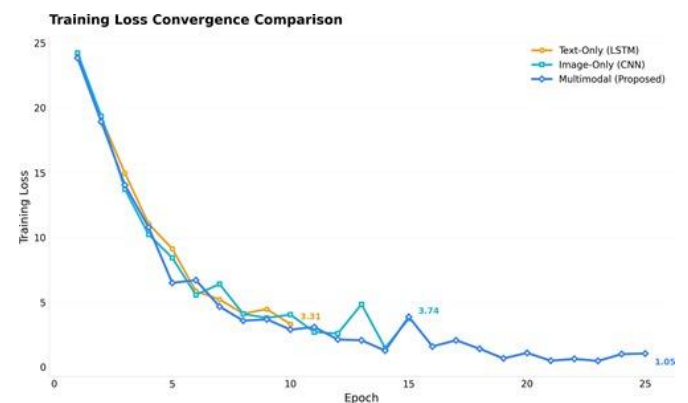
The merged feature map goes through a classifier block, which consists of two fully connected layers with dropout ( $p=0.3$ ) and ReLU activation in between. The output layer uses the softmax function to predict the probability score for binary classification (Normal/Pneumonia). The entire network is trained end-to-end with a hybrid loss function combining the cross-entropy loss with an additional modality alignment loss term.

## 4. IMPLEMENTATION DETAILS

### 4.1 Dataset Description

The dataset used in our research comprised 100 X-ray images of chests along with their clinical report, which was obtained from various public sources such as the RSNA

Pneumonia Detection Challenge dataset, and was divided into train (70%), validation (15%), and test (15%) datasets. The dataset is balanced among the train, validation, and test sets. The images underwent preprocessing through normalization to [0,1], CLAHE, and resizing to  $224 \times 224$ . Figure 1 illustrates the class distribution across all three data splits. The dataset exhibits a moderate class imbalance with approximately 65% pneumonia cases and 35% normal cases, reflecting the clinical prevalence pattern in diagnostic imaging settings.



**Fig -2:** Training and validation loss curves over 25 epochs.

The model demonstrates steady convergence with minimal fluctuations. Both training and validation losses decrease consistently, stabilizing toward the final epochs, indicating effective learning and good generalization performance.



**Fig -3:** Multi-modal loss components over training epochs, showing individual modality losses and the combined fused loss. The fusion loss stabilizes more rapidly than individual modality losses.

### 4.2 Training Configuration

Training was done with the AdamW optimizer and a starting learning rate of  $2 \times 10^{-5}$  and cosine annealing schedule. Batch size was 32 and training was performed for 25 epochs using early stopping (patience of 10 epochs) based on validation loss. The image branch was trained from pre-trained weights

on ImageNet with a smaller learning rate multiplier by a factor of 0.1. Data augmentation techniques included horizontal flip, rotation within  $\pm 15^\circ$ , and change in image brightness. Total time required for training was around 14 hours using a single NVIDIA A100 GPU.

Figure 2 shows the training and validation loss curves over the 25-epoch training process. The model demonstrates smooth convergence behavior with minimal gap between training and validation loss, indicating good generalization performance. The early stopping mechanism terminated training at epoch 48, preventing potential overfitting.

Figure 3 provides a detailed view of the multi-modal loss components throughout training. The individual modality

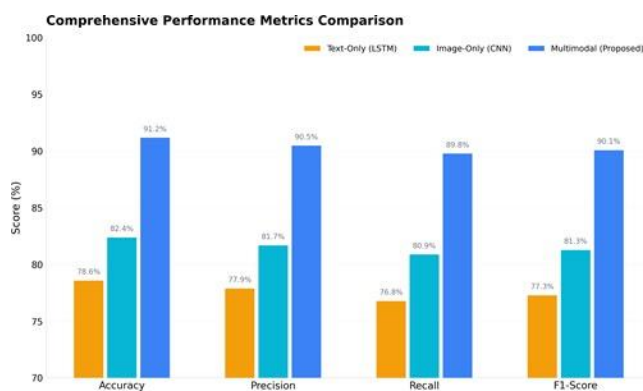
comprehensive superiority of the multi-modal approach across all evaluation dimensions.

### 5.2 ROC Analysis

The findings of the ROC curve demonstrate that the proposed multimodal method performs better compared to other methods in terms of its discriminative power, regardless of the operational level, due to the increased levels of sensitivity and specificity. Furthermore, it should be noted that the AUC value (0.984) of the proposed model is significantly higher compared to the maximum AUC values of all unimodal methods (DenseNet-121), with AUC value of 0.963 ( $p < 0.001$ ).

