Multi-Modal Deep Learning Framework for Disease Detection

Aditya Jadhav
Department of Computer Science, Virginia Tech
Blacksburg, Virginia, USA
adityasj@vt.edu

Merna Khamis
Department of Computer Science, Virginia Tech
Blacksburg, Virginia, USA
merna@vt.edu

Abstract

In modern healthcare, accurate disease diagnosis often relies on the combination of multiple information sources, such as radiology images and clinical reports. To mimic this diagnostic process, we propose a multi-modal deep learning framework that integrates these two essential data modalities to enhance diagnostic accuracy and improve the interpretability of the model's predictions. The framework is designed to leverage the complementary nature of visual features from chest X-rays and the contextual insights from clinical notes, allowing it to build a more comprehensive understanding of the underlying conditions. For this project, we adopt a multi-task approach that trains both the radiology images and doctor notes simultaneously. This approach allows the model to effectively learn correlations between the visual and textual modalities, enabling it to form richer representations for better decision-making. The framework includes two branches: a Convolutional Neural Network (CNN) for processing and analyzing chest X-rays, and a Transformer-based model (BERT) for processing clinical notes to focus on the diagnosis of diseases such as Atelectasis, Pneumonia, and Edema. These branches independently extract detailed features from their respective inputs, which are then combined into a unified feature representation. By learning from both modalities, the model is expected to improve its ability to detect and diagnose diseases by capturing the complex relationships between visual patterns in the images and medical context provided by the notes. This multi-modal approach closely mirrors how clinicians integrate multiple sources of information in their decision-making, ultimately aiming to enhance diagnostic performance and contribute to more accurate, personalized healthcare solutions.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

Keywords: Multi-Modal Learning, Medical Diagnostics, Transformer Models, Disease Detection, CNN, BERT, LSTM, PubMedBERT

ACM Reference Format:
Aditya Jadhav and Merna Khamis. 2018. Multi-Modal Deep Learning Framework for Disease Detection. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

In the domain of medical diagnostics, traditional approaches often rely on single-modality models that analyze either images or textual data independently. While effective to some extent, this approach does not replicate the integrative diagnostic process typically employed by medical professionals, who often base their diagnoses on a synthesis of diverse data types, including radiology images and textual clinical notes. Inspired by this real-world approach, our project proposes a multi-modal deep learning framework that combines these two critical types of data—radiographic and textual—to enhance the accuracy and robustness of disease detection. This framework aims to address the inherent limitations of single-modality models by leveraging complementary information from both visual and textual data, thereby mimicking the real-world diagnostic process.

This work proposes a multi-modal deep learning framework that integrates chest X-rays and clinical notes, leveraging their complementary strengths to enhance diagnostic performance. The proposed framework is designed to simultaneously extract visual features from radiology images using a Convolutional Neural Network (CNN) and semantic features from clinical notes using a Transformer-based model, BERT(PubMedBERT)[3]. These extracted features are then fused into a unified representation, enabling the model to capture correlations between the two modalities that single-modality approaches cannot achieve. This multi-modal approach not only mimics the diagnostic reasoning of clinicians but also offers the potential for more accurate, robust, and interpretable predictions.

The motivation for this project is rooted in bridging the gap between traditional diagnostic methods and advanced AI-driven solutions. By integrating imaging and textual data into a cohesive framework, we aim to create a tool that not only augments the capabilities of clinicians but also

contributes to the broader field of personalized medicine. This work seeks to advance the state-of-the-art in medical diagnostics by leveraging the power of multi-modal learning, addressing real-world challenges, and setting the stage for future innovations in AI-powered healthcare.

2 List of Contributions

2.1 Development of a Multi-Modal Framework

Our framework combines chest X-rays and doctor notes to improve diagnostic processes. A CNN is used to process the visual features from chest X-rays, while a Transformer-based model (BERT - PubMedBERT) is to handle the text data from clinical notes.

2.2 Multi-Task Learning Approach

We adopted a multi-task learning approach to train the image and text branches simultaneously, allowing the model to leverage shared representations from both visual and textual inputs. This approach enhances the model's ability to generalize across multiple data modalities, enabling it to make more robust predictions.

2.3 Feature Fusion for Joint Representation

The features extracted from the CNN and Transformer branches were combined into a unified vector space. This feature fusion step enables the model to integrate visual and textual information, resulting in a more comprehensive understanding of the diagnostic data.

2.4 Utilization of Pre-Trained Models

The CNN is based on a pre-trained model, ResNet, while the text processing utilized a Transformer model, BERT, specifically PubMedBERT. By combining these two models, we produced a comprehensive diagnosis that reflects how both images and clinical notes contribute to real-world medical decisions.

3 Related Work

Single-modality models, predominantly utilizing either Convolutional Neural Networks (CNNs) for image analysis or Transformer-based models for text processing, have shown significant efficacy in their respective domains. CNNs have been extensively applied to medical imaging tasks, demonstrating the ability to detect intricate patterns within radiological images. Conversely, Transformer models have revolutionized the processing of clinical notes by capturing complex semantic relationships within text. However, the inherent limitation of single-modality models is their inability to incorporate the complementary information available across different data types.

Recent research has begun to explore the integration of both imaging and textual data to more closely mimic clinical decision-making processes. For instance, [7] developed a transformer-based neural network architecture capable of integrating both chest radiographs and clinical parameters. Their model demonstrated improved diagnostic performance over traditional single-modality approaches, validating the efficacy of a multi-modal strategy in a clinical setting. Similarly, [12] introduced a novel architecture that aligns deep learning model explanations with radiologist eye-gaze data, effectively improving the interpretability and trustworthiness of automated diagnostic models.

