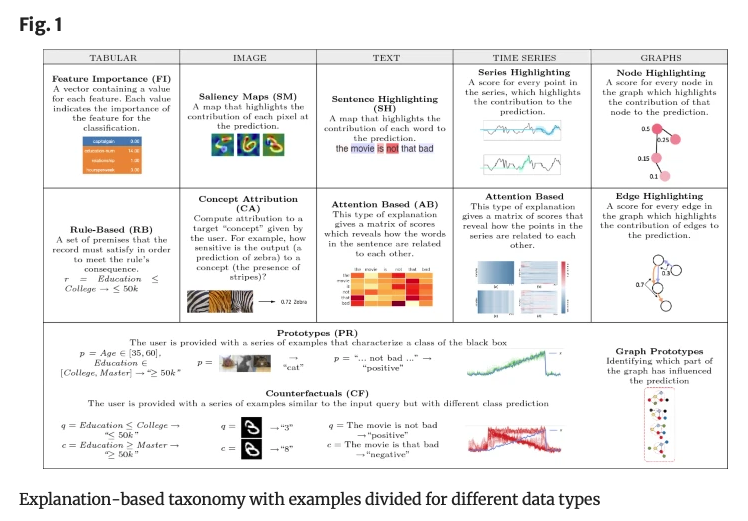
Title: Empirical Evaluation of using MultiModal data to improve explanability in Medical Imaging

**Context of research**: (the need of XAI --- > UniModal XAI --- > MultiModal XAI)

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 were 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 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 as well as 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)

UniModal XAI focuses on explaining AI models which are using single data type for their trainings such as text, image, numerical data. UniModal XAI 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) created a survey to provide a mapping of certain data type and black box model to a set of compatible explanation methods as shown in figure below.



MultiModal XAI is a newer approach that extends the principles and goals of UniModal XAI and provides the solutions for more complex problems which involve various input data and many different outputs for explanation. Adding language modality may provide more insights and valuable information to both model’s output and explanation for the decision than using the visual modality itself. Park D.H., et al. (2018) proposed an approach to provide joint textual rationale generation and attention visualisation from multimodal datasets and their findings concluded that the two modalities provide complementary explanatory strengths.

**Relevant background (**Healthcare domain:using multimodal for classifier + different interest parties for unimodal --> the need of multimodal XAI**)**

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.

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/sonographer to read the image and 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. Sonographers are keen on the important 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 performance from using MultiModal data to train the model (compared to UniModal) and the need of generating different explanations (for UniModal trained classifier) confirm the demand of using MultiModal XAI for HealthCare. This leads to my hypothesis that using MultiModal XAI can increase the explanability and interpretability in HealthCare.

**Rationale of the study**

This mini project aims to provide empirical review about how MultiModal data in XAI can be used in HealthCare/Medical Imaging, so it can confirm the hypothesis that using MultiModal XAI can increase the explanability and interpretability in HealthCare.

**Research question:**

How can MultiModal data be used to increase explanability and interpretability in Medical Imaging?

**Methodology:**

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 flowchart below

Data collection Learning process Explanation process

Data Split

Explanation generation

Model Testing

Model Evaluation

Model Selection

Metric calculation

Model Training

Data Selection

Collect

Grad-CAM method

Analysis

Pre- processing

SHAP method

The design details are as below:

1. Train model using datasets in 2 settings (UniModal and MultiModal) in which the training data can be image only (for UniModal) or image + text (for MultiModal)
2. Visual explanation will be created using either Grad-CAM (model specific for CNN based architect) or SHAP (model agnostic) .
3. The Intersection over Union metric would be measured for each task and analysed as table 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 |  |

Medical Datasets:

* + - [MIMIC-CXR v2.0](https://physionet.org/content/mimic-cxr/2.0.0/): 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.

A screenshot of a computer screen

Description automatically generated

* + - [MS-CXR](https://physionet.org/content/ms-cxr/0.1/): 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 approximately equal number of pairs for each finding.
    - [VinDr-CXR](https://www.nature.com/articles/s41597-022-01498-w.pdf): 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).

A table with numbers and letters

Description automatically generated

(Nguyen et al., 2022, “VinDr-CXR: An open dataset of chest X-rays…”)

XAI Methods for visual explanation:

* Gradient Class Activator Mapping (Grad-CAM)

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 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)

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.

Visual Explanation Metrics

|  |  |
| --- | --- |
| Metric | Details |
| [Intersection over Union](https://www.researchgate.net/profile/Hung-Nguyen-439/publication/362165633_Evaluation_of_Explainable_Artificial_Intelligence_SHAP_LIME_and_CAM/links/62d9b514f3acdd5dc20b92d1/Evaluation-of-Explainable-Artificial-Intelligence-SHAP-LIME-and-CAM.pdf) | Group segmented pixels and compare the area to bounding box |
| [Hit Rate](https://www.medrxiv.org/content/10.1101/2021.02.28.21252634v4.full) | Similar but looser metric which count 1 if heatmap overlap/contain bounding box |

**Relevant ethical issues and risks considered**

Applying AI into medical data requires several considerations about ethical issues such as privacy, confidentiality, and bias inside the data itself. For these 2 chosen 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)

There are two identified aspects that can be categories as risks of the project, such as data availability, computational resources. The datasets, MIMIC-CXR, MS-CXR and VinDr-CXR, which are publicly hosted and provided by MIT Lab, have high quality data but 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, but 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 saliency 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 vs model agnostic. Furthermore, the metrics for visual explanation in validation phase can 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 future research in broader topic “MultiModal XAI to enhance both explainability and accuracy in HealthCare”.

A diagram of a diagram

Description automatically generated

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

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

The findings from this project might provide some understandings of using MultiModal data in HealthCare toward some questions such as:

* Will bringing more data to train the model, i.e. medical findings in text, longitudinal pathology tests… increase or decrease the accuracy of classifier for Medical Imaging?
* Using the model which was trained using MultiModal data (image + text), if the input data has the image only, can the model generate the associated text for the findings with the same prediction accuracy?

**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 writing report from week#6 and keep updating with feedback from supervisors |
|  | **Total** | **18** |  |
|  |  |  |  |

**Additional information**:

For MultiModal data, both image classifier and text classifier would be used and fused by pooling method to create prediction.

Note: The pooling method is “Multimodal Compact Bilinear Pooling” referenced from the paper "[Multimodal Explanations: Justifying Decisions and Pointing to the Evidence](https://openaccess.thecvf.com/content_cvpr_2018/papers/Park_Multimodal_Explanations_Justifying_CVPR_2018_paper.pdf)".

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. H., Hendricks, L. A., 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.R., 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. C., 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). https://doi.org/10.1038/s42256-022-00536-x
11. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30, https://arxiv.org/pdf/1705.07874.pdf