AI Readiness in Healthcare through Storytelling XAI

Akshat Dubey[1,2], Zewen Yang[1] and Georges Hattab[1,2]

[1]Center for Artificial Intelligence in Public Health Research (ZKI-PH) at Robert Koch Institute, Nordufer 20, Berlin 13353, Germany
[2]Department of Mathematics and Computer Science, Free University of Berlin, Arnimallee 14, Berlin 14195, Germany

Abstract

Artificial Intelligence is rapidly advancing and radically impacting everyday life, driven by the increasing availability of computing power. Despite this trend, the adoption of AI in real-world healthcare is still limited. One of the main reasons is the trustworthiness of AI models and the potential hesitation of domain experts with model predictions. Explainable Artificial Intelligence (XAI) techniques aim to address these issues. However, explainability can mean different things to people with different backgrounds, expertise, and goals. To address the target audience with diverse needs, we develop storytelling XAI. In this research, we have developed an approach that combines multi-task distillation with interpretability techniques to enable audience-centric explainability. Using multi-task distillation allows the model to exploit the relationships between tasks, potentially improving interpretability as each task supports the other leading to an enhanced interpretability from the perspective of a domain expert. The distillation process allows us to extend this research to large deep models that are highly complex. We focus on both model-agnostic and model-specific methods of interpretability, supported by textual justification of the results in healthcare through our use case. Our methods increase the trust of both the domain experts and the machine learning experts to enable a responsible AI.

Keywords: Medical Domain, Explainable Artificial Intelligence, Artificial Intelligence, Natural Language Processing, Data Exploration, Data-Driven Storytelling, Storytelling XAI, Trust AI, XAI, AI, NLP

1. Introduction

The adoption of artificial intelligence (AI) in healthcare has been relatively slow compared to other industries, despite its immense potential benefits. Acceptance of AI by patients and healthcare providers requires building trust in these AI-powered systems in hospitals and healthcare systems [1, 2]. Explainability techniques from the field of Explainable AI (XAI) aim to increase the transparency of AI decision-making, enabling clinicians and patients to understand and validate AI results, thereby building trust [3]. With explainable AI, medical professionals can share the reasoning behind their decisions, maintaining accountability and empowering patients to make informed decisions about their care [4]. While current explainability methods provide good insight into the workings of Machine Learning (ML) and deep learning models, they are not yet suitable for domain experts who have little or no experience with such models. Trustworthy AI techniques have gained significant attention spanning various domains [5]. Explanations need to be tailored to the particular goals, concerns, and decision requirements of the non-specialist audience, highlighting the data that is most important to them [6].

There has been previous work in this area, but the research is limited in terms of audience-centric explainability through the utilization and combining currently available datasets for diverse tasks. The current research methods attempt to explain the models in terms of heat map visualizations or post hoc methods including the Local Interpretable Model-agnostic Explanations (LIME) method or SHapley Additive Explanations (SHAP) [7, 8]. Motivating works in Storytelling XAI already exist. SHAPstories and CFstories, are two new techniques that use large language models to generate narratives explaining AI predictions based on SHAP values and counterfactual explanations [9].

EXPLIMED - First Workshop on Explainable Artificial Intelligence for the medical domain - 19-20 October 2024, Santiago de Compostela, Spain

DubeyA@rki.de (A. Dubey); ZewenY@rki.de (Z. Yang); HattabG@rki.de (G. Hattab)
0009-0008-4823-9375 (A. Dubey); 0000-0001-9974-3231 (Z. Yang); 0000-0003-4168-8254 (G. Hattab)

Counterfactual explanations frequently rest on the presumptions that the features are independent and that the changes in the highlighted features match real-world activities. In actuality, though, the characteristics are not independent, and actions are likely to have simultaneous effects on several features. Post-hoc counterfactual explanations need to provide faithful changes that a human can implement in practice and should conform to the observed correlations in the training data rather than being anomalies. They may not be faithful to the original data due to problems like overfitting or excessive generalization, which results in unsatisfactory interpretability. Although ground truth plausibility of counterfactuals is acknowledged as a crucial attribute, the search results show that this characteristic is still challenging to measure in practice [10, 11]. The research also utilizes the application of the Large-Language-Models (LLMs) to enable narrative XAI, another limitation to implementing this framework in healthcare and finance. The XAINES project [12] also explores the use of narratives to explain AI systems, hypothesizing that narratives are an appropriate means to communicate explanations, particularly to users without machine learning backgrounds. Whether the intended audience is developers, domain experts, or end users who are affected by the AI's decisions, the project intends to make it possible to explain AI systems in a way that is specific to their requirements and expectations. An essential component of the XAINES approach is the narratives used to describe the AI, which facilitates the successful communication of the causal chain of events leading to the AI's behavior. The challenges of translating causal reasoning into effective explanatory narratives and interactive experiences for diverse audiences still exist. However, the existing storytelling XAI frameworks do not adequately address the effective usage of available datasets, and how to combine them and achieve end-to-end audience-centric XAI without relying on distinct models trained for specific tasks.