Additionally, [10] explored the domain shift problem in medical imaging, illustrating how models trained on data from one set of conditions often fail when applied to data from different conditions. This underscores the need for robust multi-modal approaches that can generalize across diverse clinical settings. Moreover, [5] investigated the propensity of deep learning models to exploit spurious correlations in training data. They suggested that transfer learning could be leveraged to encourage models to focus on clinically relevant features rather than biases inherent in the training data.

This approach has been further highlighted in a study by [1], which utilized deep learning models to predict self-reported gender, age, ethnicity, and insurance status from chest radiographs. Their research, conducted using over 55,000 images from the MIMIC-CXR dataset, demonstrates the capability of deep learning models to accurately predict these demographics with high degrees of accuracy, achieving AUC values up to 0.999 for gender prediction. This study not only emphasizes the potential of AI in enhancing diagnostic accuracy but also addresses the challenge of bias in medical AI applications, ensuring models perform equitably across diverse patient populations.

Moreover, the introduction of the EHRXQA dataset by [2] further exemplifies the growing interest in combining structured EHR data with imaging data to address complex medical questions through AI, thereby enhancing both the interpretability and effectiveness of diagnostic models in clinical settings.

Our proposed framework seeks to build upon these foundational studies by creating a system that not only integrates both modalities—image and text—but also incorporates the advantages of multi-task learning. This approach is inspired by the 2022 study on "Predicting Depression and Anxiety on Reddit: a Multi-task Learning Approach" by [11], which emphasized the utility of multi-task learning in effectively combining different data sources. Such an approach can be particularly beneficial in medical applications where diverse data types play a critical role in diagnostic accuracy.

By synthesizing these approaches, our project aims to bridge the gap between traditional medical diagnostics and advanced AI-driven methods, providing a more comprehensive and accurate diagnostic tool that leverages the full spectrum of available medical data.

4 Model Description

The proposed multi-modal framework is designed to integrate two complementary modalities—radiological images and clinical notes—for enhanced disease detection and diagnosis. This framework emulates the diagnostic reasoning process of medical professionals by combining visual patterns extracted from medical images with the semantic insights embedded in textual clinical reports. The model comprises two major branches: the Image Branch, responsible for processing radiology images, and the Text Branch, which handles clinical notes. The outputs of these branches are then fused into a shared latent feature space to capture the interactions between the two modalities, ultimately enabling robust multi-label disease classification [4].

4.1 Dataset Collection and Preprocessing

Our study utilizes the MIMIC-CXR dataset, a comprehensive collection of 377,110 de-identified chest radiographs from 227,835 studies conducted at the Beth Israel Deaconess Medical Center between 2011 and 2016 [6]. Each study includes both frontal and lateral views accompanied by semi-structured free-text radiology reports authored during routine clinical practice. For analysis, the radiology images were resized to 224x224 pixels and normalized. The accompanying radiology reports were tokenized and cleaned of irrelevant characters, preparing them for effective natural language processing.

An illustrative snippet from the dataset is shown to highlight the typical content and format of the data used in our analysis. Figure 1 features an example study from the MIMIC-CXR dataset: Above (a) presents the radiology report with the interpretation of the image, where any personally identifiable information (PHI) has been removed and replaced with underscores (___) to maintain privacy. Below, two chest radiographs for this study are displayed: (b) the frontal view (left image) and (c) the lateral view (right image)[6]. These images, along with their corresponding anonymized reports, are critical for our study as they provide both visual and textual data for the development of our multi-modal deep learning framework.

4.2 Multi-Task Learning Approach

We adopt a multi-task learning approach[11] to simultaneously train both the image branch and the text branch of the framework. In this approach, the model is optimized for two distinct but related tasks—one for processing the radiology images and the other for processing the clinical notes. Each task contributes to learning a shared feature space, allowing the model to leverage complementary information from both data modalities.

- **Image Task:** Focuses on identifying disease-relevant visual patterns in the X-rays. This involves detecting and analyzing specific features within the images that are indicative of various medical conditions.
- **Text Task:** Focuses on extracting diagnostic insights from the clinical notes by analyzing relevant medical language. This includes interpreting the textual descriptions provided by healthcare professionals, aligning them with the visual data, and enhancing the overall diagnostic accuracy.

The multi-task learning framework enables the model to improve its performance on both tasks by sharing representations learned from the image and text inputs. This helps the model become more robust and better equipped to handle complex medical data.

4.3 Image Branch

The Image Branch of the framework is built upon **EfficientNets**, a state-of-the-art family of convolutional neural networks known for their ability to balance model accuracy and efficiency. EfficientNets utilize a compound scaling approach that optimizes the depth, width, and resolution of the network simultaneously, making them particularly effective for medical imaging tasks where extracting fine-grained patterns from high-resolution images is critical. In this project, radiology images are preprocessed by resizing them to a standard resolution of 224×224 pixels and normalizing their pixel values to ensure compatibility with the EfficientNet architecture. These preprocessing steps ensure that the network focuses on diagnostic features without being influenced by irrelevant variations in the data. Once processed, the images are passed through the EfficientNet architecture, which consists of convolutional layers, batch normalization, and activation functions, capturing hierarchical features from low-level edges to high-level semantic patterns. The penultimate layer of EfficientNet outputs a **1280-dimensional feature vector**, encapsulating the most important visual characteristics relevant to disease detection.

