Visual Question Answering for Medical Images with Explainable AI

Deepananth K 195001027
Jayakrishnan S V 195001040

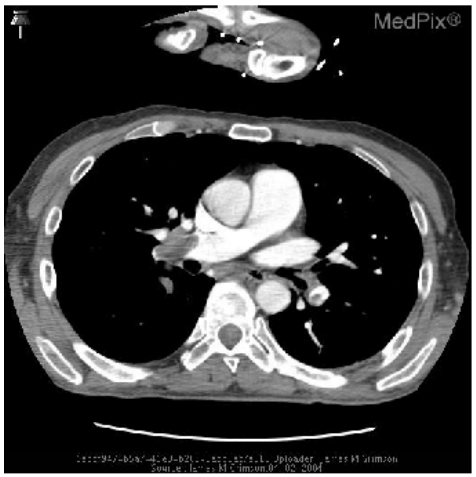
Under the Guidance of
Dr. S. Kavitha

SSN College of Engineering, Chennai

First Review
February 10, 2023

What is Visual Question Answering

- Visual Question Answering(VQA) is an emerging approach under the domains of Computer Vision and Natural Language processing, which aims to answer the user's question by analysing the given input image.
- VQA can be applied to several types of images like Natural Images, Medical Images and Cartoon Images.



(g) **Q**: which organ system is shown in the ct scan? **A**: lung, mediastinum, pleura

(h) **Q**: what is abnormal in the gastrointestinal image? **A**: gastric volvulus (organoaxial)

Figure: Sample Images and Questions with Corresponding Answers from ImageCLEF 2019 VQA-Med Dataset

