Stage 2 Proposal

Interpretable Human-Centred Multimodal Learning Framework

Chihcheng Hsieh (Student ID: N10020322)

Submitted in partial fulfillment of the requirement for the degree of
IF49 Doctor of Philosophy

School Of Information Systems
Faculty of Science
Queensland University of Technology

Thursday 20th January, 2022

Contents

1 PROPOSED THESIS TITLE AND TYPE

Thesis Title: Interpretable Human-Centred Multimodal Learning Framework
Thesis Type: Traditional Thesis by Monograph

2 PROPOSED SUPERVISORY TEAM AND THEIR CREDENTIALS

Principle Supervisor: Dr. Catarina Moreira
Dr. Moreira is a senior lecturer in the School of Information Systems and holds the role of Deputy HDR Academic Lead. She is an associate editor for the Springer Nature journal BMC Bioinformatics, section "Machine learning and artificial intelligence". She is a Computer Scientist and passionate in investigating machine learning / deep learning models, and innovative human interactive probabilistic models for explainable AI. She has submitted a grant proposal on Persuasive and Causal Probabilistic Models for Explainable AI for ARC Discovery Early Career Researcher Award (DECRA). She is currently supervising two PhD students as the principal supervisor and three PhD students and one honours student as an associate supervisor.

Associate Supervisor: Dr. Chun Ouyang
Dr. Ouyang is a senior lecturer in the School of Information Systems and has supervised four PhD students to completion. She is an active and well-established researcher and the world top 21^{st} most cited scholar in Process Mining (according to Google Scholar). Built upon her expertise in process-oriented data mining, she has developed strong research interest in explainable predictive process analytics and submitted an ARC Discovery Project on the topic as the lead investigator. She is currently supervising four PhD students and one honours student as the principal supervisor and one PhD student as an associate supervisor.

Associate Supervisor: Prof. Margot Brereton
Prof. Margot Brereton is a national and international leader in the collaborative design of new humanitarian technologies and their interfaces. She designs with real user communities whose needs are typically overlooked in technology development. She has focused on better futures for and the agency of older people, neurodiverse people, minimally-verbal children with Autism, connecting people to nature, and fostering use of endangered indigenous languages.

3 BACKGROUND AND LITERATURE REVIEW

3.1 Introductory Statement

With the outstanding performance of deep learning models, artificial intelligent (AI) has started to be adopted and affect various industries including transportation, manufacturing, advertising, finance and medical industries. Deep learning is a subset of AI that mimics the human brain and its nervous system to learn to recognise patterns. Unlike the traditional machine learning, using deep learning not only provides a high-performance result, but it also frees people from monotonous and complicated feature engineering work. Since no specific feature engineering is required, it is the perfect choice for dealing with unstructured data such as images and text. Deep learning technologies have been to have high robustness and accuracy.

One practical field adopting AI approaches is chest X-ray (CXR) diagnosis. When patients are asked to take an x-ray, a static chest X-ray image will be developed for the radiologists to identify the diseases and injuries. Reading a CXR image can be time consuming.Even experienced radiologists can have diagnostic errors caused by fatigue or extrinsic distractions [109]. In addition, there is a shortage of radiologists in many areas of the world, including in developed countries. [99, 85]. A serious shortage of radiologists can increase the already heavy burden of workload for radiologists, which exacerbates the concerns about fatigue. An interactive AI-system can assist the radiologists to perform more efficient interpretation by providing the them with pre-identified suspicious areas and lesions. Since the abnormalities have been located, radiologists can take less effort to only work on confirming the predictions and finalising the report.

Since one of the abilities of neural networks is to process image data, an enormous number of studies have been conducted to explore automated diagnosis by AI-based predictive systems. Rajpurkar et al. [77] proposed CheXNet, which consists of 121 layers of Convolution Neural Network (CNN) to detect pneumonia through chest x-ray images. CheXNet achieved 0.425 F1 score, which outperforms practicing radiologist average. Furthermore, a report text can also be generated by only providing CXR image to the deep learning model. Liu et al. [62] combine CNN encoders, RNN decoders and reinforcement learning to generate a clinical report with human-level quality. According to the works mentioned, a computer-aided system can be built to make diagnoses and provide report text, which can be beneficial to radiologists.

Deep neural networks are usually non-linear and have a large number of parameters that allow them to fit complex functions. However, due to the non-linearity of the activation functions, current neural networks cannot provide humans with an explanation for the decision-making process, which leads to the concerns about reliability. In some cases, one wrong decision made by AI system can have dire consequences. For example, a mistaken decision to buy/sell stocks can result in the company losing millions of dollars in the market. And, if a medical AI system provides an incorrect prescription, it can cause permanent harm to patients.

Knowing the serious impact on inaccurate prediction of models, governments and organisations have announced some policies to manage and regulate the use of AI systems. These policies protect the right of data subject who can be significantly affected by the predictions from AI models. A well-known, recently published regulation is Article 22 of the General Data Protection Regulation (GDPR), which states that the end-user has the right not be involved in a decision that is only made by an AI system without human intervention. Additionally, the Articles 13 to 15 require the controller to provide meaningful information about the logic involved when the data is collected from the user [29].

Given the concerns and the announcement of regulations, a new strategy has to be developed to overcome the drawbacks of deep neural networks and meet the requirements of the regulations. In addition, if the AI system wants to be adopted broadly and be trustworthy, it is necessary for the users to understand how models make predictions. People are more likely to believe the prediction of black box models when it is ensured that the output is traceable and understandable. An experiment was conducted at a NIPS conference in 2018 that asked the audience to decide which surgeon to choose for the life-saving surgery on them. The first option was a human surgeon who could explain the details of the operation to you with an 85% success rate. On the other hand, the second option was a robot that could achieve a 98% success rate but did not provide explanations and did not answer any question. At the end, the robot received just 1 vote among hundreds of audience from different domains [87]. The above example highlights the importance of explanation for reliability. Considering the issues mentioned above, we need the approach of explainable artificial intelligence (XAI) to convince the end-user to trust the decisions made by AI system.

Explainable Artificial Intelligence (XAI) is a set of approaches to extract insights of the decision-making process from a black box model. The retrieved insight provides meaningful information for users to understand how a particular output from AI system is obtained. Several approaches have currently been developed to explain the computational decision of black box models. Approaches of XAI have been divided into two categories. The first one is to use *transparent* models, which can be contemplated by human at once (such as decision trees). However, using transparent models often associated with poor performance on complex datasets. Another XAI approach is post-hoc explanation, which interprets opaque models after-the-fact. The benefit of post-hoc methods is the interpretation for opaque models (such as Neural Networks) that allows us to gain both explainability and performance. Most of the research is studying post-hoc explanation models, which try to explain the mechanisms of the black box *after* it has been trained. Some common post-hoc approaches that are widely used are LIME [83], SHAP [66] and LINDA [70]. In terms of applications of post hoc models, Ahsan et al. [1] and Teixeira et al. [105] use LIME to interpret the results obtained from neural networks. In addition, due to the research on visualising trained CNN features, [119, 93] have shown promising results. A set of activation-based and gradient-based localisation approaches have been developed and frequently used with image data to provide visualisation explanations, which includes CAM [120], [91], GradCAM++ [15] and Integrated Gradient [98]. These gradient-based explanation methods are popular on explaining image data because the generated heatmaps can overlap on original images to show the salient pixels.

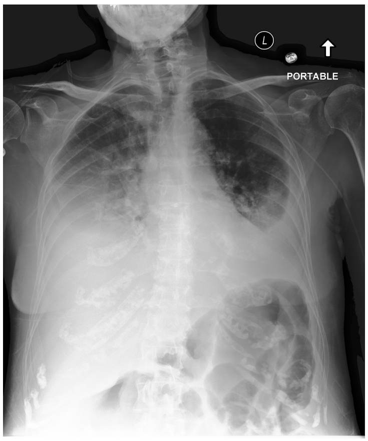
In the work of CheXNet [77] and CheXNeXt [76], Class Activation Mappings (CAMs) [120] are applied for generating heatmaps to show what are those important pixels leading to this particular diagnosis. Then they find that CAMs can successfully localise the pathology identified by the model. However, this approach to explaining the prediction can be inconsistent. Saporta et al. [89] conducted a human benchmark to evaluate current saliency-based methods, which includes GradCAM[91], GradCAM ++[15] and Integrated Gradient[98]. The evaluation result shows that the saliency methods perform significantly worse than professional radiologists. Therefore, before applying medical AI systems to the real-world cases, several problems need to be solved or mitigated to ensure reliability.

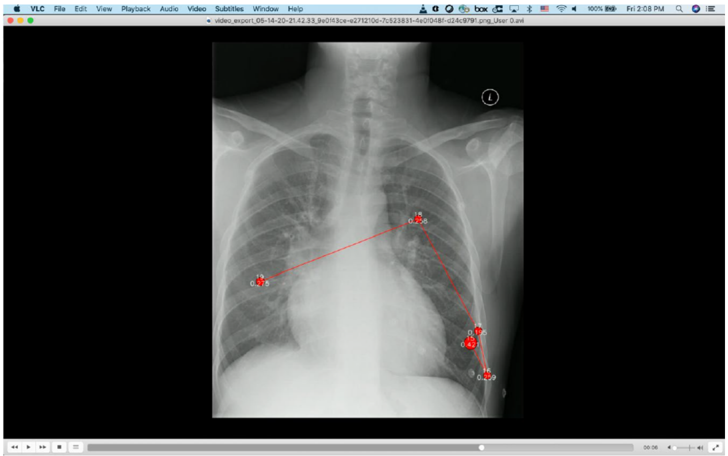
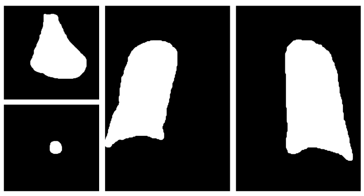
In order to close the gap in existing XAI methods in diagnosing through CXR images, we want to apply multi-modal learning [72]. Multi-modal learning is a learning strategy that includes specific architectures to allow the deep learning model to process the data from different modalities. For example, when using deep learning to make diagnoses, the model can process not only with the chest x-ray but also with the clinical information about patients simultaneously. Castillo et al. [13] found that the clinical information communicated to the radiologist can improve interpretation accuracy, clinical relevance and reporting confidence. Due to the positive influence of clinical information, it should be included as a model input to facilitate the decision making process. Instead of just providing CXR images to the model, other modalities, such as eye tracking data, report text, bounding boxes and clinical data can also be included, as shown in Table 1.

In this work, we will mainly use the data from 6 different datasets. MIMIC-IV [49] and MIMIC-ED [48] provide the clinical data about the patients. MIMIC-CXR [50] and MIMIC-CXR-JPG [51] contain the CXR images and the report texts from the hospital where patients took CXR. Eye gaze data [52] includes eye tracking data, segmentation maps, bounding boxes, dictation audios and their transcription.REFLACX dataset [9] offers another set of eye tracking information, anomaly locations ellipses, dictation audio and transcription.

