Exploring Multimodal Learning in Classifiers for Healthcare Applications

Stacy Chao STACY.CHAO@NYU.EDU
Courant Institute of Mathematical Sciences
New York University
New York, NY, USA

Aurelia Li AURELIA.LI@NYU.EDU
Center for Human Genetics & Genomics
New York University
New York, NY, USA

Abstract

The efficient integration of multi-modal data is a cornerstone of robust deep learning applications, particularly in healthcare, where diverse patient information can significantly impact diagnostic accuracy. The primary objective of this project is to develop a multi-modal model that can integrate MIMIC IV CXR imaging data with corresponding radiology notes to distinguish between 'pneumonia' and 'not pneumonia' cases. These non-pneumonia cases are the absence of any abnormal pulmonary diseases classified by the ChexPert model. A critical aspect of this project is confronting the significant class imbalance present in medical datasets, which often feature a disproportionate number of non-disease instances(Welvaars et al., 2023). Through extensive experiments on the medical dataset, we explore the behaviors and characteristics of the multi-modal models and observe relative improvements to the robustness and reliability for a more balanced learning process. We particularly focused on reducing the effects of class imbalance through sub-sampling and implementing different loss functions. Our results demonstrate that multi-modal models, which integrate both text and image data, were able to perform competitively with their single-modality counterparts across a variety of metrics in a context where class imbalance was less significant. These findings hold value for those considering the deployment of machine learning models in clinical settings, where accurate interpretation of multi-modal data is ideal for patient outcomes.

1 Introduction

From past literature, it is evident that single modality models often under-utilize patient data, which can impede the comprehensive understanding of complex health conditions (Sousa et al., 2023). The integration of diverse data sources such as lab results, diagnostic tests, medical imaging, and various screening technologies is essential to harness the full potential of health informatics. However, multi-modal models, which are designed to leverage these varied data sources, sometimes perform less optimally than their single-modality counterparts, particularly when integrating text and image data; these models often neglect one modality, leading to severely unbalanced modality-wise utilization and sub-optimal outcomes (Liu et al., 2018; Wang et al., 2020). Moreover, some models labeled as 'multi-modal'

have merely used two images from different perspectives, which does not truly capture the complexity of different data types (Wu et al., 2022).

In this paper, we address these shortcomings by developing a multi-modal model that integrates MIMIC IV CXR imaging data with corresponding radiology notes to classify cases of pneumonia versus healthy patients (Johnson et al., 2020). We acknowledge the prevalence of class imbalance in medical classification tasks and attempt to garner improvements in comparison to the uni-modal models by tackling the challenge of class imbalance. We developed a total of 12 unique models, each employing different subsampling ratios and loss functions specifically designed to enhance model performance in imbalanced datasets. This methodological diversity allows us to assess which strategies are most effective in reducing bias and improving accuracy across various metrics, such as precision, recall, and AUC. This approach not only aims to identify and analyze the imbalance in modality utilization but also ensures that performance metrics reflect a robust predictive capability rather than an overwhelming bias toward the majority class. By examining the characteristics and behaviors of these integrated models and considering strategies to mitigate significant class imbalance in medical datasets, we hypothesize that our multi-modal model will leverage a broader dataset and show potential for more nuanced learning from integrating multiple modalities. This, in turn, could make deep learning classification models more representative of the complex reality faced by physicians, thereby improving diagnostic accuracy and potentially influencing clinical practices.

2 Methods and Materials

2.1 Data

This project utilizes the MIMIC-IV (Medical Information Mart for Intensive Care) dataset, specifically the MIMIC-CXR JPG v2.0.0, which contains 227,835 imaging studies for 64,588 patients from 2011 to 2016(Johnson et al., 2020). This dataset comprises a total of 377,110 images, including both frontal and lateral views. Each chest radiograph is stored in JPEG format and has been annotated with structured free-text labels using the CheXpert labeler.

2.1.1 Processing

For our analysis, we focused exclusively on the Posteroanterior (PA) chest radiographs that were taken in an erect position using non-portable X-ray machines. We selected images where the labels for pneumonia ("Positive") and no findings ("Negative") were marked with a certainty of 1.0, indicating that the condition was definitively mentioned in the associated radiological report and confirmed in the corresponding images. Additionally, we excluded any duplicated patient records to ensure the uniqueness of our dataset.