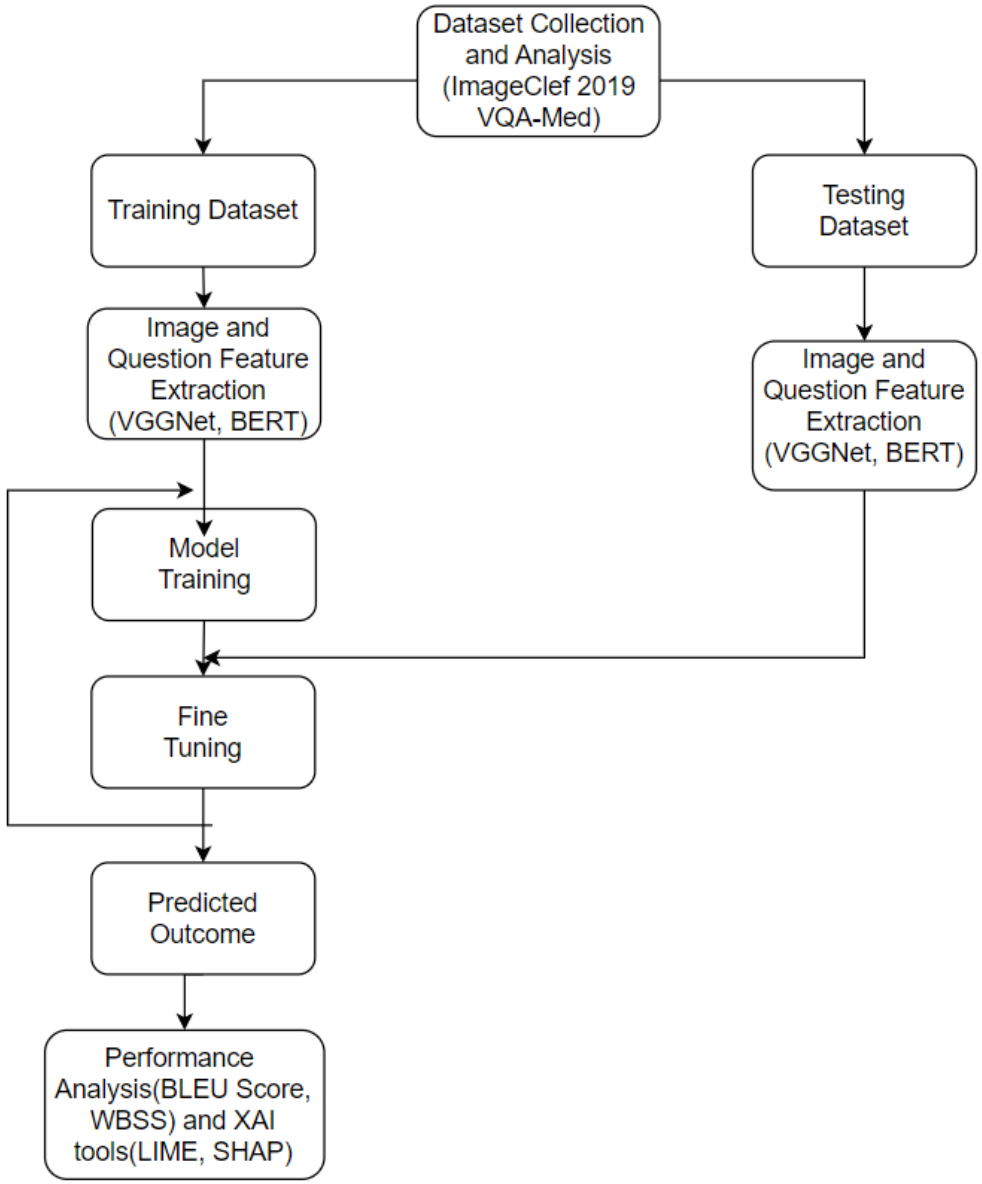
Problem Statement

- The aim of this project is to build an efficient VQA model that generates answers to questions related to medical images using deep learning techniques
- In addition, the reason behind the generated answer has to be analyzed using Explainable AI(XAI) tools such as LIME or SHAP to provide explanations on the outcome.

Objectives

- To collect and analyze the dataset of Visual Question Answering from the CLEF Forum.
- To compare and validate different transformer models for text feature extraction.
- To build an efficient VQA model that generates the answers to the questions related to the given Medical Image.
- To analyze the generated answer, Explainable AI tools have to be used for more explanations.

System Design



Module Split-up

The system design is split into four modules as follows:

- Dataset Collection and Analysis
- Feature Extraction
- VQA Model Building
- Performance Analysis and XAI

Timeline

Review	Module	Jan	Feb	Mar	Apr
Review 1	Data Collection and Analysis				
	Image Feature Extraction				
Review 2	Text Feature Extraction				
	VQA Model Building				
Review 3	Performance Analysis (BLEU Score, WBSS) and XAI (LIME, SHAP)				

Dataset

For this project, ImageClef 2019 VQA-Med dataset is used.

Dataset	**Question Category**	**No. of Questions**	**No. of Classes**
Training Dataset	Modality	3200	44
	Plane	3200	15
	Organ System	3200	10
	Abnormality	3192	1484
Testing Dataset	All	500	-

Table: Dataset Analysis

Image Feature Extraction

- In this project, CNN and pre-trained VGGNet are used for Image feature extraction
- CNN performs better without loss of information and reduced dimensionality of the image
- The training parameters for a CNN is lesser than other neural network
- VGGNet is a pre-trained CNN architecture, which is one of the best and most widely used architecture
- VGGNet extract better features from the images, which gives better results in various situations

Transformer Models for Text Processing

- A transformer model is a neural network that learns the context of the sequence of data like sentences which is a sequence of words
- They handle long-range dependencies in the sequence of data
- There are various pre-trained transformer models like BERT, XLNet, GPT, etc.,
- In this project, BERT is to be used for feature extraction and also for answer generation

Explainable AI (XAI)