We develop a storytelling XAI framework to provide end-to-end explanations customized for domain specialists, in our case, healthcare practitioners as well as ML practitioners. The storytelling XAI uses a knowledge distillation [13] approach coupled with multi-task prediction and combines it with interpretability techniques to achieve explainability. This approach helps to achieve concept-based explainability. Concept-based explainability methods aim to explain the behavior and predictions of deep neural networks (NN) using human-understandable concepts such as texts. The knowledge distillation enables us to utilize the current available datasets, combine them with different tasks, and have a single model to deal with all the different tasks. To generate explanations, our research uses both model-agnostic and model-specific methods. The model-agnostic methods use an approximation for interpretability, which allows flexibility in model selection but limits the accuracy of interpretations due to the approximations involved. Whereas the model-specific methods can interpret the results from the components of the model giving us accurate interpretation but limiting the flexibility to choose models [14]. By combining knowledge distillation, multiple datasets for diverse tasks, and interpretability techniques, we achieve explainability that is useful for domain experts and ML practitioners to decipher the decision-making process of the AI models. For this research, we have considered the use case of chest X-ray analysis to demonstrate the effectiveness of our framework in enabling XAI for healthcare professionals. This work also aims at an end-to-end application of trustworthy AI in healthcare using multi-task predictions supported by interpretation techniques. Through this research, complex deep neural networks could also be used in healthcare without compromising model trustworthiness. Storytelling XAI framework for the healthcare domain provides concept-based understanding in human terms to healthcare professionals while providing technical interpretation to ML practitioners.

2. Background

2.1. Knowledge Distillation

Knowledge distillation refers to the process of transferring knowledge from a comparatively large model to a smaller model without compromising on performance. The complex large model is called *teacher* and the smaller model is called **student**. The rationale behind model distillation is to train complex large models also called *teacher* for a specific task and then transfer the knowledge to a smaller model also called *student* using distillation loss (Equation: 1). The loss function employed in knowledge distillation to instruct the student model to imitate the teacher model's behavior is called the distillation loss. By using a "temperature" scaling function in the softmax, the logits are softened, so smoothing down the probability distribution and exposing the teacher's taught inter-class correlations. The probability p_i of class i from the logits z is calculated as:

$$p_i = exp(z_i/T) \div \sum_j exp(z_j/T) \tag{1}$$

where T is the parameter of temperature and when $T = 1$ then we get the original softmax function. The probability distribution produced by the softmax function softens with increasing T, giving more insight into which classes the teacher thought were more like the ones that were expected. The discrepancy between the instructor model's soft targets and the student model's predictions is quantified by this loss function. The student model learns the internal representation from the teacher model [13]. The student model learns not only the target outputs but also the internal representations and similarity information from the teacher model. This allows the student model to capture the same high-level concepts and reasoning as the teacher but in a more transparent and interpretable form [15, 16].

2.2. Multi-task Learning

In AI, multi-task learning is the process of teaching one model to handle several related tasks at once, as opposed to training different models for every task. The main concept is to increase overall learning efficiency and generalization performance by utilizing shared representations and commonalities across tasks Multi-task Learning can be implemented through various methods and the two most common methods are shared base network extractors with task-specific heads in which the model comprises task-specific output heads after a shared base network that extracts features. The second method is the shared decision-making layer in which task-specific layers are connected to the shared decision-making layer in the model [17].

2.3. Interpretability Techniques

Interpretability allows ML practitioners to understand the connections between the features that go into a model and the results it produces, as well as the relative importance of different features in the decision-making process. Some models, such as linear regression and decision trees, are inherently interpretable due to their simple structure, while complex models, such as deep neural networks, are often considered black boxes and require additional techniques to enhance their interpretability [18]. The two most common types of interpretation are model-agnostic methods and model-specific methods. Model-specific methods are designed to interpret specific types of models. Visualization techniques and attention mechanisms (Equation:2) are used to interpret Convolutional Neural Network (CNN) architectures and Recurrent Neural Network (RNN) or transformer architectures, respectively.

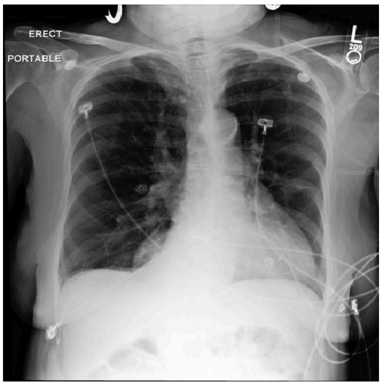
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

where,
Q :represents the query matrix.
K :represents the key matrix.
V :represents the value matrix.
d_k :represents the dimensionality of the key vectors.
This equation computes the attention weights by taking the dot product of the query and key matrices, scaling it by $\sqrt{d_k}$, applying the softmax function, and then multiplying it by the value matrix to get the attention output.

Visualization techniques include the analysis of saliency maps, and activation maps of the intermediate layers of a model, the most relevant being GradCAM (Equation:3).

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha^c_k A^k\right) \tag{3}$$

where,
$L^c_{\text{Grad-CAM}}$: represents the Grad-CAM score for class. c
$\alpha^c_k A^k$: represents the importance of the kth class.
A_k : represents the kth activation map for the last convolution layer.
ReLU denotes the rectified linear unit function, which clips negative values to zero.



Original Text

Cardiac size is normal. Moderate hiatal hernia is again noted. The lungs are hyperinflated. Patient has known emphysema. Small lung nodules, and lingular atelectasis are better seen in prior CT. There is no pneumothorax or pleural effusion.