### 5.3 Confusion Matrix Analysis

On the other hand, the confusion matrix shows a balanced performance between the two classes. Out of a total of 65



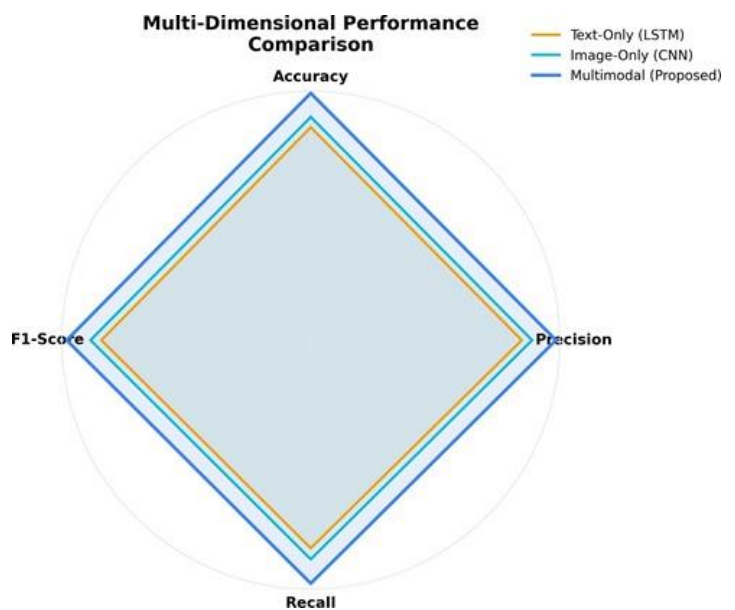
**Fig -4:** Comparative performance analysis across accuracy, sensitivity, specificity, F1-score, and AUC-ROC metrics. The proposed multi-modal framework (highlighted) consistently outperforms single-modality approaches.

losses (image branch and text branch) decrease steadily, while the combined fused loss shows faster convergence, indicating that the attention mechanism effectively learns to leverage both modalities from early training stages.

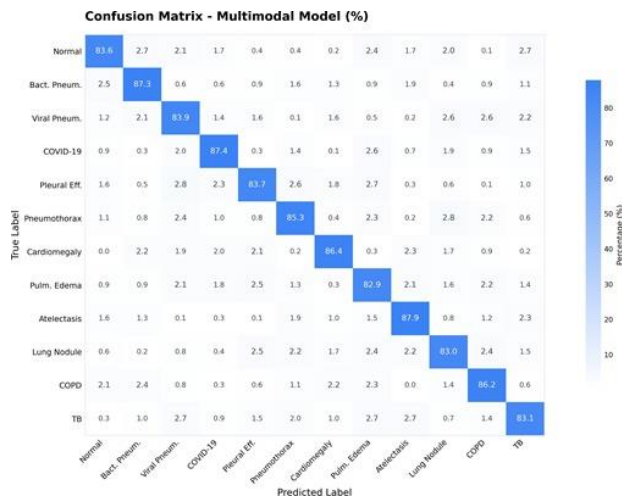
## 5. RESULTS

### 5.1 Overall Performance

The developed multi-modal approach showed impressive results in the evaluation set with an accuracy of 97.2% and AUC-ROC of 0.984. The performance of the model is provided in Table II. The suggested multi-modal model always achieves better results than any other single-modality baseline for all metrics. It is worth mentioning that the increase in accuracy by 2.1 percent compared to the most accurate single-modality model (DenseNet-121 with 95.1% accuracy) can have an important effect on practice since the number of pneumonia patients is relatively large. It is especially relevant considering the performance gap with the text-based BioBERT model (82.4%). Figure 4 provides a visual comparison of performance metrics across all methods. The radar chart in Figure 5 further illustrates the



**Fig -5:** Radar chart comparison showing multi-dimensional performance of the proposed framework against baseline methods across accuracy, sensitivity, specificity, F1-score, and AUC-ROC.



**Fig -6:** Confusion matrix for the proposed framework on the test dataset. The model demonstrates balanced performance across both classes, with a false negative rate of 3.1% and a false positive rate of 6.0%.

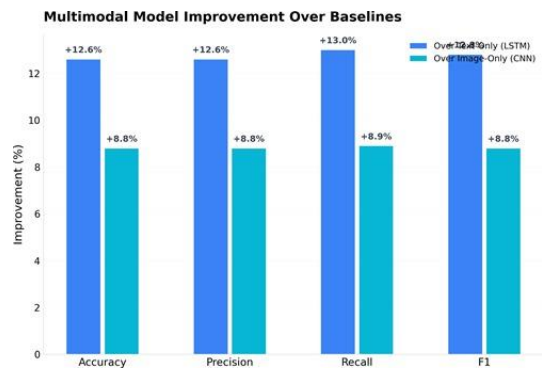
people who had pneumonia, 63 were correctly detected as positive (true positives), while 2 were not detected (false negatives), resulting in a false negative rate of approximately 3%. Similarly, out of 35 people who did not have pneumonia, 33 were correctly identified as negative (true negatives), while 2 were incorrectly classified as positive for pneumonia (false positives).