- Explainable AI(XAI) is a rapidly emerging research idea which refers to methods and techniques that helps us understand and interpret predictions made by AI models.
- There are many XAI tools available like LIME, What-If, SHAP, ELI5, etc.,
- In this project, LIME (**L**ocal **I**nterpretable **M**odel-Agnostic **E**xplanations) and SHAP (**SH**apley **A**dditive ex**P**lanations) tools are to be used
- **LIME** - outputs a list of explanations, reflecting the contribution of each feature to the prediction of a data sample
- **SHAP** - explains how the different features affect the output or what contribution do they have in the outcome of the model

Implementation & Result

DNN Techniques	**Training Accuracy**	**Testing Accuracy**
Custom CNN	0.9906	0.512
Pre-trained VGGNet	0.7125	0.681
VGGNet Trained with our Dataset	0.5706	0.586

Table: Comparison of various feature extraction techniques

Visualization

To develop an efficient feature extraction model and to choose the correct number of layers, visualization(using heatmap) techniques are used.

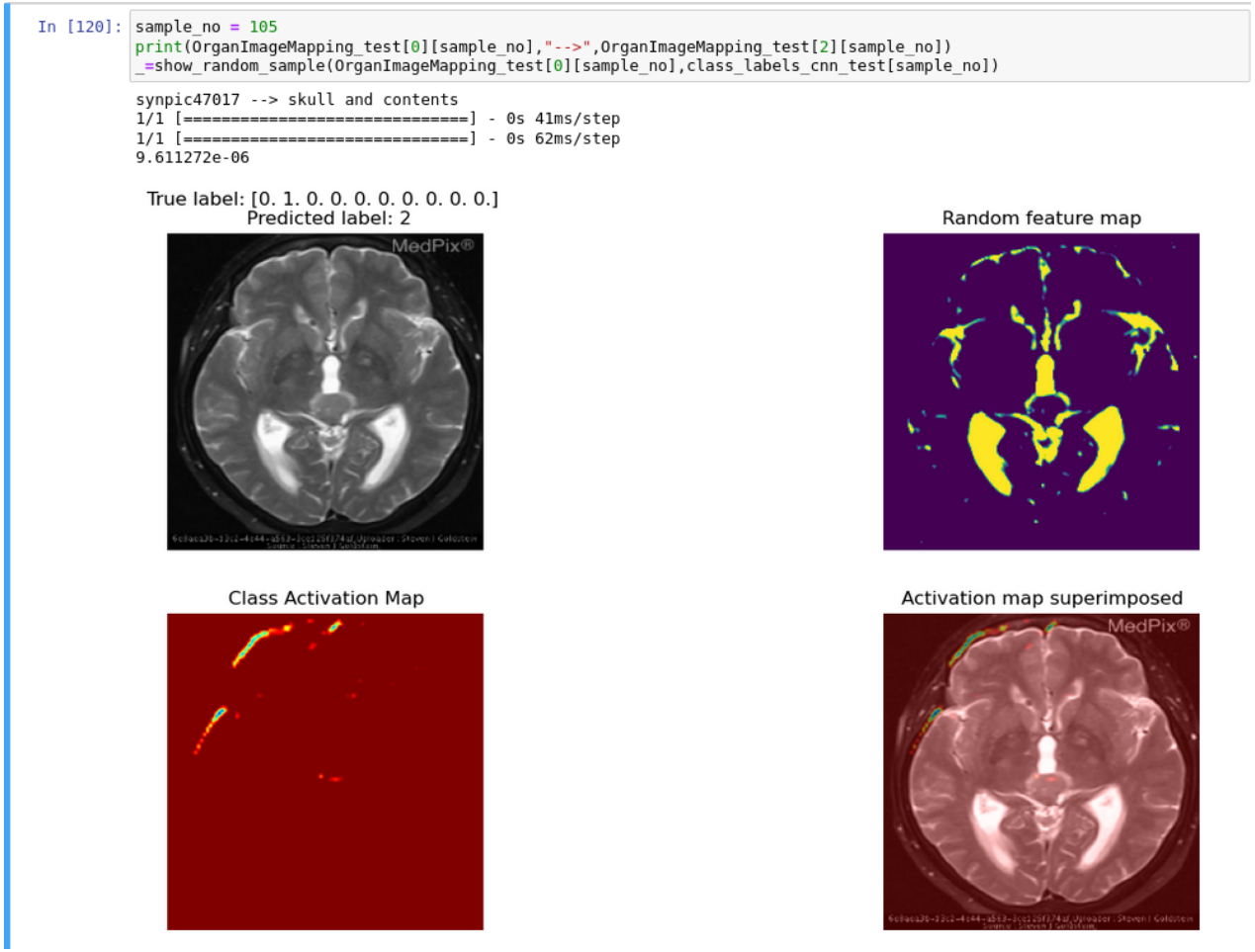


Figure: Custom CNN

Visualization contd.

```
In [124]: sample_no = 105
          print(OrganImageMapping_test[0][sample_no],"-->",OrganImageMapping_test[2][sample_no])
          _=show_random_sample(OrganImageMapping_test[0][sample_no],class_labels_cnn_test[sample_no])

synpic47017 --> skull and contents
1/1 [==============================] - 0s 77ms/step
1/1 [==============================] - 0s 116ms/step
1.4341319e-07
```