Augmented Text

['patient has known emphysema. small lung nodules, and lingular atelectasis. there is no pneumothorax or pleural effusion.']

Figure 1: A sample chest X-ray with the corresponding report with original text and extended text. Left: X-ray image of a human chest. Top right: Original text content describing the prognosis of the chest x-ray in medical jargon. Bottom right: Augmented text describing the patient's lung in a concise and clear manner.

Attention mechanisms are usually implemented to visualize text-based models. It helps the user to analyze and infer the specific part of the sentence that influences the predictions. Model agnostic methods are flexible in terms of model choice. They are a good choice for interpreting the predictions of a wide variety of models. The most common methods related to model-agnostic techniques are Local Interpretable Model-agnostic Explanations (LIME)(Equation:4) and SHapley Additive exPlanations (SHAP).

$$\text{LIME: } \hat{f}_{\text{lime}}(x) = \arg\min_{f \in \mathcal{F}} \mathcal{L}(f, \pi_x) + \Omega(f) \tag{4}$$

where,
$\hat{f}_{\text{lime}}(x)$: represents the local interpretable model produced by LIME for input. x
$\mathcal{F}$: denotes the set of possible interpretable models.
$\mathcal{L}(f, \pi_x)$: represents the loss function measuring how well the interpretable model.
f: approximates the prediction function $\pi_x$in the local neighborhood of x.
$\Omega(f)$: represents a regularization term that encourages simplicity or interpretability of the model f.
LIME approximates the behavior of the model locally using interpretable models such as linear regression by perturbing the input and correspondingly reflecting it in the predictions [19].

3. Storytelling XAI in Healthcare

Explainable Artificial Intelligence (XAI) is critical in the medical and healthcare domains because AI-driven systems can have serious and even life-threatening consequences for patients if their decisions are not transparent and interpretable. Model developers are primarily interested in the performance, stability, and robustness of the AI model. Their focus is on using explainability to debug models and improve accuracy [20]. XAI aims to shed light on the "black box" nature of complex AI models such as deep neural networks, which can be difficult for humans to understand. This is important in healthcare to build trust and acceptance of AI systems among clinicians and patients. Integrating XAI into clinical decision support systems is important to align with core ethical principles in medicine, such as autonomy, beneficence, and non-maleficence [4]. There is an increasing need for collaboration between medical and AI experts to develop appropriate frameworks for designing and implementing XAI solutions in the medical domain [21]. However, explainability in AI is highly dependent on the target audience and their specific needs and knowledge levels. Different stakeholders, such as model developers, business managers, and end users, have different requirements for AI explanations. As models become more complex, the explanations generated by AI techniques can also become more complex and difficult for non-expert users to understand [22]. Developing audience-specific explanations is challenging as AI systems become more complex. However, this is critical to building trust and acceptance of AI across different stakeholders.

By using Storytelling XAI, the gap between ML engineers and healthcare professionals is reduced. ML engineers can use Storytelling XAI to transform complex ML models into understandable information that can be shared with healthcare domain experts. ML engineers can more easily communicate the model's underlying logic and decision-making processes through storytelling, as opposed to providing raw data or technical model outputs, making it easier for domain experts to understand how the model's predictions fit into actual patient situations and scenarios. Domain experts are more likely to accept the model's suggestions and incorporate AI into clinical decision-making processes if they can follow the reasoning behind the model's outputs through a narrative.

4. Materials and Methods

For this research, we have used chest X-ray images to detect the abnormalities in the chest, identify the affected regions, and generate a report generation. The report generation task generates a report out of the chest X-image and serves as a text justification for the other task.

4.1. Dataset and Pre-processing

Our research contains two different tasks which utilize a different dataset for each task. The task and the corresponding dataset used are as follows:

1. Chest X-ray report generation: We used a pre-processed dataset provided by the authors of MedVill (Medical Vision Language Learner). The pre-processed dataset was prepared from the original MIMIC-CXR-JPG dataset [23] and Open-I dataset [24]. The MIMIC-CXR-JPG dataset contains 377,110 chest X-ray images and corresponding free-text reports. The Open-I dataset contains 3,851 reports and 7,466 Chest X-ray images, consisting of both the lateral and frontal view of the images [25]. For our ease, the lateral views were removed from the dataset. For our experimentation, we have selected 5000 image-report pairs randomly. The reports were augmented with the help of clinical T5 large which (Fig:1) is a sequence-to-sequence transformer (available on HuggingFace) to obtain the summarized version of the radiology report with the sequence length of 128 words [26]. We performed this step to obtain keywords having higher weightage in the text report through summarization.