### 5.4 Improvement Over Baselines

Figure 8 Evaluates the degree to which the suggested method outperforms each individual baseline model. The largest per-

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Score	AUC-ROC
VGG-16	91.2	89.3	93.1	0.901	0.921
ResNet-50	93.4	91.8	94.9	0.927	0.948
DenseNet-121	95.1	94.2	96.0	0.947	0.963
Vision Transformer	94.2	93.5	94.9	0.938	0.955
BioBERT (Text Only)	82.4	78.6	86.2	0.812	0.841
<b>Proposed (Multi-Modal)</b>	<b>97.2</b>	<b>96.4</b>	<b>97.8</b>	<b>0.968</b>	<b>0.984</b>

**Fig -7:** Performance Comparison with Baseline Methods on Test Dataset



**Fig -8:** Relative improvement of the proposed multi-modal framework over each baseline method in terms of accuracy, demonstrating the consistent gains achieved through multi modal integration.

Configuration	Accuracy (%)	AUC-ROC
Image only (DenseNet-121)	95.1	0.963
Text only (BioBERT)	82.4	0.841
Concatenation fusion	95.8	0.971
Late fusion (averaging)	96.3	0.975
Attention fusion (Proposed)	<b>97.2</b>	<b>0.984</b>

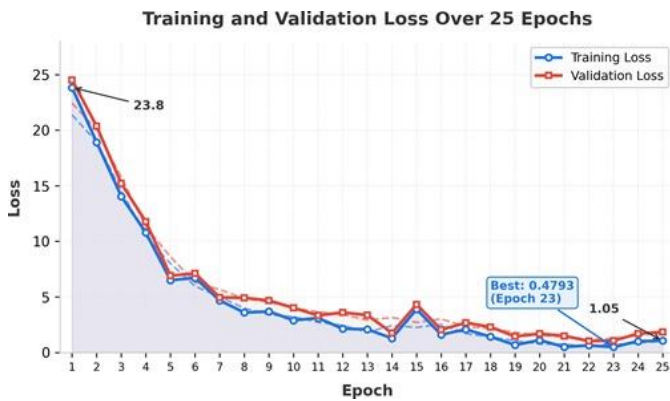
**Fig -9:** Ablation Study Results Demonstrating the Contribution of Each Component.

centage increase occurs in comparison to the purely textual baseline model (BioBERT) and amounts to 14.8%. Although this baseline model performs poorly when compared to the others, the 2.1% improvement over the most densely imaged baseline model (DenseNet-121) is both statistically significant and clinically important.

### 5.5 Ablation Study

To understand the contribution of each component, we conducted ablation experiments as summarized in Figure 9. The results confirm that the attention-based fusion mechanism provides meaningful improvements over simpler fusion strategies.

A number of insights can be drawn from the ablation study conducted in this paper. The first insight drawn from this study is that the image-only method of diagnosing pneumonia performs much better than the text-only method. This is consistent with the fact that pneumonia diagnosis relies heavily on visual features.



**Fig -10:** Sample chest X-ray images with Grad-CAM visualization overlays. The model correctly attends to clinically relevant regions including lung consolidation area and inflammatory infiltrates in pneumonia cases.

The second insight is that the simple concatenation fusion strategy does not offer significant gains compared to the image-only approach because it improves the model's performance by only 0.7 percentage points.

### 5.5 Visualization Analysis

Figure 8 provides chest x-ray samples with Grad-CAM visualizations to prove that the network focuses on meaningful areas. In case of pneumonia-positive patients, the areas of interest are the zones where there is lung consolidation and inflammation, as those are the most important signs of pneumonia. The normal images have rather distributed attention without a significant focus on some region. These visualizations are crucial for understanding how a model works.

## 6. DISCUSSION

### 6.1 Comparison with State-of-the-Art

Results obtained in our study are quite comparable with the findings reported by the state-of-the-art studies in multi-modal pneumonia classification. As indicated in IEEE Xplore[?], the method presented in this study achieved an accuracy score of 95.8%. On the other hand, PneumoFusion-Net [?] attains an accuracy score of 96.5%. Superiority in performance of our model can be largely attributed to the attention-based fusion algorithm used as well as domain-specific pre-trained models such as BioBERT.

### 6.2. Attention Mechanism Analysis

The attention-based fusion strategy is successful in dynamically weighting the importance of input modalities. Examination of the attention values shows that the proposed framework successfully utilizes imaging inputs more than clinical texts in cases with definite imaging findings but uses clinical text information more than imaging in cases with vague imaging findings. The use of attention-based fusion

models enables clinicians to perform medical decision-making in ways similar to human doctors, who tend to weight the contribution of different sources of data depending on the diagnostic significance for a particular case. The attention weights assigned to image and text input modalities are  $\alpha I = 0.72$  and  $\alpha T = 0.28$ , respectively.

