Empirical evaluation of using multimodal data to improve explanability in Medical Imaging

Anh Tuan LE

10170661

**Context of research**

Applying Machine Learning (ML) in HealthCare has shown many remarkable outcomes, especially with Deep Learning, a large-scale Neural Network, in Medical Imaging. Biswas, M., et al., (2019) created a review on using Deep Learning in medical imaging in which the benchmarks for different medical fields equalled or exceeded human output, such as in Cardiovascular, Mammography, Gastroenterology, Neurology with accuracy 93.5%, 86%, 80.06% and 75%, respectively. However, its promising performance comes with a cost of transparency since it is extremely difficult to provide the explanation for its prediction and the understanding of its algorithm in layman term for non-technical users. The insufficient transparency can be problematic, especially in critical domains, such as healthcare, finance, criminal justice where all decisions must be justified and accepted among domain experts. This results to the need for eXplainable Artificial Intelligent (XAI) methods to make the models more transparent and interpretable and to address the concerns about trust, fairness, and accountability of AI system. XAI has become a popular research topic in AI field and the popularity of the search term “Explainable AI” has increased significantly in Google Trends since 2016 (Linardatos, P., et al., 2021).

XAI methods focus on explaining AI models in which they use single data type for their trainings such as text, image, numerical data. These methods are designed to target the specific characteristics and requirements of the datatype which the model is working with. For example, Saliency Maps is often used to highlight the most contributing and affecting areas to the model’s output for the image data, Feature Importance is used to indicate the importance of the feature for the classification in tabular data, to name a few (Bodria, F., et al., 2023).

**Relevant background**

In HealthCare, multimodal architectures for AI systems are attractive since they share the similarity of using multiple data inputs from practitioners for diagnoses. Practitioners normally 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., 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 led to increased accuracy (1.2–27.7%) and AUROC (0.02–0.16) over traditional single modality model for the same task.

Using Deep Learning in Medical Imaging has shown many outstanding results as mentioned in section above, and the need of having explanation for the decision is quite essential. However, there might be more than just one type of explanation depend on the audience or interest party of the imaging result. For example, for XRAY or Ultrasound scanning, there are several parties involved such as technician to do the scanning, radiographer to read the image, provide measurements and findings, and radiologist to read the image and conclude the findings with the absence or presence of the disease. Applying Deep Learning in end-to-end AI system can provide the disease classification from the raw image, but the explanations for different interested party can be varied. Radiographers are keen on the key areas in the image associated with the findings while radiologists might want the mapping between important visual evidence to prediction result. Furthermore, general practitioners might just want the reasonable textual explanation for the disease prediction.

The improved prediction performance from using multimodal data compared to unimodal (e.g. Huang, S., et al., (2020), Soenksen, L., et al., (2022)) and the need of generating different type of explanations confirm the demand of using multimodal data with XAI for HealthCare. This leads to my hypothesis that using multimodal data with XAI methods can increase the explanability in HealthCare.

**Rationale of the study**

This mini project aims to conduct an empirical comparison of explainability capabilities in two different settings in which both unimodal and multimodal data are used with the existing XAI methods in healthcare applications. The goal is to evaluate and compare how well two data settings will help to explain the prediction results.

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.

A diagram of an explanation

Description automatically generated

**Figure 2**: Project workflow with Explanation

**Research question**

How can multimodal data be used with XAI methods to increase explanability in Medical Imaging?

**Methodology**

The project is about XAI in Medical Imaging, so it is using machine learning workflow.

The adopted methodology is empirical review in which several stages (e.g. Data Selection, Model Selection) in the common machine learning workflow/pipeline have been selected through literature review, and other stages will perform the experiment as described in Figure 3.

A diagram of a model

Description automatically generated

**Figure 3:** Project Pipeline

There are three processes in this project as displayed in the flowchart:

* Data collection: this project is using open research style, a common approach in machine learning community to use public datasets that have been provided and published by well-known research groups and organisations. These public datasets have high quality data and ready for research purpose (e.g. using de-identified patient data for medical records, data has been normalised and balanced distribution, ensure noises and bias have been removed from data), so in this project, we can just select the proper datasets for medical images that suit our needs for training models.
  + The outputs of the Data Selection task are two datasets MS-CXR and VinDr-CXR which will be described more details in dataset section in Additional Information.
  + The task for Data Split to create training, validation and test sets from main datasets will be performed in the first week of the project.
* Learning process: the model is trained with training set and evaluated using validation set. There is an optional step to perform hyperparameters tuning if the performance is below some thresholds. Then the model will be tested with test set data and the accuracy metric will be calculated.
* Explanation process: the visual explanation will be generated (using one of the two methods Grad-CAM and SHAP) and metric for explanation will be calculated. Then the metrics for different settings will be collected as Table 1 below

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Settings | Accuracy | Explanation | IoU metric |
| MS-CXR | unimodal |  | GradCAM |  |
| MS-CXR | unimodal |  | SHAP |  |
| MS-CXR | multimodal |  | GradCAM |  |
| MS-CXR | multimodal |  | SHAP |  |
| VinDr-CXR | unimodal |  | GradCAM |  |
| VinDr-CXR | unimodal |  | SHAP |  |
| VinDr-CXR | multimodal |  | GradCAM |  |
| VinDr-CXR | multimodal |  | SHAP |  |

Table 1: Output results

**Relevant ethical issues and risks**

Applying AI into medical data requires several considerations about ethical issues such as privacy, confidentiality, and bias inside the data itself. For these two public datasets, MS-CXR and VinDr-CXR (which were published by two well-known research groups, Microsoft Research and VinGroup Big Data), 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.