2. Chest X-ray abnormality detection: For this task we used a modified version of the VinDr-CXR dataset [27]. The dataset contains 15,000 samples of DICOM images labeled with 14 abnormalities namely aortic enlargement, atelectasis, calcification, cardiomegaly, consolidation, ILD, Infiltration, Lung Opacity, Nodule/Mass, Other lesion, Pleural effusion, Pleural thickening, Pneumothorax,

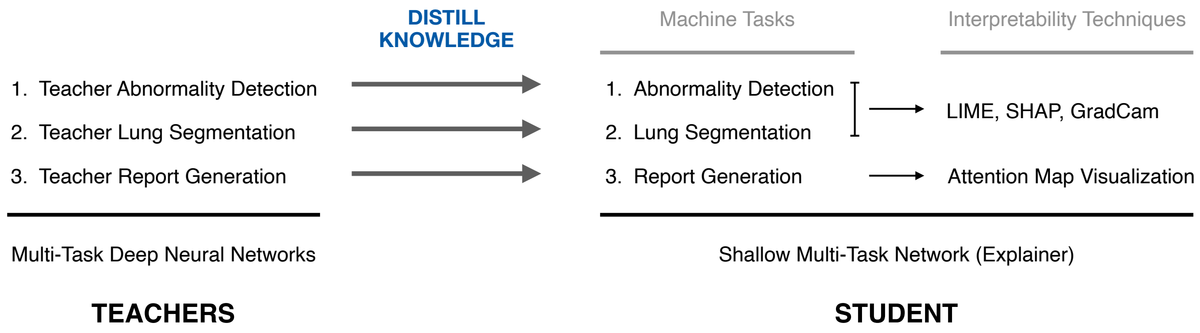

DISTILL KNOWLEDGE
Machine Tasks
Interpretability Techniques
1. Teacher Abnormality Detection
2. Teacher Lung Segmentation
3. Teacher Report Generation
1. Abnormality Detection
2. Lung Segmentation
3. Report Generation
LIME, SHAP, GradCam
Attention Map Visualization
Multi-Task Deep Neural Networks
Shallow Multi-Task Network (Explainer)
TEACHERS
STUDENT

Figure 2: Storytelling XAI Framework. Three teacher networks are trained to perform the specific tasks numbers 1, 2, and 3. Using knowledge distillation, the student model acquires knowledge to perform all the distinct tasks alone. The student model learns the underlying relationship between different features enhancing the interpretability.

and Pulmonary fibrosis along with bounding box for detection and localization of the detected anomalies. The dataset was annotated by a group of seventeen expert radiologists [28]. We selected 5,000 samples randomly without unbalancing the class ratio which could have potentially led to the problem of class imbalance.

3. Lung Segmentation: For this task, we have used the Montgomery dataset which contains images from the Department of Health and Human Services, Montgomery County, Maryland, USA. The dataset consists of 138 CXRs, including 80 normal patients and 58 patients with manifested tuberculosis (TB). The CXR images are 12-bit gray-scale images of dimension 4020 × 4892 or 4892 × 4020. Only the two lung mask annotations are available which were combined into a single image to make it easy for the network to learn the task of segmentation. To make all images of symmetric dimensions we padded the pictures to the maximum dimension in their height or width such that images are 4892 x 4892, this is done to preserve the aspect ratio of CXR while resizing. We scale all images to 512x512 pixels, which retains sufficient visual details for vascular structures.

   All the images which were not in JPEG format were converted into JPEG from DICOM format. The images were normalized to the standard range of [0-1] and resized to 512x512 pixels.

4.2. Methodology

The storytelling XAI framework (Fig:2 is divided into three parts. The priority intent of this framework is not to compete with current state-of-the-art models in terms of evaluation metrics. This framework utilizes knowledge distillation, interpretability, and datasets for a variety of tasks. Using datasets from different origins allows the framework to generalize better for real-world scenarios as well. The three parts involved are:

1. The First step involves training the complex deep neural networks for individual tasks. For the task of abnormality detection and localization, a CNN-based model with ResNet 50 backbone is trained using a categorical cross-entropy loss and mean Average Precision(mAP), which also achieves a remarkable performance on this task [29]. For the task of report generation, a CNN-RNN-Attention model with ResNet50 backbone is trained for the task using masked loss. This task utilizes the concept of image-captioning, where an image is provided as an input. The hidden layers of the architecture perform the feature extraction process. Then, these features are passed through the RNN layer coupled with the attention layer to generate a radiology report [30]. The image segmentation model uses the same model as the abnormality detection model with the last layers replaced and modified to predict the segmentation task, trained using DICE loss [31].