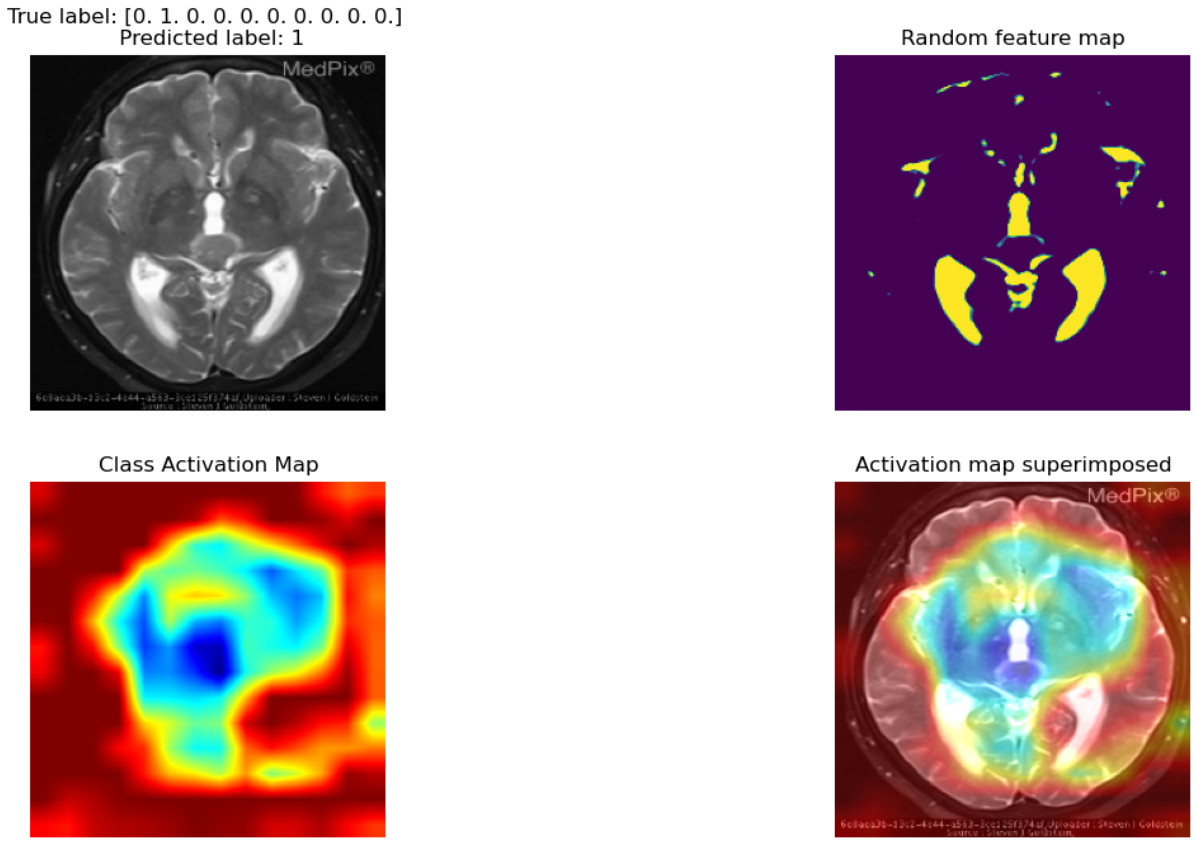


Figure: Pre-trained VGGNet

Visualization contd.

```
In [95]: sample_no = 105
         print(OrganImageMapping_test[0][sample_no],"-->",OrganImageMapping_test[2][sample_no])
         _=show_random_sample(OrganImageMapping_test[0][sample_no],class_labels_cnn_test[sample_no])

synpic47017 --> skull and contents
1/1 [==============================] - 0s 102ms/step
1/1 [==============================] - 0s 128ms/step
2.5663428e-06
```