4.3.1 Dimensionality Reduction with PCA:. The high-dimensional feature vectors generated by EfficientNet are rich in information but can pose computational challenges and increase the risk of overfitting. To address this, **Principal Component Analysis (PCA)** is applied to reduce the dimensionality from 1280 to 256 dimensions. PCA achieves this by identifying principal components that retain the most significant variance in the data while discarding redundant or less informative components.

Mathematically, PCA computes the covariance matrix:

$$C = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T$$

where μ is the mean of the dataset, and n is the number of samples. PCA then performs eigenvalue decomposition on C, retaining the top k eigenvectors corresponding to the largest eigenvalues. The reduced feature vector x_{reduced} is computed

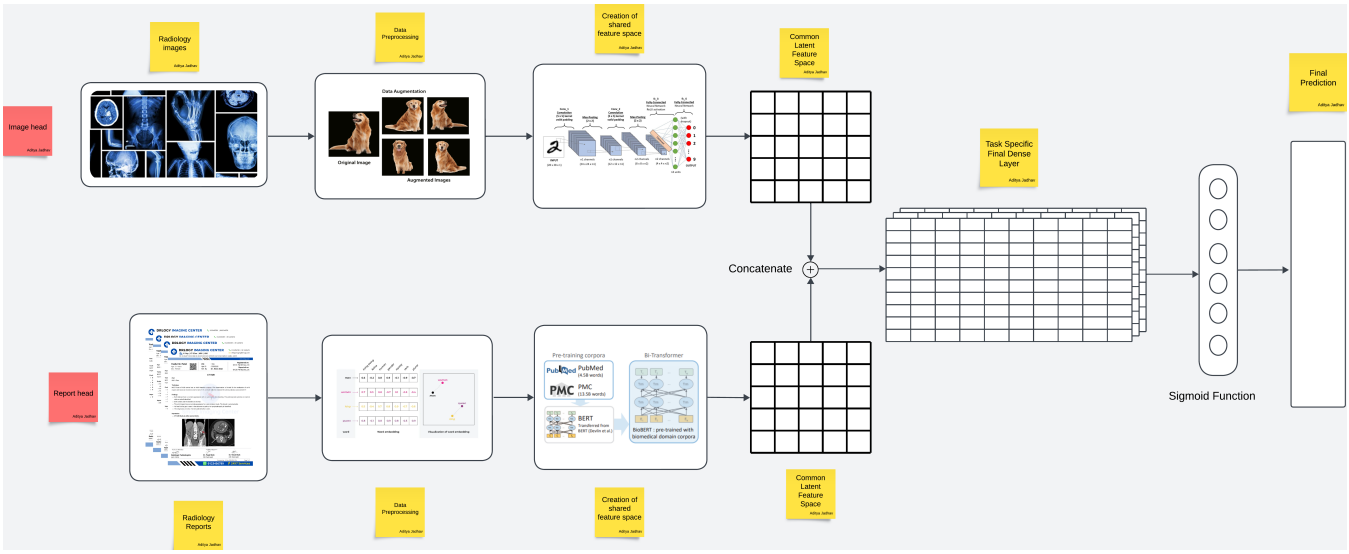

Figure 1. Flow diagram for Multi-modal Disease Prediction

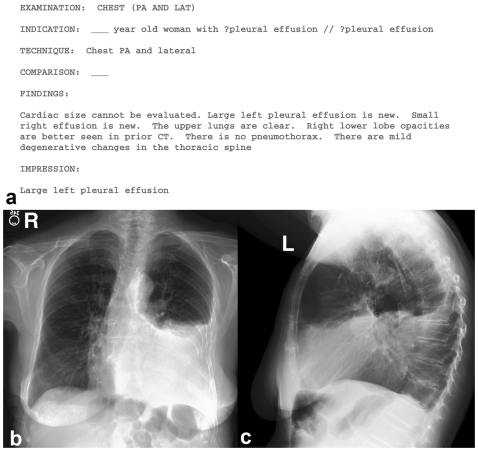

Figure 2. Example study contained in MIMIC-CXR[6]

as:

$$x_{\text{reduced}} = W_k^T (x - \mu)$$

where W_k is a matrix containing the top k eigenvectors, and $k = 256$ in this case. This dimensionality reduction ensures computational efficiency while preserving critical visual information, making the feature vectors suitable for subsequent fusion and classification tasks.

4.3.2 Loss Functions:

To optimize the training of the Image Branch, two loss functions are employed: **Binary Cross-Entropy (BCE) Loss** and **Focal Loss**. These loss functions are essential for handling the challenges posed by multi-label classification and class imbalance in medical imaging datasets.

Binary Cross-Entropy (BCE) Loss: Binary Cross-Entropy (BCE) Loss is widely used for multi-label classification tasks, where each disease label is treated as an independent binary classification problem. BCE Loss measures the error between the predicted probability $\hat{y}_i$ and the ground truth y_i for each label. The BCE Loss[13] is defined as:

$$\text{BCE Loss} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

where:

- N is the total number of samples,
- $y_i \in \{0, 1\}$ is the ground truth label for the i-th sample,
- $\hat{y}_i \in [0, 1]$ is the predicted probability for the i-th sample.