There are three main reasons why different modalities should be included. Firstly, by getting more features, the model can find a decision boundary in a higher dimension, which can help the model to identify the class more easily and perform better. In other words, richer information can facilitate the classification process and boost the robustness of model. Secondly, we believe that human recognition patterns can be extracted from eye tracking fixation points. And, the fixation information itself or the extracted pattern can help us to understand the human decision-making process and improve the model in terms of performance and explainability. At the last, some of the modalities can serve as input or output for different purposes. For instance, Karargyris et al. [53] implement two experiments in their work. One of them uses fixation information as input to promote the performance of AUC. Another experiment in the same paper uses the same data as labels in the output section, which requires the model to predict where the radiologists will pay attention to. In general, the radiologist's gaze can help us to pinpoint the abnormalities, which explains the diagnosis. The overall framework of this work has been shown in Figure 1. The modalities outlined in dashed lines can serve as input or output. In addition, the four output modalities with a blue background have the potential to provide explanations. These are: eye tracking information, report text, audio & transcription and anomaly location ellipse.

This PhD aims to develop a multi-modal learning framework to assist radiologists in making a diagnosis. The modalities that can be included in the pre-training or training phase are chest x-ray images, report texts, eye tracking information, utterance of reports, segmentation images and bounding boxes. More importantly, this framework will provide an explanatory method to demonstrate how does the model obtain certain result, which can be helpful for both patients and clinicians. Therefore, the proposed framework is a computer-aided diagnosis system with truth-worthy explanation for its diagnosis.

Modality	Example
Chest X-ray [50] [51]	

Report text [50] [51]	FINAL REPORT AP CHEST, 9:27 A.M., ___ HISTORY: Severe C. difficile colitis. Aggressive volume resuscitation. Worsening tachypnea. IMPRESSION: AP chest compared to ___: Moderate to large right and moderate left pleural effusions have both increased in size. Upper lungs are clear. Heart is obscured by the effusions, but not substantially enlarged. No free subdiaphragmatic gas. Left PIC line ends in the left brachiocephalic vein.
Eye tracking data [52] [9]	
Segmentation of 4 key anatomical structures [52]	

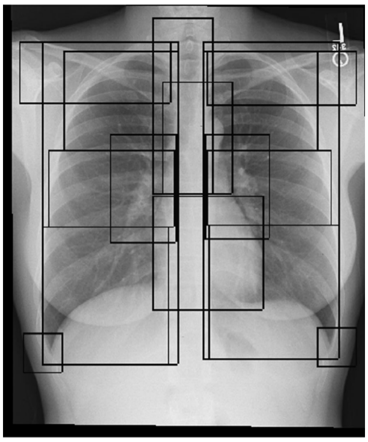
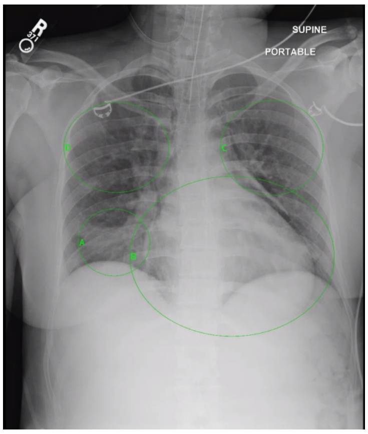
Bounding boxes of 17 anatomical structures [52]	
Anomaly location ellipses [9]	
Clinical Data [49] [48]	patients' age, gender, temperature, heat rate, respiratory rate, oxygen saturation, systolic/ diastolic blood pressure, pain level, etc.
Dictation audio [52]	The audio recorded when the radiologists is interpreting chest X-ray images.
Transcription [52] [9]	Recorded audio is automatically transcribed into a json file including word timestamp.

Table 1: **Existing tyoes of modalities for chest XRay images in the literature.**

3.2 Thesis Core Concepts

The associated basic concepts underlying this project are explained in this section to give the reader a clear understanding of this research:

- **Interpretability and explainability:**
  The difference between interpretability and explainability is subtle. When we assume that a model is interpretable, it often means that the model is a whitebox and can reasoning

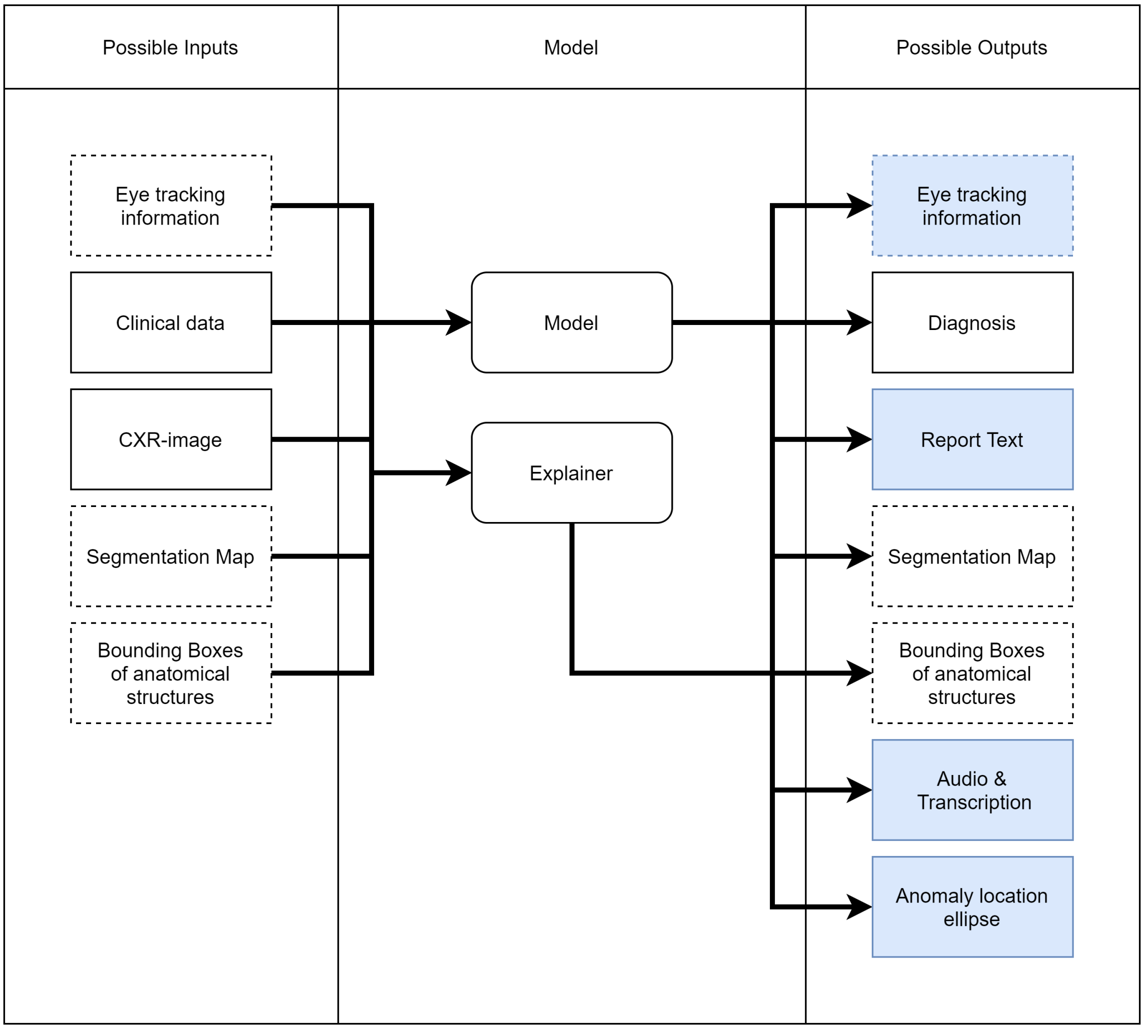


Figure 1: Overview of proposed framework.

its decisions. On the other hand, an explainable model is usually a blackbox model with an explanatory method to give insight into its decisions [24]. In most studies, however, these two words are used interchangeably. Hence, in this work, we will not emphasise the difference between them and consider them to be mutually replaceable.

- **Blackbox and whitebox models:**
  Machine learning models based on patterns, rules or decision trees are considered white-box models, such as linear regression and logistic regression. These algorithms make their decision-making process available to users in a human understandable way. On the other hand, in order to increase the capability of machine learning model, non-linearity and other complex mathematical functions have been added to the predictive algorithms, which eliminates the explainability [64]. Although blackbox models can fit more complex data and have higher performance in terms of accuracy, the loss of interpretation still raises reliability concerns and hold users back from using them in real-world cases.

- **Multi-modal learning:**
  Modality refers to the way something happens or is experienced. And the word, *sensory modalities*, stands for our primary communication and sensory channels such as seeing, hearing and touching. A research problem or a dataset is therefore referred to as multi-modal if it comprises several such modalities. Additionally, if a machine learning model has the ability to process related information from more than one modality, we consider it to be an implementation of multi-modal learning [5].

3.3 Literature Review

In order to gain a deep understanding of the existing work, a literature search is carried out. In this section, we categorise the literature into four groups including (1) Explainable AI (XAI), (2) Predictive Models for CXR Images (4) Multi-Modal Learning (5) Eye Tracking Technology.

3.3.1 Explainable AI (XAI)

Explainable AI (XAI) is a set of methods that enable people to understand and trust the decisions made by machine learning algorithms. An overview of the XAI approach was clearly presented in the work of Belle and Papantonis [8], shown as Figure 2. In XAI's research, the approaches can initially be divided into transparent and opaque models, which are referred to as whitebox or blackbox models. The transparent models consist of the algorithms that are capable of self-reasoning. These whitebox models reveal their decision-making process or their decision-making boundaries to users once they have been trained. Since transparent models can explain themselves, no additional explanation algorithm is required. On the flip side, the opaque models barely provide explanations for their decision while offering amazing performance in terms of accuracy. In order to achieve desirable predictive performance while being reliable, most research attempts to reveal how these blackbox models work. Post-hoc explainability is the method that is applied to a learned model without interfering the training process, which has the advantage of not adversely affecting the model's performance [25] [8].