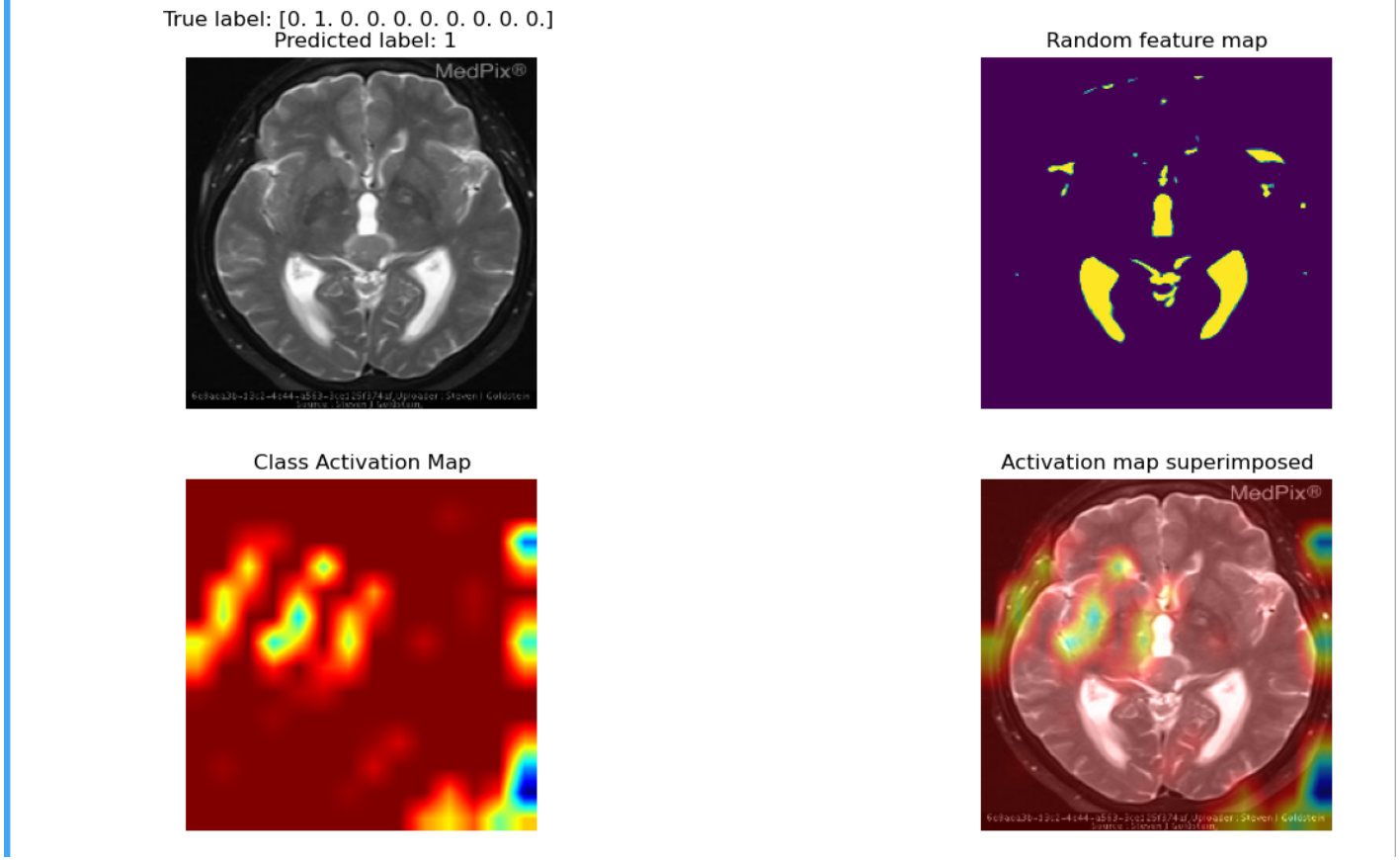


Figure: VGGNet weights Trained with our Dataset

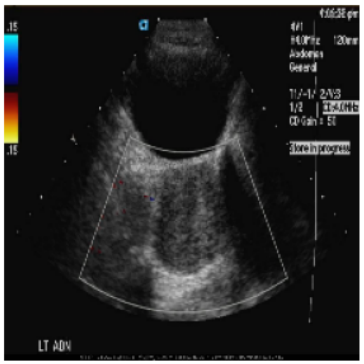
Implementation Challenges

- Dataset Complexity
- Text Present in Images

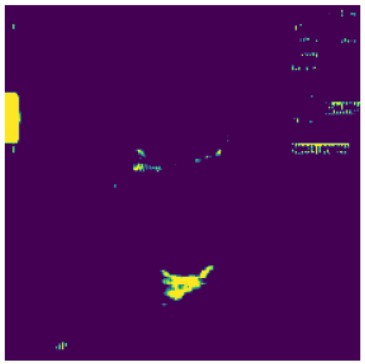
```
In [127]: sample_no = 59
          print(OrganImageMapping_test[0][sample_no],"-->",OrganImageMapping_test[2][sample_no])
          _=show_random_sample(OrganImageMapping_test[0][sample_no],class_labels_cnn_test[sample_no])

          synpic18921 --> genitourinary
          1/1 [==============================] - 0s 94ms/step
          1/1 [==============================] - 0s 140ms/step
          8.318302e-06
```

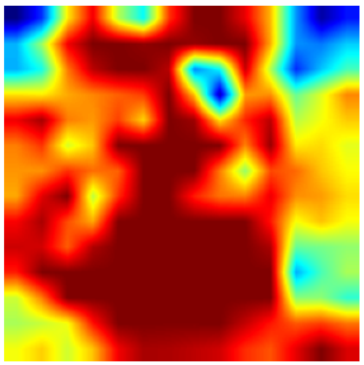
True label: [0. 0. 1. 0. 0. 0. 0. 0. 0. 0.]
Predicted label: 2



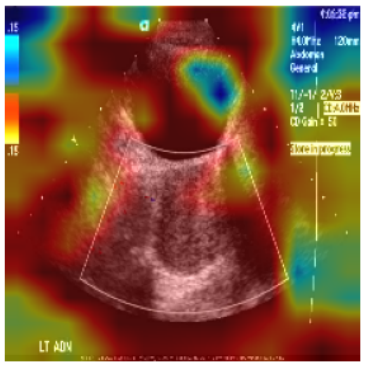
Random feature map



Class Activation Map



Activation map superimposed



Conclusion

- The dataset for the task of Visual Question Answering is collected and it is analyzed.
- To extract features that best represents the images, Deep Learning techniques such as CNN, pre-trained VGGNet and VGGNet trained on ImageClef 2019 VQA-Med datset are implemented.
- The performance of these models are compared in terms of accuracy and visualization(heatmap).
- The comparison shows that VGGNet with pre-trained weights perform better in terms of accuracy and also visualization.