Why BCE Loss? BCE Loss is particularly suited for multi-label classification because it evaluates each label independently, allowing the model to predict probabilities for multiple diseases simultaneously. In medical diagnostics, where a single patient might exhibit symptoms of multiple conditions, BCE Loss ensures that each condition is appropriately accounted for without interference from other labels.

Focal Loss: Focal Loss is designed to address class imbalance, a common issue in medical datasets where certain conditions (e.g., rare diseases) are underrepresented compared to others. It modifies the standard BCE Loss by applying a weighting factor to emphasize harder-to-classify examples. The Focal Loss[13] is defined as:

$$\text{Focal Loss} = -\frac{1}{N} \sum_{i=1}^{N} \alpha \cdot (1 - \hat{y}_i)^{\gamma} \cdot y_i \log(\hat{y}_i)$$

where:

- α is a balancing factor that adjusts the relative importance of positive and negative samples,

- γ is the focusing parameter that controls the degree of emphasis on harder examples,
- y_i and $\hat{y}_i$ are as defined above.

Why Focal Loss? Focal Loss is particularly useful when the dataset is highly imbalanced. For example, in a medical imaging dataset, common conditions might dominate the training process, causing the model to ignore rare conditions. By introducing the term $(1-\hat{y}_i)^\gamma$, Focal Loss reduces the contribution of well-classified examples (where $\hat{y}_i$ is close to 1) and focuses on harder examples (where $\hat{y}_i$ is far from 1). The parameter γ controls this focusing effect: a higher γ places more emphasis on difficult samples. The α parameter further helps by adjusting the weights for positive and negative samples, ensuring that minority classes are not overshadowed by majority classes.

4.3.3 Feature Refinement: After training, the 1280-dimensional feature vectors generated by EfficientNet are reduced to 256 dimensions using PCA, as described above. These refined feature vectors are then passed to the feature fusion layer, where they are combined with textual features extracted from the Text Branch. This dimensionality reduction improves computational efficiency, minimizes overfitting risks, and ensures that the most salient visual features contribute effectively to the unified feature representation for downstream classification tasks.

4.4 Text Branch

The Text Branch of the framework is designed to process textual clinical notes, such as radiology reports, and extract semantic features relevant to disease diagnosis. Unlike images, textual data provides contextual and descriptive information about patient conditions, enabling the model to interpret additional diagnostic cues. This branch employs a **Transformer-based model**, **BERT** specifically we used **PubMedBERT**, which are specifically fine-tuned for biomedical texts [3]. These models excel at understanding complex medical language and capturing contextual dependencies in unstructured textual data, making them particularly suitable for clinical settings.

4.4.1 Preprocessing of Clinical Notes. The input clinical notes undergo several preprocessing steps to ensure compatibility with the Transformer architecture and improve the quality of feature extraction. We began by selecting the diseases Atelectasis, Pneumonia, and Edema due to their relatively balanced distribution. Initially, labels with -1 (uncertain cases) were treated as 0 (no disease). However, this introduced bias and confusion to the model. To address this, we ultimately removed all rows with -1 labels. The text data from multiple columns (examination, indication, findings, and impression) was merged into a single column named (combined_text) for unified analysis. The cleaned data was then split into training and testing sets using an 80/20 ratio. Finally, the text was tokenized using PubMedBERT to prepare embeddings for model input.

4.4.2 Transformer-Based Feature Extraction: The Transformer-based model processes the input embeddings through its multi-layer architecture, comprising self-attention mechanisms and feed-forward networks. The self-attention mechanism enables the model to capture both local and global relationships within the text, allowing it to focus on diagnostically relevant phrases. For instance, terms such as "pneumonia," "infiltrate," or "consolidation" are contextualized based on their surrounding text, ensuring that the model understands the nuances of medical language.
At the final layer of the Transformer, a **contextualized feature vector** is generated for each token, capturing its semantic and syntactic relationships with other tokens in the input sequence. These token-level embeddings are aggregated to form a **global text representation**, typically by taking the embedding of the `[CLS]` token, which represents the entire input sequence. This global representation serves as the output of the Text Branch and is passed to the feature fusion layer for integration with visual features.

4.4.3 LSTM Classifier. The text embeddings generated by PubMedBERT are passed through a Bidirectional Long Short-Term Memory (Bi-LSTM) network. The LSTM is specifically used for its ability to capture sequential dependencies and contextual relationships within the embedding space, which is essential for analyzing clinical notes [14].

The architecture of the LSTM classifier includes:

1. **LSTM Layer:** A bidirectional LSTM with a hidden dimension of 256 processes the embeddings. This allows the model to capture both forward and backward dependencies in the sequence.
2. **Attention Mechanism:** An attention mechanism is applied to focus on the most relevant parts of the sequence. This improves the model's interpretability by identifying which features or tokens contributed most to the predictions [8].
3. **Fully Connected Layer:** The output of the attention mechanism is passed to a fully connected layer, which maps the features to the target condition labels.

The LSTM classifier outputs probabilities for each condition, which are thresholded to produce binary predictions.

4.4.4 Fine-Tuning on Clinical Data: To adapt the Transformer-based model to the domain of clinical text, it is fine-tuned on a labeled dataset of radiology reports and their corresponding diagnoses. Fine-tuning involves updating the pre-trained model's weights using a task-specific objective function, such as **Binary Cross-Entropy (BCE) Loss**. This ensures that the model learns patterns specific to the clinical domain while retaining its general language understanding capabilities. Additionally, the use of dropout

regularization during fine-tuning helps prevent overfitting, particularly when working with smaller medical datasets.