The post-hoc explanations can be further separated to two categories, including *model-agnostic* and *model-specific*. The model-agnostic treat the original model as a black box and only require the input and output of the model to explain predictions. [82, 25, 84]. Some widely used model-agnostic methods are listed and introduced below, which includes LIME[83], SHAP[66], LINDA-BN[70] and ANCHORS[84]:

- **LIME:**
  Local Interpretable Modal-agnostic Explanation (LIME) is to firstly generate a set of perturbed samples with associated labels around the input. And an interpretable model, such as decision or logistic regression, will be trained on the permutation dataset to approximate the behaviour of original model in the local area. In other words, let f and x be the original black-box model and the input, respectively. Then LIME algorithm tries to find an interpretable model, g, to ensure $g(z^{'}) \approx f(h_x(z^{'}))$ where $z^{'} \approx x^{'}$. And h_x is a mapping function $x = h_x(x^{'})$. The feature importance can therefore be extracted from

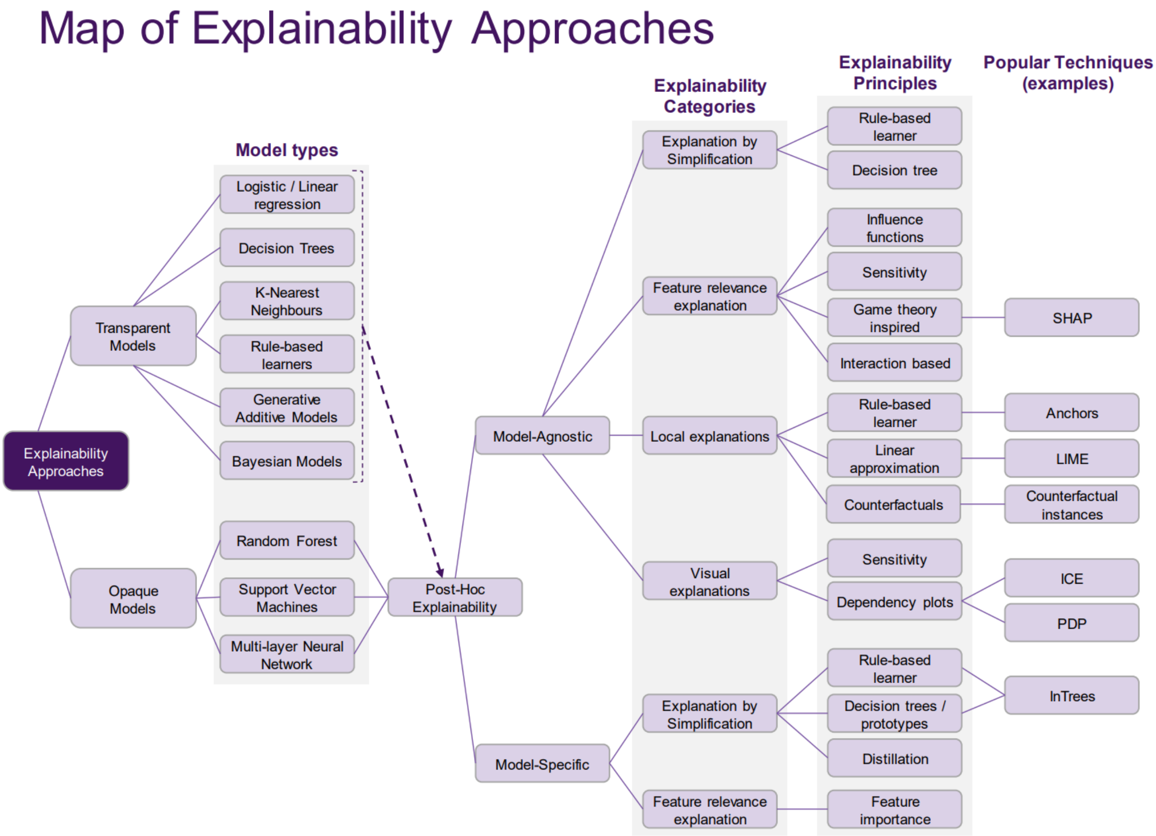


Figure 2: Overview of XAI. Belle and Papantonis [8]

the linear surrogate model to explain which important features lead to a certain decision, as the example shown in figure 3.

- **SHAP:**
  In 2017, Lundberg and Lee [66] unified 6 existing XAI methods that are using similar explanation method, including LIME, DeepLIFT, layer-wise relevance propagation, Shapley regression value, Shapley sampling value and Quantitative Input Influence. In that work, they proposed SHapley Additive exPlanations (SHAP) as a unified measure of the contribution of each feature. It uses the Shapley values from game theory and Economics [14] to assign each feature an importance value for a particular prediction.

- **LINDA-BN:**
  Local Interpretation-Driven Abstract Bayesian Networks (LINDA-BN) is a framework using the probabilistic graphical model to approximate the blackbox model in local area around input. Similar to LIME, local perturbed samples are generated with the predictions as label to form a permutation dataset. The permutation dataset will be then used to train a Bayesian Network (BN). The trained BN can explain how each feature contributes to the prediction, shown as Figure 4. Another advantage of this approach is that we can

reveal the correlations between input features due to the nature of the graphical model[70].

- **ANCHORS:**
  Ribeiro et al. [84] argue that the explainer with ability to only interpret locally is not ideal because the coverage of these explains is not clear. People tend to assume that the condition can be applied to other unseen instances, which is not always the case. The authors therefore proposed ANCHORS to fill the gap in the existing explainers. ANCHORS is an if-then rule-based algorithm to find the rule that *anchors* the prediction sufficiently locally, which means that changing other features that are not anchors does not affect the prediction result. It can explain the behaviour of complex models with highly precise rules. The most obvious benefit is the clear coverage that is applied when the anchors are untouched.

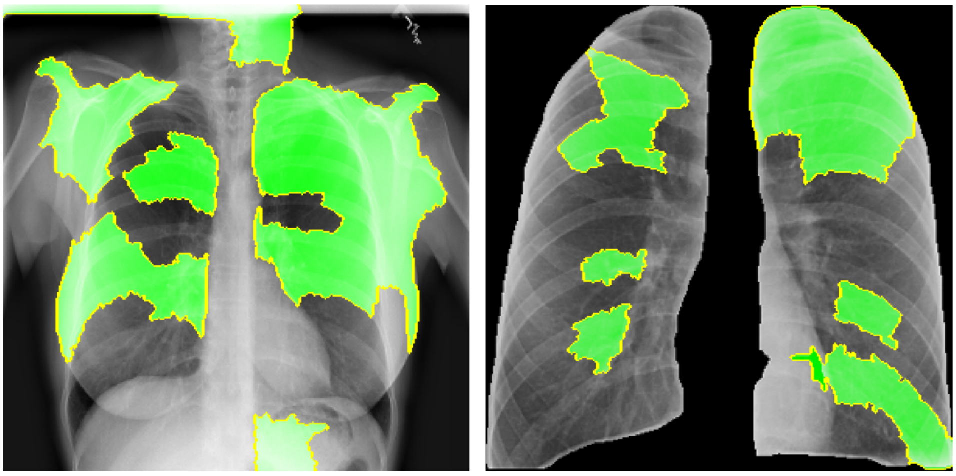


Figure 3: LIME example on Chest X-ray images. [105]

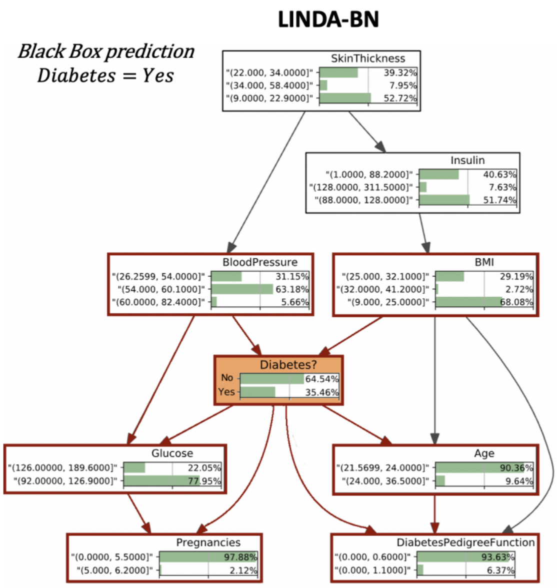


Figure 4: LINDA-BN explains why a patient is predicted not having diabetes. [70]

On the other hand, model-specific approaches are those that require specific architectures of the model in order to generate explanations. The main advantage of model specific methods is that it makes it easier to develop an efficient algorithm based on the explanations we observed [8]. Two examples of post-hoc model-specific algorithms are DeepLIFT [92] and TreeShap [65].

- **DeepLIFT:**
  Deep Learning Important FeaTure (DeepLIFT) [92] uses back-propagation to decompose the output of the model and reveal the contribution of each input feature. In addition, it discovers the dependencies between features through considering positive and negative contributions separately. In order to address the problem of saturation, the **summation-to-delta** property (Eq. 1) is proposed, where $C_{\Delta x_i \Delta t}$ is the amount of difference in output attributed to the difference of x_i.

- **TreeSHAP:**
  SHAP [66] uses unique additive features attribution method based on conditional expectation to measure the contribution of each feature. However, the complexity of calculating the SHAP value lies in exponential time, which can be computationally expensive as the number of features increases. Lundberg and Lee [65] therefore proposed Tree SHAP to estimate the SHAP value in polynomial time but only for trees ensembles of tree models. The idea of Tree SHAP is to recursively keep track of what proportion of all possible subsets flow down into each of the leaves of the tree.

At the last, we will list two intrinsic algorithms. When a model is considered intrinsically explainable, it usually means the model's explainability is embedded in the model architecture itself. Since the explainability depends on the architecture, they are model-specific and generally cannot be transferred to other models with different architectures [21]. There are two intrinsic algorithms widely used in computer vision task [95], including GradCAM [91] and Integrated Gradients [98].

- **Grad-CAM**:
  Gradient-weighted Class Activation Mapping (Grad-CAM) is an explainer designed specific for Convolutional Neural Networks (CNNs). It generalises CAM [120] to make it available on variety of CNN-based architectures. GradCAM uses back-propagation to track the gradient in the last CNN layer to visualise the semantics the model captured [91], shown as Figure 5. Another work, Grad-CAM++ [15], further improved GradCAM by taking into account the penultimate layer of CNN and updating the global average pooling algorithm. This improvement solve 2 issues on GradCAM, including (1) only part of the object can be localised and (2) object with multiple occurrences can't be localised properly.