After processing, our dataset comprised over 2,323 cases of pneumonia and 21790 negative cases. For computational efficiency and standardization, all images selected were of the original size of 2544x3056 which and were resized to a uniform resolution of 256 x 256 pixels for training. This resolution choice was informed by studies showing that resizing to 256 x 256 pixels does not significantly impact the performance of model classifiers in tasks

like Pneumonia classification within the CXR-JPG dataset.(Haque et al., 2021). We used 2 training datasets in our experiments. The full dataset described above which had an approx 90% to 10% percent negative to positive sample ratio and a sub-sampled dataset with a 75% to 25% ratio. The sub-sampled dataset had a total of 2,323 positive samples and 6,969 negative samples.

2.2 Model Architecture

2.2.1 Multi-modal Model

Our multimodal model leverages two pre-trained encoders: ResNet50 for images and BioBERT for text(He et al., 2015; Lee et al., 2019). The ResNet50 encoder, modified for grayscale input, processes X-ray images while the BioBERT encoder analyzes radiology reports. Key modifications include adapting ResNet50's initial convolutional layer to accept grayscale images and freezing its initial layers (conv1, layer1, layer2) to stabilize learning given the domain-specific nature of chest X-rays. Outputs from both encoders are transformed into 256-feature vectors and merged via a fully connected layer that predicts the class (pneumonia or no pneumonia).

2.2.2 Uni-modal Models

The uni-modal models retain the architecture from the multi-modal setup but focus on single modality input—either images or text. The image model uses the adapted ResNet50, and the text model employs BioBERT. Each model processes its respective modality and produces a prediction independently, utilizing a final fully connected layer tuned to the specific features of that modality.

2.3 Experiments

The experiments were conducted using two configurations of the dataset as previously detailed. Each model was trained with a batch size of 64, a learning rate of $1e-4$, and a linear learning rate scheduler with two steps. The training process incorporated early stopping with a patience of 3 steps and model checkpoints, both of which were based on the validation loss metric, spanning up to 15 epochs. These models were trained under different conditions, focusing on two main aspects:

1. Loss Function Evaluation: Each model was trained using either Focal Loss or Binary Cross-Entropy (BCE) Loss to investigate which loss function is more effective at handling class imbalance in our data. For the full dataset alpha was 0.85 and gamma was 1.5, while for the sub-sampled dataset alpha was 0.95 and gamma was 2.0.

2. Dataset Performance Comparison: The models were also trained on both the full dataset and a sub-sampled version to understand performance variations under different class distributions.

To assess the effectiveness of our models, we employed a comprehensive set of evaluation metrics. We tracked accuracy across training, validation, and testing phases to monitor overall performance and also tracked Precision, Recall, F1 Score, plotted AUROC and Confusion Matrices during testing phases. Additionally, the multi-modal models were subjected

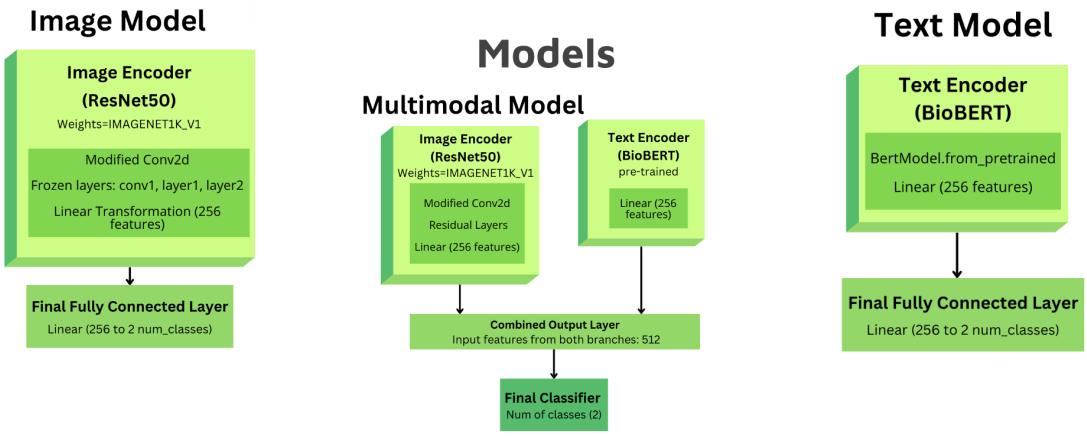

Figure 1: Architecture of the proposed models. On the left, the Image Model utilizes a grayscale-adapted ResNet50 encoder and outputs through a fully connected layer. On the right, the Text Model employs BioBERT to process radiology reports. The center diagram represents the Multi-modal Model, which combines features from both the Image and Text Models through late fusion to predict pneumonia presence.

to Bimodal Sensitivity Tests during the testing phase. In these tests, one of the modalities (text or image) in the testing dataset was intentionally shuffled to analyze the contribution of each modality to the overall performance and dependency of the model, providing insights into how each modality influences the prediction outcomes. In total, 12 unique models were developed to encompass all combinations of the described characteristics, facilitating a comprehensive understanding of how various factors affect each model type's performance in diagnosing pneumonia.