4.4.5 Dimensionality Reduction of Text Features: The global feature representation generated by the Transformer model is often high-dimensional, which can lead to computational inefficiencies and redundant information when combined with visual features. To address this, **Principal Component Analysis (PCA)** is applied to reduce the dimensionality of the textual feature vectors. By projecting the original high-dimensional representation into a lower-dimensional space, PCA retains the most critical semantic components while discarding less informative ones. The reduced textual feature vectors are then passed to the feature fusion layer, where they are concatenated with visual feature vectors from the Image Branch.

4.4.6 Importance of the Text Branch in Multi-Modal Learning. The Text Branch plays a crucial role in multimodal learning by providing contextual and descriptive information that complements the spatial patterns identified by the Image Branch. For instance, while an X-ray image might reveal visual signs of consolidation, the corresponding clinical note might specify the underlying cause, such as bacterial pneumonia or viral infection. By integrating these two modalities, the model achieves a more comprehensive understanding of the patient's condition, enabling accurate and interpretable predictions.

4.5 Feature Fusion

Feature fusion is a critical component of the proposed multimodal framework, enabling the integration of feature vectors extracted from radiology images and clinical notes into a unified latent representation. This unified representation combines the complementary strengths of both modalities—visual patterns from the images and contextual information from the text. The fusion process ensures that the model captures intricate correlations between the two data sources, leading to improved diagnostic accuracy and robustness.

After processing radiology images through the CNN (EfficientNet) and clinical notes through the Transformer-based model (BERT or BioBERT), their respective feature vectors are concatenated to form a **combined latent space**. The dimensionality of this latent space is carefully managed to ensure that it contains the most salient features from both modalities while remaining computationally efficient. For instance, the 256-dimensional feature vector from the image branch and the reduced feature vector from the text branch are concatenated to create a unified feature representation. This latent space serves as the input to a **Fully Connected Feed-Forward Neural Network (FC-FNN)**, which performs the final classification task.

4.5.1 Fully Connected Feed-Forward Neural Network (FC-FNN) for Feature Fusion: The **FC-FNN** consists of a series of fully connected (dense) layers designed to process the combined latent space and generate disease predictions. Each dense layer in the FC-FNN applies a linear transformation to the input, followed by a non-linear activation function, enabling the model to learn complex relationships between the fused features. Mathematically, for a dense layer, the transformation is represented as:

$$z = W \cdot x + b$$

where:

- z is the output vector of the layer,
- W is the weight matrix,
- x is the input vector,
- b is the bias vector.

This output is passed through an activation function such as **ReLU (Rectified Linear Unit)**, defined as:

$$\text{ReLU}(z) = \max(0, z)$$

ReLU introduces non-linearity, enabling the FC-FNN to model complex patterns in the data.

4.5.2 Output Layer with Sigmoid Activation: The final layer of the FC-FNN generates a binary classification variable for each disease. Since the task involves predicting the presence or absence of multiple diseases independently, a **sigmoid activation function** is applied to each output neuron. The sigmoid function maps the output of each neuron to a probability between 0 and 1, making it ideal for multi-label binary classification tasks. Mathematically, the sigmoid function is defined as:

$$\text{Sigmoid}(z) = \frac{1}{1 + e^{-z}}$$

where z is the input to the neuron. Each sigmoid output represents the probability of a specific disease being present, allowing the model to make independent predictions for each condition.

4.5.3 Training on Fused Features: Once the concatenated feature vector is passed through the FC-FNN, the network trains on the unified latent representation to predict disease probabilities. During training, the model optimizes a loss function (e.g., Binary Cross-Entropy Loss or Focal Loss) to minimize the error between the predicted probabilities and ground truth labels. The FC-FNN adapts its weights to capture the interactions between visual and textual features, effectively learning how the modalities complement each other. This joint learning ensures that the model can leverage correlations such as visual indications of a condition (e.g., consolidation in an X-ray) combined with textual mentions of associated symptoms or clinical findings (e.g., "shortness of breath" or "fever" in the report).

4.5.4 Why Use FC-FNN for Feature Fusion? The choice of FC-FNN for feature fusion is driven by its ability to model high-dimensional interactions between features from different modalities. Key reasons for using an FC-FNN include:

- **Flexibility**: FC-FNNs can handle inputs of varying dimensionalities, making them suitable for concatenated feature vectors from different sources.
- **Expressiveness**: Fully connected layers are capable of capturing non-linear relationships between features, allowing the model to learn intricate dependencies between image and text data.
- **Dimensionality Reduction**: The FC-FNN can reduce the dimensionality of the concatenated latent space while retaining critical information, ensuring computational efficiency in subsequent layers.
- **End-to-End Learning**: FC-FNNs enable seamless integration into the framework, allowing the entire model to be trained jointly, optimizing for both modalities simultaneously.
- **Regularization Capabilities**: Techniques such as dropout and weight decay can be applied to FC-FNNs to prevent overfitting, particularly when working with smaller datasets.

4.5.5 Benefits of Feature Fusion: The feature fusion process enables the model to integrate diverse information types into a cohesive framework, addressing the limitations of single-modality models. By combining features from images and text, the model achieves a holistic understanding of the diagnostic data, capturing insights that would be missed by analyzing each modality in isolation. For instance, visual abnormalities in an X-ray might be linked to contextual descriptions in the clinical report, such as mentions of specific symptoms or medical history. The FC-FNN leverages this synergy to enhance diagnostic accuracy, robustness, and interpretability, providing a reliable tool for medical decision-making.