- **Integrated Gradients:**
  Sundararajan et al. [98] proposed two axioms that should be satisfied by every explanatory algorithm, including *sensitivity* and *Implementation of invariance*. The axiom of sensitivity can be satisfied if an function depend on some variables, then the attribution to those variables are always zero, and vice versa. And the axiom of implementation invariance is

satisfied when the outputs of two models are equal for all inputs, regardless the difference in implementation. *Integrated gradient* is inspired by these two axioms. The most notable advantage of the integrated gradient is that it does not require a specific model architecture as long as the model is a differentiable function. Integrated gradients are defined as the path integral of the gradients along the straight line path from the baseline to the input.

$$\sum_{i=1}^{n} C_{\Delta x_i \Delta t} = \Delta t \tag{1}$$

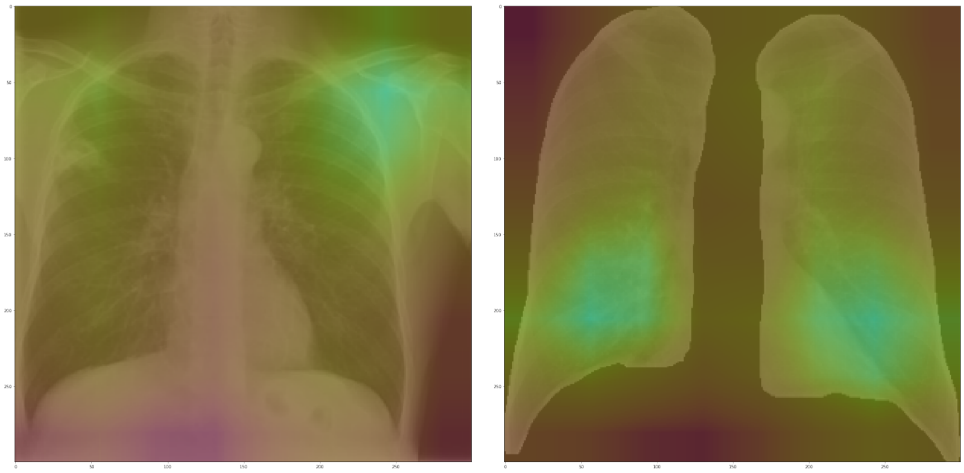


Figure 5: GradCAM example on Chest X-ray images. [105]

3.3.2 Predictive Models for CXR Images

With the rapid development of AI, the medical industry is beginning to explore the potential of machine learning to aid diagnostic process. Chest X-ray is one of the safest and most common methods of diagnosing patients; hence, the Computer-Aided Diagnosis (CAD) of CXR images is a popular field that people are investing in. Before discussing the performance, the evaluation matrix should be defined. In the normal image classification tasks, accuracy is preferred to evaluate learning algorithms since [63] have proven the ineffectiveness of AUC in some cases. On the other hand, the area under ROC (receiver operating characteristic) curve, as known as AUC, is used in medical diagnosis since 1970s [46]. The medical diagnosis requires the model to have high recall and precision in order to minimise the amount of false-negative cases, which leads to several consequences. To mitigate patient's risk, most medical predictive algorithms are evaluated by AUC [36]. Due to the different evaluation methods, the loss functions are adapted in medical fields, such as pairwise surrogate loss [33], pairwise square loss [32]and Deep AUC Maximisation (DAM) [97]. In 2020, Yuan et al. [116] proposed a new margin-based min-max surrogate loss function, which surpasses other models on the CheXpert[47] till now.

Deep learning-based approaches have been found to be efficient and effective on a variety of tasks, especially on vision-related tasks [94]. In medical field, Ting et al. [106] proposed a

CNN-based model that is able to detect diabetic retinopathy and related eye diseases by retinal images. Esteva et al. [28] shows that deep learning can reach dermatologist-level accuracy in terms of skin cancer detection. Peng et al. [75] developed a two step framework, *DeepSeeNet*, to predict the severity of age-related macular degeneration (AMD).

In some cases, the AI system can achieve or even surpass expert-level performance [77, 116, 104]. Since automated chest x-ray diagnosis can be considered as an image classification task, architectures and techniques from other fields that also process image data can be reused in the medical domain. Transfer learning is a popular and effective way to improve the generalization of the model [121]. In general, the size of available CXR datasets are smaller than existing image datasets, such as MS-COCO [61], ImageNet [23] and OpenImage [57]. For example, Yuan et al. [116] and Cohen et al. [17] use DenseNet[45] architecture. And, Dunnmon et al. [27] use AlexNet[56], ResNet [38] and DenseNet pre-trained on ImageNet database [23]. And Yates et al. [115] use pre-trained inception_v3 [102] from Google. By adopting transfer learning, those works gain noticeable improvement in performance. Despite reaching SOTA (state of the art) performance, these works rarely provide the explanations about their models predictions, which causes the concern of reliability and holds radiologists back from adopting CAD systems.

From	Modalities used	Dataset	Result
Yuan et al. [116]	CXR image	CheXpert[47]	0.9305 (AUC)
Tang et al. [104]	CXR image	ChestX-ray14[111]	0.9824 (AUC)
Cohen et al. [17]	CXR image	COVID-19 Image Data Collection[18]	0.87 (MSE)
Dunnmon et al. [27]	CXR images	Private (Not availabel)	0.96 (AUC)
Yates et al. [115]	CXR image	ChestX-ray14 [111]	0.98 (AUC)
Rajpurkar et al. [77]	CXR image	ChestX-ray14 [111]	0.435 (F1)
Rajpurkar et al. [76]	CXR image	ChestX-ray8 [111]	0.704 - 0.944 (AUC)
Brunese et al. [11]	CXR image	COVID-19 Image Data Collection[18]	0.99 (Accuracy)
Hou and Gao [42]	CXR-image	COVID-19 Image Data Collection[18]	0.95 - 0.97 (F1)
Rodin et al. [86]	Clinical Data & CXR image	MIMIC-CXR[50]	0.648 (BLUE-1)
Teixeira et al. [105]	CXR image & Segmentation image	RYDLS-20-v2[105]	0.88 (F1)

Table 2: **Existing predictive models on CXR images. (Note: more works using CheXpert dataset can be found in their leaderboard: https://stanfordmlgroup.github.io/competitions/chexpert/)**

In the medical field, most works with image data use CAM [120], GradCAM[91] or other similar algorithms that generate saliency maps to explain diagnosis [105, 76, 77, 11, 42]. However, saliency map can be unreliable in some cases. Recently, explainable methods that visualises saliency maps have been evaluated in the work of [89]. And the authors demonstrated that saliency maps consistently perform worse than human experts, regardless of the performance (AUC) of the model.

3.3.3 Multi-Modal Learning

The complementary information about a phenomenon can be obtained from various types of detectors or sensors. The term "*modality*" is then used to describe the collected information from each sensor. When several different instruments are set up to observe a phenomenon, the information collected is known as *multi-modal* data [58, 78]. As sensors and detectors become more accessible these days, deep learning can benefit from multi-modal data in terms of generalization and performance compared to the uni-modal paradigm [112]. For example, in research on autonomous driving, [30] state that most of the research using deep learning and multi-modal learning outperform other methods.

In order to form a joint representation to describe a phenomenon, a fusion operation is needed to combine the information from different modalities. Before deep learning become ubiquitous, **early fusion** and **late fusion** are the most commonly used fusion strategies. However, both of these methods require human to design the feature engineering process. Deep learning technology is therefore used to release the humans from complicated feature engineering task. Applying deep learning approaches also allows us to perform the **intermediate fusion**, which merges the representation from different modalities in any layer of the model. The fusion strategy for deep learning is another hot topic. Hori et al. [41] proposed an attention-based fusion operation that can selectively use features from different modalities. Vielzeuf et al. [108] designed CentralNet that accumulates the learned weighted sum of different modalities at each layer. At the end, the weighted sum will be used for a classifier to make final prediction. Yang et al. [114] developed Correlational Recurrent Neural Network (CorrRNN) that can handle data with temporal structure. This model consider the current state and history of each modality to form a joint representation for every timestamp. Beside combining two channels to generate a universal representation, [113] proposed another channel exchange method, which blends the information between modalities.

As the number of multi-modal datasets grows, more and more architectures and strategies are developed. Ngiam et al. [72] proposed multi-modal learning task using restricted Boltzmann machine (RBM) to process audio and video data, which shows that deep learning is capable of discovering multi-modal features. Hu et al. [43] developed an unsupervised method called Deep Multi-modal Clustering (DMC). This unsupervised learning technique achieve or even outperform human in the tasks of localisation, multi-source detection and audiovisual understanding. Hu et al. [44] designed a Scalable Deep MultiModal Learning (SDML) framework. It trains models to map the inputs of each modality to a predefined subspace, which makes the framework scalable. When introducing a new modality to this framework, instead of the whole framework, only a new network for this specific modality has to be trained. Mroueh et al. [71] developed an approach to fuse speech and visual modalities for Audio-Visual Automatic Speech Recognition (AV-ASR). The authors also showed that using another modality in addition to the model can noticeably improve classification performance. Tang et al. [103] used two image modalities generated from functional magnetic resonance imaging (fMRI) brain scans to construct an end-to-end automated autism classifier that outperforms other models that only use single modality.

In the medical field, two articles used clinical information along with x-ray images to explore the benefits of multi-modal learning. Rodin et al. [86] designed a deep learning model that use clinical data and CXR image from MIMIC-CXR[50] to generate short reports. However, instead of retrieving the patients data from correspond MIMIC-IV data tables, this paper generates the age, gender and anamnesis from *History* section of the text report by radiologists, which provides less and incomplete information compared to the tabular data. Furthermore, [4] used a private dataset from Mount Sinai Health System (MSHS) to classify hips fractures. The authors conduct several experiments to observe the performance when incorporating various clinical data. As expected, the AUC increase as more features are included. Another interesting finding of this work is that the distributions of X-ray images from different scanners are distinguishable, which implies the generalization problem.

3.3.4 Eye Tracking Technology

Eye tracking technology has been studied since a long time ago. In 1950, researchers studied pilots' eye movements through a head-mounted eye-tracking device. The findings of this research are eventually used to redesigning the cockpit, which improves usability and reduce the risk caused by human error [31, 19].

Today, most eye tracking technologies use corneal reflection, which uses a light source to illuminate the eye. This action creates a little glint under the pupil that the device can use to calibrate and track eye movement [90, 31].

Blascheck et al. [10] illustrates that A temporal order of *fixation* within an Area Of interest (AOI) can be viewed as a *gaze*. The duration of the fixation is usually represented by the radius. *Saccades* are the connections between fixations. When a saccade connects two AOIs it's considered as *transition*. Finally, *scanpath* consists of a complete sequence of fixations and saccades. The relationships between these terms are also shown in Figure 6. The authors also use eye tracking data to generate human attention map, which visualises human's interest on certain scene. And, This type of applications can assist UI workers to re-design and improve user experience [90].

When we're trying to capture where's users' attention on, it's unreliable to let the users describe the whole story by themselves. In the work of Guan et al. [35], the author demonstrated that human can reach 47% of omissions in think-aloud experiment, which means that human can't self-report precisely. Eye tracking technology is one of the most effective methods to track human attention as the eye movement and human cognition are inextricably linked. Schall and Romano Bergstrom [90] stated that eye tracking technology enables us to capture cognitive processes and visual perception. And it can potentially be used to observe people's emotions and intentions [37]. Rayner and Reingold [79] demonstrated the evidence that direct cognitive control of eye fixations in reading. Another prove that eye tracking is closely related to cognitive is that eye tracking allows us to identify disease with cognitive disorders [37], such as Parkinson's disease [96], Alzheimer's disease [20], Schizophrenia [60] and Autism [22].

Eye tracking is also a popular tool used in medical industry. Manning et al. [67] and Giovinco et al. [34] used eye tracking technology to show that radiologists with different levels of experience have noticeable difference when it comes to scanning and interpreting strategies. Since the eye tracking information can indicate the interpretation performance, McLaughlin et al. [69] proposed an algorithm to evaluate interpretation technique of the radiologists, which also supplies a valuable insight into the interpretation process. Karargyris et al. [53] implemented two experiments to demonstrate how can eye tracking information benefit deep learning based predictive system. It shows that eye tracking data can be considered as both input or label data for the model. When eye tracking data is treated as the input of the model, it can increase the accuracy of the model since the model receives more information from radiologists. When eye tracking data is considered as as the output of the model, it improves the generalisation and provides explanations. Recently, [88] also proposed a similar algorithm that uses eye tracking data as labels in a multli-task learning framework. In addition, the authors showed the gaze features from normal and abnormal CXR images differ noticeably. A set of weak labels can therefore be extracted from eye tracking data for supervised-learning tasks.