Algorithm 1 Student Model Training with Knowledge Distillation For Multi-Task Prediction

```
Initialize student model $S$ with three heads: $S_{report}$, $S_{abnormality}$, $S_{segmentation}$
Initialize teacher models $T_{report}$, $T_{abnormality}$, $T_{segmentation}$
Initialize loss functions for each task: $L_{report}$, $L_{abnormality}$, $L_{segmentation}$
Initialize hyperparameters: learning rate, temperature $T$ and distillation weights $\alpha$
Freeze parameters of $S_{abnormality}$ and $S_{segmentation}$
for epoch = 1 to $N$ do:
    for $batch$ in $training\_data$ do
        Forward pass through teacher model for report generation: $output_{report} = T_{report}(batch)$
        Compute loss for report generation: $L_{report}(S_{report}(batch), output_{report})$
        Compute knowledge distillation loss for report generation: $\mathcal{L}_{distillation}(S_{report}(batch), output_{report})$
        Compute total loss for report generation: $\mathcal{L} = (1 - \alpha) \cdot \mathrm{L}_{\text{report}} + \alpha \cdot \mathcal{L}_{\text{distillation}}$
        Backpropagation: Compute gradients $\nabla_{\theta_s}\mathcal{L}$
        Update parameters: $\theta_s \leftarrow \theta_s - \eta \cdot \nabla_{\theta_s}\mathcal{L}$
    end for
end for
Freeze parameters of $S_{report}$ and $S_{segmentation}$
for $epoch = 1$ to $N$ do
    for $batch$ in $training\_data$ do:
        Forward pass through teacher model for abnormality detection: $output_{abnormality} = T_{abnormality}(batch)$
        Compute loss for abnormality detection: $L_{abnormality}(S_{abnormality}(batch), output_{abnormality})$
        Compute knowledge distillation loss for report generation: $\mathcal{L}_{distillation}(S_{abnormality}(batch), output_{abnormality})$
        Compute total loss for abnormality detection: $\mathcal{L} = (1 - \alpha) \cdot \mathrm{L}_{\text{abnormality}} + \alpha \cdot \mathcal{L}_{\text{distillation}}$
        Backpropagation: Compute gradients $\nabla_{\theta_s}\mathcal{L}$
        Update parameters: $\theta_s \leftarrow \theta_s - \eta \cdot \nabla_{\theta_s}\mathcal{L}$
    end for
end for
Freeze parameters of $S_{report}$ and $S_{abnormality}$
for $epoch = 1$ to $N$ do
    for $batch$ in $training\_data$ do:
        Forward pass through teacher model for segmentation: $output_{segmentation} = T_{segmentation}(batch)$
        Compute loss for abnormality detection: $L_{segmentation}(S_{segmentation}(batch), output_{segmentation})$
        Compute knowledge distillation loss for report generation: $\mathcal{L}_{distillation}(S_{segmentation}(batch), output_{segmentation})$
        Compute total loss for segmentation: $\mathcal{L} = (1 - \alpha) \cdot \mathrm{L}_{\text{segmentation}} + \alpha \cdot \mathcal{L}_{\text{distillation}}$
        Backpropagation: Compute gradients $\nabla_{\theta_s}\mathcal{L}$
        Update parameters: $\theta_s \leftarrow \theta_s - \eta \cdot \nabla_{\theta_s}\mathcal{L}$
    end for
end for
```

2. The Second step involves performing knowledge distillation between a shallow CNN neural network which will be the backbone of this model. This neural network consists of different prediction heads to perform multiple tasks. The extracted feature representation of the image from the model is passed into the image classification head, text generation head, and image segmentation head to perform abnormality classification, report generation, and image segmentation, simultaneously. Knowledge distillation is performed between the backbone including the specific task head and the individual model trained for the specific task. while the other prediction heads

are frozen (C.f., Algorithm: 1). This is performed until all the prediction heads have acquired knowledge from the complex neural network through knowledge distillation. Here the shallow CNN neural network acts like a student and the individual complex models as teachers.

3. The Third and last step involves generating interpretations from the new model which has acquired knowledge from a complex deep neural network model. This step uses model-agnostic and model-specific interpretation techniques to enable interpretability. The LIME interpretations and GradCAM visualizations are incorporated to interpret the abnormality classification model and chest X-ray segmentation model. For report generation, attention visualizer [32] is used to output the results.

5. Results

The chest X-ray image was given as input to the shallow neural network, which was trained with the knowledge distillation method from three different deep neural networks for a different task. The original image was resized to a resolution of 512x512 pixels. See figure: 3. The model accurately segmented the two lungs. The abnormality detection model was able to localize and classify the localized affected regions into infiltration, consolidation, pleural effusion, and cardiomegaly. The same model generated the resulting text for radiologists and healthcare professionals. Furthermore, interpretability techniques were applied, which include LIME interpretation, attention map visualization of the generated report, and CAM analysis consisting of GradCAM and GradCAM++. As shown in the figure, the text justification asserts the output of interpretability techniques along with the prediction from the student model which will be beneficial for ML practitioners and at the same time it is able to assert the abnormality detection for healthcare professionals as well. Knowledge distillation is a well-capable method to let the shallow neural network to learn representation over a variety of tasks enabling an interpretable neural network. The benefit of the multiple tasks could be inferred from the result (Fig:3), as these multiple tasks when combined, improve the overall understanding of the predictions. The report generation task serves as a textual justification for the other two tasks, namely abnormality detection and chest X-ray segmentation. For interpretation, we have focused on LIME, GradCAM, and visualization techniques. The framework could be extended to incorporate other interpretability techniques including but not limited to SHapley Additive exPlanations (SHAP) [33], Layer-wise Relevance Propagation (LRP) [34], etc.

6. Conclusion

Our framework is a step toward enabling audience-centric explainability. We have used concept-based explainability to explain the results of the AI models to healthcare professionals. This framework serves as a baseline to demonstrate that the XAI is an amalgamation of interpretability and concept-based approaches to achieve end-to-end explainability. Note that ML practitioners can also interpret the results of the models in technical terms. This framework is also able to explain complex deep neural networks because of the knowledge distillation. The teacher models can each be a very complex deep neural network, but the student model can learn the knowledge. This framework requires multiple datasets for multiple machine tasks. The multiple data sets from different sources result in more generalized and robust models in a real-world scenario. Robust and generalized models that perform well in a variety of real-world scenarios are more likely to be trusted and understood by users. Robust and generalized models are less susceptible to adversarial manipulation, improving the reliability of their explanations.