4.5.6 Regularization and Optimization in FC-FNN:. To ensure robust learning, regularization techniques such as **dropout** are applied to the FC-FNN layers. Dropout randomly sets a fraction of the neurons to zero during training, preventing the network from relying too heavily on specific features and reducing overfitting. Additionally, optimization techniques such as the **Adam optimizer** are employed to ensure efficient convergence of the model. The learning rate is dynamically adjusted during training to balance exploration and convergence, enabling the model to fine-tune its weights for both modalities.

5 Evaluation & Experimental Results

This section describes the evaluation metrics, experimental setup, and results obtained during the testing phase of the proposed multi-modal framework. The evaluation focuses on the model's ability to accurately predict diseases using both image and text modalities, as well as its robustness and interpretability in clinical settings. We have evaluated our models' performance using the following metrics:

- Accuracy: Measures the overall correctness of predictions.
- Precision: Indicates the proportion of true positive predictions among all positive predictions, providing insight into the model's ability to minimize false positives.
- Recall: Reflects the proportion of true positive predictions out of all actual positives, assessing the model's ability to identify relevant cases.
- F1-Score: Combines precision and recall into a single metric, emphasizing their balance.

5.1 Image Model Results

The image model was evaluated using the following classification metrics, precision, recall, and F1-score—to assess its performance in predicting clinical conditions from chest X-ray images. Below is a summary of the key results:

- Atelectasis:
  - Precision: 0.8249
  - Recall: 1.0000
  - F1-Score: 0.9041
- Cardiomegaly:
  - Precision: 0.7208
  - Recall: 0.9403
  - F1-Score: 0.8161
- Consolidation:
  - Precision: 0.6882
  - Recall: 0.4129
  - F1-Score: 0.5161
- Support Devices:
  - Precision: 0.9349
  - Recall: 1.0000
  - F1-Score: 0.9663

Conditions such as Atelectasis and Support Devices achieved high F1-scores (0.9041 and 0.9663, respectively) due to a balance of high precision and recall. This suggests the model is highly reliable in predicting these conditions, which are more prevalent and well-represented in the dataset.
For conditions like Consolidation, the F1-score was lower (0.5161), primarily due to lower recall (0.4129). This indicates that the model struggled to identify all true positives for this class, possibly due to class imbalance or subtle features in the images.
Cardiomegaly showed relatively balanced precision (0.7208) and recall (0.9403), resulting in an F1-score of 0.8161. This suggests that the model is moderately effective in identifying true positives while avoiding false positives for this condition.

The image model outperformed the text branch for certain conditions (e.g., Atelectasis), demonstrating the value of visual features in identifying structural abnormalities. However, the text branch provides complementary information, particularly for conditions with subtle or ambiguous imaging features.

5.2 Clinical Notes Model Results:

The proposed model was initially evaluated using a multi-label classification framework to predict clinical conditions from textual data [4]. While this approach allowed for simultaneous predictions of multiple conditions, the evaluation metrics highlighted significant challenges. Specifically, the model struggled to generalize effectively across all conditions, particularly for less frequent classes, due to the inherent complexity of multi-label learning and class imbalance.
To address these challenges, we transitioned to a single-label classification analysis [9], where each condition was treated as an independent binary classification task. This approach provided a clearer understanding of the model's performance for each condition and identified conditions requiring further focus or rebalancing strategies.
The overall performance metrics across all conditions were as follows:

- Accuracy: 0.6050
- Precision: 0.5032
- Recall: 0.0783
- F1-Score: 0.1313

These results indicate relatively low precision, recall, and F1-scores, which can be attributed to the multi-label nature of the task and the imbalanced distribution of conditions in the dataset. The model struggled to generalize effectively across all conditions simultaneously, particularly for less frequent classes. To further investigate the model's performance, we evaluated each condition separately. The per-condition metrics were as follows:

- Atelectasis:
  - Accuracy: 0.7720
  - Precision: 0.4973
  - Recall: 0.1027
  - F1-Score: 0.1702
- Pneumonia:
  - Accuracy: 0.9209
  - Precision: 0.3939
  - Recall: 0.0105
  - F1-Score: 0.0205
- Edema:
  - Accuracy: 0.8396
  - Precision: 0.6185
  - Recall: 0.1216
  - F1-Score: 0.2032

The dataset exhibited significant class imbalance, with certain conditions (e.g., Pneumonia) appearing much less frequently than others. This imbalance led to low recall and F1-scores, especially for rare conditions.
Conditions such as Edema achieved relatively higher precision (0.6185) and F1-score (0.2032), indicating the model's ability to correctly identify positive cases. However, recall remained low across all conditions, suggesting difficulty in capturing all true positive cases.
For rare conditions like Pneumonia, the model struggled significantly, achieving a recall of only 0.0105 and an F1-score of 0.0205. This highlights the need for targeted strategies to handle such cases.