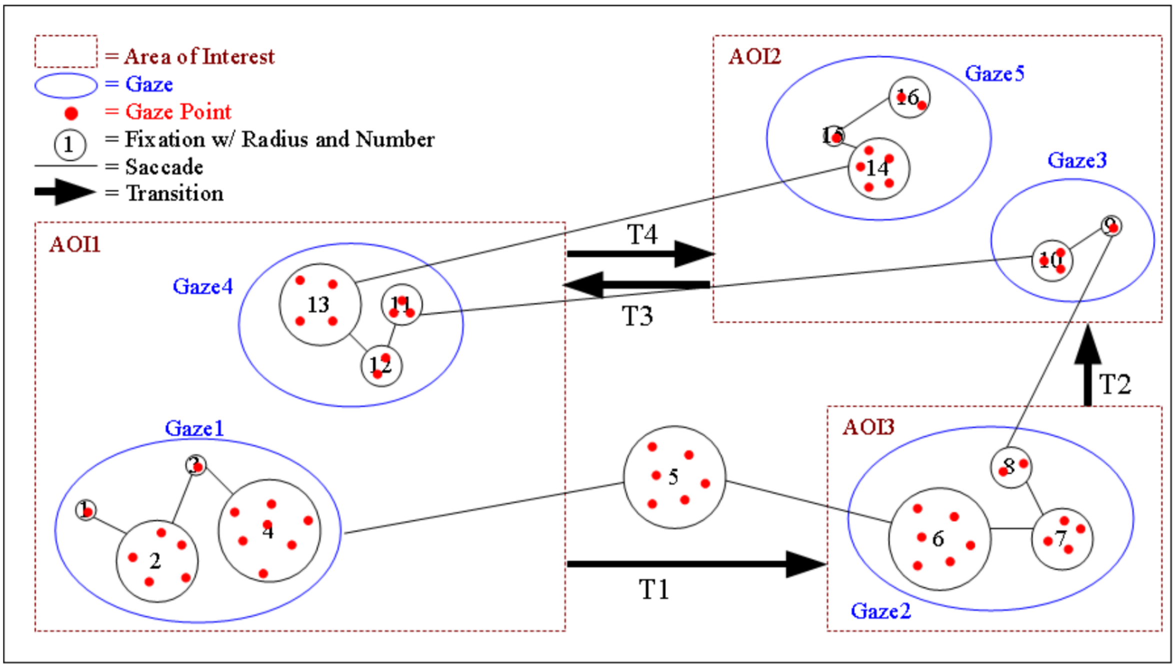


Figure 6: Terms used in eye tracking technology. [10]

3.3.5 Cognitive Load Theory

The cognitive load theory is first presented by Sweller [101] to optimise the problem solving and learning process by reducing the cognitive load. The theory is based on cognitive theory to assume that human has a limited working memory capacity. When a problem is more complicated, we can expect a greater cognitive load during solving this problem, and vice versa. When the problem a large number of items to be stored in short-term memory, an excessive cognitive load may occur to hinder the cognitive-process. When cognitive load theory is applied to foster learning process, the researchers aim to find a strategy to allow learners allocate their cognitive resource efficiently and prevent overload through improving the instructional format. In other words, if the instrumental format can consume less cognitive resources, more capacity can be used for actually learning to achieve a better learning outcome [6].

The ability of reading Chest x-ray images and giving correct interpretation is an essential problem solving skill for radiologists. By measuring the cognitive load and tracking the changes, we can understand how each factor, such as fatigue or difficulty, affects the radiologists' diagnostic process. In order to measure the cognitive load, there are various approaches, including functional near-infrared spectroscopy (fNIRS) [117], heart rate variability (HRV) [68] and electroencephalography [3]. However, in order to properly record their cognitive load, the eye tracking technology is a preferred approach because it does not need to attach devices to radiologists that may interfere and distract them when they are reading the chest x-ray image. While some studies suggest that the link between cognition and the eye is not always correct [2, 90], it's generally true that people tend to move their eyes to the stimulus that they are currently thinking about or processing [12]. And the behaviour that cognition guides the eye movement is called eye-mind link [80, 81]

Hess and Polt [39] first discovered that the pupil dilation is strongly correlated with task difficulty. When a task is more difficult, the pupillary size tends to increase. [7] also showed that Task-Evoked Pupillary Responses (TEPRs) are also related to brain activity. In general, the cognitive load is estimated through the pupil diameters changes compared to the baseline in calibration. However, this approach has two serious disadvantages. Firstly, the pupil is also sensitive to the illumination, which can vary in different environments. Another disadvantage is that the measurement of pupillary size can be affected by the head movement and optical distortions. These two factors can lead to an erroneous measurement [26]. Therefore, Index of Cognitive Activity (ICA)[16] and Index of Pupillary Activity (IPA) [26] are proposed. Instead of measuring the size of pupils compared to baseline, these two methods measure the rate of change in pupil diameter, which alleviates the problems mentioned above.

3.4 Research Problem and Research Questions

Scope. This work focuses on the creation of an interpretable human-centred multi-modal learning framework, which can receive chest X-ray images, clinical data and eye tracking data as input to make a diagnose. Figure 7 shows the 3 main areas of this work. Our analysis will be based on three public datasets [49, 50, 48, 9, 52]. And only two diseases are considered in this work, including Congestive Heart Failure (CHF) and Pneumonia.

Problem Statement. In a practical case, the radiologist commonly use clinical data to assist their process of diagnosis. Recent literature also suggests that clinical data can benefit the accuracy of diagnosis [13, 59]. However, most of the existing works in medical AI only use medical images to make a diagnosis. Focusing only on this pixel-level classification can lead to bias,

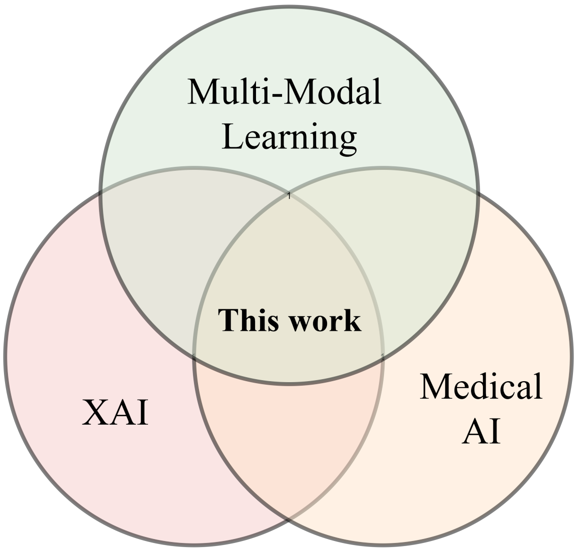


Figure 7: Venn diagram of this work.

which does not precisely substantiate the diagnosis by radiologists. Ignoring the bias in an AI model can increase the risk for minority groups in terms of discrimination and poor diagnosis. Motivated by this real-world need we identified the following research gaps, which will lead to the proposed research goals.

- Research Gap 1. **There is a lack of consideration of using clinical and eye tracking data for diagnosis prediction for CXR.** Recent literature suggests that using accurate patients' *clinical data* together with medical images can improve the performance of human radiologist in terms of accuracy. [13, 59]. Additionally, there are studies suggesting that *eye tracking data* from radiologists can shed light on how radiologists interpret Chest X-ray images and obtain diagnoses [53, 9]. However, to the best of our knowledge, no work in the existing literature has attempted to propose predictive architectures using these two modalities.
  - **Research Question 1:** What are the multi-modal architectures that are required to support different types of data, such as eye tracking data (time-series data), chest X-ray (image data) and clinical data (tabular data), for predictive diagnosis?
- Research Gap 2. **There is a lack of support to human-centred interpretability of predictive diagnosis for CXR images using eye tracking data.** There is a strong body of literature suggesting a correlation between fixation points from eye tracking data and cognitive load (such as the degree of fatigue and how hard the radiologist task is) [73, 118, 110, 55, 54]. Hence, investigating on eye tracking data can provide insights of how difficult is the radiologists tasks in multi-modal predictive architectures, leading to more human-centred interpretable frameworks.
  - **Research Question 2:** What form of eye tracking representation be used to promote the radiologists' understandability and interpretability of CXR predictions?

- **Research Question 3:** What features can be extracted from eye tracking data to measure the radiologists' cognitive load? For instance, how one can extract the degree of fatigue of the radiologists, and measure the task difficulty in CXR diagnosis?
- **Research Question 4:** How can cognitive load features be used in a multi-modal architecture to identify bias due to fatigue or task difficulty and promote human-centred learning frameworks?

4 PROGRAM AND DESIGN OF THE RESEARCH INVESTIGATION

4.1 Objectives, Methodology and Research Plan

4.1.1 Objectives

The main objective of this thesis is to propose an interpretable human-centric multi-modal predictive architecture for CXR diagnosis. It is based on the hypothesis that introducing richer information to the model, such as eye tracking data and clinical data, can be used to extract human classification patterns to improve the explainability and the accuracy of multi-modal predictive architectures.

AIM 1 (RQ 1 & RQ 2):
Gain better understanding of radiologist classification patterns of CXR images through eye tracking data that could be used for predictive diagnosis for CXR images.

AIM 2 (RQ 1 & RQ 2):
Design a multi-modal learning architecture that uses eye tracking data to capture human classification patterns in the prediction of CXR images together with patient's clinical data

AIM 3 (RQ 3 & RQ 4):
Extend the multimodal learning architecture to incorporate cognitive load features to support interpretability and human-centered explanations.

AIM 4 (RQ 1 & RQ 2 & RQ 3 & RQ 4):
Build an interface to assess the proposed interpretable human-centric multi-modal architecture and evaluate it with experienced radiologists from *Hospital Group Lusíadas*, Portugal.

4.1.2 Methodology

This research will follow the guidance from Design Science Research Methodology (DSRM) [40]. With DSRM, 7 guidelines can be listed specifically for this research, including (1) Designing the artifact of this project, which is the explainable medical AI framework. (2) The objective of this framework is to better diagnosis system in terms of performance and explainability. (3)

The result of this framework will be evaluated in 2 phases. The performance of the predictive model will firstly be evaluated by AUC. And the explanation for the prediction will be evaluated by a professional human radiologist in Portugal (4) The novelty of this work is the exploration of using different modalities during diagnosis, which may result a more robust and explainable framework. (5) The framework itself must be rigorously defined, formally represented, coherent, and internally consistent whether in developing or evaluating phase. (6) the framework should be designed in an iterative and effective manner. (7) The result of this framework should be presented to both technology-oriented and management-oriented audiences effectively. Moreover, according to the DSRM model [74] proposed, our process model is shown as Figure 8.