There are two aspects that can be categories as risks of the project, such as data availability and computational resources. The datasets, MS-CXR and VinDr-CXR, which are publicly hosted and provided by MIT Lab, require account registration and data training before access is granted. They might take couple weeks to complete the entire process, and it imposes a risk for data availability for this mini project. Training the model for Neural Network (NN) (Convolutional NN for image classifier and Recurrent NN for text classifier) may require intensive computational resources. Although these resources are provided and shared by ECU AI Lab, there is still a risk of availability for resource allocation for this project. All those risks will be carefully monitored and controlled throughout entire 6-month duration of the project.

**Significance and Impact**

In the mini project, the model will be trained with medical datasets using unimodal (image only) and multimodal (image + text) settings and the metrics for visual explanations will be measured by the intersection over union (IoU) area between the heat maps generated by XAI methods and bounding boxes from human annotations. Comparing the IoU metrics in unimodal and multimodal settings will answer the question whether using multimodal can increase the explanability compared to unimodal . Also, comparing the metrics for explanations generated by Grad-CAM (model specific) and SHAP (model agnostic) can provide some findings about using different XAI methods, in this case model specific versus model agnostic.

Furthermore, the metrics for visual explanation in validation phase might be used to feed it back into the training phase so it can help to steer the training toward higher accuracy. The insights from this experiment might lead to some broader idea about increasing both accuracy and explanability using multimodal data. It will become foundation for my future research in broader topic “multimodal XAI to enhance both explainability and accuracy in HealthCare.”

The model for this project is using Neural Network architect (CNN for image classifier and RNN for text classifier) and fusion method to combine the results from two classifiers and based on the paper from Park, D., et al. (2018). The details will be explained more in Prediction module section in Additional Information. This project can be generalised with many additional modalities by replacing/adding the textual data and textual classifier by other data type su ch as tabular or time-series data and their associated classifiers. The insights from this project can guide further exploration studies using different architectures for image classifier, such as encoder or transformer.

**Timeline**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Task** |  | **Start from** | **Duration (weeks)** | **Details** |
| Data preparation |  | Week #1 | 1 | Create training/validation/test set |
| Test classifiers (originally designed for multimodal ) in unimodal settings |  | Week #2 | 2 | Validate and fine-tune multimodal classifiers for unimodal setting |
| Test visual explanation methods |  | Week #4 | 2 | Test Grad-CAM and SHAP in CXR data |
| Perform 8 tasks to populate metrics for Result Table |  | Week #6 | 12 | Generate visual explanation and calculate IoU metric for each task |
| Data analysis |  | Week #18 | 1 |  |
| Write report |  | Week #6 |  | Start drafting report from week#6 and keep updating with feedback from supervisors |
|  |  | **Total** | **18 weeks** |  |
|  |  |  |  |  |

**Additional information**

***Dataset:***

The data type for this project is Chest XRAY in DICOM format, and for each data point in the dataset, there will be a pair of DICOM image for the Chest XRAY and an associated free-text report including all findings and conclusion for disease.

|  |  |
| --- | --- |
| X-ray of a person's chest  Description automatically generated | A screenshot of a computer screen  Description automatically generated |
| DICOM image with findings and bounding boxes (VinDr-CXR, Nguyen H.Q, et al., 2022) | Associated report in free-text (MIMIC-CXR, Johnson, A. E., et al., 2019) |

**Figure 4**: CXR Dataset examples from VinDr-CXR and MIMIC-CXR

* + - MIMIC-CXR v2.0:[[1]](#footnote-2) Chest X-RAY with associated radiology reports, 227835 imaging studies for 64588 patients, each study contains at least 2 images for front and lateral view. CXR reports are semi-structured and normally has ‘findings and ‘impression’ sections which are for assessment details from radiologist and summary of the most pertinent findings, respectively. Findings section can be used in training text classifier while impression will be used as class label for image classifier.
    - MS-CXR:[[2]](#footnote-3) Subset Chest XRAY based on MIMIC-CXR v2.0 which contains 1162 image–sentence pairs of bounding boxes and corresponding phrases, collected across eight different cardiopulmonary radiological findings, with an equal number of pairs for each finding.
    - VinDr-CXR:[[3]](#footnote-4) 18000 postero-anterior view CXR scanned from 2 hospitals in Vietnam with annotations from 17 experienced radiologists (> 8 years of experience) for the presence of 22 critical findings (local labels with bounding box) and 6 diagnoses (global label).