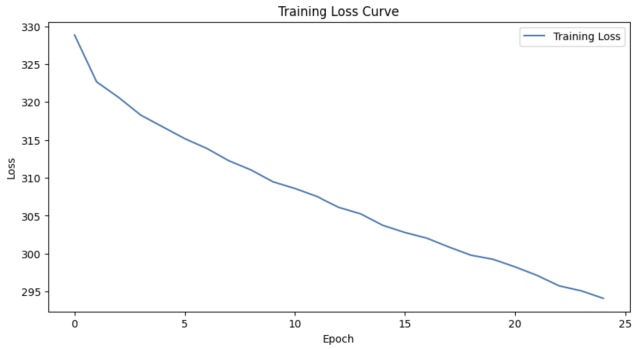

Figure 3. The training loss curve for the clinical notes mode

Also, figure 3 shows the training loss curve illustrates the model's convergence over 25 epochs. The gradual decrease in loss shows that the model is effectively learning from the training data without overfitting. The steady decline in loss indicates a stable optimization process, while small oscillations suggest potential areas for further fine-tuning.
Finally, the LSTM classifier also played an important role in modeling the sequential relationships in the text embeddings. Its ability to process and aggregate information from sequences contributed significantly to the overall performance metrics, particularly for conditions where subtle contextual patterns in the notes were critical for accurate prediction.

5.3 Multi-modal approach results:

In this study, we employed a Fully Connected Neural Network (FCNN) architecture as a baseline model to explore its capability to handle the given dataset. The FCNN consisted of multiple dense layers with ReLU activations, interspersed with dropout layers to mitigate overfitting. An embedding layer of size 512 was used to transform the input features into dense representations, which were then passed through the network for prediction. The architecture was intentionally kept straightforward to establish a baseline for performance evaluation and allow ease of interpretability.
The training process revealed the following results:

- Validation accuracy across folds remained consistently low, with values below 30%.
- The highest validation accuracy recorded was approximately 27.5%.
- Validation loss decreased over epochs, indicating that the model was learning, but without substantial accuracy improvements.
- The ReduceLROnPlateau callback effectively reduced the learning rate when validation loss stagnated, aiding convergence in later epochs.

Several challenges contributed to the model's underperformance. Firstly, due to storage restrictions associated with the image dataset, we were compelled to train the model on a very small subset. This limitation in dataset size affected the diversity and representativeness of the training data, severely constraining the model's ability to generalize to unseen data. Secondly, the embedding layer size of 512, while functional, might not have been sufficient to represent the high-dimensional features effectively. Increasing this size could provide a richer representation and potentially improve the model's performance. Furthermore, while the FCNN is effective for tasks requiring simple transformations, its lack of architectural complexity might have hindered its ability to capture intricate patterns present in the data. Metrics also pointed to issues such as overfitting to the training data and insufficient preprocessing, which likely compounded the observed performance bottlenecks.

6 Broader Impacts and Discussion

The proposed multi-modal framework for disease detection has far-reaching implications across clinical, societal, and technological domains. By integrating radiology images and clinical notes into a unified diagnostic model, this approach demonstrates significant potential to enhance diagnostic accuracy, reduce healthcare disparities, and advance artificial intelligence methodologies in critical domains such as healthcare. The ability of the model to combine visual patterns from images with contextual information from text emulates the integrative decision-making process of medical professionals, providing a tool that could improve patient outcomes by enabling earlier and more accurate disease detection.

However, one of the most significant challenges encountered in this work was the high computational power required to train the multi-modal framework effectively. Training deep learning models that combine convolutional and Transformer-based architectures inherently demands extensive memory and processing power, and the Fully Connected Feed-Forward Neural Network (FC-FNN) responsible for feature fusion was particularly affected by these limitations. For optimal performance, the framework ideally requires 20 to 40 training epochs to fully converge and generalize. Due to resource constraints, we could only train the model for 10 epochs, and even during this limited training period, issues such as insufficient storage and computational capacity often disrupted the process. Adjustments to batch sizes and other hyperparameters were necessary to accommodate the available resources, which likely impacted the FC-FNN's ability to learn complex interactions between the image and text modalities. Despite these challenges, the framework demonstrated promising results, suggesting that better computational resources could unlock its full potential.

Another limitation was the dimensionality of the combined latent space produced during feature fusion. While dimensionality reduction techniques such as Principal Component Analysis (PCA) were employed, the fused feature vectors still required significant resources for training. These challenges highlight the need for advanced hardware, such as high-memory GPUs or TPUs, for real-world applications of this framework.

7 Conclusion

In this work, we proposed a multi-modal deep learning framework that integrates radiology images and clinical notes for disease detection, emulating the holistic diagnostic approach used by medical professionals. By leveraging convolutional neural networks (EfficientNets) for image processing and Transformer-based models for text analysis, the framework effectively combines complementary information from both modalities to enhance diagnostic accuracy and robustness. Despite facing significant computational challenges, such as limited training epochs and resource constraints, the model demonstrated promising results, highlighting its potential to improve healthcare delivery. The integration of interpretability features further ensures that the model's predictions are transparent and trustworthy, making it a viable tool for clinical adoption. While this framework provides a strong foundation for multi-modal learning in healthcare, future efforts should focus on addressing computational limitations, improving scalability, and exploring additional modalities to expand its applicability. This work underscores the transformative potential of AI in advancing medical diagnostics and bridging gaps in healthcare access worldwide.

8 Work Distribution

In this project, each team member plays a crucial role in developing different aspects of the multi-modal deep learning framework, ensuring the integration of both radiology images and clinical reports to enhance disease detection.