References

[1] A. Lubna, S. Kalady and A. Lijiya, *MoBVQA: A Modality based Medical Image Visual Question Answering System*, TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), Kochi, India, 2019, pp. 727-732, doi: 10.1109/TENCON.2019.8929456.

[2] Allaouzi, Imane et al. *An Encoder-Decoder Model for Visual Question Answering in the Medical Domain*, Conference and Labs of the Evaluation Forum (2019).

[3] Al-Sadi, Aisha & Al-Ayyoub, Mahmoud & Jararweh, Yaser & Costen, F.. (2021). *Visual Question Answering in the Medical Domain Based on Deep Learning Approaches: A Comprehensive Study. Pattern Recognition Letters.* 150. 10.1016/j.patrec.2021.07.002.

[4] Sharma, D., Purushotham, S. & Reddy, C.K. *MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain.* Sci Rep 11, 19826 (2021).

References contd.

[5] Yangyang Zhou, Xin Kang, Fuji Ren. *Employing Inception-Resnet-v2 and Bi-LSTM for Medical Domain Visual Question Answering.* CLEF (Working Notes) 2018.

[6] Knapič S, Malhi A, Saluja R, Främling K. *Explainable Artificial Intelligence for Human Decision Support System in the Medical Domain*. Machine Learning and Knowledge Extraction. 2021; 3(3):740-770.

[7] S. H. P. Abeyagunasekera, Y. Perera, K. Chamara, U. Kaushalya, P. Sumathipala and O. Senaweera, *LISA : Enhance the explainability of medical images unifying current XAI techniques*. 2022 IEEE 7th International conference for Convergence in Technology (I2CT), Mumbai, India, 2022, pp. 1-9, doi: 10.1109/I2CT54291.2022.9824840.

[8] Lin Z, Zhang D, Tac Q, Shi D, Haffari G, Wu Q, He M, Ge Z. *Medical visual question answering: A survey*. arXiv preprint arXiv:2111.10056. 2021 Nov 19.

Thank You

Literature Survey

Paper Title	Methodology	Limitations
MoBVQA: A Modality based Medical Image Visual Question Answering System [**MoBVQA**]	A CNN is trained for the different modalities like X-Ray, MRI, CT, Ultrasound **Dataset**: ImageCLEF 2019 VQA-Med **Answer Generation**: CNN Classifier **Analysis**: Accuracy-60.8, BLEU Score-63.4	Only modality based questions are considered
Visual question answering in the medical domain based on deep learning approaches: A comprehensive study [**Al-Sadi**]	The Questions are classified into 4 categories and multiple models are trained for each type of question **Dataset**: ImageCLEF 2019 VQA-Med **Image Feature Extraction**: VGGNet16 **Answer Generation**: Ensemble of Classification models **Analysis**: Accuracy-60.8, BLEU Score-63.4	All models built for each question categories are classification models which is completely a black-box approach

Literature Survey Contd.

Paper Title	**Methodology**
Explainable Artificial Intelligence for Human Decision Support System in the Medical Domain [**XAI-CNN**]	**Dataset:** Red Lesion Endoscopy data **XAI tools:** LIME, SHAP, CIU A CNN is trained using the dataset. XAI tools are then used for visualization in terms of heatmap. The result of visualization is then compared
LISA : Enhance the explainability of medical images unifying current XAI techniques [**XAI-LISA**]	**Dataset:** COVID-19 Dataset **XAI tools:** LIME, SHAP, Anchors **Other XAI techniques:** Integrated Gradients Transfer Learning is used for the detection of COVID-19. The XAI tools LIME, SHAP Anchor and Integrated Gradient techniques' results are combined to give explanations.