In this research, we have limited our investigation to some specific tasks of abnormality detection, report generation, and chest X-ray segmentation. Other tasks could be added if the datasets for these tasks are available and the inputs are common to all tasks. We have demonstrated a healthcare use case, but this framework could be extended to other domains given the available datasets. Multi-task models are a great way to exploit the relationships between different commonalities. This technique allows the model to learn shared feature representations. By leveraging the shared representations across related tasks, multi-task learning can learn more efficiently from limited data compared to Learning a shared representation that performs well across multiple related tasks can lead to more stable and meaningful predictions. Our Storytelling XAI framework can be implemented in critical-stakes domains to enable audience-centric explanations with robust predictions and generalized models, without compromising the performance of even the most complex deep neural networks.

7. Discussions

A potential concern is that providing an excessive number of explanations or information may lead to an overload of data, hindering comprehension and user satisfaction. However, the presentation of multiple explanations can enhance transparency and facilitate trust in AI models by addressing the diverse needs of users. By offering a range of explanations, users can gain a more nuanced understanding of the decision-making process, which can ultimately enhance their satisfaction and comprehension [35]. One might also inquire as to the rationale behind the use of shallow models. The use of shallow models is justified by their reduced susceptibility to overfitting and superior generalization capabilities when processing new data sets, particularly in the context of limited data. Training shallow models for multi-tasking allows for the exploitation of relationships between different input modalities, which is not feasible for single-task models. The majority of interpretability methods employ approximation techniques. As the depth of a neural network model increases, the reliance on approximation increases and reliability decreases. Consequently, shallow methods align well with the aforementioned criteria [36].

8. Limitations

Our framework currently lacks the ability of a user to interact with the XAI results freely. This may be resolved with a user interface that focuses on integrating Attention Maps, LIME, SHAP, and GradCam visualizations. When extending this framework to other domains, it may be somewhat problematic to collect different datasets for different tasks. There is limited support for teacher models. The current possibilities are to pick either CNN or RNN-based architectures. For transformer architectures, it may be difficult to perform knowledge distillation. The process of knowledge distillation of transformer models is inherently challenging due to the intricate architectural complexity, the considerable number of parameters, and the difficulty in transferring the intricate patterns learned by self-attention mechanisms. Furthermore, the process is further complicated by the necessity to balance multiple objectives and the risk of overfitting. We have experimented with tasks where the input is solely an image. There is indeed not only a limitation in terms of multimodal support but also textual-only data. Despite the limitations, the work lays down the foundation for future work.

9. Future Work

In future work, we would develop a user interface with a human-in-the-loop, preferably domain experts which satisfies the nested model for AI design and validation [37]. Then we would focus on conducting a design study and include a new metric to calculate an agreement score between the framework and the domain experts. There would also be an extensive effort to integrate the XAI Question Bank (XAI-QB) into our framework [38]. The XAI-QB is critical to the development of explainable AI systems and is designed to help XAI practitioners understand the explanation needs of different end users. It represents common user expectations and covers major categories of questions for standard supervised ML systems.