Aditya is responsible for developing and optimizing the image branch of the framework. Their primary focus is on implementing a Convolutional Neural Network (CNN) to process and analyze the chest X-ray images. Aditya leveraged state-of-the-art pre-trained CNN models, fine-tuning them on the dataset to capture intricate patterns and features from the radiology images. Through careful application

of data augmentation, advanced image preprocessing techniques, and model tuning, he ensured that the image branch accurately extracts meaningful visual features that can contribute significantly to the overall predictive performance of the model.

Merna focused on building and refining the text branch of the framework. This branch processes clinical notes using a Transformer-based model (BERT - PubMedBERT), extracting relevant medical insights from the text data. Also, she fine-tuned the transformer model to handle the specific domain of clinical language, ensuring that the text branch can accurately interpret doctor notes and medical reports. By applying natural language processing (NLP) techniques, including tokenization, embedding, and advanced text preprocessing, Merna ensured that the textual information was captured and transformed into valuable features for the model.

Together, we collaborated to unify the feature representations extracted from both the image and text branches. This crucial step involves integrating the learned features from the CNN and the transformer model into a unified vector space, enabling the framework to fully leverage the complementary nature of the visual and textual data. We worked closely to refine the joint feature representation and ensure that the model can learn the complex correlations between the two modalities.

Finally, we both jointly evaluated the performance of the multi-modal framework, using rigorous testing and validation to assess how well the model performs in detecting diseases based on the combined inputs. We analyzed the results, fine-tune the system, and collaborated on interpreting the findings to optimize the model's accuracy and robustness. Through our combined efforts, we ensured that the framework can deliver meaningful, interpretable predictions that enhance diagnostic accuracy and contribute to personalized healthcare.

References

[1] Jason Adleberg, Amr Wardeh, Florence X. Doo, Brett Marinelli, Tessa S. Cook, David S. Mendelson, and Alexander Kagen. 2022. Predicting Patient Demographics From Chest Radiographs With Deep Learning. *Journal of the American College of Radiology* 19 (2022), 1151–1161. https://doi.org/10.1016/j.jacr.2022.06.008

[2] Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jeongwoo Oh, Lei Ji, Eric I-Chao Chang, Tackeun Kim, and Edward Choi. 2023. EHRXQA: A Multi-Modal Question Answering Dataset for Electronic Health Records with Chest X-ray Images. *Conference on Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks* (2023). https://github.com/baeseongsu/ehrxqa

[3] M Hasny, A Vasile, M Gianni, A Bannach-Brown, M Nasser, M Mackay, D Donovan, J Šorli, I Domocos, M Dulloo, et al. [n. d.]. A review of BERT models for systematic review screening in medicine. ([n. d.]).

[4] Francisco Herrera, Francisco Charte, Antonio J Rivera, María J Del Jesus, Francisco Herrera, Francisco Charte, Antonio J Rivera, and María J del Jesus. 2016. *Multilabel classification.* Springer.

[5] Sarah Jabbour, David Fouhey, Ella Kazerooni, Michael W. Sjoding, and Jenna Wiens. 2020. Exploiting and Preventing Shortcuts in Deep Learning Applied to Chest X-Rays. *Proceedings of Machine Learning Research* (2020). https://doi.org/10.1145/9876543.9876544

[6] Alistair E.W. Johnson, Tom J. Pollard, Seth J. Berkowitz, et al. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data* 6 (2019), 317. https://doi.org/10.1038/s41597-019-0322-0

[7] Firas Khader, Gustav Müller-Franzes, Tianci Wang, et al. 2023. Multimodal Deep Learning for Integrating Chest Radiographs and Clinical Parameters: A Case for Transformers. *Journal of Medical Imaging* 10, 3 (2023), 123–134. https://doi.org/10.1145/1234567.1234568

[8] Gang Liu and Jiabao Guo. 2019. Bidirectional LSTM with attention mechanism and convolutional layer for text classification. *Neurocomputing* 337 (2019), 325–338.

[9] Lisa Michielan, Lothar Terfloth, Johann Gasteiger, and Stefano Moro. 2009. Comparison of multilabel and single-label classification applied to the prediction of the isoform specificity of cytochrome p450 substrates. *Journal of chemical information and modeling* 49, 11 (2009), 2588–2605.

[10] Eduardo H P Pooch, Pedro L. Ballester, and Rodrigo C. Barros. 2020. The Impact of Domain Shift in Chest Radiograph Classification. *Machine Learning in Health Care* 8, 2 (2020), 200–215. https://doi.org/10.1145/7654321.7654322

[11] Alkulaib Lu Sarkar, Alhamadani. 2022. Predicting Depression and Anxiety on Reddit: a Multi-task Learning Approach. *Journal of Behavioral Data Science* 2, 1 (2022), 100–115. https://doi.org/10.1145/1234567.1234569

[12] Matthew Watson, Bashar Awwad Shiekh Hasan, and Noura Al Moubayed. 2023. Using Model Explanations to Guide Deep Learning Training in Medical Imaging. *Journal of Computational Medicine* 15, 1 (2023), 45–59. https://doi.org/10.1145/8765432.8765433

[13] Michael Yeung, Evis Sala, Carola-Bibiane Schönlieb, and Leonardo Rundo. 2022. Unified Focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Computerized Medical Imaging and Graphics* 95 (2022), 102026. https://doi.org/10.1016/j.compmedimag.2021.102026

[14] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. 2019. A review of recurrent neural networks: LSTM cells and network architectures. *Neural computation* 31, 7 (2019), 1235–1270.