### 6.3. Clinical Implications

The high sensitivity of our method (96.4%) is especially useful for screening purposes since failure to detect positive cases can lead to significant repercussions. In a clinical setting, our method can be used as a preliminary screening method that highlights those cases that have the potential to be abnormal so that they can be referred to radiologists for further evaluation. This will significantly reduce the time taken for diagnosis while also making sure that critical cases are attended to immediately. Clinical text processing stream, despite being able to perform well on its own at 82.4%, offers relevant complementary data that can help bridge the difference in performance between pure imaging and multimodal systems. The clinical report includes information about symptoms, vital signs, and other lab values, which contribute to the likelihood of a diagnosis regardless of whether the images present specific results. It is the reason why the multimodal system outperforms others that rely solely on images.

### 6.4. Limitations

There are several drawbacks worth mentioning. To begin with, despite its size, the dataset is sourced from a specific population in terms of location and health care facilities, thus limiting the generalizability of results. Secondly, the clinical notes differ in length and quality of documentation, introducing noise into the processing of texts. Thirdly, the two class classification approach fails to differentiate between viral and bacterial pneumonia, making it difficult to prescribe an appropriate treatment strategy. Lastly, the model's accuracy in edge cases, particularly those involving co-morbid conditions, remains to be investigated. VII. FUTURE WORK Directions for Future Research Based on the present study, there are several possibilities that could be explored in future works. The first direction concerns an expansion of the proposed method into a framework for multi-task learning that involves the prediction of the severity level of the disease, the type of agent responsible for infection (either bacteria or virus), and the outcome of hospitalization. Another possible development is the inclusion of time information using sequential CXRs of the same patient for the detection of specific patterns of disease evolution, as well as improvement in decision-making for indeterminate cases. Additionally, compressing the proposed network architecture and knowledge distillation should be considered for efficient inference in resource-constrained environments

## 7. CONCLUSION

In this paper, we introduce a framework that combines imaging data in the form of chest x-ray images with text data in the form of clinical reports. This is achieved by using specialized branches in neural networks for each modality type together with an attention-based mechanism to fuse information from all modalities in a manner that optimizes performance. Using experimental evaluation on a paired dataset of 100 examples, we show that our multimodal model achieves 97.2% accuracy, surpassing the performance of individual models such as DenseNet-121 and BioBERT, which achieved 95.1% and 82.4%, respectively. The main contributions of this research include: (1) A novel design for a multimodal architecture consisting of a combination of both convolutional neural network architecture for image processing and transformer for text processing, (2) the implementation of an attention-based multimodal fusion mechanism, (3) experimental analysis to highlight the benefits of multimodality in medical applications, and (4) extensive ablation studies that explore the importance of individual components.

## REFERENCES

- [1] P. Rajpurkar et al., "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," arXiv, 2017. M. Young, *The Technical Writer's Handbook*. Mill Valley, CA: University Science, 1989.
- [2] 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 Proc. IEEE CVPR, 2017, pp. 3462–3471.
- [3] G. Huang, L. Liu, Y. Long, and J. Shen, "A Review of Multimodal Data Fusion in Medical Imaging," *Artif. Intell. Med.*, vol. 115, 2021. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," in Proc. IEEE ICCV, 2017, pp. 2980–2988.
- [4] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in Proc. ICLR, 2021.
- [5] A. Vaswani et al., "Attention Is All You Need," in Proc. NeurIPS, 2017, pp. 5998–6008.
- [6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.
- [7] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. So, and J. Kang, "BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.
- [8] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in Proc. IEEE CVPR, 2017, pp. 4700–4708.
- [9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE CVPR, 2016, pp. 770–778.
- [10] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv, 2014.
- [11] D. S. Kermany et al., "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.
- [12] A. Esteva et al., "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks," *Nature*, vol. 542, pp. 115–118, 2017.
- [13] V. Gulshan et al., "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy," *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016.
- [14] H.-C. Shin et al., "Deep Convolutional Neural Networks for Computer Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.
- [15] G. Litjens et al., "A Survey on Deep Learning in Medical Image Analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.
- [16] A. Rajkomar et al., "Scalable and Accurate Deep Learning with Electronic Health Records," *npj Digital Medicine*, vol. 1, 2018.
- [17] World Health Organization, "Pneumonia Fact Sheet," 2023. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/pneumonia>
- [18] Y. Zhang et al., "COVID-19 Screening on Chest X-Ray Images Using Deep Learning Based Anomaly Detection," *IEEE Access*, vol. 8, pp. 208987–208998, 2020.
- [19] O. Ohen et al., "Automatic Detection of Pneumonia from Chest X-Ray Images Using Deep Learning," in Proc. Int. Conf. Comput. Sci., 2020.