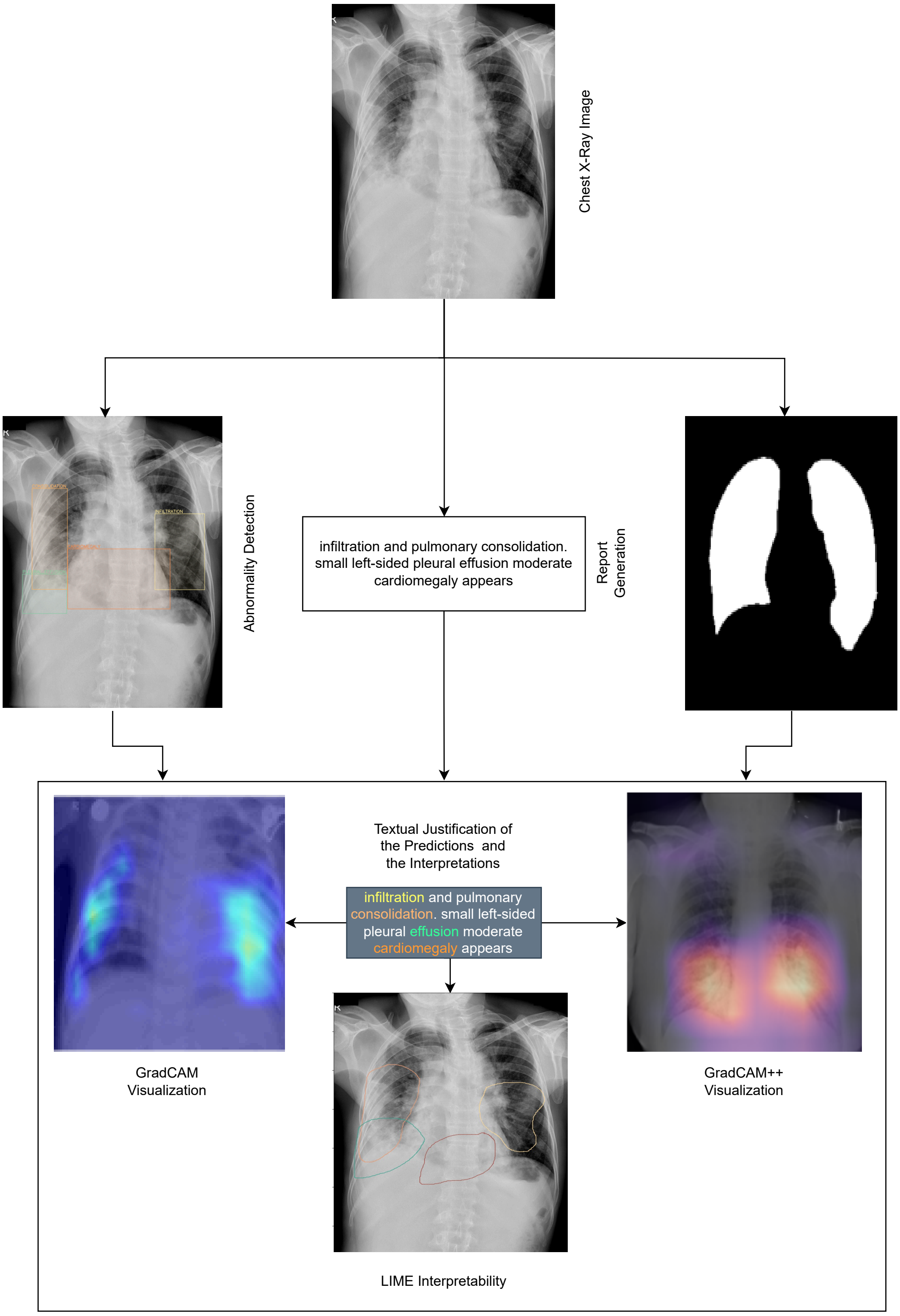

Chest X-Ray Image
Abnormality Detection
infiltration and pulmonary consolidation. small left-sided pleural effusion moderate cardiomegaly appears
Report Generation
Textual Justification of the Predictions and the Interpretations
infiltration and pulmonary consolidation. small left-sided pleural effusion moderate cardiomegaly appears
GradCAM Visualization
GradCAM++ Visualization
LIME Interpretability

Figure 3: Overview of the Storytelling XAI framework applied in the medical imaging domain. The input image is provided as an input to the student model. The student model performs abnormality detection, lung segmentation, and report text generation. The individual results are provided as input to the interpretability module to generate interpretations. The attention map visualization shows different detected abnormalities.

References

[1] L. A. Celi, B. Fine, D. J. Stone, An awakening in medicine: the partnership of humanity and intelligent machines, The Lancet Digital Health 1 (2019) e255–e257.

[2] T. H. Davenport, J. P. Glaser, Factors governing the adoption of artificial intelligence in healthcare providers, Discover Health Systems 1 (2022) 4.

[3] S. Reddy, Explainability and artificial intelligence in medicine, The Lancet Digital Health 4 (2022) e214–e215.

[4] J. Amann, A. Blasimme, E. Vayena, D. Frey, V. I. Madai, P. Consortium, Explainability for artificial intelligence in healthcare: a multidisciplinary perspective, BMC medical informatics and decision making 20 (2020) 1–9.

[5] Z. Yang, X. Dai, A. Dubey, S. Hirche, G. Hattab, Whom to trust? elective learning for distributed gaussian process regression, in: Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, 2024, pp. 2020–2028.

[6] W. Jin, J. Fan, D. Gromala, P. Pasquier, G. Hamarneh, Invisible users: Uncovering end-users' requirements for explainable ai via explanation forms and goals, arXiv preprint arXiv:2302.06609 (2023).

[7] H. Panwar, P. Gupta, M. K. Siddiqui, R. Morales-Menendez, P. Bhardwaj, V. Singh, A deep learning and grad-cam based color visualization approach for fast detection of covid-19 cases using chest x-ray and ct-scan images, Chaos, Solitons & Fractals 140 (2020) 110190. URL: https://www.sciencedirect.com/science/article/pii/S0960077920305865. doi:https://doi.org/10.1016/j.chaos.2020.110190.

[8] S. prasad Koyyada, T. P. Singh, An explainable artificial intelligence model for identifying local indicators and detecting lung disease from chest x-ray images, Healthcare Analytics 4 (2023) 100206. URL: https://www.sciencedirect.com/science/article/pii/S2772442523000734. doi:https://doi.org/10.1016/j.health.2023.100206.

[9] D. Martens, C. Dams, J. Hinns, M. Vergouwen, Tell me a story! narrative-driven xai with large language models, arXiv preprint arXiv:2309.17057 (2023).

[10] A. Asemota, G. Hooker, Longitudinal counterfactuals: Constraints and opportunities, arXiv preprint arXiv:2403.00105 (2024).

[11] S. Baron, Explainable ai and causal understanding: Counterfactual approaches considered, Minds and Machines 33 (2023) 347–377.

[12] M. Hartmann, H. Du, N. Feldhus, I. Kruijff-Korbayová, D. Sonntag, Xaines: Explaining ai with narratives, KI-Künstliche Intelligenz 36 (2022) 287–296.

[13] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv preprint arXiv:1503.02531 (2015).