3 Results

Our study commenced with training the Multimodal, Text, and Image Models using the full dataset under Binary Cross-Entropy (BCE) Loss. Looking at Figure 2, it can be seen that the Text Model displayed superior performance across all metrics other than Recall, suggesting a significant advantage when processing textual data alone in this context. Further investigations through the Bimodal Sensitivity Test revealed a marked dependency of the multimodal model on text data. As indicated in Figure 2, shuffling text inputs resulted in dramatic performance drops, unlike shuffling image inputs, which showed less of a performance impact. This suggests that text data played a more critical role in the model's decision-making processes. However, we can still see when testing the sensitivity of the multi-modal model to each modality the model depends on both modalities despite the Multi-Modal model performing less optimally than its text only counterpart.

Model Type	Test Accuracy	Precision	Recall	F1 Score	AUROC
Multimodal Full BCE loss	0.9934	0.9732	0.9934	0.9594	0.9966
Image Full BCE loss	0.9123	0.2212	0.0517	0.0793	0.7735
Text Full BCE loss	0.9948	0.9869	0.9558	0.9688	0.9995

Bimodal Sensitivity Tests	Test Accuracy	Precision	Recall	F1 Score	AUROC
Multimodal Full BCE loss	0.9934	0.9732	0.9934	0.9594	0.9966
Multimodal Full BCE loss Image Shuffle	0.9921	0.9805	0.9306	0.9549	0.9644
Multimodal Full BCE loss Text Shuffle	0.8451	0.113	0.1065	0.1097	0.5121

Figure 2: Testing performance scores for Multimodal and Unimodal (Image and Text) models using BCE loss are depicted in the top table. The bottom table illustrates the dependency of the Multimodal model on each modality, as demonstrated by the performance impact of shuffling text and image data during Bimodal Sensitivity Tests.

Model Type	Test Accuracy	Precision	Recall	F1 Score	AUROC
Multimodal Full BCE loss	0.9934	0.9732	0.9934	0.9594	0.9966
Multimodal Full Focal loss	0.9911	0.9339	0.9728	0.9492	0.9991
Image Full BCE loss	0.9123	0.2212	0.0517	0.0793	0.7735
Image Full Focal loss	0.835	0.2762	0.5318	0.352	0.7812
Text Full BCE loss	0.9948	0.9869	0.9558	0.9688	0.9995
Text Full Focal loss	0.9917	0.9272	0.9868	0.9522	0.9997

Figure 3: Green= overall winner, Yellow= winner between loss functions. This table displays the test accuracy, precision, recall, F1 score, and AUROC for multi-modal and uni-modal models using both Binary Cross-Entropy (BCE) and Focal Loss. It highlights the differences in performance metrics across different model types and loss functions, illustrating how multi-modal integration impacts diagnostic accuracy in comparison to single modality models.

The next phase involved assessing the efficacy of Focal Loss in addressing class imbalance. Figure 3 shows that there were no significant performance improvements with Focal Loss over BCE Loss on the full dataset. This led us to experiment with a sub-sampled dataset, which moderately reduced class imbalance and is detailed in Figure 4. Here, we observed improved balance between precision and recall, particularly in models trained with Focal Loss. Multi-modal models consistently outperformed single-modality models in this setup, particularly in AUROC scores, underscoring the effectiveness of integrating both modalities and robust generalization overall. Overall, our findings support the utility of multi-modal approaches, the decision to use sub-sampling helped to balance the dataset, which seems to allow the Multi-modal models to effectively integrate diverse data types underscoring the value of integrating both text and image data, which likely provides a richer, more comprehensive feature set for making predictions.