***Prediction module:***

This project is using Neural Network architect (CNN for image classifier and RNN for text classifier) for the model and fusion method to combine the results from two classifiers and based on the paper from Park, D., et al. (2018).

A diagram of a diagram

Description automatically generated

**Figure 5:** Prediction and Explanation Models with multimodal data

(Park, D., et al., 2018, “multimodal explanations: Justifying decisions and pointing to the evidence”)

***Explanation module:***

XAI Methods for visual explanation:

* Gradient Class Activation Mapping (Grad-CAM) [[4]](#footnote-5)

Saliency methods are the most common techniques to create visual explanation which produce heatmap to highlight the area in the image that might influence to the prediction model. Saporta, A., et al. (2022) created a benchmark of using saliency methods in XAI for chest X-RAY, and among 7 common methods, Grad-CAM produced better results than other methods such as DeepLIFT, Integrated Gradients, Layer-Wise Relevance Propagation and Occlusion.

* Shapley Additive exPlanations (SHAP)[[5]](#footnote-6)

Lundberg et al. (2017) proposed a framework using Shapley values to provide explanation for the prediction by calculating individual feature contributions. It was model agnostic and can support both local and global categories. It has shown its high performance in tabular data, especially for feature importance, but still could be applied in images for super pixel groups, like in LIME technique.

***Metric for Visual Explanation:***

* Intersection over Union (IoU)[[6]](#footnote-7)

Intersection over Union, also known as Jaccard Index, is the most popular metric in machine learning for object detection benchmark in which the shape properties (e.g. width, heigh, location) of the two comparing objects are measured with overlapped area over the union area, and it can be computed as formular below:

A close-up of a logo

Description automatically generated

where A and B are the prediction and ground truth bounding boxes

(Rezatofighi H, et al., 2019)

|  |  |
| --- | --- |
| A black and green rectangle  Description automatically generated | X-ray of a person's chest  Description automatically generated |
| IoU example (Rezatofighi, H., et al., 2019) | Heat map with ground truth bounding box (Preechakul, K., et al., 2022) |

**Figure 6:** IoU example and Heatmap example for IoU calculation

References:

1. Biswas, M., Kuppili, V., et al., 2019. State-of-the-art review on deep learning in medical imaging. Front. Biosci. (Landmark Ed) 2019, 24(3), 380–406. https://doi.org/10.2741/4725
2. Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy (Basel, Switzerland), 23(1), 18. https://doi.org/10.3390/e23010018
3. Bodria, F., Giannotti, F., et al. Benchmarking and survey of explanation methods for black box models. Data Min Knowl Disc 37, 1719–1778 (2023). https://doi.org/10.1007/s10618-023-00933-9
4. Park, D., Hendricks, L., et al. (2018). multimodal explanations: Justifying decisions and pointing to the evidence. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 8779-8788).
5. Soenksen, L., Ma, Y., Zeng, C. et al. Integrated multimodal artificial intelligence framework for healthcare applications. npj Digit. Med. 5, 149 (2022). https://doi.org/10.1038/s41746-022-00689-4
6. Huang, S., Pareek, A., Seyyedi, S., Banerjee, I., & Lungren, M. P. (2020). Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. NPJ digital medicine, 3(1), 136.
7. Johnson, A. E., Pollard, T. J., et al. (2019). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data, 6(1), 317.
8. Boecking, B., Usuyama, N., et al. (2022, October). Making the most of text semantics to improve biomedical vision–language processing. In European conference on computer vision (pp. 1-21). Cham: Springer Nature Switzerland.
9. Nguyen, H. Q., Lam, K., Le, L. T., Pham, H. H., Tran, D. Q., Nguyen, D. B., ... & Vu, V. (2022). VinDr-CXR: An open dataset of chest X-rays with radiologist’s annotations. Scientific Data, 9(1), 429.
10. Saporta, A., Gui, X., et al. Benchmarking saliency methods for chest X-ray interpretation. Nat Mach Intell 4, 867–878 (2022).
11. Selvaraju, R. R., Cogswell, M., et al (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).
12. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30,
13. Rezatofighi, H., Tsoi, N., et al., (2019). Generalized Intersection over Union. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
14. Preechakul, K., Sriswasdi, S., et al. (2022). Improved image classification explainability with high accuracy heatmaps. iScience. 25. 103933. 10.1016/j.isci.2022.103933

1. https://physionet.org/content/mimic-cxr/2.0.0/ [↑](#footnote-ref-2)
2. https://physionet.org/content/ms-cxr/0.1/ [↑](#footnote-ref-3)
3. https://www.nature.com/articles/s41597-022-01498-w.pdf [↑](#footnote-ref-4)
4. https://arxiv.org/abs/1610.02391 [↑](#footnote-ref-5)
5. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html [↑](#footnote-ref-6)
6. https://giou.stanford.edu [↑](#footnote-ref-7)