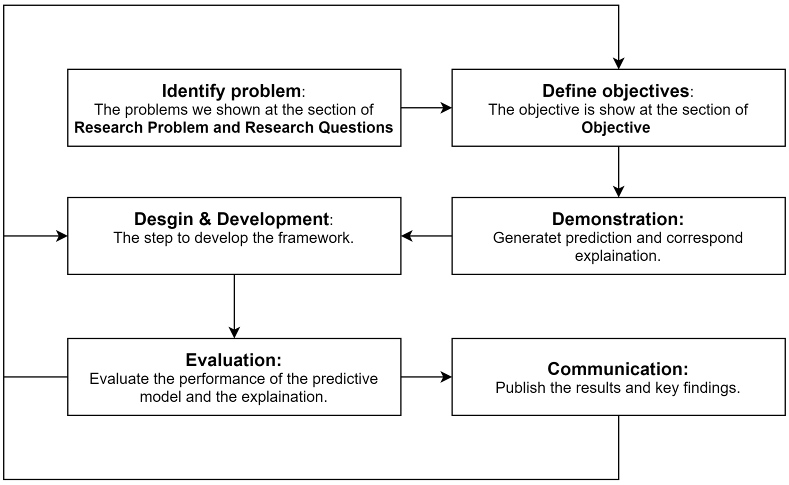


Figure 8: DSRM Process model for this research.

4.1.3 Research Plan

In this section, the overall plan of this research will be shown. In order to achieve the 3 objectives we listed in the section 4.1, this research will be divided to 3 corresponded phases.

Phase I: Data exploration, radiologists interview (Aim 1)

Description: In order to generate an explainable result for the radiologist, we first have to know how the chest x-ray image is interpreted by radiologists. This understanding can be gained through two approaches. First, we will examine the existing eye tracking data to identify the radiologists' classification patterns. Next, we will interview the radiologists so they can tell us how they actually make a diagnosis based on the *chest x-rays images* and *clinical data*. And the eye tracking data and classification patterns we observed are also shown to help understand how professional radiologists perceive this type of information. Since we already have an established research collaboration with an experience radiologist from the main hospital in Portugal, the result from each phase will be evaluated with the help of the radiologist.

- **Activity 1:** Explore eye tracking and clinical dataset to see what type of feature or pattern we can extract.

- **Activity 2:** Identify the possible classification pattern data by observing the difference in eye movements between a normal and an abnormal case in terms of fixation, saccades and scanpaths.

- **Activity 3:** Interview with radiologists to know how they diagnosis based on clinical data and Chest X-ray image, which can help us to understand what are those data necessary for radiologists to make a correct diagnosis.

- **Activity 4:** Interview with radiologists to find out what kind of explanation they expect to have from a computer-aided system.

- **Activity 5:** Interview with radiologists to find out how they perceive the eye tracking data and the extracted classification pattern.

Expected Outcomes: By understanding the chest x-ray interpretation process and the classification pattern of eye tracking information, we can design the feature engineer process for the eye tracking information and clinical data, which will be beneficial for the predictive model. The perception of radiologists on eye tracking data can guide us to use the eye tracking information realistically. According to the radiologists' expectation on a computer-aided system, we can design a framework that generates desirable explanations.

Phase II: Multi-model Architecture Development (Aim 2 & Aim 3)

In the first phase, we will have known how which features or patterns in the dataset will be useful for creating a predictive model that can provide reasonable diagnosis. In this phase, we will firstly decide what tasks we desire the framework to perform. And, based on the decided tasks, we will make the decision on the model architecture considered available input and output data. For instance, if we expect the framework not only to predict the disease but also provide text explanation, an architecture that supports text generation is needed, such as RNN decoder [100] or Transformer [107]. If we expect the model to generate bounding boxes for lesions, an architecture that can perform object detection task must be adopted. Overall, the pipeline of this framework will hierarchically consist of 2 stages. The first stage will be responsible for providing accurate diagnoses. And, the second stage will generate explanations to support the diagnoses, such as bounding boxes and saliency maps.

- **Activity 1:** Based on available data and the suggestions from radiologists, we will determine what information from dataset should be considered as input and output.

- **Activity 2:** According to the expected input and output for the predictive model, we decide what functionalities (learning tasks) we will provide in this framework.

- **Activity 3:** Based on the learning tasks we decided, we survey existing literature to find out optimum architecture for each modality. And, according to the architecture, we determine which fusion strategy and loss function to use for the framework. When we have multiple labels in the output, a method of combining losses from different tasks is needed.

- **Activity 4:** Conduct training and parameter searching on the framework to optimise performance and generated explanations.
- **Activity 5:** The performance of the model will be evaluated by AUC. And the generated explanation will be examined by human radiologist.

Expected Outcomes: The outcome of this phase is the design of framework architecture and the trained model that can classify the disease accurately and provide insightful explanation for its predictions.

Phase III: Model Evaluation (Aim 4)

In the previous phase, we run the model to obtain prediction through terminal and scripts. However, this approach can be inefficient and confusing to radiologist. Since the radiologist is the end-user that will be helping us on evaluating the predictions and diagnoses, an interface specially designed for radiologists is required. The activities in this phase can be iterative until the end-user are satisfied and a desired prototype can be delivered.

- **Activity 1:** Conduct an initial interview with radiologists to discuss the design of user interface. To make the application interactive and intuitive, we also have to discuss how radiologist prefer to import the data and export the results.
- **Activity 2:** According to the discussion we have in previous activity, the requirements will be listed as implementable features using MoSCoW prioritisation method.
- **Activity 3:** Follow the MoSCoW list, we will implement the interface and functionalities to meet the requirements.
- **Activity 4:** Once the application has been completed and tested internally, we will conduct another interview with radiologists to run the tests and collect feedback from them. When no improvement is asked by radiologists, we can proceed to next activity. However, if a change is requested, we can redirect to Activity 2 to make adjustments.
- **Activity 5:** Eventually, we will deploy this prototype to make it available online. Another system for collecting feedback will be put in place to look for ways to improve the interface. Regular maintenance and further update will also be planned.

Expected Outcomes: At the last phase, the deliverable outcome will be an interactive and robust interface specifically designed for the needs from radiologists. This prototype will be able to access the model we trained in last phase to take clinical information and chest x-ray image as input and make diagnoses for patients. In addition, the explanation of the diagnosis decisions will also be generated and shown on the screen for radiologists to evaluate the results. The interface of the framework will be intuitive and user-friendly to enhance the experience. A future work can also base on this prototype to develop a fully functional system.

4.2 Resources and Funding Required

4.3 Time for Completion

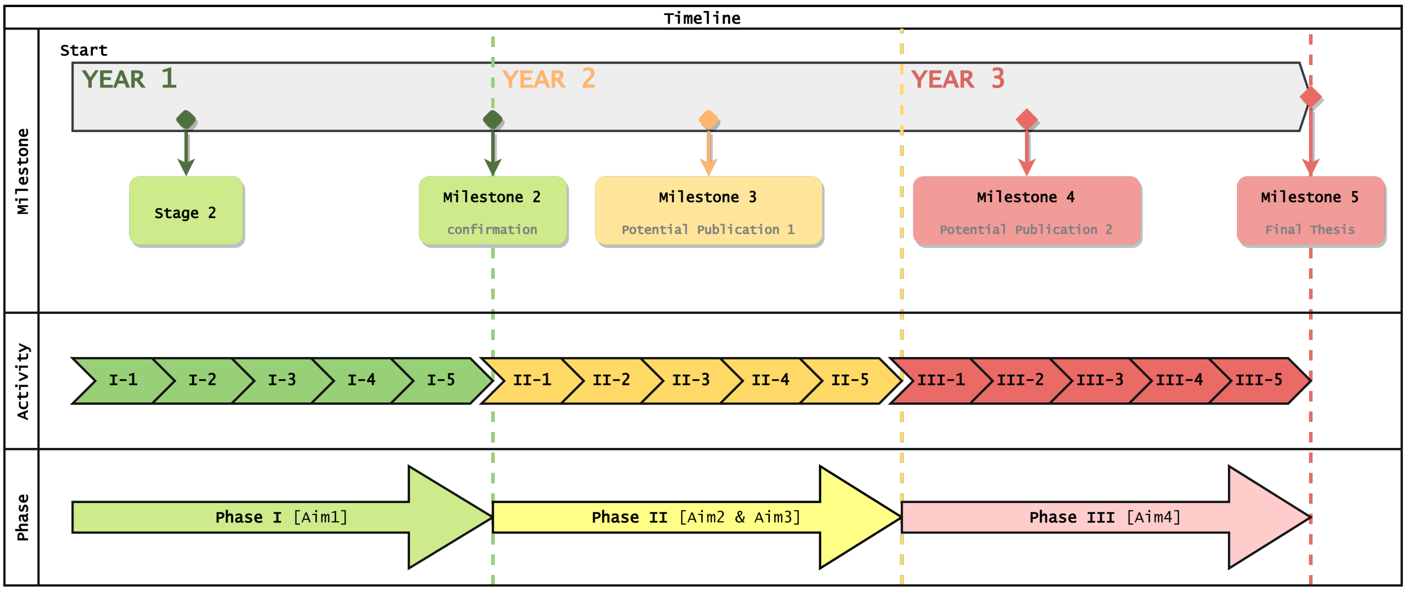


Figure 9: Project Timeline.

References

[1] Ahsan, M. M., Gupta, K. D., Islam, M. M., Sen, S., Rahman, M. L., and Hossain, M. S. (2020). COVID-19 symptoms detection based on NasNetMobile with explainable AI using various imaging modalities. *Machine Learning and Knowledge Extraction*, 2(4):490–504.

[2] Anderson, J. R., Bothell, D., and Douglass, S. (2004). Eye movements do not reflect retrieval processes: Limits of the eye-mind hypothesis. *Psychological Science*, 15(4):225–231.

[3] Antonenko, P., Paas, F., Grabner, R., and van Gog, T. (2010). Using electroencephalography to measure cognitive load. *Educational Psychology Review*, 22(4):425–438.

[4] Badgeley, M. A., Zech, J. R., Oakden-Rayner, L., Glicksberg, B. S., Liu, M., Gale, W., McConnell, M. V., Percha, B., Snyder, T. M., and Dudley, J. T. (2019). Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ digital medicine*, 2(1):1–10.

[5] Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

[6] Bannert, M. (2002). Managing cognitive load—recent trends in cognitive load theory. *Learning and instruction*, 12(1):139–146.

[7] Beatty, J., Lucero-Wagoner, B., et al. (2000). The pupillary system. *Handbook of psychophysiology*, 2(142-162).

[8] Belle, V. and Papantonis, I. (2020). Principles and practice of explainable machine learning. *CoRR*, abs/2009.11698.

[9] Bigolin Lanfredi, R., Zhang, M., Auffermann, W., Chan, J., Duong, P.-A., Srikumar, V., Drew, T., Schroeder, J., and Tasdizen, T. (2021). Reflacx: Reports and eye-tracking data for localization of abnormalities in chest x-rays.

[10] Blascheck, T., Kurzhals, K., Raschke, M., Burch, M., Weiskopf, D., and Ertl, T. (2014). State-of-the-art of visualization for eye tracking data. In *EuroVis (STARs)*.

[11] Brunese, L., Mercaldo, F., Reginelli, A., and Santone, A. (2020). Explainable deep learning for pulmonary disease and coronavirus covid-19 detection from x-rays. *Computer Methods and Programs in Biomedicine*, 196:105608.

