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

1. Abstract (250 words)

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

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 system would use Chest xray images and the associated clinical text reports to train the models, then compare the accuracy metric for each disease classifier in validation and test datasets with the result from model trained with Chest xray images only.

The methodology used in this project is quasi-experimental design where the dataset is created by selecting the frontal view scans from patients and ignore those with lateral views only. The entire dataset with 200.000 samples would be used to train model in unsupervised learning to find out the optimal representations. Those representations then are filtered by 4 common pulmonary diseases, e.g. Cardiomegaly, Pneumonia, Pulmonary Edema, Pleural Effusion with the minimum of 40000 samples for each disease, and they could be used to train the model for that disease classifier. In validation dataset, the accuracy for each disease classifier trained by MultiModal data is 2 to 3.2% higher than models trained by UniModal data while in test dataset for unseen samples, the accuracy from Multimodal trained models are 2.2 higher than Unimodal trained models.

Keywords: Multimodal data, representation learning

With the same training epochs, deep learning architect for image encoders, and hyperparameters, the MultiModal settings can yield better results, more accurate for prediction, and this provides a use-case to confirm the hypothesis of using MultiModal data can improve performance in Medical Imaging.

**Table of Content**

**List of tables and figures**

1. Introduction/ Background (2000 words)
   1. Introduction (500)

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.

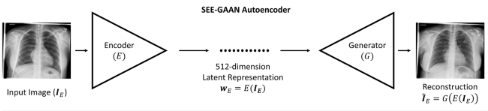
* 1. Background: (1500)
     1. **General approach and challenges for MultiModal in machine learning (500)**
     2. **Fusion methods (500)**
     3. **Representation learning (500)**

1. Project justification: (500)
   1. Justification: (300)
   2. Research problem, hypothesis (200)
2. Methodology (1000 words)
   1. Methodology: Quasi-experiments (400)
   2. Ethical considerations (300)
      1. Use public dataset:

Applying AI into medical data requires several considerations about ethical issues such as privacy, confidentiality, and bias inside the data itself. For this public dataset, MIMIC-CXR provided by MIT research group (add reference here), they both have de-identified clinical patient data and balanced distribution in term of patient gender, normal vs abnormal result, among ‘findings’ items (for local label with bounding boxes) and among ‘diagnosed’ items (for global labels as diseases), so there is no ethical issue when using these public datasets

* + 1. Follow provider constraints and ECU compliant: to store, access, process and share
  1. Significance and impact (300)
     1. Provide fair comparison between Multimodal and Unimodal
     2. Experiment in big dataset
     3. Create foundation works for PhD research about how to interpret/explain the multimodal data
     4. Provide code to public with end-to-end training for unsupervised learning (Mutual Information for image/text Encoders and Unimodal with AutoEncoder ) and supervised learning for image classifiers and metric generation for validation

1. Architecture details (2000)
   1. Unsupervised training for representation learning: (1200)
      1. UniModal: AutoEncoder (300)

****

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

1. Findings (1000 words)
   1. Presented (300)
   2. Interpreted (300)
   3. Discussed (400)
      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

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)

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

The purpose of this assignment is to:  
1. Present the research problem in a logical way in relation to the field of knowledge and clearly  
define the research questions.  
2. Justify the methodology and methods in relation to answering the research questions. This  
would include addressing briefly any ethical issues.  
3. Present and interpret the findings of your research, and place them within the context of other  
work, as well as your future PhD research.  
4. Consider the contribution your research will make to knowledge in your discipline and in a  
broader context - the potential impact beyond academia.  
5. Demonstrate your ability to: critically evaluate the literature; select, justify and apply appropriate  
approach to collect, produce and analyse data; organise, evaluate, interpret and present  
findings, in a logical, succinct, and cohesive manner.