[14] A. Carrillo, L. F. Cantú, A. Noriega, Individual explanations in machine learning models: A survey for practitioners, arXiv preprint arXiv:2104.04144 (2021).

[15] H. Han, S. Kim, H.-S. Choi, S. Yoon, On the impact of knowledge distillation for model interpretability, arXiv preprint arXiv:2305.15734 (2023).

[16] X. Liu, X. Wang, S. Matwin, Improving the interpretability of deep neural networks with knowledge distillation, in: 2018 IEEE International Conference on Data Mining Workshops (ICDMW), IEEE, 2018, pp. 905–912.

[17] M. Crawshaw, Multi-task learning with deep neural networks: A survey, arXiv preprint arXiv:2009.09796 (2020).

[18] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

[19] P. Linardatos, V. Papastefanopoulos, S. Kotsiantis, Explainable ai: A review of machine learning interpretability methods, Entropy 23 (2020) 18.

[20] B. Hadji Misheva, D. Jaggi, J.-A. Posth, T. Gramespacher, J. Osterrieder, Audience-dependent explanations for ai-based risk management tools: A survey, Frontiers in Artificial Intelligence 4

(2021) 794996.

[21] N. Prentzas, A. Kakas, C. S. Pattichis, Explainable ai applications in the medical domain: a systematic review, arXiv preprint arXiv:2308.05411 (2023).

[22] Y. Rong, T. Leemann, T.-T. Nguyen, L. Fiedler, P. Qian, V. Unhelkar, T. Seidel, G. Kasneci, E. Kasneci, Towards human-centered explainable ai: A survey of user studies for model explanations, IEEE Transactions on Pattern Analysis and Machine Intelligence (2023).

[23] A. Johnson, M. Lungren, Y. Peng, Z. Lu, R. Mark, S. Berkowitz, S. Horng, Mimic-cxr-jpg-chest radiographs with structured labels, PhysioNet (2019).

[24] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, C. J. McDonald, Preparing a collection of radiology examinations for distribution and retrieval, Journal of the American Medical Informatics Association 23 (2016) 304–310.

[25] J. H. Moon, H. Lee, W. Shin, Y.-H. Kim, E. Choi, Multi-modal understanding and generation for medical images and text via vision-language pre-training, IEEE Journal of Biomedical and Health Informatics 26 (2022) 6070–6080.

[26] M. Chizhikova, M. Diaz-Galiano, L. A. U. Lopez, M. T. Martín-Valdivia, Sinai at radsum23: Radiology report summarization based on domain-specific sequence-to-sequence transformer model, in: The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, 2023, pp. 530–534.

[27] DungNB, H. Q. Nguyen, J. Elliott, KeepLearning, NguyenThanhNhan, P. Culliton, Vinbigdata chest x-ray abnormalities detection, 2020. URL: https://kaggle.com/competitions/vinbigdata-chest-xray-abnormalities-detection.

[28] H. Nguyen, et al., Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations.(2022) doi: 10.48550, arXiv (2012).

[29] S. Shivadekar, B. Kataria, S. Hundekari, K. Wanjale, V. P. Balpande, R. Suryawanshi, Deep learning based image classification of lungs radiography for detecting covid-19 using a deep cnn and resnet 50, International Journal of Intelligent Systems and Applications in Engineering 11 (2023) 241–250.

[30] J. Aneja, A. Deshpande, A. G. Schwing, Convolutional image captioning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5561–5570.

[31] R. Rashid, M. U. Akram, T. Hassan, Fully convolutional neural network for lungs segmentation from chest x-rays, in: Image Analysis and Recognition: 15th International Conference, ICIAR 2018, Póvoa de Varzim, Portugal, June 27–29, 2018, Proceedings 15, Springer, 2018, pp. 71–80.

[32] F. Liu, C. Yin, X. Wu, S. Ge, Y. Zou, P. Zhang, X. Sun, Contrastive attention for automatic chest x-ray report generation, arXiv preprint arXiv:2106.06965 (2021).

[33] Y. Arslan, B. Lebichot, K. Allix, L. Veiber, C. Lefebvre, A. Boytsov, A. Goujon, T. F. Bissyandé, J. Klein, Towards refined classifications driven by shap explanations, in: International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer, 2022, pp. 68–81.

[34] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, K.-R. Müller, Layer-wise relevance propagation: an overview, Explainable AI: interpreting, explaining and visualizing deep learning (2019) 193–209.

[35] W. Yang, Y. Wei, H. Wei, Y. Chen, G. Huang, X. Li, R. Li, N. Yao, X. Wang, X. Gu, et al., Survey on explainable ai: From approaches, limitations and applications aspects, Human-Centric Intelligent Systems 3 (2023) 161–188.

[36] S. Xu, Y. Song, X. Hao, A comparative study of shallow machine learning models and deep learning models for landslide susceptibility assessment based on imbalanced data, Forests 13 (2022) 1908.

[37] A. Dubey, Z. Yang, G. Hattab, A nested model for ai design and validation, arXiv e-prints (2024) arXiv–2407.

[38] Q. V. Liao, D. Gruen, S. Miller, Questioning the ai: informing design practices for explainable ai user experiences, in: Proceedings of the 2020 CHI conference on human factors in computing systems, 2020, pp. 1–15.