[12] Carter, B. T. and Luke, S. G. (2020). Best practices in eye tracking research. *International Journal of Psychophysiology*, 155:49–62.

[13] Castillo, C., Steffens, T., Sim, L., and Caffery, L. (2020). The effect of clinical information on radiology reporting: A systematic review. *Journal of Medical Radiation Sciences*, 68(1):60–74.

[14] Charnes, A., Golany, B., Keane, M., and Rousseau, J. (1988). *Extremal Principle Solutions of Games in Characteristic Function Form: Core, Chebychev and Shapley Value Generalizations*, pages 123–133. Springer Netherlands, Dordrecht.

[15] Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. (2018). Grad-CAM: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE.

[16] Chen, S. and Epps, J. (2014). Using task-induced pupil diameter and blink rate to infer cognitive load. *Human–Computer Interaction*, 29(4):390–413.

[17] Cohen, J. P., Dao, L., Roth, K., Morrison, P., Bengio, Y., Abbasi, A. F., Shen, B., Mahsa, H. K., Ghassemi, M., Li, H., and Duong, T. Q. (2020a). Predicting COVID-19 pneumonia severity on chest x-ray with deep learning. *Cureus*, 12(7):e9448.

[18] Cohen, J. P., Morrison, P., and Dao, L. (2020b). Covid-19 image data collection.

[19] Cooke, L. (2005). Eye tracking: How it works and how it relates to usability. *Technical Communication*, 52(4):456–463. Name - Association for Computing Machinery; Copyright - Copyright Society for Technical Communication Nov 2005; Document feature - ; Last updated - 2019-11-23; CODEN - TLCMBT; SubjectsTermNotLitGenreText - New York.

[20] Crawford, T. (2015). The disengagement of visual attention in alzheimer's disease: a longitudinal eye-tracking study. *Frontiers in Aging Neuroscience*, 7:118.

[21] Das, A. and Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*.

[22] de Belen, R. A. J., Bednarz, T., and Sowmya, A. (2021). *EyeXplain Autism: Interactive System for Eye Tracking Data Analysis and Deep Neural Network Interpretation for Autism Spectrum Disorder Diagnosis.* Association for Computing Machinery, New York, NY, USA.

[23] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

[24] Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning.

[25] Došilović, F. K., Brčić, M., and Hlupić, N. (2018). Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 0210–0215.

[26] Duchowski, A. T., Krejtz, K., Krejtz, I., Biele, C., Niedzielska, A., Kiefer, P., Raubal, M., and Giannopoulos, I. (2018). *The Index of Pupillary Activity: Measuring Cognitive Load ¡i¿ Vis-à-Vis¡/i¿ Task Difficulty with Pupil Oscillation*, page 1–13. Association for Computing Machinery, New York, NY, USA.

[27] Dunnmon, J. A., Yi, D., Langlotz, C. P., Ré, C., Rubin, D. L., and Lungren, M. P. (2019). Assessment of convolutional neural networks for automated classification of chest radiographs. *Radiology*, 290(2):537–544.

[28] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118.

[29] European Commission (2016). 2018 reform of eu data protection rules.

[30] Feng, D., Haase-Schütz, C., Rosenbaum, L., Hertlein, H., Gläser, C., Timm, F., Wiesbeck, W., and Dietmayer, K. (2021). Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1341–1360.

[31] Fitts, P. M., Jones, R. E., and Milton, J. L. (2005). Eye movements of aircraft pilots during instrument-landing approaches. *Ergonomics: Psychological mechanisms and models in ergonomics*, 3:56.

[32] Gao, W., Jin, R., Zhu, S., and Zhou, Z.-H. (2013). One-pass auc optimization. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 906–914, Atlanta, Georgia, USA. PMLR.

[33] Gao, W. and Zhou, Z. (2012). On the consistency of AUC optimization. *CoRR*, abs/1208.0645.

[34] Giovinco, N. A., Sutton, S. M., Miller, J. D., Rankin, T. M., Gonzalez, G. W., Najafi, B., and Armstrong, D. G. (2015). A passing glance? differences in eye tracking and gaze patterns between trainees and experts reading plain film bunion radiographs. *The Journal of Foot and Ankle Surgery*, 54(3):382–391.

[35] Guan, Z., Lee, S., Cuddihy, E., and Ramey, J. (2006). The validity of the stimulated retrospective think-aloud method as measured by eye tracking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, page 1253–1262, New York, NY, USA. Association for Computing Machinery.

[36] Hajian-Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian J Intern Med*, 4(2):627–635.

[37] Harezlak, K. and Kasprowski, P. (2018). Application of eye tracking in medicine: A survey, research issues and challenges. *Computerized Medical Imaging and Graphics*, 65:176–190. Advances in Biomedical Image Processing.

[38] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

[39] Hess, E. H. and Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science*, 143(3611):1190–1192.

[40] Hevner, A. R., March, S. T., Park, J., and Ram, S. (2004). Design science in information systems research. *MIS quarterly*, pages 75–105.

[41] Hori, C., Hori, T., Lee, T.-Y., Zhang, Z., Harsham, B., Hershey, J. R., Marks, T. K., and Sumi, K. (2017). Attention-based multimodal fusion for video description. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[42] Hou, J. and Gao, T. (2021). Explainable DCNN based chest x-ray image analysis and classification for COVID-19 pneumonia detection. *Scientific Reports*, 11(1).

[43] Hu, D., Nie, F., and Li, X. (2019a). Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[44] Hu, P., Zhen, L., Peng, D., and Liu, P. (2019b). Scalable deep multimodal learning for cross-modal retrieval. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 635–644, New York, NY, USA. Association for Computing Machinery.

[45] Huang, G., Liu, Z., and Weinberger, K. Q. (2016). Densely connected convolutional networks. *CoRR*, abs/1608.06993.

[46] Huang, J. and Ling, C. (2005). Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17(3):299–310.

[47] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., and Ng, A. Y. (2019). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison.

[48] Johnson, A., Bulgarelli, L., Pollard, T., Celi, L. A., Mark, R., and Horng, S. (2021a). Mimic-iv-ed.

[49] Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., and Mark, R. (2021b). Mimic-iv.

[50] Johnson, A. E. W., Pollard, T., Mark, R., Berkowitz, S., and Horng, S. (2019a). The mimic-cxr database.

[51] Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C., Mark, R. G., and Horng, S. (2019b). MIMIC-CXR-JPG: A large publicly available database of labeled chest radiographs. *CoRR*, abs/1901.07042.

[52] Karargyris, A., Kashyap, S., Lourentzou, I., Wu, J., Tong, M., Sharma, A., Abedin, S., Beymer, D., Mukherjee, V., Krupinski, E., and Moradi, M. (2020). Eye Gaze Data for Chest X-rays (version 1.0.0).

[53] Karargyris, A., Kashyap, S., Lourentzou, I., Wu, J. T., Sharma, A., Tong, M., Abedin, S., Beymer, D., Mukherjee, V., Krupinski, E. A., and Moradi, M. (2021). Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for AI development. *Scientific Data*, 8(1).

[54] Klingner, J. (2010). *Measuring Cognitive Load During Visual Tasks by Combining Pupillometry and Eye Tracking.* PhD thesis. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2021-09-28.

[55] Krejtz, K., Duchowski, A. T., Niedzielska, A., Biele, C., and Krejtz, I. (2018). Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PLOS ONE*, 13(9):1–23.

[56] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

[57] Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., and Ferrari, V. (2020). The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*.

[58] Lahat, D., Adali, T., and Jutten, C. (2015). Multimodal data fusion: An overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477.

[59] Leslie, A., Jones, A. J., and Goddard, P. R. (2000). The influence of clinical information on the reporting of CT by radiologists. *The British Journal of Radiology*, 73(874):1052–1055.

[60] Levy, D. L., Sereno, A. B., Gooding, D. C., and O'Driscoll, G. A. (2010). *Eye Tracking Dysfunction in Schizophrenia: Characterization and Pathophysiology*, pages 311–347. Springer Berlin Heidelberg, Berlin, Heidelberg.

[61] Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

[62] Liu, G., Hsu, T.-M. H., McDermott, M., Boag, W., Weng, W.-H., Szolovits, P., and Ghassemi, M. (2019). Clinically accurate chest x-ray report generation. In Doshi-Velez, F., Fackler, J., Jung, K., Kale, D., Ranganath, R., Wallace, B., and Wiens, J., editors, *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pages 249–269. PMLR.

[63] Lobo, J. M., Jiménez-Valverde, A., and Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17(2):145–151.

[64] Loyola-Gonzalez, O. (2019). Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, 7:154096–154113.

[65] Lundberg, S. M. and Lee, S. (2017a). Consistent feature attribution for tree ensembles. *CoRR*, abs/1706.06060.

[66] Lundberg, S. M. and Lee, S. (2017b). A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874.

[67] Manning, D., Ethell, S. C., and Crawford, T. (2003). Eye-tracking AFROC study of the influence of experience and training on chest x-ray interpretation. In Chakraborty, D. P. and Krupinski, E. A., editors, *Medical Imaging 2003: Image Perception, Observer Performance, and Technology Assessment*, volume 5034, pages 257 – 266. International Society for Optics and Photonics, SPIE.

[68] McDuff, D., Gontarek, S., and Picard, R. (2014). Remote measurement of cognitive stress via heart rate variability. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2957–2960.

[69] McLaughlin, L., Bond, R., Hughes, C., McConnell, J., and McFadden, S. (2017). Computing eye gaze metrics for the automatic assessment of radiographer performance during x-ray image interpretation. *International journal of medical informatics*, 105:11–21.

[70] Moreira, C., Chou, Y.-L., Velmurugan, M., Ouyang, C., Sindhgatta, R., and Bruza, P. (2021). LINDA-BN: An interpretable probabilistic approach for demystifying black-box predictive models. *Decision Support Systems*, 150:113561.

[71] Mroueh, Y., Marcheret, E., and Goel, V. (2015). Deep multimodal learning for audio-visual speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2130–2134.

[72] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). Multimodal deep learning. In *ICML*, pages 689–696.

[73] Palinko, O., Kun, A. L., Shyrokov, A., and Heeman, P. (2010). Estimating cognitive load using remote eye tracking in a driving simulator. In *Proceedings of the 2010 Symposium on Eye-Tracking Research amp; Applications*, ETRA '10, page 141–144, New York, NY, USA. Association for Computing Machinery.

[74] Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3):45–77.

[75] Peng, Y., Dharssi, S., Chen, Q., Keenan, T. D., Agrón, E., Wong, W. T., Chew, E. Y., and Lu, Z. (2019). Deepseenet: A deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs. *Ophthalmology*, 126(4):565–575.

[76] Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C. P., Patel, B. N., Yeom, K. W., Shpanskaya, K., Blankenberg, F. G., Seekins, J., Amrhein, T. J., Mong, D. A., Halabi, S. S., Zucker, E. J., Ng, A. Y., and Lungren, M. P. (2018). Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLOS Medicine*, 15(11):1–17.

[77] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D. Y., Bagul, A., Langlotz, C. P., Shpanskaya, K. S., Lungren, M. P., and Ng, A. Y. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *CoRR*, abs/1711.05225.