Model Type	Test Accuracy	Precision	Recall	F1 Score	AUROC
Multimodal Full BCE loss	0.9934	0.9732	0.9934	0.9594	0.9966
Multimodal 75 BCE loss	0.9828	0.9905	0.9444	0.966	0.9971
Model Type	Test Accuracy	Precision	Recall	F1 Score	AUROC
Multimodal Full Focal loss	0.9911	0.9339	0.9728	0.9492	0.9991
Multimodal 75 Focal loss	0.9925	0.9802	0.9881	0.9835	0.9981

Model Type	Test Accuracy	Precision	Recall	F1 Score	AUROC
Multimodal 75 BCE loss	0.9828	0.9905	0.9444	0.966	0.9971
Image 75 BCE loss	0.7595	0.7465	0.1567	0.2541	0.7636
Text 75 BCE loss	0.9489	0.9588	0.8486	0.8975	0.9933
Model Type	Test Accuracy	Precision	Recall	F1 Score	AUROC
Multimodal 75 Focal loss	0.9925	0.9802	0.9881	0.9835	0.9981
Image 75 Focal loss	0.624	0.3988	0.7307	0.5113	0.7417
Text 75 Focal loss	0.9817	0.9377	0.9957	0.9652	0.998

Figure 4: Performance metrics of models trained on the sub-sampled dataset, illustrating improved balance between precision and recall, particularly for models trained with Focal Loss. Multi-modal models outperformed single-modality models, particularly in AUROC scores.

4 Discussion

In this study, we investigated the performance of multimodal and single-modality models in classifying pneumonia, with a particular focus on reducing the effects of class imbalance through subsampling and exploring loss functions. Our results demonstrate that multimodal models, which integrate both text and image data, were able to perform competitively with their single-modality counterparts across a variety of metrics. This was particularly evident when models were trained on a subsampled dataset designed to reduce class imbalance, where multimodal models exhibited superior generalization capabilities and higher precision-recall balance. Despite the promising results, the study's limitations need to be acknowledged. The reliance on text data, which may not always be available or accurately annotated in clinical settings, could restrict the practical application of our findings. Moreover, the sub-optimal performance of the image-only models may indicate challenges in extracting meaningful information from visual data, potentially due to the inherent complexity, ambiguity or perhaps be due to the modality being less informative of the images in our setup.

Future work would entail optimizing data selection and pre-processing so that the tasks accurately represent the diversity of clinical scenarios encountered in practice, developing techniques to extract more meaningful features from medical imagery , and exploring more sophisticated methods for integrating text and image features in order to improve the contribution balance of both data modalities. This will ensure a more clinically practical classifier can be considered for healthcare tasks.

Contributions:

Stacy Chao trained all full dataset versions of the Multi-modal, Text and Image models with either focal or BCE loss. 6 Models trained total.

Aurelia Li trained all sub-sampled dataset versions of the Multi-modal, Text and Image models with either focal or BCE loss. 6 Models trained total.

Pre-processing was a collective effort as was the analysis and the writing of this report.

References

Md Inzamam Ul Haque, Abhishek K. Dubey, and Jacob D. Hinkle. The effect of image resolution on automated classification of chest x-rays. *medRxiv*, 2021. doi: 10.1101/2021.07.30.21261225. URL https://www.medrxiv.org/content/10.1101/2021.07.30.21261225v1.full.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Alistair E. W. Johnson, Tom J. Pollard, Roger G. Mark, et al. MIMIC-IV-CXR: Medical information mart for intensive care, chest x-ray dataset. PhysioNet, 2020. URL https://physionet.org/content/mimic-cxr-jpg/2.0.0/.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, September 2019. ISSN 1367-4811. doi: 10.1093/bioinformatics/btz682. URL http://dx.doi.org/10.1093/bioinformatics/btz682.

Kuan Liu, Yanen Li, Ning Xu, and Prem Natarajan. Learn to combine modalities in multimodal deep learning, 2018.

J. V. Sousa, P. Matos, F. Silva, P. Freitas, H. P. Oliveira, and T. Pereira. Single modality vs. multimodality: What works best for lung cancer screening? *Sensors (Basel, Switzerland)*, 23(12):5597, 2023. doi: 10.3390/s23125597.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard?, 2020.

Koen Welvaars, Jacobien H F Oosterhoff, Michel P J van den Bekerom, Job N Doornberg, Ernst P van Haarst, OLVG Urology Consortium, and the Machine Learning Consortium. Implications of resampling data to address the class imbalance problem (ircip): An evaluation of impact on performance between classification algorithms in medical data. *JAMIA Open*, 6(2):ooad033, 05 2023. doi: 10.1093/jamiaopen/ooad033.

Nan Wu, Stanisław Jastrzebski, Kyunghyun Cho, and Krzysztof J. Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks, 2022.