Title: Empirical Evaluation of using MultiModal data to improve performance in Medical Imaging compared to single modal data

1. Abstract (500 words)

Background: 2 sentences, Research question: 1,Methodology:3, Key findings: 3, Implications: 1

Traditional machine learning (ML) systems in healthcare have usually adopted common approaches using statistical methods and applied in single datatype, such as image or text only. With recent advances in deep learning for image and transformer-base attention mechanism for text processing, combining multimodal data in medical has become feasible solutions which can utilise all advantages of data from multiple sources. In medical imaging, using Multimodal data in which each modality can capture different aspects of pathology might create a more comprehensive view of the anatomy and build up the clinical context for diagnosis. This mini project provides the empirical evaluation of using multimodal data to improve performance in Medical Imaging.

The dataset using in this project is from MIMIC CXR, provided by MIT lab, including Chest Xray images and the associated clinical text reports. There are 2 settings in this project, UniModal and MultiModal. Unimodal is using deep learning architecture with AutoEncoder for image modality and Multimodal is combining deep learning for image encoder, attention mechanism for text encoder and neural estimator for fusing method. The output of these settings, image embeddings or fused image and text embeddings would be used in the training of the classifier for each common pulmonary disease. The evaluation metrics are accuracy for classifier and separability for feature extraction, where embeddings are grouped in positive and negative classes and distance between these classes are measured. In general, for each disease, the accuracy for classifier in MultiModal is from 2 to 5% higher than UniModal and separability is 4% higher. For cross disease in MultiModal setting, Cardiomegaly has much higher accuracy than Pneumonia (76.85% compared to 63.29%), but its separability metric is lower than Pneumonia’s, around 15% (1.286 compared to 1.508) which might be a good starting point for follow-up research. With the same training epochs, deep learning architect for image encoders, and hyperparameters, the MultiModal settings can yield better result, more accuracy for prediction, and this provide a use-case to confirm the hypothesis of using MultiModal data can improve performance in Medical Imaging.

1. Introduction/ Background (3000 words)
   1. Literature review (2000)

**General approach for MultiModal**

MM in Machine Learning:<https://ieeexplore.ieee.org/abstract/document/8269806>

MM in Medical image fusion:

<https://www.sciencedirect.com/science/article/pii/S0010482522000452>

In HealthCare, multimodal architectures for AI systems are attractive since they share the similarity of using multiple data inputs from practitioners for diagnoses. Practitioners commonly use the combination of patient electronic health record in tabular format (e.g. age, demographics, history), medical images (e.g. XRAY, MRI, CT), time-series data (e.g. SpO2, blood pressure, ECG ), un-structured data (e.g. notes, reports, voice recordings) for disease diagnosis. Many research papers have performed evaluation and compared results between using unimodal and multimodal in HealthCare. Soenksen L., et al. (2022) proposed a framework to facilitate the generation and testing of AI system that leverage multimodal inputs from four data modalities (i.e., tabular, time-series, text, and image). Their findings have shown that the models trained with multimodal inputs can out-perform similar single-source approach (unimodal) across various healthcare demonstrations and settings. Similarly, Huang S.C, et al. (2020) conducted a systematic review and implementation guidelines of using fusion methods for medical imaging and electronic health records using deep learning. Their findings confirmed that multimodality fusion models generally led to increased accuracy (1.2–27.7%) and AUROC (0.02–0.16) over traditional single modality models for the same task.

**Common methods for feature extraction/dimension reduction(500):** CNN, Deep Learning for Image, Transformer for Text

ML in bio and health:

<https://link.springer.com/article/10.1186/s40537-025-01108-7#Sec4>

**Architecture for fusion methods (500):** early, intermediate and late fusion

* 1. Project justification (500)

**Rationale of the study**

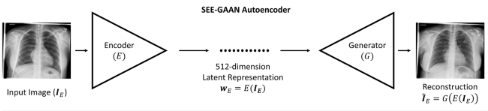
This mini project aims to conduct an empirical comparison of two different settings in which both unimodal and multimodal data are used with the existing deep learning architect in healthcare applications. The goal is to evaluate and compare how well two data settings will help to improve the prediction results from downstream classifier.

Figure 2 demonstrates the workflow of the comparison framework in which the metrics for visual explanation in two settings, unimodal data and multimodal data, will be calculated and compared to confirm the hypothesis that Metric2 for multimodal will be bigger than Metric1 for unimodal . The details of Dataset, Prediction module, Explanation module and Metric calculation will be provided in Additional Information section.

* 1. Research problem, hypothesis (500)

Additional modality will create more separated clusters for feature extraction (using un-supervised learning), and will help to increase accuracy in classifier (supervised learning)

1. Methodology (2000 words)
   1. Detailed (1000)
      1. Resnet
      2. AutoEncoder

****

* + 1. Bert transformer
    2. Mutual Information
    3. Multi Layer Perceptron (MLP)
  1. Dataset (200); MIMIC-CXR
  2. Metrics (300): accuracy for MLP classifier and separability for Resnet encoder

Benchmark for Image-Text in medical: <https://proceedings.neurips.cc/paper_files/paper/2024/file/0cb35e10bf7bb73d10c12414edbd63fd-Paper-Datasets_and_Benchmarks_Track.pdf>

* 1. Justified (500)

fair comparison between UniModal and MultiModal where they all use same architecture of ResNet CNN with 6 layers for image encoder and produce the same shape of embeddings for down stream classifier

1. Findings (2000 words)
   1. Presented (500)
   2. Interpreted (1000)
   3. Discussed (500)
      1. Dependency: dataset, network architect, hyperparameters
      2. Fair comparison for 2 settings and stable/reproducible benchmarks
2. Conclusion (500)  
   • summary, implications, further research, link to main PhD
3. References
4. Appendices