[78] Ramachandram, D. and Taylor, G. W. (2017). Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108.

[79] Rayner, K. and Reingold, E. M. (2015). Evidence for direct cognitive control of fixation durations during reading. *Current Opinion in Behavioral Sciences*, 1:107–112. Cognitive control.

[80] Reichle, E. D., Pollatsek, A., and Rayner, K. (2012). Using e-z reader to simulate eye movements in nonreading tasks: A unified framework for understanding the eye–mind link. *Psychological Review*, 119(1):155 – 185.

[81] Reichle, E. D., Reineberg, A. E., and Schooler, J. W. (2010). Eye movements during mindless reading. *Psychological Science*, 21(9):1300–1310. PMID: 20679524.

[82] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016a). Model-agnostic interpretability of machine learning.

[83] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016b). "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938.

[84] Ribeiro, M. T., Singh, S., and Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

[85] Rimmer, A. (2017). Radiologist shortage leaves patient care at risk, warns royal college. *BMJ*, page j4683.

[86] Rodin, I., Fedulova, I., Shelmanov, A., and Dylov, D. V. (2019). Multitask and multimodal neural network model for interpretable analysis of x-ray images. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1601–1604.

[87] Rudin, C. and Radin, J. (2019). Why are we using black box models in ai when we don't need to? a lesson from an explainable ai competition. *Harvard Data Science Review*, 1(2). https://hdsr.mitpress.mit.edu/pub/f9kuryi8.

[88] Saab, K., Hooper, S. M., Sohoni, N. S., Parmar, J., Pogatchnik, B., Wu, S., Dunnmon, J. A., Zhang, H. R., Rubin, D., and Ré, C. (2021). Observational supervision for medical image classification using gaze data. In de Bruijne, M., Cattin, P. C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., and Essert, C., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 603–614, Cham. Springer International Publishing.

[89] Saporta, A., Gui, X., Agrawal, A., Pareek, A., Truong, S. Q., Nguyen, C. D., Ngo, V.-D., Seekins, J., Blankenberg, F. G., Ng, A. Y., Lungren, M. P., and Rajpurkar, P. (2021). Benchmarking saliency methods for chest x-ray interpretation.

[90] Schall, A. and Romano Bergstrom, J. (2014). Introduction to eye tracking. In Romano Bergstrom, J. and Schall, A. J., editors, *Eye Tracking in User Experience Design*, pages 3–26. Morgan Kaufmann, Boston.

[91] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[92] Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3145–3153. PMLR.

[93] Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. In *In Workshop at International Conference on Learning Representations*.

[94] Singh, A., Sengupta, S., and Lakshminarayanan, V. (2020). Explainable deep learning models in medical image analysis. *Journal of Imaging*, 6(6).

[95] Spinner, T., Schlegel, U., Schäfer, H., and El-Assady, M. (2020). ¡bold¿explainer¡/bold¿: A visual analytics framework for interactive and explainable machine learning. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1064–1074.

[96] Stuart, S., Alcock, L., Godfrey, A., Lord, S., Rochester, L., and Galna, B. (2016). Accuracy and re-test reliability of mobile eye-tracking in parkinson's disease and older adults. *Medical Engineering Physics*, 38(3):308–315.

[97] Sulam, J., Ben-Ari, R., and Kisilev, P. (2017). Maximizing auc with deep learning for classification of imbalanced mammogram datasets. In *VCBM*, pages 131–135.

[98] Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. *CoRR*, abs/1703.01365.

[99] Sunshine, J. H., Maynard, C. D., Paros, J., and Forman, H. P. (2004). Update on the diagnostic radiologist shortage. *American Journal of Roentgenology*, 182(2):301–305.

[100] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

[101] Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285.

[102] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2015). Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567.

[103] Tang, M., Kumar, P., Chen, H., and Shrivastava, A. (2020a). Deep multimodal learning for the diagnosis of autism spectrum disorder. *Journal of Imaging*, 6(6).

[104] Tang, Y.-X., Tang, Y.-B., Peng, Y., Yan, K., Bagheri, M., Redd, B. A., Brandon, C. J., Lu, Z., Han, M., Xiao, J., and Summers, R. M. (2020b). Automated abnormality classification of chest radiographs using deep convolutional neural networks. *npj Digital Medicine*, 3(1).

[105] Teixeira, L. O., Pereira, R. M., Bertolini, D., Oliveira, L. S., Nanni, L., Cavalcanti, G. D. C., and Costa, Y. M. G. (2021). Impact of lung segmentation on the diagnosis and explanation of COVID-19 in chest x-ray images. *Sensors*, 21(21):7116.

[106] Ting, D. S. W., Cheung, C. Y.-L., Lim, G., Tan, G. S. W., Quang, N. D., Gan, A., Hamzah, H., Garcia-Franco, R., San Yeo, I. Y., Lee, S. Y., Wong, E. Y. M., Sabanayagam, C., Baskaran, M., Ibrahim, F., Tan, N. C., Finkelstein, E. A., Lamoureux, E. L., Wong, I. Y., Bressler, N. M., Sivaprasad, S., Varma, R., Jonas, J. B., He, M. G., Cheng, C.-Y., Cheung, G. C. M., Aung, T., Hsu, W., Lee, M. L., and Wong, T. Y. (2017). Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. *JAMA*, 318(22):2211–2223.

[107] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

[108] Vielzeuf, V., Lechervy, A., Pateux, S., and Jurie, F. (2018). Centralnet: a multilayer approach for multimodal fusion. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops.*

[109] Waite, S., Scott, J., Gale, B., Fuchs, T., Kolla, S., and Reede, D. (2017). Interpretive error in radiology. *American Journal of Roentgenology*, 208(4):739–749.

[110] Wang, Q., Yang, S., Liu, M., Cao, Z., and Ma, Q. (2014). An eye-tracking study of website complexity from cognitive load perspective. *Decision Support Systems*, 62:1–10.

[111] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. (2017). Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[112] Wang, Y., Huang, W., Sun, F., Xu, T., Rong, Y., and Huang, J. (2020a). Deep multimodal fusion by channel exchanging. *Advances in Neural Information Processing Systems*, 33.

[113] Wang, Y., Huang, W., Sun, F., Xu, T., Rong, Y., and Huang, J. (2020b). Deep multimodal fusion by channel exchanging. *Advances in Neural Information Processing Systems*, 33.

[114] Yang, X., Ramesh, P., Chitta, R., Madhvanath, S., Bernal, E. A., and Luo, J. (2017). Deep multimodal representation learning from temporal data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[115] Yates, E., Yates, L., and Harvey, H. (2018). Machine learning "red dot": open-source, cloud, deep convolutional neural networks in chest radiograph binary normality classification. *Clinical Radiology*, 73(9):827–831.

[116] Yuan, Z., Yan, Y., Sonka, M., and Yang, T. (2020). Robust deep AUC maximization: A new surrogate loss and empirical studies on medical image classification. *CoRR*, abs/2012.03173.

[117] Yuksel, B. F., Oleson, K. B., Harrison, L., Peck, E. M., Afergan, D., Chang, R., and Jacob, R. J. (2016). Learn piano with bach: An adaptive learning interface that adjusts task difficulty based on brain state. In *Proceedings of the 2016 chi conference on human factors in computing systems*, pages 5372–5384.

[118] Zagermann, J., Pfeil, U., and Reiterer, H. (2016). Measuring cognitive load using eye tracking technology in visual computing. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, BELIV '16, page 78–85, New York, NY, USA. Association for Computing Machinery.

[119] Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, pages 818–833. Springer International Publishing.

[120] Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., and Torralba, A. (2015). Learning deep features for discriminative localization. *CoRR*, abs/1512.04150.

[121] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. (2019). A comprehensive survey on transfer learning. *CoRR*, abs/1911.02685.

5 Appendix

5.1 Research Integrity Online

The completion certificate is attached below.


OREI QUT

This is to certify that

Chihcheng Hsieh

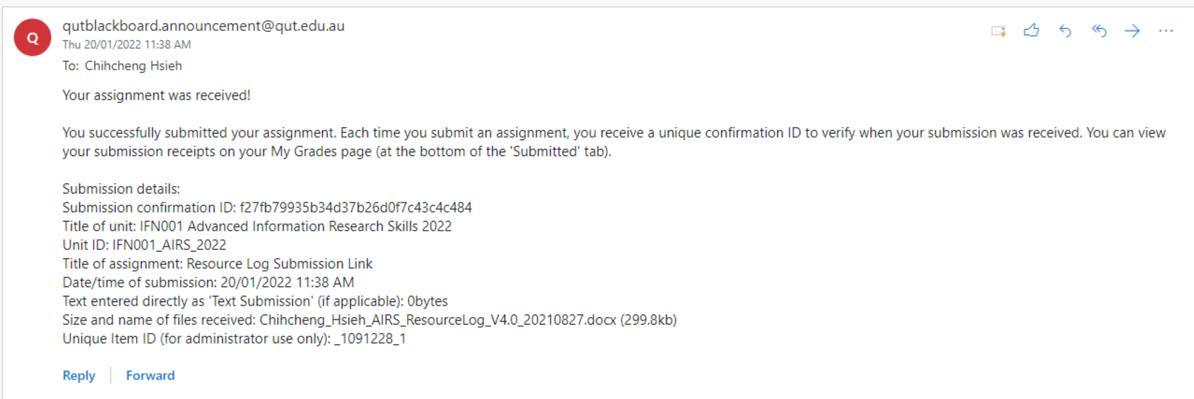
has successfully completed

Research Integrity Online for Students

on

1 December 2021

5.2 Proof of AIRS submission


qutblackboard.announcement@qut.edu.au
Thu 20/01/2022 11:38 AM
To: Chihcheng Hsieh

Your assignment was received!

You successfully submitted your assignment. Each time you submit an assignment, you receive a unique confirmation ID to verify when your submission was received. You can view your submission receipts on your My Grades page (at the bottom of the 'Submitted' tab).

Submission details:
Submission confirmation ID: f27fb79935b34d37b26d0f7c43c4c484
Title of unit: IFN001 Advanced Information Research Skills 2022
Unit ID: IFN001_AIRS_2022
Title of assignment: Resource Log Submission Link
Date/time of submission: 20/01/2022 11:38 AM
Text entered directly as 'Text Submission' (if applicable): 0bytes
Size and name of files received: Chihcheng_Hsieh_AIRS_ResourceLog_V4.0_20210827.docx (299.8kb)
Unique Item ID (for administrator use only): _1091228_1

Reply | Forward

5.3 Coursework

I am required to the following coursework as part of the of the completion of this course:

- **IFN001 Advanced Information Retrieval Skill (AIRS)**: Assessment submitted prior to submission of Stage 2 Report.

5.4 Ethical Clearance

I am still discussing ethical clearance with my supervisory team.

5.5 Health and Safety

I have completed complete Health & Safety training.

5.6 Intellectual Property

I do not need to sign an IP Assignment Agreement

5.7 Collaborative Arrangement

I do not require a Collaborative Agreement.

5.8 Data Management

Only public data will be considered for this research